Beyond Differential Expression: Methods and Tools for Mining the Transcriptomic Landscape of Human Tissue and Disease

by

Patrick Raphael Schmid

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2012

Author . . . .
Department of Electrical Engineering and Computer Science
February 3, 2012

Certified by . . . .
Dr. Bonnie Berger
Professor of Applied Mathematics and Computer Science
Thesis Supervisor

Accepted by . . . .
Professor Leslie A. Kolodziejski
Chair of the Committee on Graduate Students

Beyond Differential Expression: Methods and Tools for Mining the Transcriptomic Landscape of Human Tissue and Disease

by

Patrick Raphael Schmid

Submitted to the Department of Electrical Engineering and Computer Science
on February 3, 2012, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Although there are a variety of high-throughput technologies used to perform biological experiments, DNA microarrays have become a standard tool in the modern biologist's arsenal. Microarray experiments provide measurements of thousands of genes simultaneously, and offer a snapshot view of transcriptomic activity. With the rapid growth of public availability of transcriptomic data, there is increasing recognition that large sets of such data can be mined to better understand disease states and mechanisms. Unfortunately, several challenges arise when attempting to perform such large-scale analyses. For instance, public repositories to which the data is being submitted to were designed around the simple task of storage rather than that of data mining. As such, the seemingly simple task of obtaining all data relating to a particular disease becomes an arduous task. Furthermore, prior gene expression analyses, both large and small, have been dichotomous in nature, in which phenotypes are compared using clearly defined controls. Such approaches may require arbitrary decisions about what are considered "normal" phenotypes, and what each phenotype should be compared to.

Addressing these issues, we introduce methods for creating a large curated gene expression database geared towards data mining, and explore methods for efficiently expanding this database using active learning. Leveraging our curated expression database, we adopt a holistic approach in which we characterize phenotypes in the context of a myriad of tissues and diseases. We introduce scalable methods that associate expression patterns to phenotypes in order to assign phenotype labels to new expression samples and to select phenotypically meaningful gene signatures. By using a nonparametric statistical approach, we identify signatures that are more precise than those from existing approaches and accurately reveal biological processes that are hidden in case vs. control studies. We conclude the work by exploring the applicability of the heterogeneous expression database in analyzing clinical drugs for the purpose of drug repurposing.

Thesis Supervisor: Dr. Bonnie Berger
Title: Professor of Applied Mathematics and Computer Science

Acknowledgments

I would like to thank Dr. Bonnie Berger and Dr. Isaac Kohane for many years of guidance and support. In addition, I would like to thank my close collaborator and friend, Dr. Nathan Palmer, for his time and effort in tackling shared research problems over countless cups of coffee. I would also like to thank my family and friends who have always provided me with a great deal of support. Last, but definitely not least, I would like to thank my loving wife, Candice, for her continued love, patience, and support. Without all of these people, none of this work would have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The science-fiction film *GATTACA* depicts a world in which a person's susceptibility to different diseases is known at birth based on an analysis of the newborn's genetic code. Although the bleak outlook of the future presented in this film is plagued by the detrimental use of genetic information to form social castes, imagine a rosier view in which a person's genetic information can be used not only to prevent certain diseases, but also to provide personalized treatment that is attuned to an individual's exact biological and environmental properties. For example, imagine the amount of pain and suffering that can be avoided if a surgeon where to be able to conclusively determine the origin and exact subtype of a tumor and compare the treatment outcomes of patients with similar biological and clinical properties such that the most efficacious treatment can be implemented. This will become the norm in the near future. In order for this to become a reality, however, a vast amount of data needs to be leveraged and combined to produce accurate predictors for the wide array of clinical outcomes.

While there are many types of biological data that can be used to aid in answering the question of what makes a certain tissue different from another, or a certain disease similar to some other seemingly dissimilar disease, *gene expression* analyses have become standard in high-throughput analyses of tissues and diseases. Simply stated, a gene expression experiment (also known as a *microarray* experiment) provides a snapshot view of thousands of genes and denotes whether they are turned "on" and "off" (see Section 1.1.3 for more details). Such snapshots can be used to compare different types of tissue (e.g. lung vs. brain tissue) or different states of a tissue (e.g. normal vs. diseased). For example, Alizadeh et al. [6] performed an analysis of a large B-cell lymphoma, a malignancy of the lymphatic system, by analyzing which genes were turned "on" and "off" in the resected lymphatic tissue of patients. Based solely on the gene expression patterns, they were able to find two distinct clusters of patients. What made these two sub-populations different? A dramatic difference in mortality rate. By "merely" looking at the genes that were expressed in lymphatic tissue they were able to generate a diagnosis with great clinical relevance. Imagine if we could perform such analyses for all types of diseases.

To make these sorts of analyses and potential subsequent clinical applications routine, however, we require a large curated database of thousands, or even hundreds of thousands, of samples across multiple phenotypes. Leveraging the data in such a database, we can then not only examine the outcomes of a single disease, but rather, begin to understand the biological underpinnings of hundreds of diseases and their subtypes. Furthermore, it becomes imperative not to perform these analyses in isolation, but rather in the context of other tissues and diseases from various types of patients. For instance, the treatment course for the same disease may be markedly different for two individuals based on other diseases they may also have. With rapidly growing repositories of public microarray data (see Figure 2-1), the notion of using hundreds of samples spanning various tissues and diseases to perform detailed gene expression based analyses has become feasible. Similarly, with the constant decrease in price and complexity of performing microarray experiments, the clinical application of microarrays is within reach. Unfortunately, without a so-called "black box" that a clinician can use to test a given patient's gene expression data against, gene expression data cannot be used as a diagnostic tool.

Other recent work utilizing large disparate datasets by Butte et al. [19] and Segal et al. [116] show that it is possible to find genes and gene modules that are significantly associated with various phenotypes. Alternatively, Dudley et al. [28] recently showed how the genes that are expressed in various diseases can be used for repurposing drugs.

Commercial ventures such as NextBio [67] and Oncomine [103] have also begun to take the results from disparate biological experiments to elucidate novel insights. Building upon the foundation of the ideas and insights of these large-scale analyses, we show how we can build a large, curated gene expression database (Chapter 2) and then how it can be used to accurately label previously unseen expression samples with their phenotypic labels (Chapter 3), elucidate sets of phenotype specific "marker genes" (Chapter 4), expand an expression database through active learning (Chapter 5), and how it can be applied to analyze drugs (Chapter 6).

1.1 Biology and terminology

Before delving deeper, let us review (or, for some, learn for the first time) some introductory biology. For those of you who are familiar with transcriptional biology and the workings of microarray technology, feel free to skip to the next chapter.

1.1.1 Basic biology

At the most basic level, living organisms are made up of individual cells. Some very simple organisms, such as bacteria, are unicellular and are called prokaryotes[1]. Humans, on the other hand, are eukaryotic organsism and are not only multicellular, but are comprised of cells that have a nucleus. Although there are many different types of cells in complex multicellular organisms (liver cells, brain cells, blood cells, etc.), each cell contains the entire blueprint, or genetic code, for that particular organism. As such, it could theoretically be possible to make a whole new organism by taking any cell from that organism and copying it (just like they did in the book, and later movie, *Jurassic Park*). This genetic information is stored in the form of DNA (deoxyribonucleic acid) and is primarily found in the nucleus of the cell[2]. When one refers to an organism's "genetic code," one generally means the arrangement of

[1] More accurately, organisms that are comprised of cells that lack a cell nucleus are called prokaryotes. Eukaryotes are organisms that are made up of cells that have a cell nucleus.

[2] There is also a small amount of mitochondrial DNA (mtDNA) in the energy producing structures called the mitochondria.

the four chemical bases (also called nucleotides) adenine (A), guanine (G), cytosine (C), and thymine (T) that make up the DNA (Figure 1-1). While all humans share about 99% of the 3 billion bases, the differences in the arrangement of the A, C, T, and Gs for the remaining 1% is what differentiates you from me [84]. Importantly, in the double helix of DNA, adenine always pairs with thymine, and guanine always pairs with cytosine. Although outside the scope of this introduction, it is vital that these pairings remain constant, as during cell replication, it is imperative that each daughter cell can make a full double helix of DNA from just one strand of DNA.

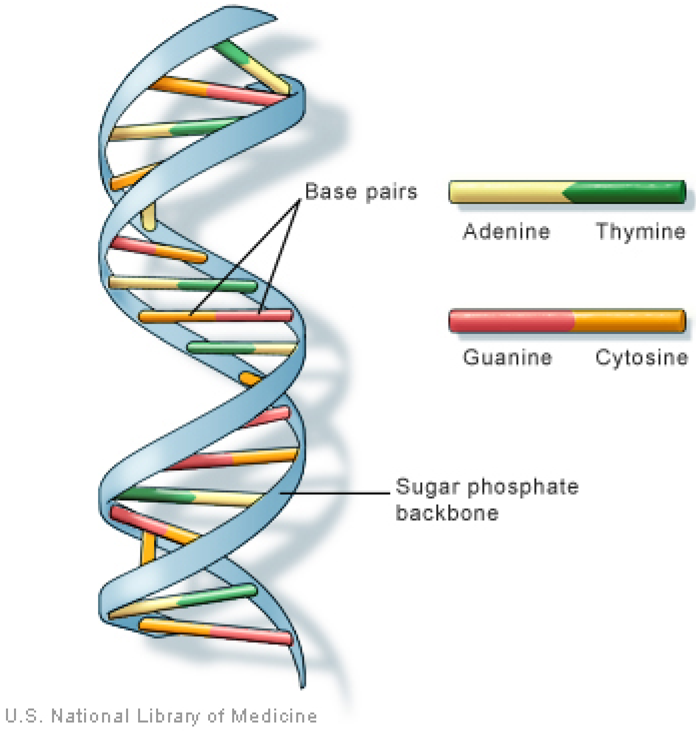


Figure 1-1: Adenine pairs with Thymine, and Guanine pairs with Cytosine to make the familiar double helix of DNA [84].

If the DNA is considered the blueprint document of an organism, the genes that are encoded in this DNA can be considered to be the individual specifications for the rooms, stairwells, and so forth. The 3 billion bases of DNA are subdivided into smaller regions known as genes. Currently it is estimated that humans have between 20,000 and 25,000 genes [84]. Each of these genes, which can be as short as a few hundred DNA bases to over 2 million bases, are the instructions for building molecules known as proteins. Proteins are the workhorses of the cells and are required for the function, structure, and regulation of the tissues and organs in the body.

1.1.2 Transciptional biology

As the contents of this work deals with gene expression of humans[3], let us explore the process of how proteins are synthesized from DNA. As aforementioned, the DNA of a eukaryotic cell is located in the nucleus. Most of the work that is performed by the cell, however is undertaken by proteins in the cytoplasm outside of the nucleus. The genetic code of the gene[4] located on the DNA is not directly converted into protein in the nucleus, but first converted to RNA (ribonucleic acid) that then moves out of the nucleus and is used as a "carbon copy" of the DNA blueprint to create the protein. Just like DNA, RNA is comprised of four nucleotides. Unlike DNA, however, uracil is used in the place of thymine (the RNA alphabet is A, C, U, and G). This process of converting DNA to RNA is called *transcription* and the specific type of RNA that is produced is called messenger RNA (mRNA)[5].

Once the mRNA has been exported out of the nucleus into the cytoplasm, it makes its way to the ribosomes, the protein factories of the cell. Here, the protein is built as a chain (polymer) of amino acids where the sequence of amino acids is determined by the template provided by the mRNA. Unlike the one-to-one translation of DNA to RNA (except for the T that becomes a U), the nucleotides of the RNA are processed

[3] The data that we use for this work is all human data, but it could just as easily be applied to any other organism.

[4] Although colloquially one says that genes are what become proteins, it is actually the open reading frame (ORF) within the gene that is transcribed to RNA. As a gene is any heritable piece of DNA it also includes other information, such as promoter regions, that are not directly used in the creation of a protein. Thus, the mRNA that is produced starts from the 5' (read five-prime) region of the ORF that begins with a start codon, and goes until a stop codon is reached in the 3' area. Bits of DNA before the start codon are considered "upstream" of the ORF and are known to be located in the 5' untranslated region (UTR). Similarly, DNA past the stop codon are in the downstream 3' UTR. It is well known that there are many proteins (known as transcription factors) that bind to specific promotor regions in the UTR and activate or deactivate the transcription of the downstream ORF.

[5] Other types of RNA include transfer RNA (tRNA) that bring amino acids to the site of protein synthesis, and ribosomal RNA (rRNA) that is the catalytic component of ribosomes.

in groups of three. Although there are 64 possible combinations of of trinucleotides (commonly known as *codons*) there are only 20 common, naturally occurring amino acids. Thus, there are several codons that code for the same nucleotide[6]. Also, a few of these codons do not represent amino acids, but rather the start (or initiation) and stop (or termination) codons that aptly describe the location to start and stop converting the mRNA into the protein. This entire process of using mRNA as a blueprint for generating a new protein molecule is called *translation*.

Both programmed events within the cell and external events can cause the initiation of transcription and translation. For example, the genetic machinery for circadian rhythm includes transcriptional events that happen approximately every 24 hours without any external stimuli. The model of rhythm generation in *Drosophila* is detailed in the work of Wilsbacher and Takahashi [139]. Alternatively, pathological events within the cell can start transcriptional activity. For instance, self-destruction (apoptosis) can be triggered by self-repair or damage-detection programs internal to the cell when something "breaks" the DNA within the nucleus. On the other hand, the external piezoelectric forces[7] generated in the bones caused by walking can gradually cause bone remodeling by stimulating transcriptional activity of certain bone cells[8]. An "in-between" example is where hormones secreted from distant organs bind to the receptors on the cell, triggering the transcriptional process.

1.1.3 Gene expression experiments

The term *gene expression experiment* (also known as a *microarray experiment*) has been previously used but never clearly defined. In essence, a microarray experiment is a snapshot view that simultaneously measures the *expression levels* of thousands of genes in a sample. The higher the expression level, the more "turned on" the gene, and the lower the expression level, the more "turned off." Although they are called "gene" expression experiments, they actually measure the quantity of mRNA that is

[6] A biological instance of the famed "pigeonhole principle."

[7] Piezoelectricity is the charge that builds up in bone and DNA (and other solid materials) caused by the application of mechanical pressure or stress.

[8] Osteoblastic and osteoclastic cells, to be exact.

present (expressed) in the sample. The assumption is that if more mRNA is present, more proteins corresponding to that mRNA will be generated in the cell. In this manner, we can compare the quantity of mRNA corresponding to thousands of genes across different phenotypes. By analyzing what genes are “turned on” and “turned off” (i.e. which genes are being transcribed and translated into proteins) in different phenotypic conditions, we can hope to identify what causes brain tissue to be brain tissue and not skin tissue. It is important to note that microarray technology is not special because it can uniquely measure gene expression, but rather because it can do it in a high-throughput manner. Instead of measuring the expression of one gene at a time, microarrays allow researchers to analyze the expression of thousands of genes simultaneously.

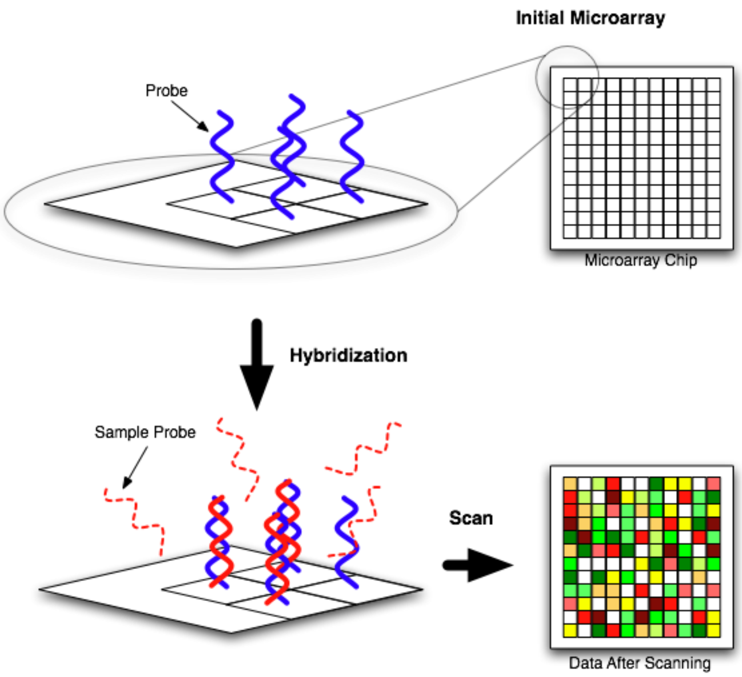


Figure 1-2: The basics of microarray technology. Fluorescence-tagged cDNA *sample probes* for a tissue or system of interest are hybridized to a microarray chip containing cDNA *probes*. After the hybridization process, the chip is scanned using a laser, and the intensity levels at each probe location are measured to determine the expression level for a particular gene.

For most common microarrays, a scientist starts by extracting mRNA from a tissue or system of interest (e.g. brain) and creates a fluorescence-tagged *complimentary DNA* (*cDNA*) copy of this mRNA[9] (Figure 1-2). These *sample probes* are then hybridized to a microarray *chip* (also known as a *platform*)) that have cDNA *probes* attached to the surface in a predetermined grid pattern. The underlying idea behind this process is that a sample probe will only bind to its complementary probe, thus allowing a scientist to measure the quantity of the sample probe present. After leaving the microarray chip submerged in the solution containing the sample probes for several hours, any excess unhybridized sample probes are washed off. The microarray is then scanned using laser light and a digital image scanner records the brightness level at each probe location. The brightness at a particular spot is correlated with the RNA level in the original tissue or system of interest [112] and is thus used as the expression level for that gene. Since the probes that are on the sample chip are the same for the different conditions being tested (i.e. exact duplicates of the chip are used) in a single "dataset" generated by a researcher, the differences in the expression levels for the genes can be attributed to the biological differences and not technical differences (Figure 1-3).

Throughout this work, the following definitions will be used unless explicitly stated otherwise. A microarray *dataset* (*series*) will be a set of microarray *experiments* (*samples*) that were conducted by a specific lab for a specific purpose. For example, if a group of scientists were studying lung cancer and performed ten microarray experiments, five disease state experiments and five control experiments, then the set of these ten experiments is a dataset. Each experiment will also have associated with it a *sample chip* (*platform* or *array*). The platform is the actual chip that the microarray experiment was conducted on, for example the Affymetrix HGU-133A chip. Figure 1-3 shows a pictorial representation.

There are multiple different forms of microarray technologies, the two major ones being *spotted cDNA arrays* and *oligonucliotide arrays.* While both of them measure gene intensity levels, the approach of how they are created and the way in which the

[9] Recall that adenine always pairs with thymine, and guanine always pairs with cytosine. Because this always is true, we can create the complementary DNA (i.e. if it was an A it becomes a T, if it was a T it becomes an A).

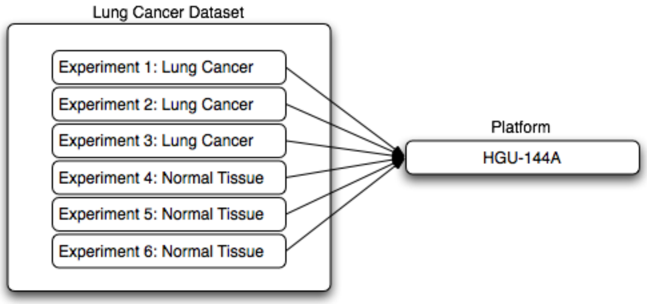

Lung Cancer Dataset
Experiment 1: Lung Cancer
Experiment 2: Lung Cancer
Experiment 3: Lung Cancer
Experiment 4: Normal Tissue
Experiment 5: Normal Tissue
Experiment 6: Normal Tissue
Platform
HGU-144A

Figure 1-3: The relationship of a *dataset*, an *experiment*, and a *platform*. For a single dataset there are multiple different samples produced (in this case 6), all of which are performed on a single chip (platform) type (in this case the Affymetrix HGU-133A).

intensities are measured differ. The former was introduced by Mark Shena et al. [112] in 1995 and is also known as a cDNA microarray. Typically, a robotic spotter picks up cDNA that has been amplified using *Polymerase chain reaction* (PCR) and places it on a glass slide. When performing the experiment, two conditions are actually tested simultaneously, each with a different fluorescent color. The intensity levels are then measured as a ratio of the two conditions. On the other hand, oligonucleotide arrays are generated by a photolithographic masking technique first described by Stephen Fodor et al. [37] and made popular by Affymetrix. Unlike the cDNA arrays, oligonucleotide arrays only measure one condition at a time. One therefore needs to perform multiple experiments in order to compare multiple conditions. A more in-depth explanation about microarray technology and the various types of microarrays can be found in *Microarrays for an Integrative Genomics* [65]. Our work will exclusively deal with oligonucleotide array data performed on the Affymetric HG-U133 Plus 2.0 array.

Difficulties in dealing with microarrays

Although microarray technology enables one to get a genome-wide snapshot of the quantity of RNA levels in a sample, there are many factors that make this data difficult to deal with. Simply put, the data is *noisy*. For example, a replicate experiment that uses exactly the same experimental setup can, and often does, report different expression levels. While this may seem disconcerting, this irreproducibility of data is not limited to microarray technology, but also occurs in most types of experiments in which miniscule quantities are measured with a highly sensitive device. The standard approach to dealing with this problem is to make many replicates and hope that the intensity values of the repeats converge to the true measure (this is one of the reasons why generating a large curated database of expression data is useful). Unfortunately, not only are microarray experiments very expensive, but these sort of repeats tend to eliminate noise caused by measurement errors and not the biological variation inherent in the samples being studied.

Another major obstacle in dealing with microarray technology is the lack of cross platform reproducibility. As detailed in [65], high intensity levels in a cDNA experiment did not correspond well with high levels in oligonucleotide experiments. In light of these findings, the current work only uses single channel data. Furthermore Hwang et al. [57] performed a study where they compared two human muscle biopsy datasets that used two generations of the Affymetrix arrays (HG-U95Av2 and HG-U113A) and showed that they obtained differences in both cluster analysis and the differentially expressed genes. While this is an unfortunate conclusion, this sort of noise is inevitable and cannot be countered. For this reason, we only use gene expression data from a single gene expression platform (Affymetrix HG-U133 Plus 2.0).

Chapter 2

Concordia: The system and its application to GEO

The widespread adoption of electronic storage media throughout the medical and biomedical research communities presents significant new challenges and opportunities. The American Recovery and Reinvestment Act of 2009 will invest $19 billion in a program to promote the adoption of information technology throughout the American health care infrastructure in the coming years. In particular, the Act emphasizes widespread implementation of electronic health record (EHR) systems. By recent estimates, only 17% of doctors and 10% of hospitals are currently utilizing such systems [16]. The financial incentive schedule included in the program, valued at approximately $17 billion, is intended to motivate doctors and hospitals to adopt technologies that interoperate with other parts of the healthcare system by 2015, or face financial penalty in subsequent years [16]. The volume of data generated by this mandate over the coming years will be tremendous.

In addition to the imminent proliferation of electronic medical records, a variety of high-throughput biomedical assays have been refined over the past decade, and more continue to be developed today. It is expected that the data derived from these assays will eventually be brought to bear on clinical diagnostics as well as therapeutic drug design. The volume of data available from some of these sources (e.g., NCBIs Gene Expression Omnibus repository [31, 13], the European Bioinformatics Institute's ArrayExpress [97]) has already outstripped our ability to perform large-scale, automated discovery of relevant patterns among records with shared phenotype. Moreover, at present, there exist no systems capable of associating these assay records in a standardized and meaningful way with relevant EHRs or other clinical narrative. Such cross-pollination would enable sophisticated quantitative clinical diagnostic systems, as well as accelerate the pace of therapeutic innovation.

In addition, there are no open, scalable, standardized systems for cataloging and searching large volumes of medical data that leverages existing expert knowledge. Many institutions have developed proprietary in-house solutions that tend to be ad hoc, lack portability between problem domains (e.g., systems designed for retrieving medical records cannot be easily adapted to the task of retrieving medical literature) and require a major technical undertaking. The applications that consume such services must interact with several different systems that cannot interoperate with one another in any natural, meaningful way.

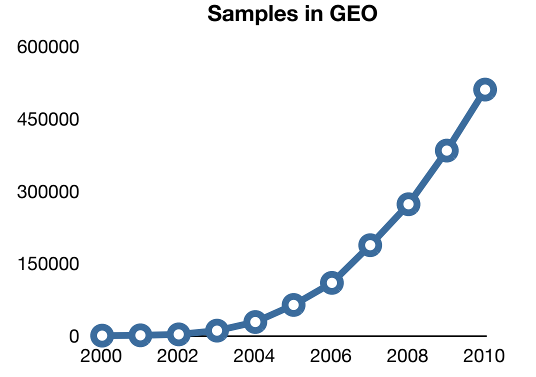


Figure 2-1: The number of gene expression samples has been growing at a dramatic rate since the inception of NCBI's Gene Expression Omnibus 10 years ago.

To this end, we have developed a scalable, standards-based infrastructure for searching multiple disparate databases by mapping their corresponding textual contents onto a structured medical ontology. Although we only present several targeted use cases for this system, the framework can be leveraged against any database where free-text attributes are used to describe the constituent records (for example, medical images might be associated with a short description, or clinical lab results with doctor's notes). Similar to the spirit in which a traditional search engine allows one free text-query to search for multiple content types (web pages, images, maps, etc.) through an open API, the system likewise provides a platform built to open standards, able to support a diverse suite of applications that need to query a variety of clinically relevant content (EHRs, biomedical assays, journal publications) using Web 2.0 methodologies. Such a system would form the cornerstone backend search tool required to build portable applications that leverage the wide variety of data-rich resources that are becoming available, thus addressing one of the core challenges in personalized healthcare practice: identifying clinically distinct subgroups to which a particular patient belongs [64].

We envision the utility of such a query tool to increase over time as the volume of biological assay data and "traditional" medical information converted to electronic form grows. Rather than simply providing persistent storage of such documents (as is the case microarray databases such as GEO and ArrayExpress), a unified, generic search and retrieval tool will give the practitioner of medical, biological, or information sciences the ability to query a wide variety of document sources, and navigate the results in an intuitive and meaningful way. As previous endeavors to mine narrative text associated with biological experiments [19] and medical records [109, 108] have shown, there is a substantial amount of useful information that is readily available. In a clinical setting, applications of data mining projects include identification of populations for recruitment and for sample acquisition, observational studies married to sophisticated time-series analysis for pharmacovigilance, quality improvement and biosurveillance [72]. Furthermore, deeper understanding of the systems biological processes can be gleaned by incorporating the vast amount of publicly available data. For example, Lukk, et al. used gene expression experiments of various phenotypes from ArrayExpress and depicted a map of human gene expression [77].

2.1 The Concordia framework

Concordia is a framework for mapping both queries and source documents onto a structured ontology. This enables users to leverage both the textual information inherent in the document and the ontological associations among the relevant keywords. More concretely, we take the free-text associated with a given record (the description of the contents of a medical image, for example) and use a natural language processing (NLP) program (see 2.1.1) that maps this free-text to the predefined concepts in the ontological vocabulary. For instance, the text associated with an x-ray of broken bone may read, “Compound transverse fracture of tibia caused by skiing accident.” We then insert this record in an ontological index such that a query for all of the concepts that it directly was mapped to (e.g. “tibia” and “compound transverse fracture”) by the NLP program *and* any of the ancestor concepts (e.g. “leg” or “fracture”) would return the record. Queries to this system can either be performed using one or more of the concepts in the ontological vocabulary or via free-text that is then converted to a set of keywords automatically. When the query is in the form of free-text, the same NLP program used to index the documents is used to obtain the concepts for the provided input. Using this framework, therefore, it is possible to perform arbitrarily specific queries for uses such as

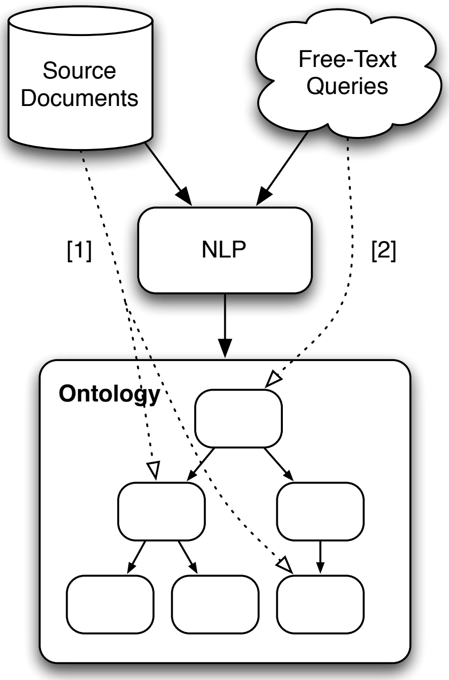


Figure 2-2: Both the text from the source documents [1] and the free-text queries [2] get mapped to UMLS concepts. Querying for the parent concept [2] will return all documents relating to child concepts as they relate to the more specific concepts.

data mining or patient recruitment for a particular study. For a further example that depicts the mapping of a “Lung adenocarcinoma” gene expression sample into a structured medical ontology see Figure 2-3.

In addition to simple queries based on single concepts, the system can efficiently aggregate documents that match arbitrarily complex logical combinations of concepts. This has been implemented as a standard stack-based algorithm [91] for evaluating infix set logic expressions. Here, the operands will be set operators (union, intersection, difference) and the arguments will be UMLS concepts. Conceptually, the algorithm works by replacing the stack entry for each UMLS concept in the expression with the set of database records that reference it, then proceeding with the logical evaluation as usual. This will enable the user to perform free-text queries such as “anemia and cancer” or “lung cancer and metastasis but not smoking” against the library of documents.

2.1.1 Why use an ontology? What ontology should we use?

With the growing argument for letting the data drive the associations between related concepts [51], why are we relying on a manually curated ontology to drive the associations between concepts? First, and foremost, unlike traditional text-based domains such as web-search or document retrieval, the aim of the Concordia framework is not only to query for documents related to concepts, but also to enable the integration of various sources of possibly non-textual data. As others have previously noted, the conceptual representation of data using an ontology allows such disparate databases to be linked in a transparent way to facilitate data analysis [136]. Furthermore, there are two major challenges that arise when searching free-text medical literature as it appears in electronic medical records, medical reference volumes or other relevant documents: resolving synonyms and identifying conceptual relationships between medical terms.

Multiple synonymous phrases are often used to describe one common medical or biological concept. For example, the terms “malignant neoplasm of the lung” and “lung cancer” both describe the same medical concept, but there is no agreement

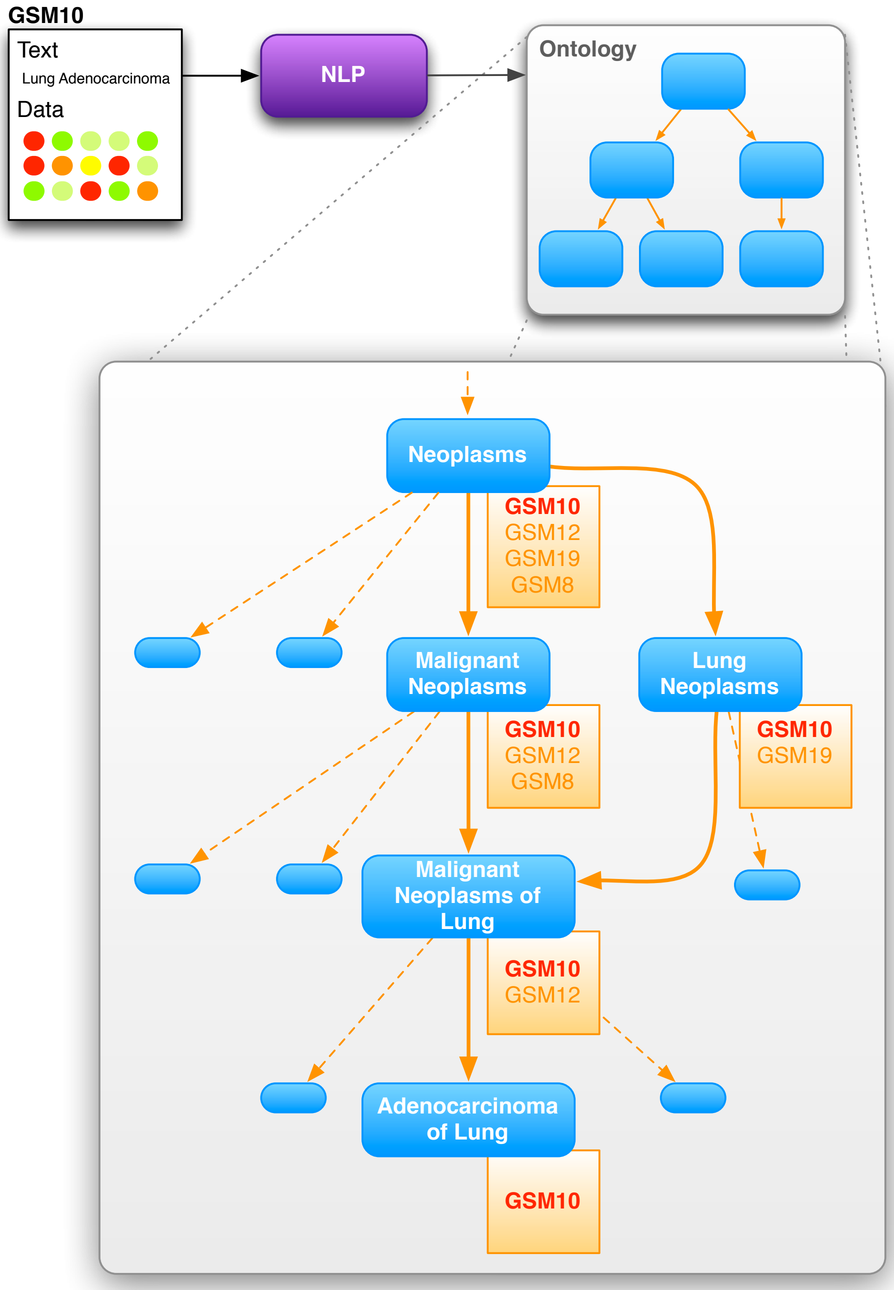


Figure 2-3: The free-text associated with a record is analyzed using a natural language processing program that maps the free-text to the predefined concepts in the ontological vocabulary. Using this model, we can combine existing expert knowledge (in the form of the associations in the ontology) and the information inherent in the text of the records. In this example, therefore, we can associate the data of GSM10 with the concept "Adenocarcinoma of the lung," and all of its ancestors in the hierarchy.

on which term should be used to describe the one underlying concept, a malignant cancerous growth appearing in the lung. To see where this becomes a challenge, consider searching a database for the phrase “lung cancer” where all of the constituent documents refer to “malignant neoplasm of the lung.” Searching the database by simple string matching will fail to find the documents related to the query. The use of a controlled vocabulary, however, mitigates this issue as there is one “correct” concept for “lung cancer.”

As for the case of potentially complex associations between various concepts, the relationships between concepts are clearly defined by the ontological structure of the controlled vocabulary. As depicted in Figure 2-3, for example, we see the clear relationship between the concept “Neoplasm” and “Adenocarcinoma of Lung.” While this link may be relatively trivial as both terms reference a word related to “cancer,” the relationship between “Inflammatory disorder” and “Asthma” is more opaque. Furthermore, the expert associations provided by an ontology allow queries to be made for concepts that may not have been directly mentioned in any of the source text of the corresponding data records. Continuing with the previous example, it may be the case that there are only records for “Asthma” and “Arthritis” in the database. Due to the hierarchic relationships in the ontology, however, we can return all records associated with “Asthma” and “Arthritis” when a user queries for “Inflammatory disorder.” Thus, this hierarchical index allows us to efficiently traverse the ontology and retrieve records related to a particular concept and its descendants (or ancestors).

Although it may be possible to generate a *de novo* taxonomization of the medical vocabulary with a large enough corpus of medical data, both of these challenges can be addressed by employing the cumulative expert knowledge that is represented in well-established ontologies of a controlled vocabulary. While countless ontologies exist, and the Concordia framework can employ any one of them, the National Library of Medicines Unified Medical Language System (UMLS) [87] provides the ideal hierarchically structured controlled vocabulary for generating a database that allows users to insert and query documents along the lines of medically relevant concepts. Using the MetaMorphosys tool provided by the National Library of Medicine, we created a custom ontology, known as a Metathesaurus, built from the expert curated thesauri of UMLS, SNOMED and MeSH.

Mapping documents and queries onto UMLS Metathesaurus

In order to be able to use the UMLS medical ontology, the Metathesaurus, we first have to map the free-text associated with each record to the set of standardized concepts. To do this, we employ the the MetaMap [7] tool that matches syntactic noun phrases from an input text to UMLS concepts, effectively "standardizing" the text to a set of unique concepts. The method is comprised of the five following steps:

1. **Parsing**: The text is parsed into noun phrases using the SPECIALIST minimal commitment parser [83].

2. **Variant Generation**: Variants are generated for each phrase using the SPECIALIST lexicon and a database of synonyms.

3. **Candidate Retrieval**: The "candidate set" of all strings in the Metathesaurus that match at least one of the variants is generated.

4. **Candidate Evaluation**: Each of the candidates in the candidate set is evaluated against the input text.

5. **Mapping Construction**: Candidates from disjoint parts of each input phrase are combined and are then scored. The combined candidate mappings with the highest scores correspond to MetaMap's best interpretation of the input text.

In our setting, the application of MetaMap to the textual portions of data records allows us to resolve the problems of synonyms. One of the major benefits of this approach is that when we later query the database, we can apply the same standardization to the input query as was used to transform the original source text. In this manner, we can search for database entities matching the query in the structured space of standardized UMLS concepts rather than free-text. In addition, when a practitioner later wishes to perform large-scale data mining on such a database, we can treat the UMLS concepts associated with the database entities as a discrete labeling thereof, without applying ad-hoc text searches to identify groups of related records.

MetaMap, however, only provides the direct mappings from the free-text to the exact UMLS concepts that are referenced in that text. To leverage the full potential of the UMLS ontology, we map each of the directly hit concepts (the concepts that MMTx actually labeled the free-text with) up the hierarchy in order to provide the aforementioned functionality of returning records referencing "Asthma" and "Arthritis" when a user queries for "Inflammatory disorder." The downside of performing this mapping is that nodes high up in the hierarchy can become severely bloated as they contain record IDs for all records that its descendant nodes contain. However, empirical testing showed that the dramatic speed increase obtained from not having to recursively traverse descendants of a node to obtain all record IDs made this a worthwhile tradeoff.

2.1.2 Software infrastructure

As depicted in Figure 2-4, the Concordia framework acts as a piece of middle-ware between user interfaces and the underlying data repositories. All communications to, from, and within the framework are via standards based protocols. Open to the public are a set of Simple Object Access Protocol (SOAP) methods that allow a user to query for information such as all record IDs in the database, the set of record IDs corresponding to a given concept, the set of record IDs corresponding to an arbitrary logical combination of concepts, the set of ancestor (or descendant) concepts for a given concept, and so forth. For a detailed user-interface example, see Section 2.2. The current implementation has this SOAP service implemented in Microsoft's C# and is running on a Windows 2000 Server[1].

[1] This server has been virtualized and currently is merely a virtual Windows 2000 Server running on the same hardware as the remaining parts of the system.

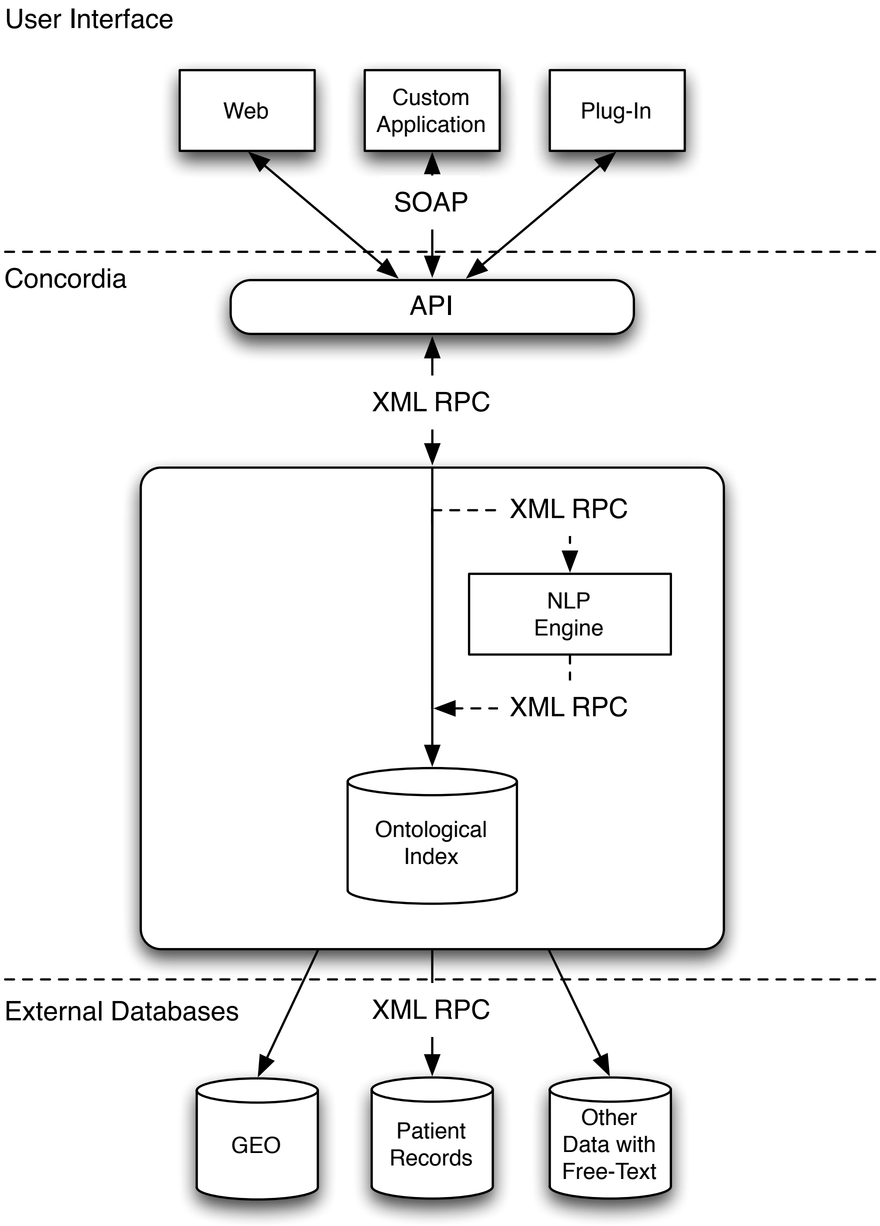


Figure 2-4: The Concordia framework acts as a piece of standards based middleware between user interfaces and traditional data repositories to provide the functionality of querying the data along the lines of concepts (and their relationships) as defined by some arbitrary ontology. To allow for maximum portability and scalability, Interactions from the user interface(s) are sent to the framework via SOAP which then interacts with Concordia over XML-RPC. Once the record indicies have been identified in the ontology, XML-RPC requests are sent to the underlying databases that contain the source documents.

The SOAP interface interacts with the Concordia framework via XML Remote Procedure Calls (RPC). Within the framework itself, we also employ XML RPCs for the communication between NLP engine and the ontological index. If the user wishes to obtain the actual data records, the system will then communicate with the underlying source database(s) to obtain the records. Although the system allows for making queries to the underlying source databases (which may be located on different servers of different organizations) via XML RPCs, it is also capable of directly communicating to underlying databases without the use of XML RPCs. If only the record IDs are requested, they are simply returned without interacting with any (possibly) outside database. These results, regardless of whether they are just the IDs or the full records, will be passed back to the user via the SOAP interface.

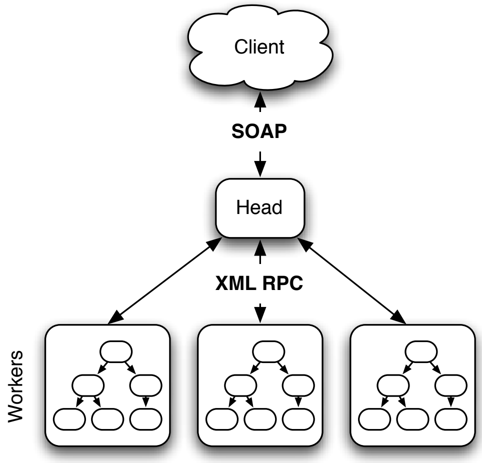


Figure 2-5: Scalability of the Concordia system architecture. Due to the use of XML-RPC calls between all parts of the system, the system can be extended to include multiple worker nodes that fulfill the request of a head node.

The persistent hierarchical database in the Concordia framework is written in Java and utilizes the Oracles's BerkeleyDB JE package. Although there is a longstanding debate [89, 79, 59] as to whether hierarchic database models (e.g. the IBM Information Management System, the Microsoft Windows Registry, and XML) offer better performance than relational databases (e.g. MySQL, Microsoft SQL Server, Postgres, etc.) we find that the ability to efficiently store and retrieve large blocks of data outweigh the benefits of the flexibility provided by a traditional relational database. Furthermore, the in-core nature of the BerkeleyDB allows us to easily serialize the data structures manipulated by our search algorithms without the communication overhead incurred when interacting with an out-of-core database service.

The use of XML RPC based communication between the various parts of the framework allows for a scalable, federated system. Similar in spirit to Googles MapReduce methodology [24], queries can be processed by a head node which in turn requests that multiple worker nodes perform the database search in parallel (see Figure 2-5). Each of these worker nodes will be capable of searching a separate portion of the database. Results can then be returned to the head node, aggregated, and returned to the client. In addition, this infrastructure enables us to scale to meet future needs by simply adding additional worker nodes. Although the example federated structure in Figure 2-5 only depicts a single layer of worker nodes, it is entirely possible to have worker nodes make XML-RPC requests to other worker nodes that are responsible for different parts of the database. Furthermore, this system can be made fault tolerant in a mission-critical environment by replicating worker nodes or dynamically reassigning the responsibilities of a failed node.

An example browser interface for gene expression data that has been processed using the Concordia framework is detailed in Section 2.2.5.

2.2 Concordification of GEO

2.2.1 GEO in a nutshell

Although there are a large variety of biological and medical data sources that could be indexed using Concordia, we limited the scope of this work to the gene expression samples from the Gene Expression Omnibus (GEO) [13]. GEO is a public database containing gene expression and molecular abundance provided by the National Center for Biotechnology Information (NCBI). GEO data is divided into GEO Data Sets (GDS), GEO Series (GSE), GEO Samples (GSM), and GEO Platforms (GPL) files (Figure 2-6). GDS and GSE files are datasets, GSM files are individual samples, and GPL files are the microarray platforms (arrays) on which the samples were prepared.

The difference between a GDS and GSE file is that a GDS file contains additional meta information that the curators of GEO added to the original GSE file that was uploaded. For example, GDS files contain *subset* information about each experiment such that one can see what condition a given experiment has in the dataset. The dataset with the identifier GDS1, for instance, was an experiment conducted to find genes related to reproductive tissue in *Drosophila melanogaster*. The various subset information provided includes information such as gender of the fly for the given sample and the tissue the sample was created from. Another important difference between GDS and GSE files is that a GDS may only contain experiments that were conducted on a single GPL platform. It is possible for a GSE to contain experiments with multiple platforms because there are instances when an experimenter compared multiple microarray platform technologies or performed a cross-species study. It is important to note that there are many more GSE files in GEO than GDS files, as there are many datasets which have yet to be manually annotated. Due to the large size of the GEO database, we only downloaded the human microarray data performed on the Affymetrix HG-U133 Plus 2.0 array. A complete list of the 192 series and 3030 samples that were downloaded can be found in Appendix A.

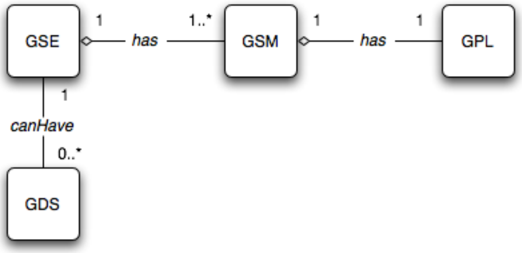


Figure 2-6: The relationship of GEO files as represented by a UML diagram.

2.2.2 Normalizing the gene expression samples

Our database is comprised of 3030 gene expression samples belonging to 192 distinct series performed on the Affymetrix HG-U133 Plus 2.0 arrays that were obtained from GEO (Appendix A). The original CEL files were downloaded from GEO and

MAS 5.0 normalization was performed on each sample before summarizing all probe specific values to gene specific values using a trimmed mean. MAS 5.0 was chosen over other more “aggressive” normalization methods because it can be performed on a per sample basis unlike other methods that require the entire dataset (or in our case entire database) to be used for normalization.

2.2.3 Concordification of GEO

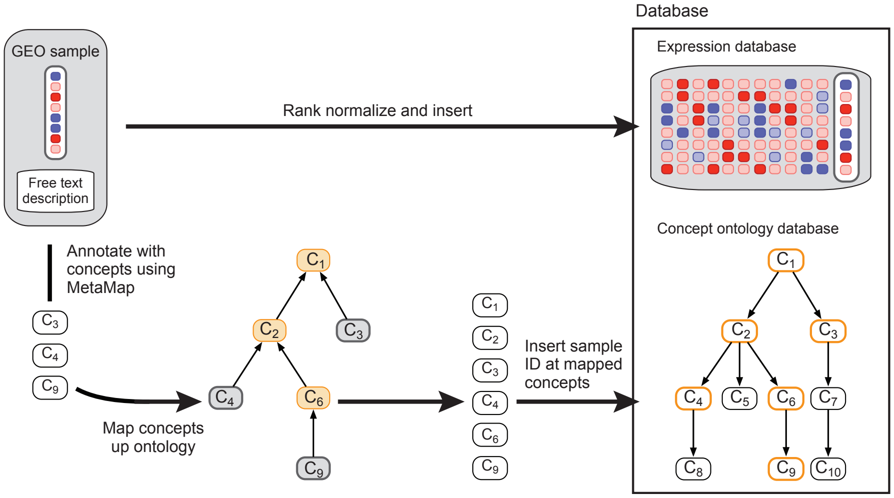


Figure 2-7: The Concordia database for GEO is comprised of a database of gene expression samples mapped to Unified Medical Language System (UMLS) concepts that is used to classify new input microarray samples. The free-text associated with each sample is processed using the National Library of Medicines MetaMap program to map each sample to a set of UMLS concepts. These concepts are then mapped up the ontology so that all ancestor concepts of the ones deemed relevant by MetaMap are also included as correct annotations for each respective sample. The gene expression values for these samples are then normalized and inserted into the Concordia database. Unlike previous endeavors, new data can be added to this system continually, without causing any interruption to the classification engine.

As aforementioned, a major obstacle to recovering signal from biological data (in this case transcriptional signals from microarray array samples) lies in the inconsistent ways in which the samples are described through their associated free-text metadata. Furthermore, there is no easy way to download a large set of disparate experiments and perform large-scale analysis without substantial effort. We follow the lead of

Butte, et al. [19] and extracted the title, description, and source fields from each of the 3030 expression samples and annotated them using the Java implementation of the National Library of Medicines (NLM) MetaMap program, MMTx [7]. A custom Unified Medical Language System [17] (UMLS) thesaurus was generated using NLMs MetaMorphosys program that only contained the concepts from the UMLS, MeSH, and SNOMED ontologies. These automated annotations were then verified by hand (see 2.2.4) such that we were left with 672 distinct UMLS concepts. Since these concepts only represented the most detailed level of annotation, we mapped the concepts back up the ontology such that a sample labeled with a very specific concept also received labels corresponding to all of its ancestor concepts. Due to the domain of the data, we filtered the concepts down to only those that are descendants of either "Disease" or "Anatomy," resulting in a total of 1489 unique concepts. The full list of UMLS concepts that were used can be found in Appendix A.

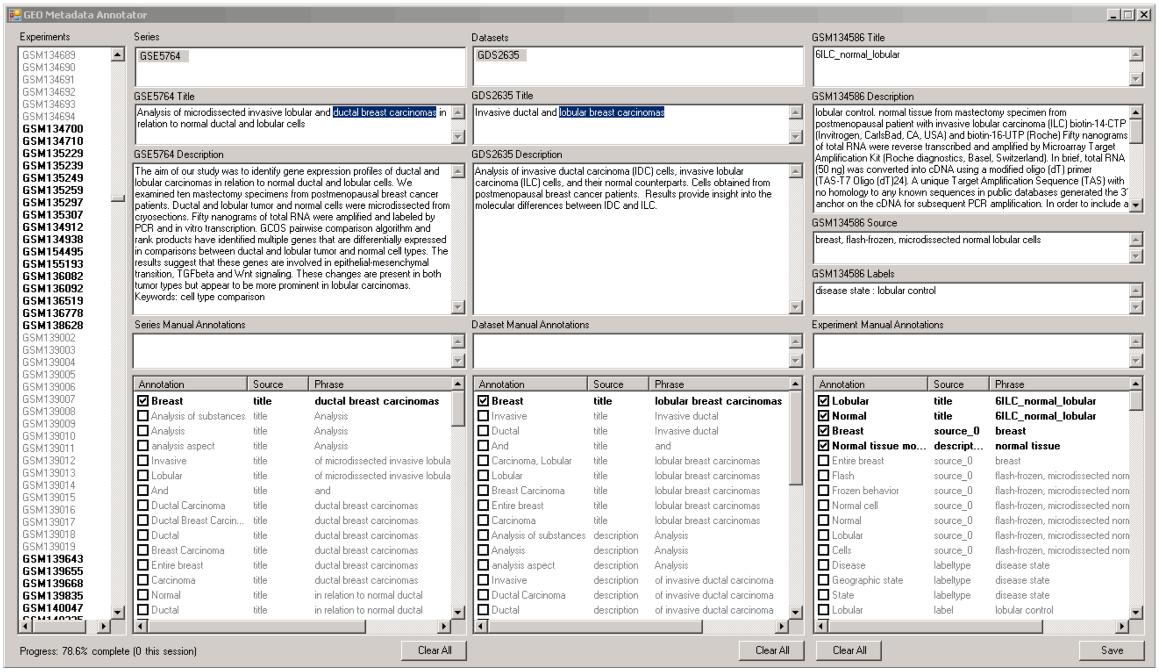
2.2.4 UMLS noise filtering

A major shortcoming of the approach of indexing biological and medical literature with concepts from the Metathesaurus using MetaMap (and many other natural language processing techniques), is the overabundance of false-positive results. This problem has been cited in the literature over the past several years [87]. Butte et al. [19] point out that poor text formatting, poor choice of identifiers, irrelevant text, and spelling errors all contributed to mis-annotations. For example, running MetaMap on the series description of GEO series 2230 (GSE2230), the abbreviation "PD" erroneously maps to the concept "Parkinson's Disease." When we examine the original text we see that the author intended no association with the concept "Parkinson's Disease":

> Analysis of gene expression by Affymetrix microarray in a CD4+ T lymphocyte clone transduced with hTERT-GFP vector after after 44 and 80 population doublings (PDs). The untransduced (32 PDs) and GFP-control vector transduced (47 PDs) T cell clone populations served as

controls.

The MetaMap method simply operates on syntactic fragments and cannot discern the context from which the abbreviation was taken, and hence cannot infer the meaning of the "PD" abbreviation. To overcome such problems of over-sensitivity, we performed manual validation of the annotations automatically generated by MetaMap. We developed a simple C# based application that obtained the raw annotation results from MetaMap, and then allowed us to manually indicate the correct set of concepts for each record (Figure 2-8). In Chapter 5 we delve into more detail about how to efficiently curate a large database using the results from the NLP software along with leveraging the expression signal provided by each sample.



Experiments: GSM134689, GSM134690, GSM134691, GSM134692, GSM134693, GSM134694, **GSM134700**, **GSM134710**, **GSM135229**, **GSM135239**, **GSM135249**, **GSM135259**, **GSM135297**, **GSM135307**, **GSM134912**, **GSM134938**, **GSM154495**, **GSM155193**, **GSM136082**, **GSM136092**, **GSM136519**, **GSM136778**, **GSM138628**, GSM139002, GSM139003, GSM139004, GSM139005, GSM139006, GSM139007, GSM139008, GSM139009, GSM139010, GSM139011, GSM139012, GSM139013, GSM139014, GSM139015, GSM139016, GSM139017, GSM139018, GSM139019, **GSM139643**, **GSM139655**, **GSM139668**, **GSM139835**, **GSM140047**

Series: GSE5764

GSE5764 Title: Analysis of microdissected invasive lobular and ductal breast carcinomas in relation to normal ductal and lobular cells

GSE5764 Description: The aim of our study was to identify gene expression profiles of ductal and lobular carcinomas in relation to normal ductal and lobular cells. We examined ten mastectomy specimens from postmenopausal breast cancer patients. Ductal and lobular tumor and normal cells were microdissected from cryosections. Fifty nanograms of total RNA were amplified and labeled by PCR and in vitro transcription. GCOS pairwise comparison algorithm and rank products have identified multiple genes that are differentially expressed in comparisons between ductal and lobular tumor and normal cell types. The results suggest that these genes are involved in epithelial-mesenchymal transition, TGFbeta and Wnt signaling. These changes are present in both tumor types but appear to be more prominent in lobular carcinomas.
Keywords: cell type comparison

Series Manual Annotations

Annotation	Source	Phrase
☑ **Breast**	**title**	**ductal breast carcinomas**
☐ Analysis of substances	title	Analysis
☐ Analysis	title	Analysis
☐ analysis aspect	title	Analysis
☐ Invasive	title	of microdissected invasive lobula
☐ Lobular	title	of microdissected invasive lobula
☐ And	title	and
☐ Ductal Carcinoma	title	ductal breast carcinomas
☐ Ductal Breast Carcin...	title	ductal breast carcinomas
☐ Ductal	title	ductal breast carcinomas
☐ Breast Carcinoma	title	ductal breast carcinomas
☐ Entire breast	title	ductal breast carcinomas
☐ Carcinoma	title	ductal breast carcinomas
☐ Normal	title	in relation to normal ductal
☐ Ductal	title	in relation to normal ductal

Progress: 78.6% complete (0 this session)

Datasets: GDS2635

GDS2635 Title: Invasive ductal and lobular breast carcinomas

GDS2635 Description: Analysis of invasive ductal carcinoma (IDC) cells, invasive lobular carcinoma (ILC) cells, and their normal counterparts. Cells obtained from postmenopausal breast cancer patients. Results provide insight into the molecular differences between IDC and ILC.

Dataset Manual Annotations

Annotation	Source	Phrase
☑ **Breast**	**title**	**lobular breast carcinomas**
☐ Invasive	title	Invasive ductal
☐ Ductal	title	Invasive ductal
☐ And	title	and
☐ Carcinoma, Lobular	title	lobular breast carcinomas
☐ Lobular	title	lobular breast carcinomas
☐ Breast Carcinoma	title	lobular breast carcinomas
☐ Entire breast	title	lobular breast carcinomas
☐ Carcinoma	title	lobular breast carcinomas
☐ Analysis of substances	description	Analysis
☐ Analysis	description	Analysis
☐ analysis aspect	description	Analysis
☐ Invasive	description	of invasive ductal carcinoma
☐ Ductal Carcinoma	description	of invasive ductal carcinoma
☐ Ductal	description	of invasive ductal carcinoma

GSM134586 Title: 6ILC_normal_lobular

GSM134586 Description: lobular control. normal tissue from mastectomy specimen from postmenopausal patient with invasive lobular carcinoma (ILC) biotin-14-CTP (Invitrogen, CarlsBad, CA, USA) and biotin-16-UTP (Roche) Fifty nanograms of total RNA were reverse transcribed and amplified by Microarray Target Amplification Kit (Roche diagnostics, Basel, Switzerland). In brief, total RNA (50 ng) was converted into cDNA using a modified oligo (dT) primer (TAS-T7 Oligo (dT)24). A unique Target Amplification Sequence (TAS) with no homology to any known sequences in public databases generated the 3' anchor on the cDNA for subsequent PCR amplification. In order to include a

GSM134586 Source: breast, flash-frozen, microdissected normal lobular cells

GSM134586 Labels: disease state : lobular control

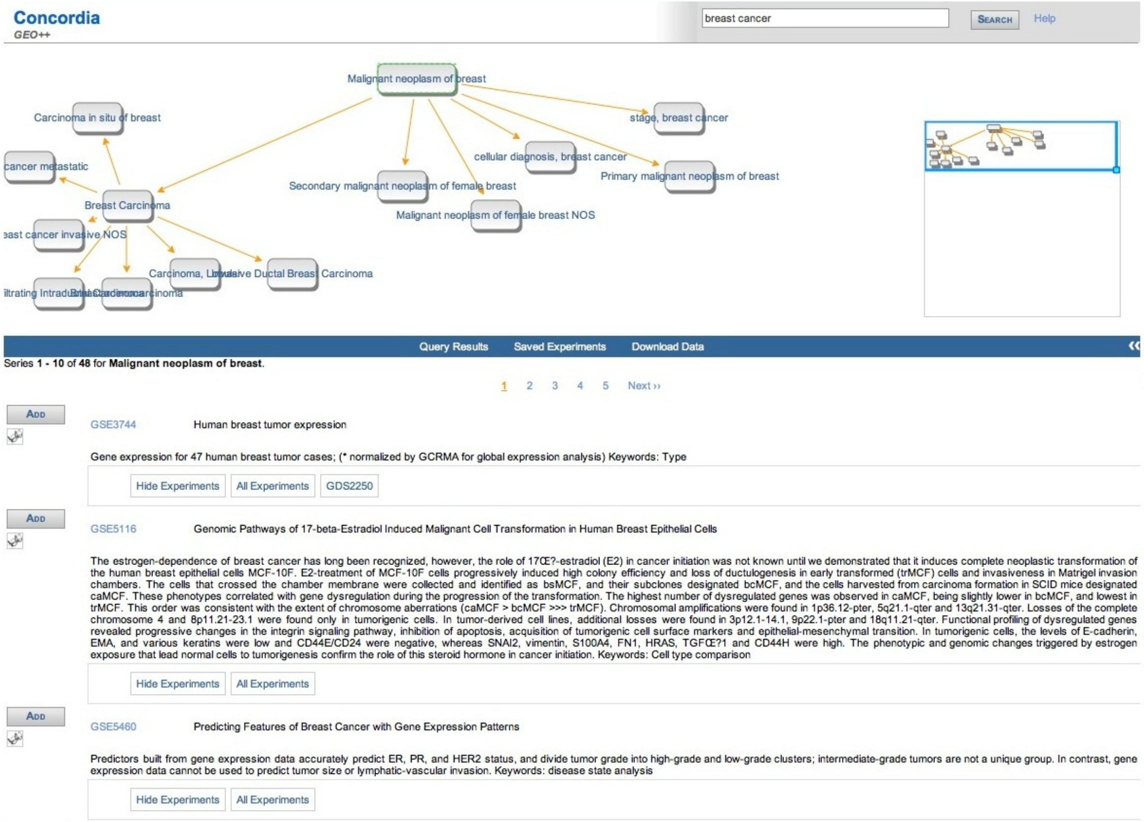
Experiment Manual Annotations

Annotation	Source	Phrase
☑ **Lobular**	**title**	**6ILC_normal_lobular**
☑ **Normal**	**title**	**6ILC_normal_lobular**
☑ **Breast**	**source_0**	**breast**
☑ **Normal tissue mo...**	**descript...**	**normal tissue**
☐ Entire breast	source_0	breast
☐ Flash	source_0	flash-frozen, microdissected norn
☐ Frozen behavior	source_0	flash-frozen, microdissected norn
☐ Normal cell	source_0	flash-frozen, microdissected norn
☐ Normal	source_0	flash-frozen, microdissected norn
☐ Lobular	source_0	flash-frozen, microdissected norn
☐ Cells	source_0	flash-frozen, microdissected norn
☐ Disease	labeltype	disease state
☐ Geographic state	labeltype	disease state
☐ State	labeltype	disease state
☐ Lobular	label	lobular control

Figure 2-8: A screen shot of the application that was used to perform manual curation of UMLS concepts. Through this application one can select the concept(s) that are relevant to a given GEO series, dataset, and sample. It is also possible to add concepts manually that were missed by the NLP program.

2.2.5 Ontology based browsing of GEO

We also developed a sample front-end to the Concordia framework in an AJAX based web application that allows a user to browse the UMLS hierarchy and view the gene expression samples that have been mapped to the concepts (Figure 2-9). The top panel allows the user to navigate through the library of experiments based on the hierarchical organization of UMLS concepts. The lower panel allows the user to view and interact with the data for experiments that were labeled at or below any particular location in the concept hierarchy. The user can select the experiments of interest and then download a large matrix of the expression intensity values for all of the experiments along with their respective UMLS concepts.


Concordia
GEO++
breast cancer
SEARCH
Help
Malignant neoplasm of breast
Carcinoma in situ of breast
stage, breast cancer
cellular diagnosis, breast cancer
cancer metastatic
Primary malignant neoplasm of breast
Secondary malignant neoplasm of female breast
Breast Carcinoma
Malignant neoplasm of female breast NOS
east cancer invasive NOS
Invasive Ductal Breast Carcinoma
Query Results
Saved Experiments
Download Data
1
2
3
4
5
Next ››
ADD
Hide Experiments
All Experiments
GDS2250

Series **1 - 10** of **48** for **Malignant neoplasm of breast**.

GSE3744 Human breast tumor expression

Gene expression for 47 human breast tumor cases; (* normalized by GCRMA for global expression analysis) Keywords: Type

GSE5116 Genomic Pathways of 17-beta-Estradiol Induced Malignant Cell Transformation in Human Breast Epithelial Cells

The estrogen-dependence of breast cancer has long been recognized, however, the role of 17Œ?-estradiol (E2) in cancer initiation was not known until we demonstrated that it induces complete neoplastic transformation of the human breast epithelial cells MCF-10F. E2-treatment of MCF-10F cells progressively induced high colony efficiency and loss of ductulogenesis in early transformed (trMCF) cells and invasiveness in Matrigel invasion chambers. The cells that crossed the chamber membrane were collected and identified as bsMCF, and their subclones designated bcMCF, and the cells harvested from carcinoma formation in SCID mice designated caMCF. These phenotypes correlated with gene dysregulation during the progression of the transformation. The highest number of dysregulated genes was observed in caMCF, being slightly lower in bcMCF, and lowest in trMCF. This order was consistent with the extent of chromosome aberrations (caMCF > bcMCF >>> trMCF). Chromosomal amplifications were found in 1p36.12-pter, 5q21.1-qter and 13q21.31-qter. Losses of the complete chromosome 4 and 8p11.21-23.1 were found only in tumorigenic cells. In tumor-derived cell lines, additional losses were found in 3p12.1-14.1, 9p22.1-pter and 18q11.21-qter. Functional profiling of dysregulated genes revealed progressive changes in the integrin signaling pathway, inhibition of apoptosis, acquisition of tumorigenic cell surface markers and epithelial-mesenchymal transition. In tumorigenic cells, the levels of E-cadherin, EMA, and various keratins were low and CD44E/CD24 were negative, whereas SNAI2, vimentin, S100A4, FN1, HRAS, TGFŒ?1 and CD44H were high. The phenotypic and genomic changes triggered by estrogen exposure that lead normal cells to tumorigenesis confirm the role of this steroid hormone in cancer initiation. Keywords: Cell type comparison

GSE5460 Predicting Features of Breast Cancer with Gene Expression Patterns

Predictors built from gene expression data accurately predict ER, PR, and HER2 status, and divide tumor grade into high-grade and low-grade clusters; intermediate-grade tumors are not a unique group. In contrast, gene expression data cannot be used to predict tumor size or lymphatic-vascular invasion. Keywords: disease state analysis

Figure 2-9: A screen shot of a web application built in front of a the Concordiafied gene expression data from GEO. The top panel allows the user to navigate through the library of experiments based on the hierarchical organization of UMLS concepts. The lower panel allows the user to view and interact with the data for experiments that were labeled at or below any particular location in the concept hierarchy.

Having data available in this format, it becomes easy for a researcher to quickly download various types of phenotypic data and perform analyses. Examples of the types of analyses that can be performed with a curated database of gene expression data will be the topic of the remaining chapters.

Chapter 3

Beyond differential expression: Localizing expression samples in a heterogeneous transcriptomic landscape

Although gene expression microarrays have been a standard, widely-utilized biological assay for many years, we still lack a comprehensive understanding of the transcriptional relationships between various tissues and disease states. When microarray technology first became available, the high cost of performing these gene expression experiments was a likely cause for the small number of samples in early microarray studies. However, today, even with the hundreds of thousands of expression array data sets available through public repositories such as NCBIs Gene Expression Omnibus (GEO) [13], the lack of standardized nomenclature and annotation methods has made large-scale, multi-phenotype analyses difficult. Furthermore, it is often challenging to obtain the appropriate number of tissue samples from humans [65], and thus new studies are limited in the number of replicates for a given tissue or in the number of types of tissues. Thus, expression analyses have typically used the decade old approach of comparing expression levels across two states (e.g., case vs. control) or a limited number of phenotype classes [30, 48, 133]. Even recent large-scale gene expression investigations, whether they have attempted to elucidate phenotypic signals [73, 93, 103] or applied those signals for downstream analyses such as drug repurposing [68, 122], involved comparisons between two states or classes.

Comparative analyses, where transcriptional differences are directly measured between two phenotypes, inherently impose subjective decisions about what constitutes an appropriate control population. Importantly, such analyses are fundamentally limited in scope and cannot differentiate between biological processes that are unique to a particular phenotype or part of a larger process that is common to multiple phenotypes (e.g. a generic "cancer pathway"). Moreover, the results of such comparative analyses can be limited in generalizability as they make assumptions about the phenotypes being compared [102]. Alternatively, in a data-rich environment, we can take a holistic view of gene expression analyses.

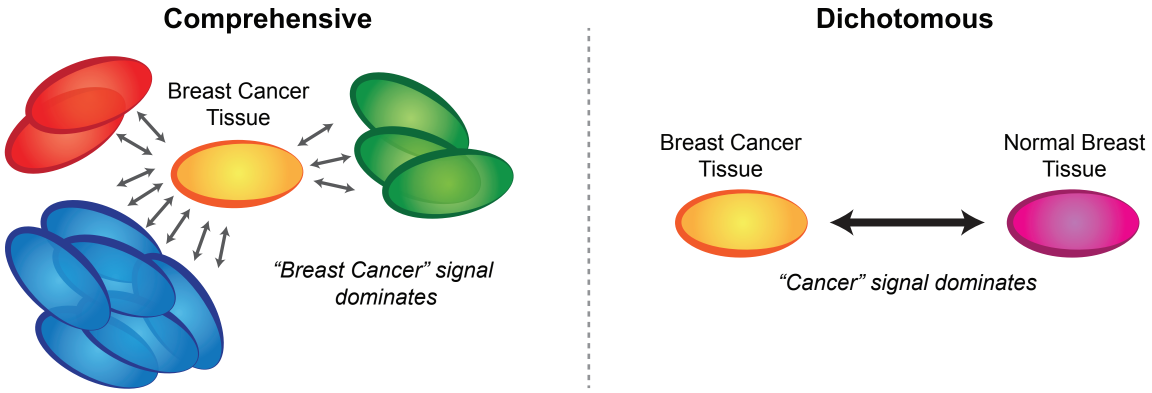


Figure 3-1: A comprehensive perspective on expression analysis enables the elucidation of biological signals that are thematically coherent but provide an alternative view to traditional dichotomous approaches. For example, the gene-signature for "breast cancer" is enriched for breast specific development and carbohydrate and lipid metabolism in our comprehensive approach, as opposed to being dominated by a more general "cancer" signal.

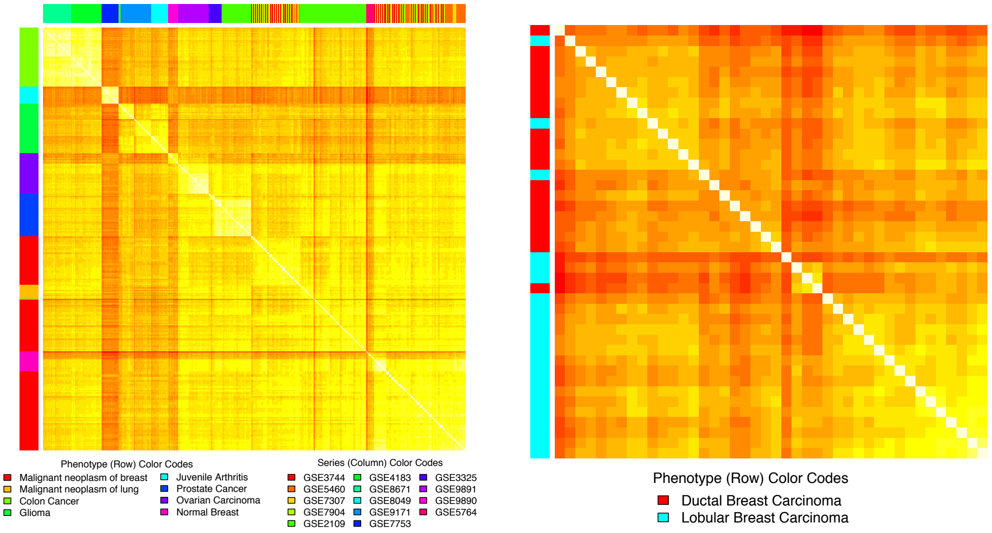
We introduce a scalable and robust statistical approach that leverages the full expression space of a large diverse set of tissue and disease phenotypes to accurately perform and glean biological insights. By viewing a given phenotype in the context of this comprehensive transcriptomic landscape, we circumvent the need for predefined control groups and presupposed relationships between phenotypes (Figure 3-1). We devise, implement and validate the accuracy of an enrichment statistic that provides detailed phenotypic information for new samples when they are mapped onto and compared with the transcriptomic landscape (`http://concordia.csail.mit.edu`).

3.1 Sample correlation as a distance metric

As a practical example, supervised learning (classification) on gene expression data has long held the promise of improved clinical diagnostics. Indeed, many analyses over the last decade have noted that a variety of human diseases are associated with aberrant transcriptional activity ([19, 48, 55, 65, 135, 141] to name but a few). In this setting, a large, diverse “training” database of microarray data would be assembled where each sample is labeled according to phenotype (e.g., “squamous cell lung cancer”, “lobular breast carcinoma”, “type II diabetes”). New unlabeled samples (e.g., hybridized from the peripheral blood of a patient with a tumor of unknown primary origin) could then be compared to the database of training data, allowing the system to generate a “best guess” about the phenotype of the new sample. In our example, such a system would provide an additional and significant piece of evidence for aphysician determining a course of treatment.

One of the major challenges associated with building such a system revolves around generating coherent labeling of the training data against which the unlabeled samples are compared. Using the Unified Medical Language System (UMLS) [17] labels produced by annotating the free-text descriptions associated with gene expression samples from the Gene Expression Omnibus (GEO) [13] as explained in Section 2.2, we see that the Concordia system is capable of recovering these coherent labelings for a large database of gene expression studies. Furthermore, we see that there is strong agreement between these labels and the transcriptional signal encoded in the array data.

A subset of the samples available from GEO was indexed using our prototype system. Figure 3-2(a) shows a clustering of experiments from 14 distinct GEO series



(a) Correlation clustering of 8 phenotypes

(b) Correlation clustering of 2 types of breast carcinoma

Figure 3-2: (a) A clustering of gene expression experiments extracted from the database. Eight different disease states broadly cluster together, even across data series. (b) Here, the expression data for two subtypes of breast cancer cluster according to the breakdown of their UMLS concept labelings, as retrieved by the Concordia representation of GEO.

based on a nonparametric Spearman correlation statistic that measures similarity between expression profiles for each experiment.[1] The experiments were extracted from this database by searching for 8 different phenotypes (glioma, breast cancer, lung cancer, arthritis, etc.). The column of colors down the left-hand side of the plot indicates the UMLS concept associated with each experiment; the row of colors across the top of the plot indicates the data series (logical grouping of experiments submitted to GEO as a batch) from which the experiments were derived. Of particular interest, experiments that were returned by querying our prototype system for each concept clearly clustered together, and this clustering is coherent between data series. Figure 3-2(b) shows the clustering of the lobular and ductal breast cancer experiments from GEO Series GSE2109. Here we see that with only a few exceptions, the two

[1] The Spearman correlation is equivalent to the Pearson correlation between the rankings of the data. In other words, the raw gene expression intensities X_i and Y_i of the two expression samples X and Y are ranked to obtain x_i and y_i. The correlation, ρ, is then computed as $\frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{(x_i-\bar{x})^2(y_i-\bar{y})^2}}$ where $\bar{x}$ and $\bar{y}$ are the means of x and y.

subtypes of breast cancer are grouped according to their respective type. Thus, not only can we cluster experiments across significantly different phenotypes, but we can also differentiate different subtypes of cancers.

This provides evidence that there are strong transcriptional signals that describe the phenotype of the samples in this database and that, when properly processed with our proposed infrastructure, those signals are immediately apparent. Concordia, thus provides the missing link between large, data-rich but loosely-curated resources (such as GEO) and the enormous potential that they hold.

3.2 Making sense of the transcriptomic landscape

As a first step towards a holistic approach to gene expression analysis, we must make sense of the substructure of the global transcriptomic landscape. As mentioned in Chapter 2, we constructed a curated gene expression database of 3030 diverse samples (from 192 distinct series) obtained from NCBIs Gene Expression Omnibus [13] (GEO). These samples were annotated with their phenotypes (tissue of origin, disease state, etc.) using the anatomical and disease concepts in a custom subset of the Unified Medical Language System [17] (UMLS) concept ontology via both natural language processing (NLP) and manual validation (see Section 2.2 for details).

3.2.1 The transcriptomic landscape

While visualizing the full transcriptomic landscape encompassing all genes is not feasible, the first two principal components (PCs) of the centered and scaled expression level of 20252 genes across the database provide a representation of the phenotypic relationships that captures roughly 20% of the variance in the data. The phenotypic clusters portrayed by shaded convex hulls were created by iteratively using the convex hull function (`chull`) in the R statistical language package. Although others have suggested that the primary factors driving the organization of the global transcriptomic landscape can largely be attributed to hematopoietic and malignant programming [77], we alternatively see the cell and tissue specific signatures of blood,

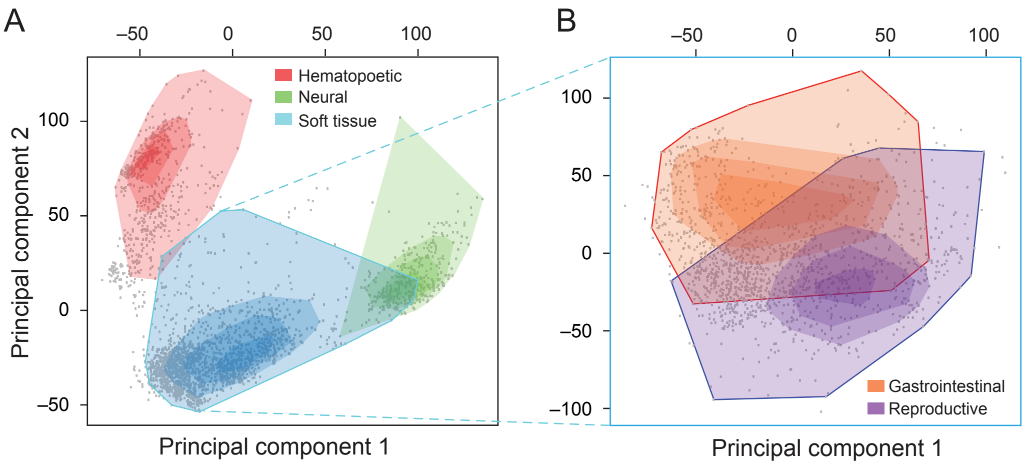


Figure 3-3: The gene expression landscape, as represented by the first two principal components of the expression values of 20252 genes from 3030 microarray samples separates into three distinct clusters: blood, brain, and soft tissue (A). The shading of the regions corresponds to the amount of data located in that particular region of the landscape such that the darker the color, the more data exists at that location. Interestingly, the area where the soft tissue intersects the blood tissue corresponds to bone marrow samples, and where it intersects the brain tissue, mostly corresponds to spinal cord tissue samples. There is a clear separation of reproductive and gastrointestinal tissue samples when the analysis is limited to just the soft tissue cluster (B).

brain, and soft tissue are dominant (Figure 3-3). Indeed, when analyzing the tissue specific characteristics of these clusters, we observe the over-expression of fibrillar and epithelial genes such as COL3A1, COL6A3, KRT19, KRT14, and CADH1 in the soft tissue cluster and neural genes such as GFAP, APLP1, GRIA2, PLP1, and SLC1A2 in the brain cluster. The density estimate plots of the expression intensity values for the top 20 over-expressed tissue specific genes and the GO enrichment analysis of the top 250 tissue specific genes for each cluster further points to over-enrichment for terms related to each of the three tissue types (Appendix B)

These tissue specific genes were selected by performing permutation t-tests comparing, for example, the log-normalized expression values for the blood samples for a given gene to the log-normalized expression values of the samples associated with brain and soft tissue. Each permutation run consisted of computing the t statistic for the actual labeling of the samples and comparing it to the t statistics produced when the labels were randomly permuted 200 times while keeping the sample size distribution constant. To counter the potential influence of sampling bias, this entire procedure was performed 100 times, each time using only a random 75% of the data for each tissue type. Genes that were deemed significant were those that had a false discovery rate corrected p-value of 0.05 or lower in all 100 runs. The genes were then sorted such that a gene that had a larger difference in means between the phenotypes was ordered before those that had a smaller difference. GO enrichment was performed on the top 50, 100, and 250 genes for each tissue type using FuncAssociate 2 [15]. We report only the GO terms that had a resampling-based p-value less than 0.05.

By additionally performing principal component analysis on 1065 soft tissue samples (all non-cancerous samples that are also not blood or brain), it becomes apparent that the notion of phenotypic grouping seen in the full landscape is conserved on multiple levels of phenotypic granularity. Not only are individual tissue samples in confined regions, they are also organized by functionality. Tissues that are sensitive to reproductive hormones (ovary, uterus, myometrium, endometrium, prostate, penis, and breast) group together to form a distinct sub-region in the smooth landscape (Figure 3-3). Juxtaposed to them are primarily gastrointestinal tract samples from tissues such as colon, stomach, intestine, liver, and esophagus.

3.2.2 Tissue similarity network

While the aforementioned transcriptomic landscape provides a view of the various phenotypes by placing them in a two dimensional plane, a tissue "network" can be constructed to further detail the relationship between the various phenotypes. This similarity network was generated by computing correlations of a representative sample of a tissue type to all other representatives of the other tissues. The representative was chosen to be the sample that was closest to the centroid in the set of samples for that phenotype. To contend with sampling bias, the correlations were computed 100 times; the centroid for each phenotype having been chosen from a random 75% subset of the samples for that phenotype. The similarity network was then created based on the tissue-tissue relationships with an average correlation greater than 0.8

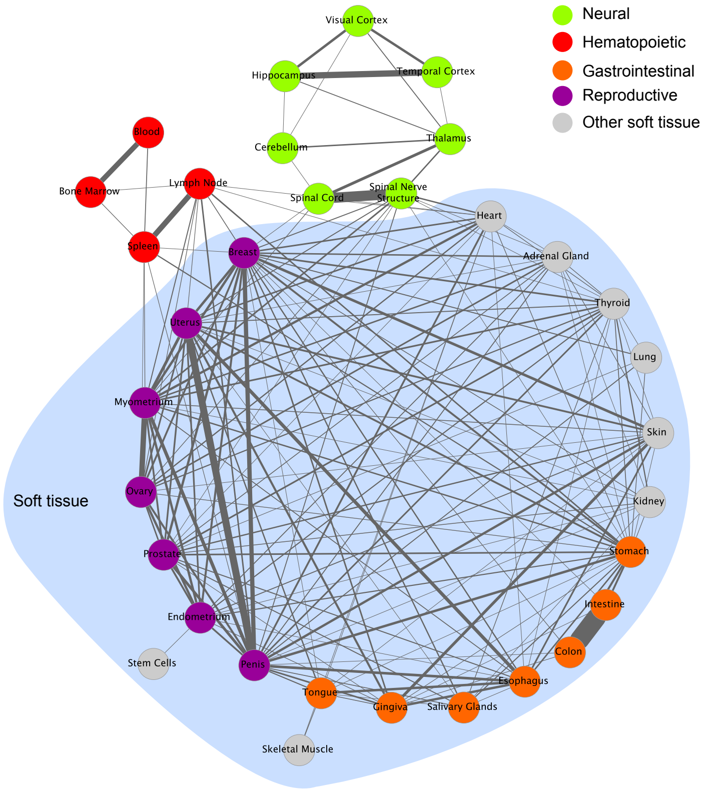


Figure 3-4: Tissue correlation network recapitulates gene expression landscape. A tissue network constructed from the correlations that averaged greater than 0.8 across 100 random sub-samplings runs between the various tissues mirrors the structure of the larger expression continuum while simultaneously showing more fine-grained relationships between various phenotypes. The thickness of the line indicates the strength of the correlation, while the color of the nodes corresponds to the higher-level biological groupings of brain, blood, gastrointestinal, and reproductive. The grey nodes indicate tissues that do not belong to the aforementioned types. Similar to the view provided by the visualization of the transcriptomic landscape, we see the distinct grouping of brain, blood, and soft tissues. Here we also see strong intra-relationships between the gastrointestinal tissues and the reproductive tissues.

across all 100 subsampling runs. The colors of the nodes denote the general tissue class (blood, brain, gastrointestinal, reproductive, and other). The figure itself was rendered using Cytoscape [118].

This tissue similarity network, for example, identifies the functionally related tissues of the myometrium, endometrium and ovary, and also highlights the interconnectedness of blood, the lymph nodes, and spleen. It even depicts the relationship between tongue and other skeletal muscle tissue while still showing the similarity tongue tissue has with the other gastrointestinal phenotypes. In contrast to previous efforts that illustrated this phenomenon using known genetic associations [75], this sort of de novo quantification of tissue specificity is relevant to understanding the biological similarities and differences of various phenotypes, the development of large-scale clinical prognostication engines, the quantification of diagnostic accuracy of putative biomarkers [66], and for developing suitably broad-spectrum or targeted therapeutics.

3.3 Phenotypic concept enrichment

Although correlation analyses and the visual representation of the transcriptomic landscape provide insight into the broad relationships between various phenotypes, our ability to harness these expression signals to map new, previously unseen samples into a database of expression samples is compelling. Beginning with our customized UMLS concept annotation of the 3030 samples, we restricted the set of UMLS concepts to the 1489 anatomy and disease concepts that mapped to at least three expression samples (Figure 2-7). We developed a sample-centric method based on the Kolmogorov-Smirnov statistic to label new samples with UMLS concepts that are over-represented in their local expression neighborhoods (Section 3.3.1). It is noteworthy that no hard boundaries are drawn when a new input sample is labeled, but rather the concepts pertinent to the transcriptomic neighborhood for the input sample are reported. Furthermore, as it is often difficult to define an appropriate comparator (how does one define a truly "normal" tissue sample in a clinical setting [86, 99]?), this approach has the advantage that it does not require case-control type input but, rather, just a single microarray sample. To illustrate its function, we provide a web-based resource (`http://concordia.csail.mit.edu`) that allows users to submit their own microarray samples performed on the popular Affymetrix HG-U133 Plus 2.0 array and obtain their over-enriched tissue and disease concepts (see Section 3.3.7 for details).

3.3.1 Enrichment score calculation

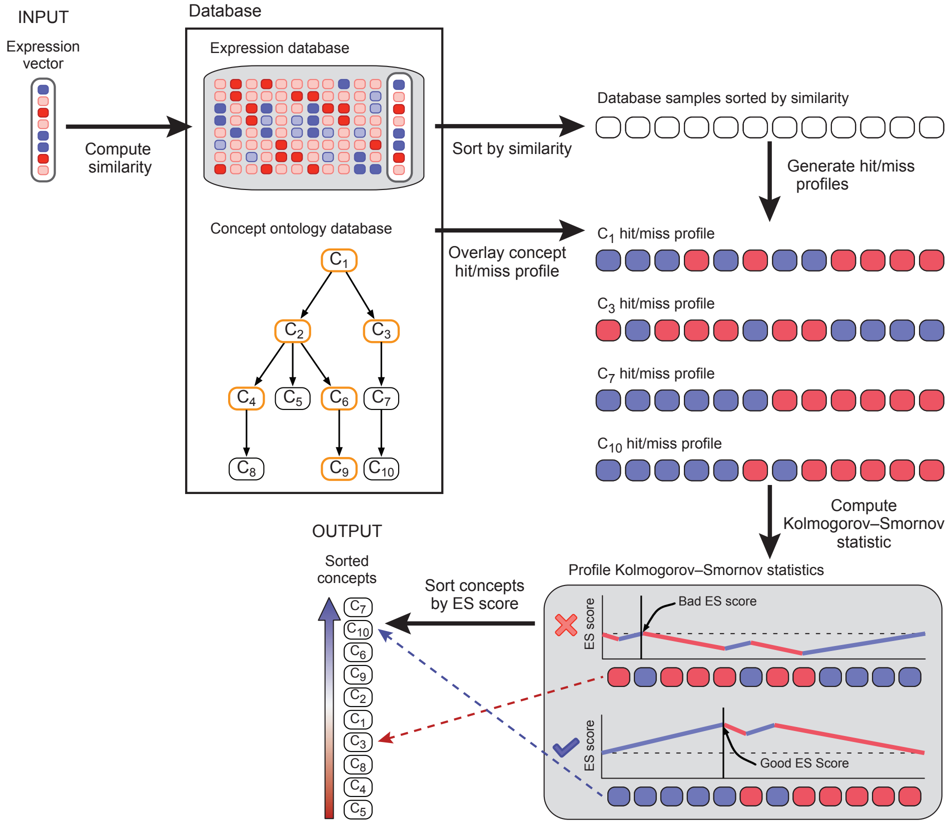


Figure 3-5: A user submits a gene expression profile to the database that then computes the similarity to all other samples in the database. Based on the similarity, an enrichment score is computed for each UMLS concept for which data exists in the database and the concepts are returned to the user in order of statistical significance.

We use the database of gene expression samples to assess over-enrichment for particular disease- and tissue-specific signals. Given a new expression profile, for each concept represented in the database, we calculate a statistic that measures the strength of association between the sample and concept, as implied by its similarity to the labeled database samples.

We measure the similarity of the new expression profile to those contained in the database by computing the Spearman rank correlation, ρ, between the profile and all database samples. For a particular concept, we then calculate an enrichment score that measures the difference between the distributions of correlation coefficients for the database samples that map to the concept versus those that do not map to the concept (Figure 3-5).

Algorithmically, the statistic is calculated as follows: First, the database consisting of n curated expression samples $\{s_1, s_2, s_3, \ldots, s_n\}$ is sorted (in decreasing order) according to each observations Spearman correlation with the new profile. Let s'_1, s'_2, s'_3, …, s'_n represent the samples ordered according to their correlation coefficients $\rho'_{s'_1}$, $\rho'_{s'_2}$, $\rho'_{s'_3}$, …, $\rho'_{s'_n}$. For a given concept c in the set C, the set of all UMLS concepts in our database, let S_c be the set of all database samples associated with the concept. That is, $S_c = \{s_i | s_i \text{ is associated with } c\}$. We define an ordered list of x_i values:

$$x_i = \begin{cases} \frac{\frac{1+\rho'_{s_i}}{2}}{\sum_{s'_j \in S_c} \frac{1+\rho'_{s_j}}{2}} & \text{if sample } s'_i \text{ is associated with concept } c, \\ \frac{-1}{(n-|S_c|)} & \text{if not.} \end{cases} \tag{3.1}$$

Intuitively, when s_i is associated with the concept in question, the x_i value corresponds to the fraction of total correlation between the new sample and all database samples associated with the concept. All of the x_i values for the concept "hits" sum to 1, and all of the x_i values for the concept "misses" sum to -1.

Then we compute a running sum of x_i across all n database samples and take the maximum value achieved by this running sum as our enrichment score (ES) for the concept in question:

$$ES_c = \max_{1 \leq j \leq n} \sum_{1 \leq j \leq n} x_i \tag{3.2}$$

This sum across all n samples is zero. We are interested in concepts where there is strong positive deviation from 0. These are the concepts whose associated samples are more highly correlated with the new profile than those samples that are not associated with the concept.

Figure 3-6 provides a pictorial example of the enrichment score calculation process for a breast cancer sample (GSM175794) for the concepts "breast" and "brain." The x-axis refers to the index in the sorted list of samples (index 0 is the sample that is most correlated to GSM175794) and the y-axis refers to the enrichment score. The black line depicts the running sum of Equation 3.1 while the red line provides an indication of the maximal score obtained (Equation 3.2). Since the score quickly increases in Figure 3-6(a) we can infer that the samples most highly correlated with GSM175794 are other breast samples (that is why they are first in the sorted list). On the other hand, the samples that are associated with the concept "brain" are very far away (in correlation space) from the breast cancer sample. Thus we obtain a concept enrichment score of 0.58 for "breast" while we barely get a positive score for "brain."

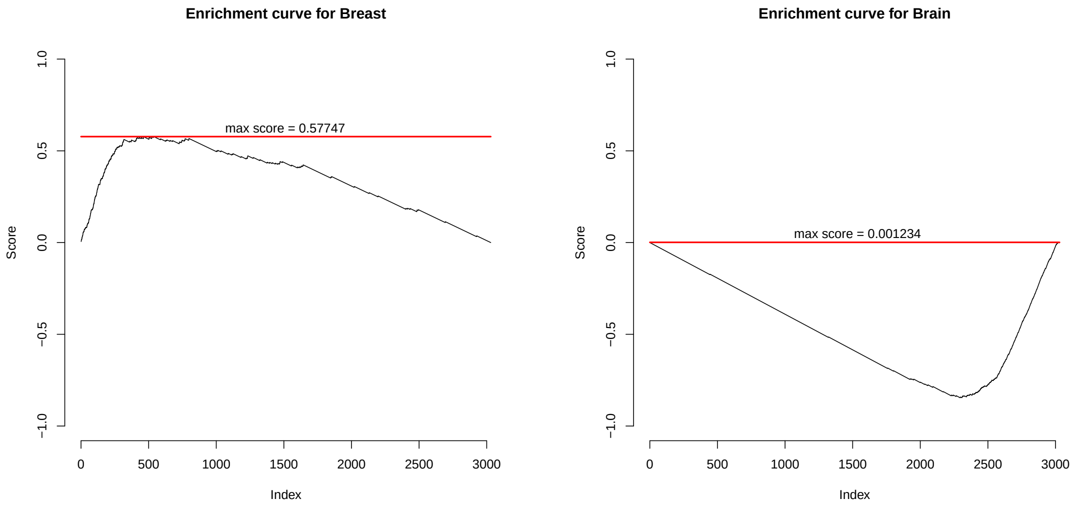


Figure 3-6: Two example enrichment score plots for a breast cancer sample (GSM175794) for the concepts "breast" and "brain." The x-axis refers to the index in the sorted list of samples (index 0 is the sample that is most correlated to GSM175794) and the y-axis refers to the enrichment score. The black line depicts the running sum of Equation 3.1 while the red line provides an indication of the maximal score obtained (Equation 3.2). As is to be expected from a breast cancer sample biopsied from breast tissue, we obtain a concept enrichment score of 0.58 for "breast" while we barely get a positive score for "brain."

3.3.2 Quantifying performance

We performed leave-one-sample-out cross-validation to validate the accuracy of our method for correctly assigning an unknown sample to the correct phenotype (i.e., UMLS concepts for both anatomy and disease). The receiver operating characteristic (ROC) curve was computed for each of the 1489 UMLS concepts, and the standard measure of area under the curve (AUC) that summarizes both the true-positive and false-positive rates was used as a measure of accuracy represent the full spectrum of performance characteristics available under a variety of enrichment score thresholds.

We compute the ROC curves for each concept as follows (see Figure 3-8 for an example ROC curve). For each concept c in the database, we iteratively leave out each sample s, and compute ss enrichment score for c using the remaining samples in the database. The samples are then sorted from highest to lowest by their enrichment score for c. By walking down this list of sorted samples we calculate the running true- and false-positive counts. The true-positive (TP) count is incremented if the i^{th} sample in the list is actually labeled with concept c. If the sample is not actually labeled with concept c, the false-positive (FP) count is incremented. Dividing the TP and FP respectively by the number of known positives and negatives at each position i, we obtain the true-positive (TPR) and false-positive rates (FPR). By plotting the TPR vs. FPR we obtain the ROC curve. The larger the area under the ROC curve (AUC), the greater the gene expression signal for that concept as the samples with the highest enrichment scores for the concept were truly labeled with that concept.

When using this method to label a new sample, we compute its ES (w.r.t. the entire database) for each concept. We then report the systems estimated FPR for each concept at the samples observed concept-specific enrichment score. These FPR values are derived from the running statistics used to generate the ROC plots: simply look up the new samples score position in the list of sorted scores, and report the FPR at that position (or the next-lowest score, i.e., next-worst FPR, if there is not an exact match among the database scores).

We use the area under the curve (AUC) and an empirical false positive rate (FPR) to characterize the systems ability to recover signal rather than random sampling or permutation testing (as performed by another Kolmogorov-Smirnov statistic based method, GSEA [129] for several reasons. If we work with the null hypothesis that the samples ES for a given concept looks like the ES of a random permutation of the database samples (e.g., the ordering prescribed by the correlation scores between this sample and the rest of the database are the result of random shuffling), then we have not accounted for the correlation structure among the database samples themselves. Because the expression values of samples for a given concept (assuming the concept has some signal in gene expression space) will be highly coordinated, they will appear grouped together regardless of the phenotype of the new sample, resulting in a localized “bump” in the running enrichment score. This localized “bump” is often large enough to cause us to reject the null hypothesis, even when the new sample shouldnt be associated with the concept in question.

If instead we attempt to randomize the input and reject the null hypothesis that the new samples concept-specific ES looks like the ES of a random point in gene expression space for this concept, we run into a different problem: how do we parameterize such a sampling procedure? Because in vivo gene expression programs contain highly correlated subprograms [116], there are large portions of gene expression space that are unavailable to a living cell (i.e., there are relationships among the genes expression intensities that one never observes in nature). Surely these “impossible” expression inputs should not be considered when generating the null distribution.

If we try to overcome this sampling problem by using real human gene expression observations, we arrive at the cross-validation strategy described above. Rather than set a threshold learned from this data for accepting or rejecting a concept outright, we find it more informative to let the user understand the overall amount of signal present in the data for a given concept (ROC plots), and report an expected false positive rate for the concept at the ES observed for the new sample.

3.3.3 Performance results

We see an average accuracy of 92.8% (AUC value of 0.928) after restricting the set of UMLS concepts to the 1209 that have samples from two or more expression series in GEO to ensure that a diverse set of data is used. Even when we restrict the concepts to the 450 that have at least 50 distinct samples originating from at least five different data series, the average accuracy is approximately 89.8%. Table 3.1 contains the performance of a selection of UMLS concepts, along with the number of samples and series that were associated with that concept. Unsurprisingly, “broader” con-

Concept	**AUC**	**Num Series**	**Num Samples**
Malignant Neoplasms	0.82	74	855
Malignant neoplasm of breast	0.97	9	69
Malignant neoplasm of ovary	0.99	4	51
Malignant neoplasm of lung	0.97	4	98
Leukemia	0.99	13	151
Soft Tissue	0.69	98	1513
Breast	0.93	13	195
Ovary	0.95	8	103
Lung	0.95	9	131
Inflammatory disorder	0.79	13	91
Rheumatoid Arthritis	0.93	7	31
Inflammatory Bowel Diseases	0.99	2	24

Table 3.1: Cross-validation performance for a selected subset of UMLS concepts. The complete table can be found in Table C.1 in Appendix C.

cepts have poorer performance compared to the more specific concepts, as the former encompass a much more diverse expression signal. The complete list of performance values can be found in Appendix C. Note that many of these concepts are similar and thus have many database samples in common; as a direct consequence, many of the concepts have similarly high (low) AUC values.

In general, deeper concepts in the hierarchy have both fewer samples associated with them and have higher accuracies (Figure 3-7). Although it may be tempting to draw the conclusion that fewer samples mean higher accuracies, it is actually that the deeper concepts correspond to gene expression samples that have greater biological similarities. For example, the deeper concept Malignant neoplasm of breast has a higher predictive power with 69 samples than the broader concept Primary malignant neoplasm with 833 samples.

3.3.4 Quantification of the “batch” effect

There have been several reports that data from different datasets are not comparable as the dataset (aka “batch”) signal is dominant [102, 94] . While the localization of phenotypes as seen in the expression landscape (Figure 3-3), regardless of series of origin, depict the lack of a dataset effect in principal component space, the

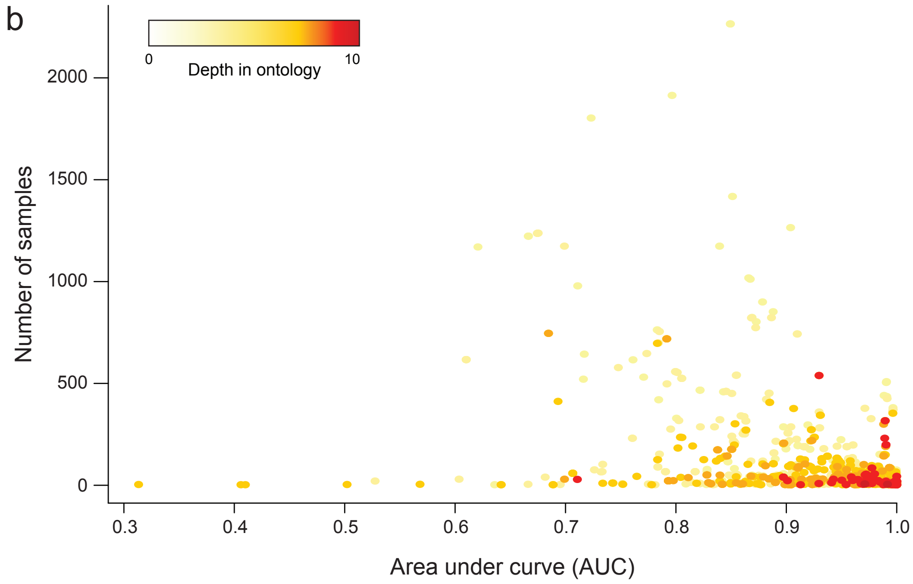


Figure 3-7: A plot depicting the relationship between the number of samples, the depth in the UMLS hierarchy and the AUC such that each point represents one of the disease or tissue concepts. Notice that the deeper in the ontology a concept is (i.e. more specific), the higher its AUC.

cross-validation performance shows that this phenomenon holds true when all gene expression data is considered. Although the area-under-the-curve (AUC) and receiver operating characteristic (ROC) curves are generally used to quantify the performance of classifier, they can also be used as a proxy to quantify the significance of a batch effect. As high AUC values can only be attained through accurate identification of phenotypes in cross-validation, it is a necessary precondition for samples associated with a given phenotype to be more closely related to each other than those associated with another phenotype.

In addition, by associating the series of origin for each sample used to generate the ROC plot, we can visually inspect the degree of the batch effect by the clustering of the samples from these series. For instance, the ROC curve for the concept "leukemia" (Figure 3-8) shows that: 1) samples with the phenotype, regardless of dataset, are closer to the other samples with the same phenotype, and 2) samples from various datasets are intermingled. The leukemia samples were more closely related to other leukemia samples with a mean intra-phenotype, inter-series correlation of 0.1 higher compared to other samples within their own dataset that were non-

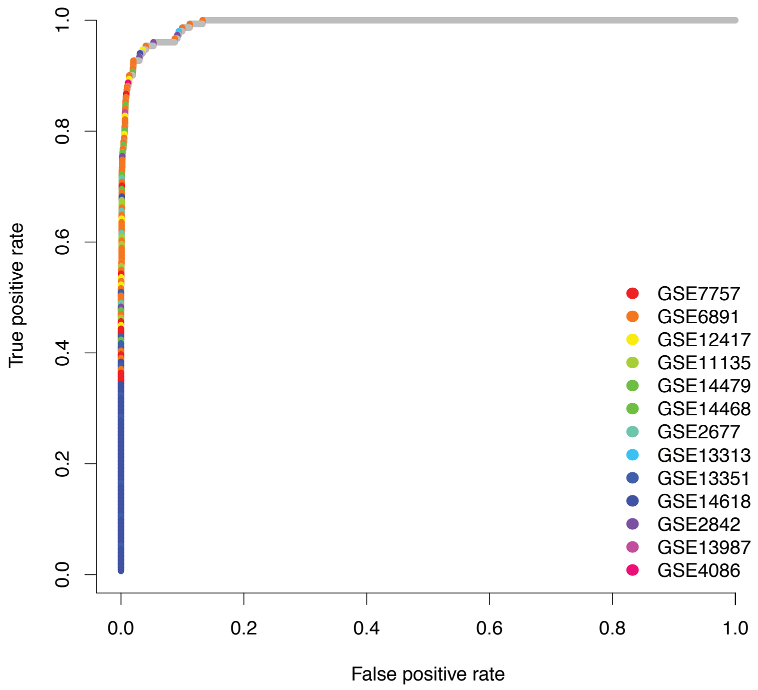


Figure 3-8: The ROC curve for leukemia. The colors plotted along the curve correspond to the series of origin for each of the samples used to generate the curve. The intermingling of series points to the robustness of the phenotypic signal: samples with the same phenotype cluster together before all other phenotypes, and samples from different data series are intermingled within a phenotype.

leukemia samples (inter-phenotype, intra-series). We see that this trend is evident in the ROC curves across all types of phenotypes. Intuitively, if this were not the case, not only would the AUC values for concepts that have samples from multiple series have to be substantially lower than those with fewer series, but also the phenotypic localization evident in the transcriptome landscape would have been overshadowed by dataset localization.

In an effort to quantify the dataset effect (DE) from the correlation structure of the gene expression samples used in the construction of the transcriptome landscape, we compared the mean difference in correlation between all samples in a series with the phenotype to all other samples in other series with that phenotype to the mean difference in correlation of samples with a given phenotype in a series against all other samples in that series without the phenotype. In the event that the signal from the data series is greater than that of the phenotype, one would expect that the

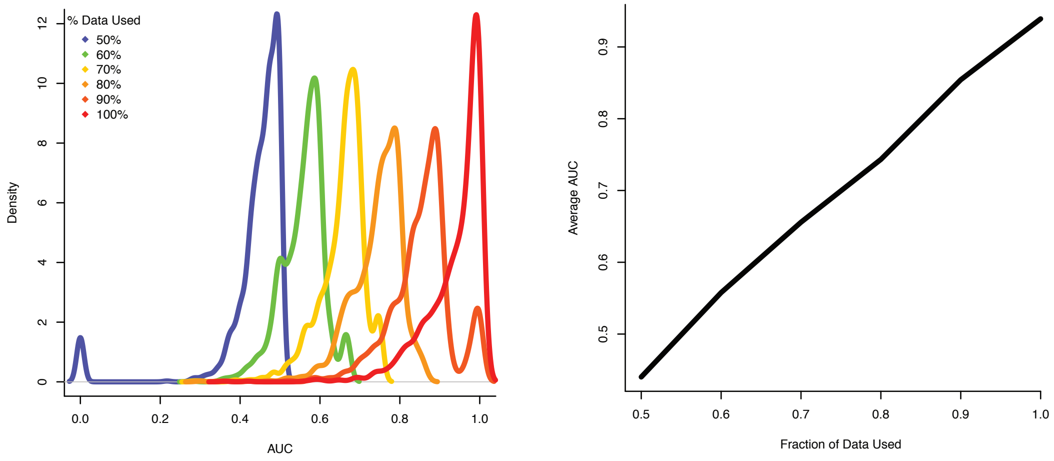
Tissue	Dataset Effect	P Value
Spleen	-0.22	0
Esophagus	-0.2	0
Salivary Glands	-0.2	0
Cerebellum	-0.18	0
Prostate	-0.17	0
Lymph Node	-0.17	0
Myometrium	-0.14	0
Tongue	-0.14	0
Liver and/or Biliary Structure	-0.14	0
Kidney	-0.13	0
Skeletal Muscle	-0.12	0
Spinal Cord	-0.11	0
Stomach	-0.11	0
Endometrium	-0.11	0
Spinal Nerve Structure	-0.1	0
Heart	-0.1	0
Brain	-0.08	0
Adrenal Gland	-0.08	0
Lung	-0.06	0
Colon	-0.05	0
Penis	-0.05	0.06
Gingiva	-0.05	0
Skin	-0.04	0
Ovary	-0.04	0
Hippocampus	-0.03	0
Breast	-0.02	0
Intestine	-0.02	0
Bone Marrow	-0.01	0
Stem Cells	0	0
Thyroid	0	0.46
Uterus	0.04	0.98
Blood	0.06	0.34
Epithelial	0.07	0
Bone	0.09	0

Table 3.2: The dataset effect was measured the mean difference in correlation between all samples in a series with the phenotype to all other samples in other series with that phenotype to the mean difference in correlation of samples with a given phenotype in a series against all other samples in that series without the phenotype. A negative dataset effect value implies that the phenotypic signal dominated the dataset signal while a positive value implies that the dataset signal was greater.

intra-series correlation between differing phenotypes is greater than the inter-series correlation between samples corresponding to identical phenotypes. The p-values (Pv) were computed by randomly shuffling the phenotype labels on the samples and computing the dataset effect 100 times for each tissue type. The empirical p-value was determined by finding the position in the sorted list of sampled dataset effect values. The majority of the tissues for which sufficient data was available (at least two series with the phenotype, and at least one series containing both the phenotype of interest and at least one other phenotype), do not exhibit the existence of a batch effect. For example, across 6 series with normal prostate tissue, the correlation of prostate samples to other prostate samples in other series is on average 0.17 higher than the correlation of those samples to other samples within their own series. In the few instances where the correlation within the dataset is higher, it generally is due to the highly similar nature of the samples and that the tissue signal dominates the disease signal. In the case for the blood series, for instance, normal blood is being compared to diseased blood. Table 3.2 provides these numbers the tissues that are represented in the tissue similarity network (Section 3.2.2) such that a negative batch effect implies that the phenotypic signal dominated the dataset signal.

3.3.5 Scalability

Due to the non-parametric data-driven nature of this method, we are not constrained by the amount of gene expression samples that are present in the database. However, the question arises as to whether or not adding more samples to the transcriptomic landscape provides a higher resolution picture, or if it merely muddles the picture. In an effort to resolve this question quantitatively, we computed the classification accuracy of each concept when the number of samples that were used to compute the enrichment score for that given concept was set to 50%, 60%, 70%, 80%, and 90%. For example, using all 69 samples for “Malignant neoplasm of breast” yields an accuracy of 96.5%. Then, keeping all else constant, we randomly removed half of the “Malignant neoplasm of breast” samples and recomputed the enrichment score. This random re-computation was performed five times for each concept at each threshold. In the



(a) Density estimates of performance with varying amounts of data

(b) Average AUC values with varying amounts of data

Figure 3-9: Improvement of accuracy of the enrichment statistic with the increase of data in the database. (a) Density estimate of the performance of the method over various amounts of data. (b) The average AUC values over all concepts when varying the amount of data used to compute the enrichment scores. For example, when using only 50% of the data for a given concept, the average AUC drops down to 42%.

case of “Malignant neoplasm of breast,” for instance, the average accuracy across the five runs using only 34 samples is a mere 37%. We see that the average accuracy across all concepts drastically increases from 44% to roughly 93% when increasing the amount of data used (Figure 3-9). It is also noteworthy that the concepts that are the most susceptible to change are specific concepts (e.g., “Pluripotent stem cells” and “Myeloid Leukemia”), while the classification accuracy of the broad topics (e.g., “soft tissue” and “disorders”) are unaffected by the quantity of data as the underlying gene expression values are so vastly different.

This implies that the power of this type of macroscopic analysis increases with the amount of underlying data, as the signal of the phenotype becomes more apparent. Since our approach employs a non-parametric enrichment statistic that only requires the concept annotation of the samples in the original gene expression database, it can be updated in real-time without having to “retrain” the database. A system such as this could thus be deployed in a research or clinical setting where new samples are continually being added and analyzed, with minimal alteration of normal protocols. As new samples are added, the system would continually improve its understanding of the biological signals that constitute individual phenotypes.

With our database primed with the 3030 labeled samples ranging from normal breast to blood from children with septic shock, we applied Concordia to 15904 other GEO samples performed on the Affymetrix HG-U133 Plus 2.0 array and mapped each sample onto the transcriptomic landscape. In this manner, we are able to provide the concept enrichment scores for 1489 anatomy and disease related concepts for other samples based on the current biological "knowledge-base" of Concordia. These concept enrichment scores can thus be used as an additional source of biological information when performing future large-scale gene expression analyses. For example, if a researcher is looking for expression samples relating to "breast" tissue, he could both examine the text that is associated with each sample, and look at the expression similarity of that particular sample and the concept for "breast."

3.3.6 Specificity of the conventional classification of tissue and disease

Employing the classification accuracies of the conventional clinical categories as defined by the UMLS hierarchy allows us to systematically estimate the classification robustness of conventional clinical labels as compared to molecular pathophenotypes [75]. The sub-tree of the ontology rooted at "Inflammatory disease," is a striking illustration of the faithful reflection of specificity as a function of depth in the tree (Figure 3-10). As conventional wisdom would dictate, concepts relating to broad phenotypic topics that span multiple tissue or disease categories have lower classification potential than specific concepts located deeper in the ontology that have a more conserved gene expression pattern. For instance, we see the classification accuracy of the more specific concept, "Chronic arthropathy" (98%), is significantly higher than that of "Inflammatory disorder" (78.9%). In general, we see that the conventional clinical classification of tissue and disease mirrors the underlying gene expression signature. If, for example, the opposite effect were observed, such that concepts higher in the hierarchy had higher accuracies, the structure of clinical nomenclature would

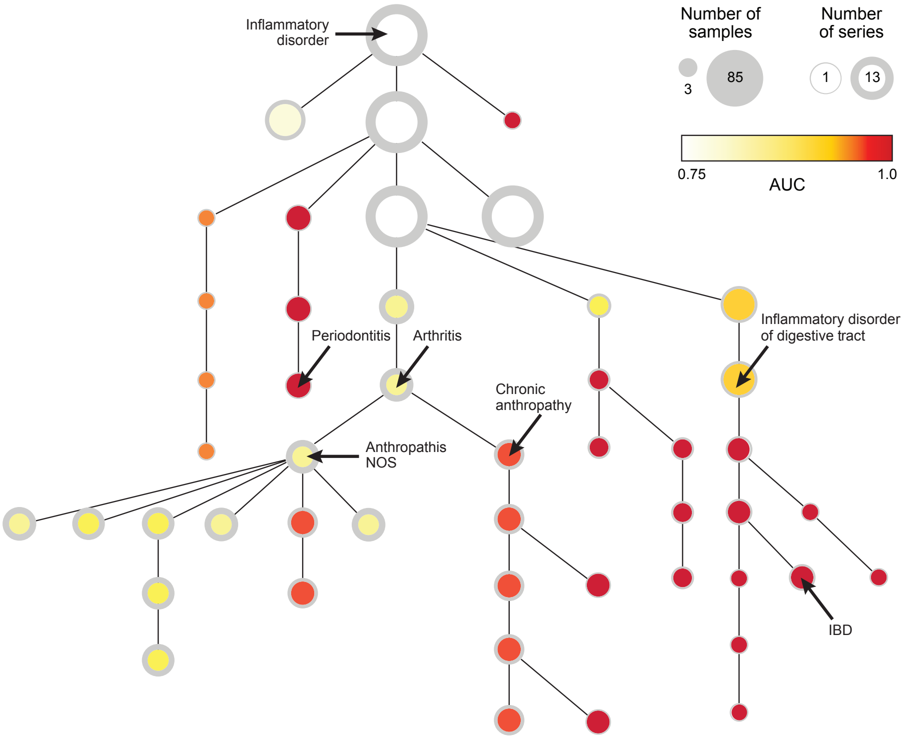
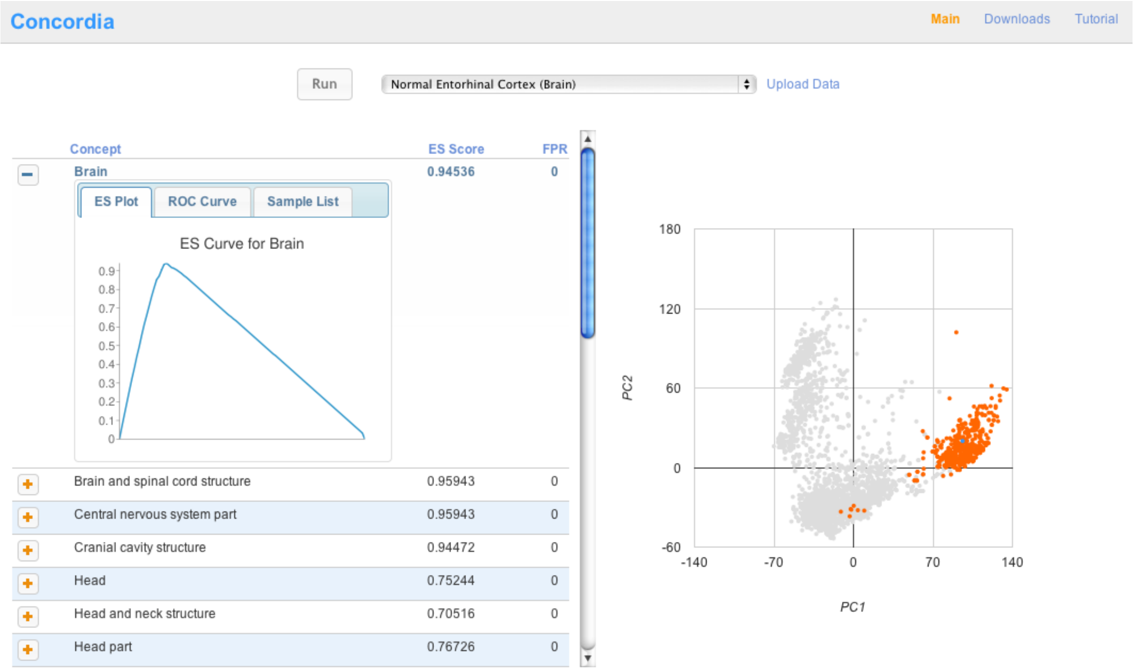


Figure 3-10: Specificity of conventional clinical classification of disease. Part of the UMLS ontology rooted at "Inflammatory disease" in which the color of the node indicates the classification accuracy; the size, the number of samples in the database with that concept; and the thickness of the line, the number of different datasets to which the samples belong. As conventional wisdom would dictate, labels corresponding to broad phenotypic topics appear higher in hierarchy and have lower predictive power.

be put into question. It is important to note that the ordering based on depth in the UMLS hierarchy is not global, but a local phenomenon. For example, as shown in Figure 3-10, "Arthritis" splits into two sub-trees in which the side rooted at "Chronic arthropathy" has a high predictive value all the way down the sub-tree, while the other sub-tree that has a wider variance in predictive accuracies.

3.3.7 Concept enrichment web interface

In order to provide concept enrichment statistics for new samples performed by other researchers, we have developed an online resource (`http://concordia.csail.mit.edu`) that allows users to submit their own expression samples performed on the



Normal Entorhinal Cortex (Brain)

Concept	ES Score	FPR
Brain	0.94536	0
Brain and spinal cord structure	0.95943	0
Central nervous system part	0.95943	0
Cranial cavity structure	0.94472	0
Head	0.75244	0
Head and neck structure	0.70516	0
Head part	0.76726	0

Figure 3-11: A screenshot of concept enrichment web interface. In this example, a gene expression sample corresponding to normal entorhinal cortext (brain) tissue was submitted. The blue dot in the midst of the orange dots represents the location of this sample in the transcriptomic landscape. After performing the concept enrichment using the database primed with 3030 gene expression samples, we see that this new sample is (correctly) over-enriched for concepts relating to the brain. When selecting all samples in our database associated with the concept “Brain” (the orange dots), we see that this new sample indeed lies within their midst.

Affymetrix HG-U133 Plus 2.0 array. Once the sample has been uploaded to the server, the expression values are ranked normalized and the concept enrichment is computed as detailed in Section 3.3.1. These scores are then reported to the user (along with the enrichment score curve, the ROC curve for each concept as obtained through the cross validation procedure, and all of the samples in our database corresponding to that concept). In addition to these concept enrichment scores, we also provide the user with the visual representation of the transcriptomic landscape and the location of their submitted sample in the landscape.

For example, in Figure 3-11 we see a gene expression sample corresponding to normal entorhinal cortex (brain) tissue that was submitted. The blue dot in the midst of the orange dots represents the location of this sample in the transcriptomic landscape. After performing the concept enrichment using the database, we see that this new sample is (correctly) over-enriched for concepts relating to the brain. When selecting all samples in our database associated with the concept “Brain” (the orange dots), we see that this new sample indeed lies within their midst.

3.4 Tissue specific signal of tumor metastases

The clinical problem of distinguishing whether a cancerous lesion represents a primary tumor, or a metastasis from a distant malignancy, presents a test case for our ability to localize a sample to the appropriate phenotypic group within the transcriptomic landscape. By combining the aforementioned sample- and gene-centric methods, we are able to map new tumor metastasis tissue samples onto the gene expression landscape, providing an unbiased measure of their phenotypic predisposition based on gene expression. It is commonly known by pathologists that tumor metastasis tissue biopsies viewed under the microscope resemble the tissue of the primary site rather than that of a tissue in the metastasized location. Indeed, we find that metastatic tissue samples localize in the vicinity of their tissue of origin in the transcriptomic landscape (Figure 3-12), even without the use of specially-tuned primary site detection methods [18, 111].

For instance, using mapping metastasized breast cancer samples (GSE14107) on to the transcriptomic landscape, we see that all of the metastases, regardless of whether they were removed from the lung, brain, or bone, more closely resemble breast tissue than their biopsy locations (A). Some of the over-enriched UMLS concepts include White Adipose Tissue, Subcutaneous Fat, Subcutaneous Tissue, Lactiferous duct, Mammary lobe, and Glandular structure of breast. The 15 of the 17 colorectal cancer samples from GSE10961 (B) were all labeled with Rectum and sigmoid colon, Colonic Diseases, Functional, and Colon carcinoma with a false positive rate of below 0.05; the other two samples had a FPR of 0.06 for Colon Carcinoma. A table of other identified metastatic samples and their corresponding top concepts can be found in Appendix C

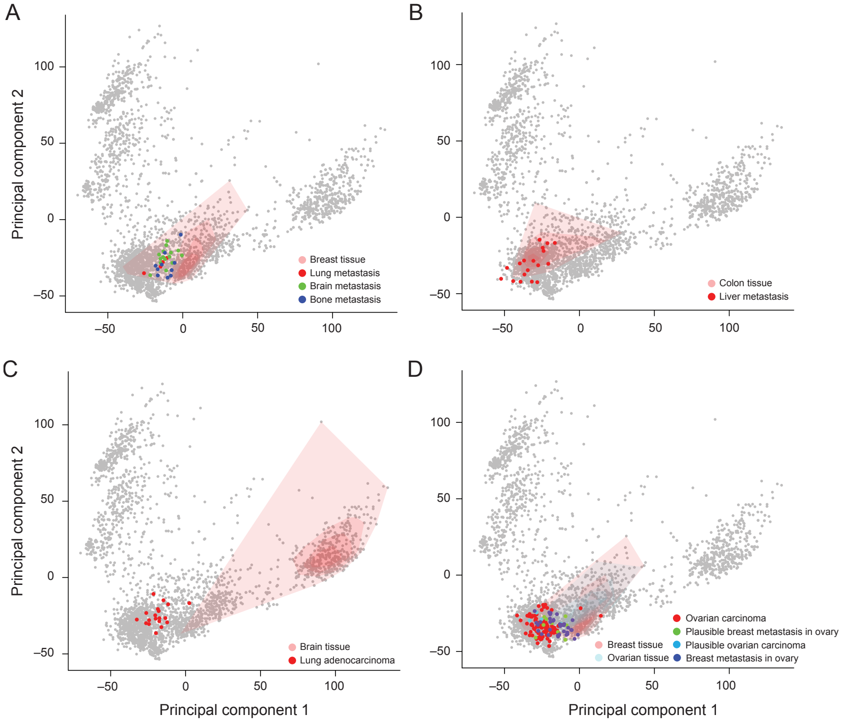


Figure 3-12: Principal component analysis shows that metastatic samples more closely resemble their primary sites. Along with the concept enrichment, the first two principal components of the gene expression data show that the gene expression signature of tumor metastases more closely resembles that of their primary site location than that of their metastasized sties. (A) Breast tumors that metastasized to the lung, brain, and bone still appear to be more closely related to other breast samples than to their metastasis sites. (B) Colon tumors that metastasized to the liver lie proximal to colon tissue and are enriched for concepts such as Rectum and sigmoid colon and Colon carcinoma. (C) While we were not able to correctly identify the exact primary site location, the lung adenocarcinoma samples that metastasized to the brain look nothing like brain tissue that is located in the top right cluster (see Figure 3-3). (D) In the context of the entire transcriptome landscape, there is significant overlap in breast and ovarian tumor and tissue samples; this makes it difficult to properly distinguish between them.

Those metastases that were mislabeled provide a measure of the unbiased degree of overlap between these metastases, reflecting the lack of hard boundaries in the continuum of the transcriptomic landscape. This is particularly evident within the soft-tissue cluster (bottom left Figure 3-3), in which the tissue specific signal can be dwarfed by the larger variances caused by the blood and brain tissue samples. Although the use of a supervised learning approach, such as in Schaner et al. [111], could mitigate these issues and be used to identify the tissue of origin, these approaches minimize the significant biological overlap of some of these samples, which may have implications for therapeutic selection [29]. Thus, for example, our approach appropriately does not label brain metastasis samples (GSE14108) with brain descriptors, yet a transcriptome-wide approach such as ours that encompasses samples ranging from normal tissue to blood samples from autistic patients, cannot label them with the correct primary epithelial site other than correctly identifying it as a form of adenocarcinoma (C). Similarly, due to the close proximity of breast and ovarian tissue samples in the transcriptomic landscape, we had trouble distinguishing between breast and ovarian metastases (GSE20565) (D).

3.5 Conclusion

With the ever-growing repositories of data, both public and private, it has become not only possible, but also imperative to embrace the full transcriptomic continuum of tissue and disease. Employing a comprehensive, non-case vs. control approach and making use of the multi-dimensional nature of gene expression data, we capture biological processes that are typically overshadowed in traditional analyses. Furthermore, we are able to recapitulate the biological and medically relevant concepts relating to merely a single new expression sample. Indeed, as the power of this macroscopic analysis increases with the amount of underlying data, this approach has the potential to more fully leverage large, public databases with biological data, and to benefit further as more data are added. Although we have presented our sample- and gene-centric methods utilizing medically relevant concepts, the data-driven nature of these methods implies that by changing the scope or domain of these labels, they can be applied to analyses in different contexts with relative ease.

As suggested by some [75], systematic application of molecular pathology measurements will allow a useful shifting of the conventionally employed diagnostic classification boundaries to include the notion that there are intermediate pathotypes that cross the boundaries of the conventional medical classifications. These intermediate pathotypes are more closely coupled to the actual underlying pathology, thus revealing not only shared pathology but also opportunities for development of shared treatment [29, 35]. It may be the case that the gene expression signatures of disease provide clues to a disease network [12] other than what classical medical knowledge dictates, thus providing new insights into relationships between diseases that previously were unknown.

It has been proposed that the future of personalized medicine, and the proper application of genomic and genetic data, requires an understanding of both who the patient is and the characteristics of the subpopulation to which the patient belongs [64]. Clinical applications of our approach, in conjunction with other genetic, environmental and phenotypic information, could more accurately and consistently annotate clinical samples and provide an impartial view of the landscape of clinico-pathological classification. As well as pointing out differences between tissues and diseases, our approach provides an unbiased perspective on the overlapping biology of diseases with the attendant implications for therapy selection in personalized medicine. In addition, it could also be used to error check human annotations by analyzing the concordance of the gene expression signatures of the patient with the textual information provided by the clinician. Furthermore, as it makes use of an enrichment statistic that only requires the usual standard of care in the labeling of samples, this system could be deployed in a clinical setting with minimal alteration of normal procedures. By shifting away from a dichotomous view and employing the global transcriptomic landscape, we hope to address one of the key requirements of personalized medicine and begin to answer one of its fundamental questions, what other samples am I most similar to so that the most effective treatment can be administered?

Chapter 4

Beyond differential expression: Marker genes in a non-dichotomous world

The notion of a *marker gene* is nothing new. By definition, a marker gene, is a gene that provides some information about a phenotype. For example, genes BRCA1 (breast cancer 1, early onset) and BRCA2 (breast cancer 2, early onset) are known to be tumor suppressors and that mutations to to these genes can cause increased susceptibility to breast and ovarian cancer [34]. While BRCA1 and BRCA2 are individual genes that are linked to certain phenotypes, it is often the case that no single gene uniquely that sets of genes. So called *marker gene sets* are sets of genes that are used to define a phenotype. For example, MammaPrint [45] is a clinical test that measure the expression of 70 genes to determine the risk that a breast tumor will metastasize to other parts of the body. Similarly, OncotypeDX [95] is a 21 gene predictor for women who have ER$^+$ breast cancer to select those at higher risk for recurrence and those more likely to benefit from aromatase inhibitor or tamoxifen treatment.

Using the curated database of 3030 gene expression samples from GEO that was introduced in Chapter 2, we shift from the sample centric analyses of Chapter 3 to gene centric analyses. As in the previous chapter, we shy away from the traditional case vs. control methodology adopted by others, but view the expression patterns of genes in a holistic manner. Our new perspective on interpreting gene expression space helps uncover phenotype-specific marker genes beyond those discovered by traditional dichotomous views of gene expression. When we apply our method to identify marker genes for various cancers, we find that the marker genes are highly specific to the particular cancer as opposed to generic cancer processes such as cell-cycle and cell adhesion that are found by others that employ case vs. control experiment design [104]. Furthermore, capitalizing on the hierarchical nature of the phenotypic labels associated with our samples[1], we also demonstrate that genes previously linked to specific types of carcinomas may actually be part of a broader "carcinoma" process. Finally, we illustrate how metastasized tumor samples are transcriptomically more proximal to other cancer samples from their respective primary sites, as opposed to cancerous tissue from the metastasis sites from which the samples were resected.

4.1 Marker gene finding: Finite impulse response filter

We developed a method to identify marker genes that characterize a specific phenotype in the context of broad transcriptomic landscapes, and not in the context of dichotomous classes. Instead of defining a marker gene as one that is over- or under-expressed in a case vs. control study, we define a marker gene as a gene that has a "localized" expression signature for a phenotype; i.e., how grouped together are all of the samples corresponding to that phenotype for that gene. If all of the samples for a phenotype have a very similar expression level (all very high, all very low, etc.), the gene may be considered a marker gene for that phenotype. We employ a standard signal processing tool, a finite impulse response filter (FIRF) [82], on each genes expression values across the entire database of 3030 diverse expression samples to quantify the degree of expression level localization for a given phenotype.

[1] Recall, we are using the UMLS ontology and such we have parent and child relationship information for all phenotypic labels. For details see Chapter 2.

In contrast to a standard t-test based approach, this approach does not require us to define a specific control phenotype against which we test for separation, a poorly defined task when comparing against such a diverse database. Moreover, this method identifies genes with expression levels that are highly localized for a given phenotype, thereby allowing for the diverse population of other samples to express these genes at simultaneously higher and lower levels (something for which a t-test cannot directly account). For example, the gene DBC1 exhibits a highly specific range of expression across the stem cell samples, and ranked highly (among the top 0.5% of all genes) in its ability to localize the stem cell samples by the described method. However, the non-stem cell samples demonstrate both higher and lower expression levels of this gene, causing a standard Students t-test (treating all non-stem cell samples as the control group) to rank this gene at only the 24.6% strongest among all genes.

A gene's *marker gene score*[2], the level of localization of the expression intensities, is computed using a FIRF. For each gene g phenotype p pair we first sort all of the expression samples by their expression intensities for g. Using a "sliding window" of size equal to the number of samples corresponding to p, we compute the fraction of samples in that window that are associated with p. If all samples in the window are associated with p, then the score for that window is 1, if none are associated with p, then the score is 0. This window is iteratively moved across the sorted list of samples so that we obtain a score for all possible positions. The marker gene score for a particular gene-phenotype pair is the maximum score that is achieved in any of the windows. A p-value is computed for each score based on a binomial distribution.

Figure 4-1 portrays a pictorial representation of the scores produced by the the FIRF for two genes LRRTM2 (26045) (Figure 4-1(a)) and OR1J4 (26219) (Figure 4-1(b)). As outlined above, the samples are sorted in order from lowest to highest expression value for the respective genes and the enrichment of brain samples in the sliding window is calculated. The x-axis shows the index in this sorted list. The red line corresponds to the score of FIRF (the fraction of the samples in the sliding

[2] We may alternatively use the term *marker gene localization score* or *expression localization score* interchangeably with *marker gene score*.

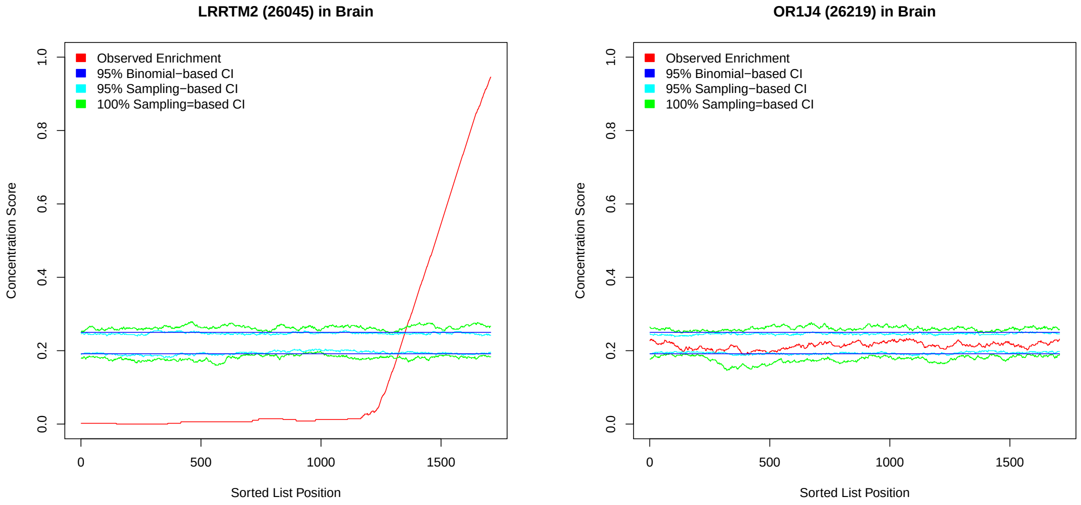


Figure 4-1: The marker gene FIRF score plots for genes (a) LRRTM2 (26045) and (b) OR1J4 (26219). The samples are sorted in order from lowest to highest expression value for the respective genes and the enrichment of brain samples in the sliding window is calculated. The red line corresponds to the score of FIRF for each of the genes for brain tissue. The green and blue lines depict the binomial and sampling based confidence intervals. The former gene (leucine rich repeat transmembrane neuronal 2) is considered a marker gene for brain as its score is significantly outside the confidence intervals. On the other hand, the score for the latter (olfactory receptor, family 1, subfamily J, member 4) is wholly contained within the confidence intervals and thus is (correctly) not a marker gene for brain tissue.

window that are brain samples) for each of the genes. The green and blue lines depict the binomial and sampling based confidence intervals. The former gene (leucine rich repeat transmembrane neuronal 2) has a marker gene score of 0.95 (i.e. there was a window in which 95% of the samples were brain samples) and is thus considered a marker gene for brain. We see that this score is significantly outside the confidence intervals. Furthermore, as to be expected of a neuronal gene, we see that the marker gene score peaks at the right hand side of the plot, the side where all of the samples with high expression intensity values for LRRTM2 are located. On the other hand, the score for the latter gene (olfactory receptor, family 1, subfamily J, member 4) is wholly contained within the confidence intervals and thus is (correctly) not a marker gene for brain tissue.

4.1.1 Specificity of marker genes

It has been suggested that the so-called “incidentalome” of incidental findings is a threat that has yet to be addressed in either biological or clinical settings [66]. The consequences of non-comprehensive views of biomarkers, such as prostate specific antigen, continue to cause needless harm and costs [125]. As an example of the utility of the marker gene scores used in conjunction with our database of expression samples that have highly curated ontological phenotypic labels, we can show that many genes are not specific to a single disease. This sort of quantification of phenotype specificity is of course relevant to the diagnostic accuracy of putative biomarkers and for developing suitably broad-spectrum or targeted therapeutics.

To illustrate this, we took the 459 carcinoma tissue samples in our database and computed the “carcinoma” marker gene localization scores by comparing them to the 270 other tumor samples. As the UMLS concepts are in a structured ontology, we computed the marker gene scores for the 13 concepts subordinate to “carcinoma” (such as “adenocarcinoma,” “Adenosquamous carcinoma,” etc.) for which we had at least three expression samples. From the list of genes sorted by their carcinoma marker gene score p-value, we removed all genes that had a better p-value in any of the 13 subordinate concepts. This yielded a list of 5805 genes that had better p-values at the more general concept “carcinoma” than at any of the more specific subordinate carcinoma types. Functional enrichment analyses of the top 10, 20, 50, 100, and 150 genes in this list reveals processes such as ”regulation of cell adhesion,” ”response to growth factors,” and various other morphogenesis and development terms. Furthermore, within the sorted list of carcinoma genes, we see genes previously implicated in carcinomas such as COL1A1 [78, 100, 143] and ELF3 [22] in the top 5. As such, we see that these genes that have previously been implicated in particular types of carcinomas may instead be part of a larger “carcinoma” process, rather than specific to breast or colorectal cancer.

4.2 Phenotype marker gene sets

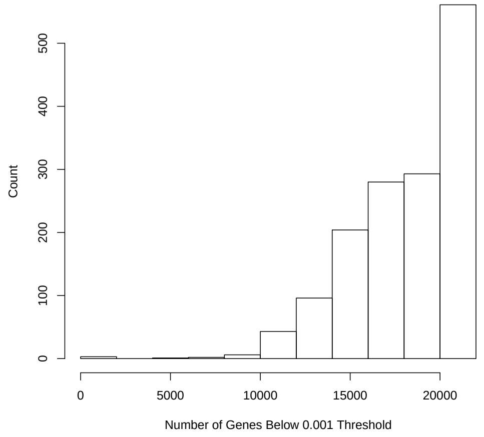


Figure 4-2: The histogram of the number of the number of genes that were below a 0.001 binomial p-value cutoff across 1489 UMLS concepts. The total heights of the histogram adds up to 1489. A count is added to, for example, bin 500 if there was a concept which had 500 genes with a p-value of below 0.001.

An astute reader may have noticed that that marker gene scores for each phenotype merely provide a ranking of genes. In other words, genes with a high marker gene score have a high degree of expression intensity localization while those that have a low score do not. Although computing a p-value based on either the binomial distribution or randomized trials as depicted in Figure 4-1 is a good start at keeping only the "good" ones, a significant number of genes have low binomial p-values. For instance, if we set a p-value threshold at 0.001, and count the number of genes that are below that threshold for the 1489 UMLS concepts available in the Concordiafied version of GEO (see Chapter 2 for details), we see that a large number of concepts have at least half of the genes below the threshold (Figure 4-2). To go beyond the simple rankings of all genes for a phenotype as provided by the marker gene score, we generate gene sets that optimally describe that phenotype. We identify the cutoff for the number of genes to include in the set by balancing the gene sets ability to accurately classify samples of its own phenotype while minimizing the presence of non-phenotype specific signal (Methods). Not only does this method sidestep the requirement of defining the appropriate comparator phenotypes, but it also facilitates the identification of thematically coherent gene signatures that reveal very different aspects of biology from traditional ones.

4.2.1 Generating phenotype specific gene sets

Algorithm 1: Computing marker gene AUCs

```
Input: $M$: The list of genes sorted from high to low by their marker gene
       score.
Input: $S$: The set of all samples such.
Input: $S_{pheno}$: The set of samples $\in S$ for a given phenotype.
Input: $E$: The matrix of gene expression values with gene IDs in the rows and
       sample IDs in the columns.
Output: $A$: A matrix of AUC values with sample IDs in the rows and index
        into $M$ in the columns.
for $i$ in 1...length($M$) do
    $E_i \leftarrow$ E[1:i, ]
    for $i$ in 1...length($S$) do
        $s \leftarrow S[i]$
        $S' \leftarrow S \backslash \{s\}$
        $corrs \leftarrow$ The correlations of vector $E_i[, s]$ to sub-matrix $E_i[, S']$
        $S'' \leftarrow S'$ sorted by the correlations, $corrs$, from highest to lowest
        $v \leftarrow generateHitMissIndicatorVector(S'', S_{pheno})$
        $A[s, i] \leftarrow computeAUCFromHitMissVector(v)$
```

To generate phenotype specific marker gene sets, we use an additional method to determine the appropriate cut-off for the number of genes required to describe a particular phenotype p. As detailed in Algorithm 1, first, the genes are sorted according to their marker gene score from highest to lowest. We then iteratively examine the quality of the top i genes, balancing their positive predictive capability with the amount of noise that they add. Starting with the first two highest scoring genes, we iteratively remove each sample s and compute its correlation to all other samples (S') using only those two genes. We generate a receiver operating characteristic (ROC) curve for s and use the area under the curve (AUC) of as a summary statistic. The ROC curve is generated by sorting all samples by their correlation to s, and incrementing the true-positive (TP) count when that sample is associated with p, and increment the false-positive (FP) count when that sample is not associated with p. If, all samples of the phenotype of interest are all more correlated to the sample being studied, then we achieve an area under the curve (AUC) of 1. For the sake of clarity, let us assume that we store each of these AUC values in matrix A such that $A[s, i]$ contains the AUC value for sample s when we use the top i genes to compute the correlations. Once all AUCs are computed for two genes, we add the next highest scoring gene, and re-compute all AUC values.

As portrayed in Algorithm 2, once we have the matrix A containing all of the AUC values for each sample for all number of marker genes, we determine the optimal number of genes by computing the ratio of the mean AUC values for all samples that indeed are associated with the current phenotype to the mean AUC of all samples that are not associated with the current phenotype. Intuitively this ratio works as follows. The mean "hit" AUCs is the average AUC that we get when attempting to classify samples of the current phenotype. If the marker genes truly do represent the signal of the current phenotype, then using the top i marker genes should make samples of the phenotype have higher correlations resulting in higher AUC values. The more genes that we use (the greater i is) the more of the signal we can capture, and thus, hope to increase the average AUC value. On the other hand, the more genes that we include, the more likely we are to include signal that is not associated with the phenotype in question. Therefore, the more genes we include, the greater the mean AUC for the samples that are not associated with the current phenotype becomes. By taking the maximal value of the ratio of these two means, we find the number of genes that is required that not only does a great job of classifying the phenotype in question, but also does a poor job of classifying others.

Algorithm 2: Computing optimal number of marker genes

```
Input: $M$: The list of genes sorted from high to low by their marker gene score.
Input: $S$: The set of all samples.
Input: $S_{pheno}$: The set of samples $\in S$ for a given phenotype.
Input: $S_{other}$: The set of all other samples $\in S$ ($S_{pheno} \in S_{other} = \emptyset$).
Input: $A$: A matrix of AUC values with sample IDs in the rows and index into $M$ in the columns as computed in Algorithm 1.
Output: $M_{optimal}$: The list of optimal number of marker genes.
$\mu_{hit} \leftarrow mean(A[S_{pheno},])$
$\mu_{miss} \leftarrow mean(A[S_{other},])$
$ratio \leftarrow \frac{\mu_{hit}}{\mu_{miss}}$
$M_{optimal} \leftarrow M[0 : max(ratio)]$
```

This is exactly the behavior that we see in Figure 4-3. Here we plot the “hit” and “miss” AUC curves for the phenotype “breast.” As to be expected, we see that the “hit” curve (in black) is improving as more genes are added. Similarly, after an initial decrease, the “miss” AUC curve (in red) begins to increase as well. When we take the ratio of the two curves, it becomes apparent that using the top 164 genes provides the optimal gene set of the phenotype “breast.”

Figure 4-4 shows the same behavior depicted in Figure 4-3 but without taking the mean. Although we use the mean hit and miss AUCs to determine the optimal number of genes to use in the gene set, this heatmap shows the AUC values for each sample (in the rows) for each number of genes used in the gene set (in the columns) such that high AUC values are yellow and low AUCs are red. The AUC values for all breast samples (denoted by the blue color on the bottom left of the plot) is high across all ranges of top number of genes to include. However, the AUCs that are initially low for the non-breast samples improve as more genes are added to the gene set.

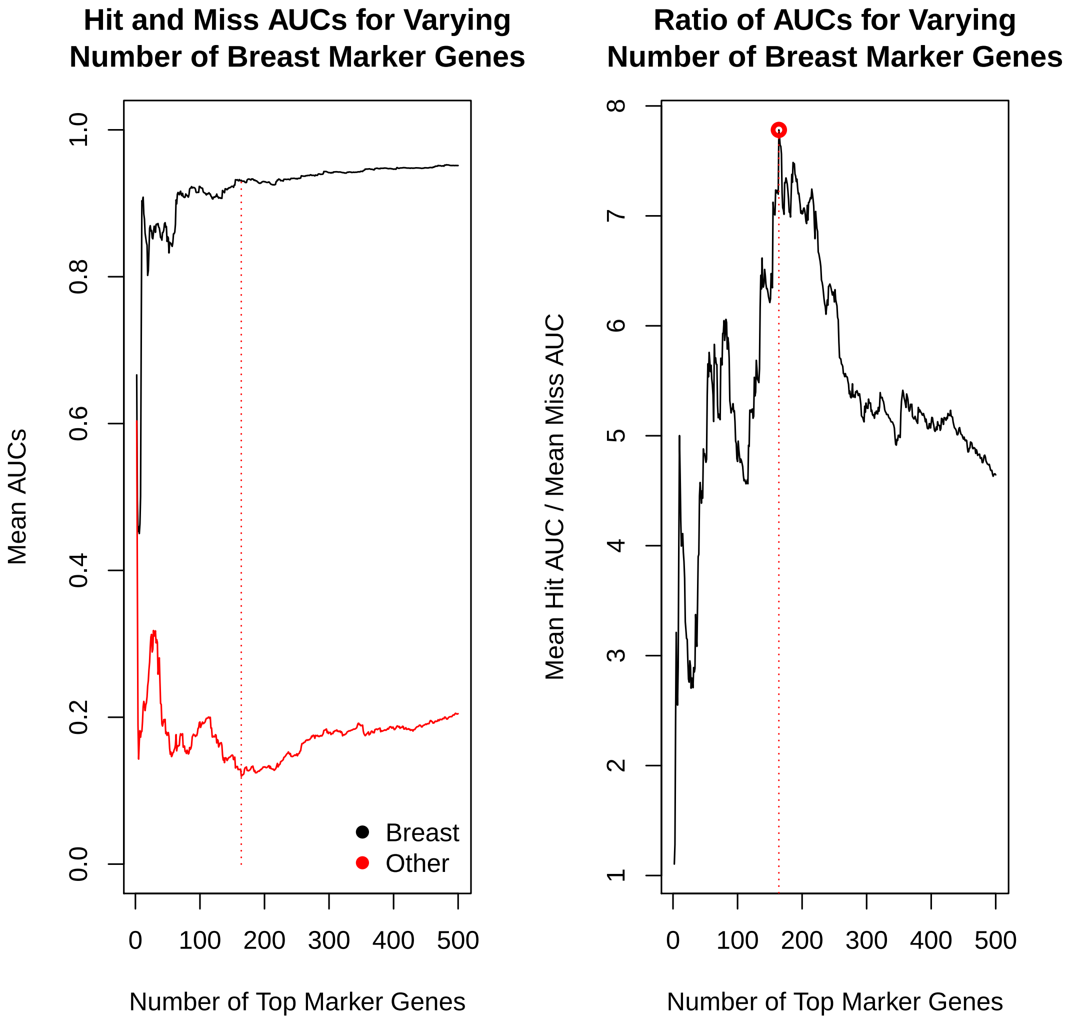


Figure 4-3: The plot on the left depicts the mean AUCs for all of the hit (black line) and miss (red line) samples for generating a marker gene set for breast tissue. The hit line is generated by taking the average of all AUCs for each sample that is a hit (breast tissue) at each of the number of top marker genes (from 0 to 500). The red line, is generated the same way for all samples that are not breast tissue. The plot on the right depicts the ratio of the average hit and miss AUC curves. The circled location is where the ratio is maximized and represents the optimal number of top marker genes to use to create a breast marker gene set.

4.2.2 Breast cancer gene set

Using the aforementioned marker gene set generation process, we derived the breast cancer gene set from a landscape of 673 samples representing 17 different cancerous tissues. The 74 genes (Table 4.1) that comprise this set are functionally en-

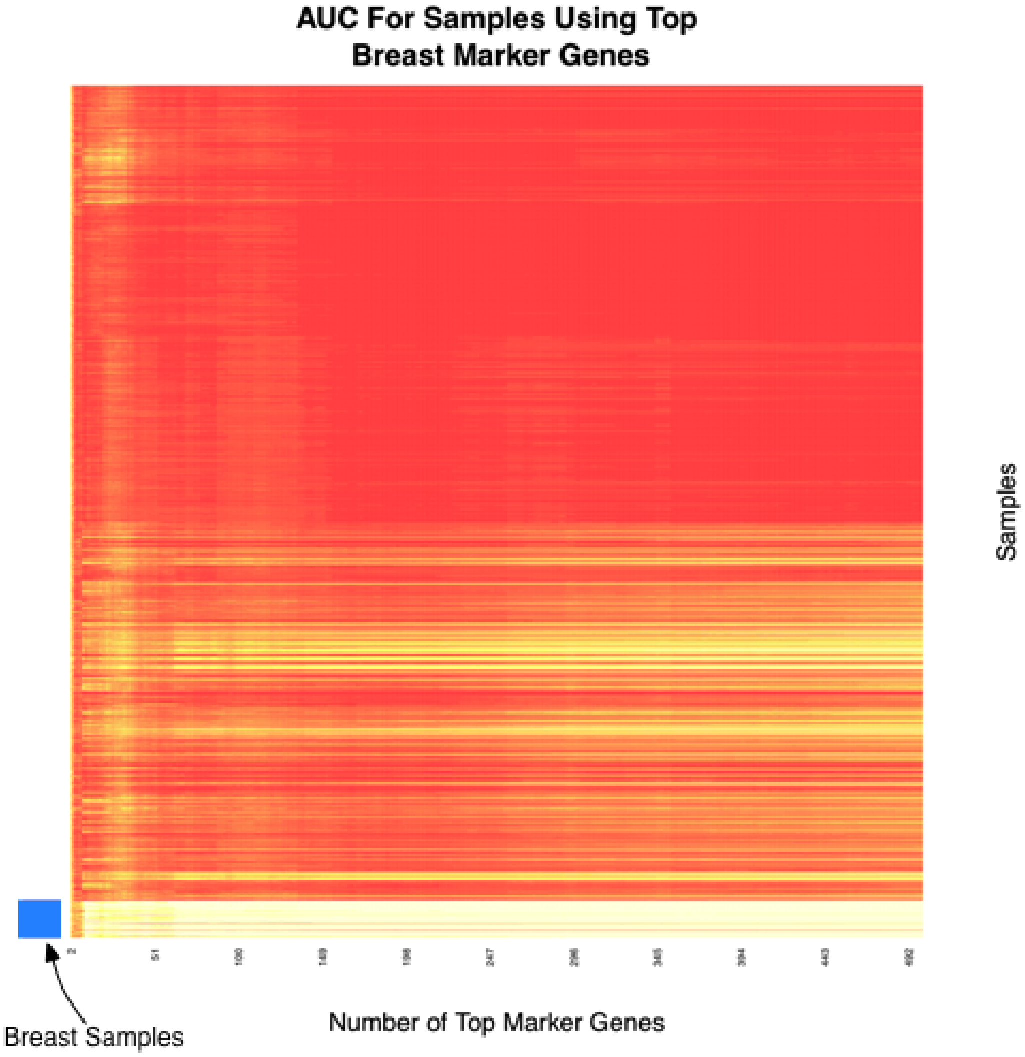


Figure 4-4: Although we use the mean hit and miss AUCs to determine the optimal number of genes to use in the gene set, this heatmap shows the AUC values for each sample (in the rows) for each number of genes used in the gene set (in the columns) such that high AUC values are yellow and low AUCs are red. The AUC values for all breast samples (denoted by the blue color on the bottom left of the plot) is high across all ranges of top number of genes to include. However, the AUCs that are initially low for the non-breast samples improve as more genes are added to the gene set.

riched for processes related to breast specific development, and carbohydrate and lipid metabolism. The complete list of over-enriched GO terms can be found in Appendix D. These pathways, revealed through gene expression, are consistent with independent clinical and genetic data suggesting an important role for carbohydrate and lipid metabolism in breast cancer. For example, women with type 2 diabetes may have higher susceptibility to breast cancer [85]. Three genes specifically implicated in this analysis, ENPP1, ADIPOQ and PPARA, are of particular interest. ADIPOQ is expressed in adipose tissue exclusively. Variants in the ADIPOQ gene and protein levels are implicated in prostate cancer [26] and breast cancer [61]. Similarly, ENPP1 levels have been correlated to progression-free survival in tamoxifen-treated patients with breast cancer [137]. PPARA is one of a family of nuclear transcription factors that has been found to stimulate both adipocyte (fat cell) differentiation and fatty acid oxidation [70]. Moreover, the PPARA signaling pathway has been implicated in breast cancer progression [119], and in a case-control study a polymorphism of PPARA was identified to be associated with a two-fold increase in breast cancer [47].

Notably missing from this list of enriched pathways are processes commonly associated with cancer, such as cell-cycle and cell-adhesion [104]. We can recreate this conventional perspective by selecting the set of candidate marker genes using a method based on a permutation t-test. We performed a t-test for each gene and computed the empirical p-value based on 1000 random permutations of the phenotype labels. As many of the p-values were 0, we sorted the list of genes by the z score of the actual t statistic as compared to the 1000 t statistics generated by the random permutations. Using this metric, we were able to sort the genes even if they had equal p values. Enrichment analysis of gene ontology (GO) [8] terms was then performed using the Bioconductor GOstats [32] library in R. This method based on the traditional permutation t-test reveals enrichment for processes that are associated with cancer in general, but not specific to breast cancer, such as "cellular response to tumor necrosis factor," "induction of apoptosis," and other tumor related processes. Furthermore, according to the permutation t-test method, PPARA is less significant than nearly 17% of the other genes (ADIPOQ is in the top 2% and ENPP1 is in the top 0.5%). In comparison, using the FIRF-based method, the tumor necrosis related genes, such as RIPK1, TRADD, and TNFRSF25, do not appear until, respectively, 18%, 54%, and 97% of the other more breast cancer-specific genes appear first.

To ascertain the "cancer" gene set using our method based on expression local-

Breast Tissue	ANKRD30A, hCG_25653, VTCN1, TBC1D9, TRPS1, SCUBE2, STC2, CCL28, KRT14, ROPN1, OXTR, SFRP1, FIGF, NFIB, ELF5, INHBB, IRX2, KRT6C, CYP4Z1, PROL1, DSG3, KRT5, IRX3, LYPD3, IRX5, PLIN, EGR2, MGP, TSHZ2, IRX1, FABP4, GABRP, MIA, SEMA3C, SAV1, TFAP2B, SERPINB5, SFN, SLC39A6, PI15, CTSO, DSC3, CX3CL1, TFAP2C, KCNMB1, DUSP4, XBP1, ANO1, ADIPOQ, AZGP1, KLK5, LEP, SCGB2A2, FXYD3, ADAMTS5, SAA2, AMIGO2, GATA3, TNN, TRIM29, RERG, GLYATL2, ALB, RPS4P13, TAT, MUCL1, FOXA1, KRT7, MUC15, PPL, SCGB3A1, FMO2, C1orf226, RPL3P7, ITGB6, KIT, PER2, LTF, C4orf7, PLAT, CIDEC, RLBP1L1, CD300LG, GRP, PLEKHG4, NTN4, SERPINA3, ZNF750, MMP7, AMOTL2, C4orf32, S100A2, AGR3, KRT6B, CITED4, TM4SF1, C10orf81, EGR3, FGF10, GRHL1, ARHGDIB, SRPX, NA, MAB21L1, KIAA1881, FMO1, GHR, EFCAB4A, C1orf116, TP63, TMC5, MYLK, AGR2, COL8A2, CPB1, CRABP2, RPL3, TAGLN, NA, ACTA2, MAPT, CREB3L4, CITED1, CRNDE, COL6A6, SCGB1D2, BNIPL, RBBP8, RPS8, SFRP2, FAT2, THRSP, NA, MPZL1, VPS8, RPL13A, CNN1, RPS10, SCN2A, ESR1, TGFBR3, IL6ST, KRT17, KLHL13, C9orf152, MEIS3P1, WFDC2, SLC16A4, SLC34A2, TM4SF18, PTPRZ1, RPS3, FOXI1, TFF3, STARD4, FAM46B, LGR6, MB, RPL10A, CRISPLD1, PIP, PTHLH, TUSC5, C16orf61
Breast Cancer Tissue	ANKRD30A, EFHD1, SCGB2A2, hCG_25653, TRPS1, PIP, CYP4Z2P, TBC1D9, PRLR, GATA3, COX6C, TFAP2B, AZGP1, SERPINA3, FLJ45983, XBP1, SPDEF, CYP4Z1, NA, NME3, MAGED2, PLIN, MUCL1, SCUBE2, TFAP2A, NAT1, DCAF10, MB, SYCP2, CCDC74B, RPS6KA3, FOXA1, RNF128, MAPT, MGP, CREB3L4, IRX5, ARSG, RABEP1, TPRG1, ENPP1, WWP1, RET, CUX1, RMND5B, FSIP1, TBX3, ESR1, ABCC11, TFAP2C, AR, SLC39A6, ACOT4, PM20D2, PIK3R3, METRN, ACADSB, C6orf211, LRRC15, ODC1, ADIPOQ, HSD17B11, COL10A1, CPB1, TMEM25, THRSP, CCDC82, HDAC11, RBM7, TTC39A, KDM4B, ERP44, PBX1, PPARA

Table 4.1: The 164 breast tissue and 74 breast cancer marker genes selected using the finite impulse response filter (FIRF) followed by setting the cutoff at the number of genes to as to maximized the sets ability to properly predict the members of its own phenotype class while minimizing the presence of non-phenotype specific signal.

ization, however, we expanded the landscape of data to include not only 17 cancers, but also 2187 samples across 30 non-cancerous tissue types. By comparing all cancers against all non-cancers, we unsurprisingly then find that the most significant genes are functionally enriched for processes that are typically associated with tumors: “cell division,” “cell cycle,” and “DNA repair,” to name but a few. Taken together, landscape-based gene signature discovery can recapitulate canonical cancer pathways, but also can identify a complementary set of gene signatures with distinct biological implications.

4.3 Tissue specific signal of tumor metastases revisited

The clinical problem of distinguishing whether a cancerous lesion represents a primary tumor, or a metastasis from a distant malignancy, presents a test case for our ability to localize a sample to the appropriate phenotypic group within the transcriptomic landscape. By combining the sample-centric concept enrichment method (Chapter 3) and gene-centric methods detailed above, we are able to map new tumor metastasis tissue samples onto the gene expression landscape, providing an unbiased measure of their phenotypic predisposition based on gene expression. It is commonly known by pathologists that tumor metastasis tissue biopsies viewed “under the microscope” resemble the tissue of the primary site rather than that of a tissue in the metastasized location. Indeed, we find that metastatic tissue samples localize in the vicinity of their tissue of origin in the transcriptomic landscape (Figures 3-12 and 4-5), even without the use of specially-tuned primary site detection methods [18, 111].

As was show in Section 3.4, when we analyze the 29 metastasized breast cancer samples resected from lung, brain, and bone (GSE14107), they more closely resemble breast tissue than their biopsy locations (Figure 4-5) in the expression landscape of all 3030 samples. UMLS concepts from Concordia that are over-enriched in the metastasized samples include “White Adipose Tissue,” “Subcutaneous Fat,” “Subcutaneous Tissue," "Lactiferous duct," "Mammary lobe," and "Glandular structure of breast." Furthermore, when we restrict the analysis to use only the 164 genes in the breast gene set identified using our aforementioned FIRF-based method, we observe that these metastasized breast samples lie within the context of other primary breast cancer samples in the database, which in turn are juxtaposed to normal breast tissue.

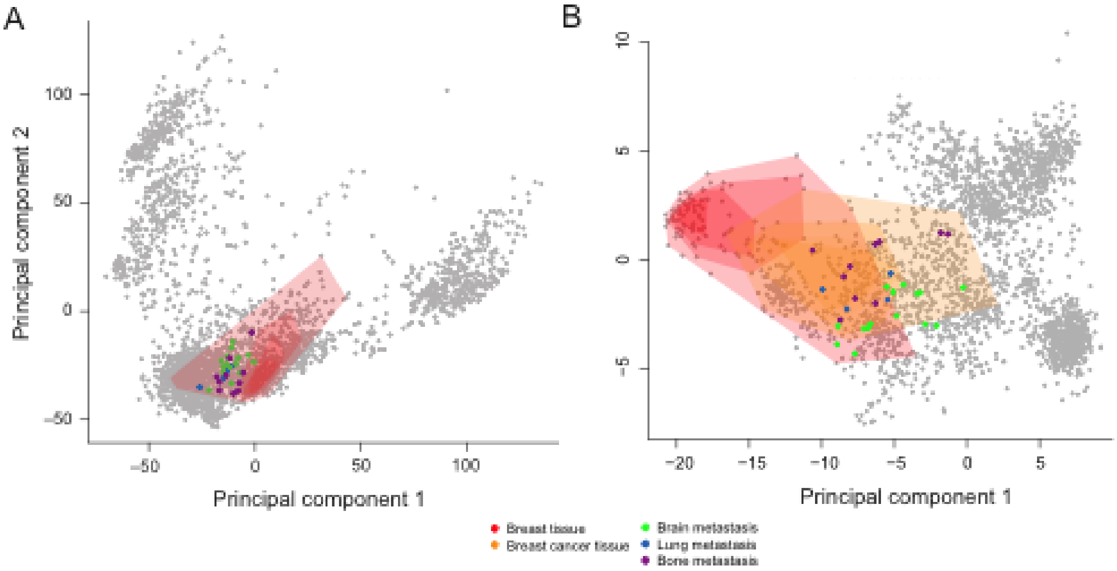


Figure 4-5: Sample- and gene-centric expression analyses show that metastasized samples more closely resemble their primary sites than their biopsy site. (A) Breast tumors that metastasized to the lung, brain, and bone (GSE14107) still appear to be more closely related to other breast samples than to their metastasis sites when placed in the transcriptomic landscape of 3030 other expression samples. (B) Recomputing the PCs using only the 164 genes of the breast gene set, as opposed to all 20252 genes, recapitulates the proximity of the metastasized breast cancer samples to breast tissue samples, and shows that they lie within the confines of the other breast cancer samples in the database.

4.4 Stem cell marker genes

Using the marker gene scores as defined before, we now turn to finding the genes that are most highly associated with stem cell activity. There have been numerous investigations, from a variety of perspectives, into the relationship between normal organogenesis programs and malignancy, particularly with respect to the stem cell properties of self-renewal and pluripotentiality [107, 114, 128]. At the molecular level, certain malignant tumors and developing tissues have been shown to exhibit shared transcription factor activity, regulation of chromatin structure and signaling characteristics [90]. Stem cell-like enrichment patterns for well-characterized gene sets have been observed in breast cancers as well as bladder cancers and poorly differentiated glioblastomas [14]. Stem cell populations have been identified that are specific to individual tissues, yet share some of the same gene expression characteristics of embryonic stem cells [140]. Similarly, diverse malignancies have been shown to share broad developmental gene expression programming [90]. Multiple controversies continue to circulate around the role of particular genes in stem cells vs. differentiated tissues (e.g. N-cadherin [71]), and the extent to which the activation of various stem cell-like programs and pathways occurs across various tissues and diseases.

The cancer stem cell hypothesis asserts a model of tumorigenesis that may tie some of these observations together. The theory suggests that only a small fraction of tumor cells (cancer stem cells) maintain the ability to self renew, with the majority of a tumors mass composed of the progeny of these stem cells, themselves lacking proliferative potential [138]. This model implies a hierarchical organization of tumorigenesis that closely reflects normal tissue development, thus accounting for the high degree of functional heterogeneity observed in solid tumors [27, 54]. Under these assumptions, expression profiles derived from resected tumor samples (comprising both the hypothesized cancer stem cells and their progeny) would be expected to broadly resemble those of the normal tissue of origin, with a degree of stem cell like activity also apparent.

Originally identified in hematopoietic cancers, leukemic stem cells were observed to express several markers in common with normal stem cells [36]. Subsequently, analogous models have been developed for a number of solid tumors (e.g., brain [121], breast [5], skin [33], ovarian epithelial [11], prostate [23], bone [43], and colon [105] cancers), primarily through the identification of a small population (typically less than 5%) of tumor cells that were unique both in their expression of a set of specific surface markers as well as their ability to induce phenocopies of their original tumors in xenograft and transplant models.

Although the cancer stem cell model and the experimental approach to identifying cancer stem cell populations have been replicated across a variety of tissues, the exact molecular signatures derived from the proliferative cells have varied widely. As yet, the extent to which there exist any molecular fingerprints commonly attributable to multiple types of cancer stem cells remains unclear. For example, leukemia stem cells have been identified by a CD34$^+$CD38$^-$ phenotype shared with hematopoietic stem cells [74], while brain cancer and colon cancer stem cells have been isolated among CD133$^+$ cells [105, 121]. Breast cancer stem cells have been defined by a CD44$^+$CD24$^-$ phenotype [5], while prostate cancer stem cells have been isolated from minority CD44$^+$/$\alpha_2\beta_1^{hi}$/CD133$^+$ populations [23]. Bone sarcoma cells with proliferative potential have been shown to express activated Stat3 [43]. These cells also expressed a subset of the embryonic stem cell-associated genes (Oct3/4, Nanog), but again, the degree to which these trends may be apparent across other populations of cancer stem cells is unknown [142].

4.4.1 Creating the stem cell marker gene set

Using the method based on the finite impulse response filter (FIRF) explained in Section 4.1, we identified a set of genes with highly specific stem cell expression intensities. Previous studies have examined the expression patterns of literature-curated gene sets relating to embryonic stem cell-like activity among a variety of malignancies [14]. In contrast, we have constructed a gene set in silico that reflects only those transcriptional signals with the greatest ability to localize the stem cell samples within the spectrum of human tissues and diseases.

A variety of thresholds were evaluated according to the ability of the implied gene sets to differentiate between stem cell samples and the other phenotypes in the dataset via an analysis of variance (ANOVA). For each possible number of top-scoring stem genes from 3-502 (displayed at the top of Figure 4-6), we project all of the samples in the database into the first two principal components of gene space (panel on top right), and highlight in color 6 relevant phenotypes (as in Figure 4-7): embryonic /

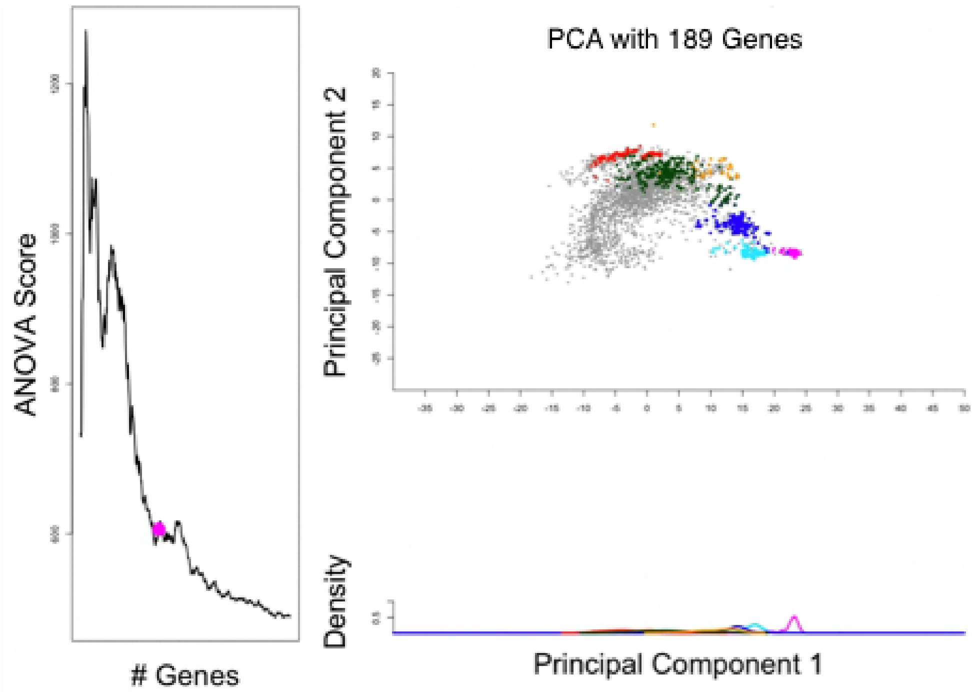


Figure 4-6: A variety of thresholds were evaluated according to the ability of the implied gene sets to differentiate between stem cell samples and the other phenotypes in the dataset via an analysis of variance (ANOVA). Here we show the result when the 189 genes that are part of our stem cell gene set are used. For each possible number of top-scoring stem genes from 3-502 (displayed at the top of the figure), we project all of the samples in the database into the first two principal components of gene space (panel on top right), and highlight in color 6 relevant phenotypes (as in Figure 4-7): embryonic / induced pluripotent stem cells in magenta; mesenchymal stem cells in cyan; immortalized cell line samples in blue; blood precursor cells in orange; leukemia samples in green; normal blood in red. The panel below the PCA scatter plot shows the distribution of stemness index values (PC1 projection coordinates) for each highlighted phenotype. The plot on the left of the frame shows the ANOVA score (including all highlighted phenotypes) for the clustering defined by the current stemness index highlighted by a magenta dot on the curve showing all ANOVA scores for all of the depicted FIRF thresholds. Higher ANOVA scores indicate better multi-way separation of the individual phenotypes along the stemness index. The genes presented here represent a set capable of simultaneously separating the pluripotent, multipotent, progenitor, malignant and normal samples, while also retaining tissue-specific features (e.g., clearly separating normal blood, neural and epithelial tissues)

induced pluripotent stem cells in magenta; mesenchymal stem cells in cyan; immortalized cell line samples in blue; blood precursor cells in orange; leukemia samples in green; normal blood in red. The panel below the PCA scatter plot shows the distribution of *stemness index* values (PC1 projection coordinates) for each highlighted phenotype. The plot on the left of the frame shows the ANOVA score (including all highlighted phenotypes) for the clustering defined by the current stemness index highlighted by a magenta dot on the curve showing all ANOVA scores for all of the depicted FIRF thresholds. Higher ANOVA scores indicate better multi-way separation of the individual phenotypes along the stemness index. The genes presented here represent a set capable of simultaneously separating the pluripotent, multipotent, progenitor, malignant and normal samples, while also retaining tissue-specific features (e.g., clearly separating normal blood, neural and epithelial tissues). Here we used an ANOVA as opposed to the previously introduced "hit" vs. "miss" AUC ratio as we were trying to maximize the difference in not just two clusters (hit and miss) but rather six.

After various iterations, we found that using the top 189 stem cell marker genes yielded the best results. Henceforth, these 189 genes will be referred to as the stem cell gene set and the complete list of genes can be found in Tables D.3 - D.6 in Appendix D. While we will not delve into the topic here, we shall see how these stem cell marker genes can be used to find potential drugs that effect cell-cycle (Section 6.3.1).

4.4.2 Stem-like signature stratifies a diverse expression database by pluripotentiality and malignancy

Via principal component analysis (PCA), we examined the transcriptional profile of the stem cell marker genes across the entire collection of normal tissues, cancers and stem cells. Performing PCA across only the stem cell marker genes (including all samples in the data set) allowed us to measure the extent to which the specific transcriptional activity observed in the stem cell population was apparent in each of the other phenotypes.

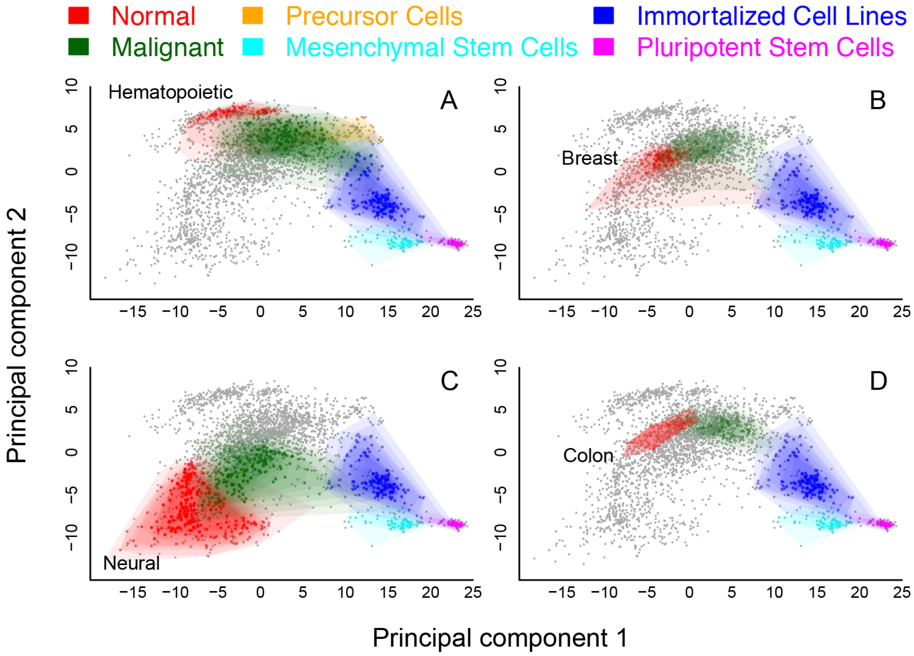


Figure 4-7: The stem cell signature genes stratify a phenotypically diverse database according to pluripotentiality. Each panel shows the entire expression database plotted on the principal coordinates defined by the stem cell signature genes. PC1 is represented on the x-axis of each plot, while PC2 is on the y-axis. In each plot, the pluripotent stem cells (IPS and ES) are clustered on the extreme right-hand side (magenta), followed by mesenchymal stem cells (cyan) and immortalized cell lines (blue). Taken together, the panels demonstrate that, across tissue types, this stem cell signature draws a coherent picture of pluripotentiality and differentiation. While the distinction between the pluripotent stem cells and normal tissues represents the predominant signal (PC1) in the data, the contrast in the expression profiles of hematopoietic and neural tissues apparently defines the second strongest signal (PC 2). Even so, both tissues respective malignancies show a common tendency to exhibit greater stem-like activity, as demonstrated by their closer proximity to the pluripotent stem cell cluster (A, B, C, D) Blood, breast, neural and colon all demonstrate the same enhanced stem-like expression activity among their respective malignancies.

This analysis revealed a striking trend apparent in the first two principal components (PCs) of the gene set; most importantly, PC1 captured a measure of cellular potency, while PC2 reflected the broad transcriptional differences between hematopoietic, neural and epithelial tissues. These trends are demonstrated in Figure 4-7. Each panel highlights in color the PCA region occupied by a particular normal tissue population (red) and its associated malignancies (green), as well as any related precursor cells (orange), immortalized cell line samples (cyan), multipotent (blue) and pluripotent stem cells (magenta) (PCA was computed jointly across all samples; each cancer is highlighted individually for clarity). The pluripotent stem cells included in this analysis were a combination of both embryonic stem cells and induced pluripotent stem cells. The locations of all other samples in the data set are shaded gray to provide context.

The dominant characteristic of PC 1 is its ability to separate the pluripotent stem cells from the normal tissue samples (e.g., the normal tissues shown in Figure 4-7 blood, breast, brain, colon, shaded red, consistently lie on the extreme left side of the plots, whereas the pluripotent stem cells, shaded magenta, lie on the extreme right). Moreover, PC1 apparently reflects a finer-grained continuum of cellular potency: the multipotent stem cells are clustered near the pluripotent stem cells, with the hematopoietic progenitors (the only progenitors in our dataset) slightly farther away.

Further, the hematopoietic, neural and epithelial cancers (shaded green in Figure 4-7) contained in our data all clustered directly between the stem cell populations and their associated normal non-malignant samples. This suggests that the stem cell marker genes captures a kernel of stem cell-like transcriptional activity that is concurrently apparent in a variety of malignancies. These findings build on previous observations that genes associated with stem cell-like activity demonstrate differential expression in a variety of epithelial cancers with respect to their normal tissue counterparts [140]. Our analysis reveals that stem-like expression profiles are observable not only in epithelial cancers, but also in neural and hematopoietic malignancy as well.

We will use the coordinates of an expression profiles projection into the first principal component of the gene space defined by the stem cell marker genes as a relative measure of “stemness”, our *stemness index.*

4.4.3 Functional diversity of the stem cell gene set

Hierarchical clustering of these genes transcriptional activity in a population of pluripotent stem cells revealed four distinct coexpression modules. For each module, we then identified a set of over-enriched GO biological processes [8].

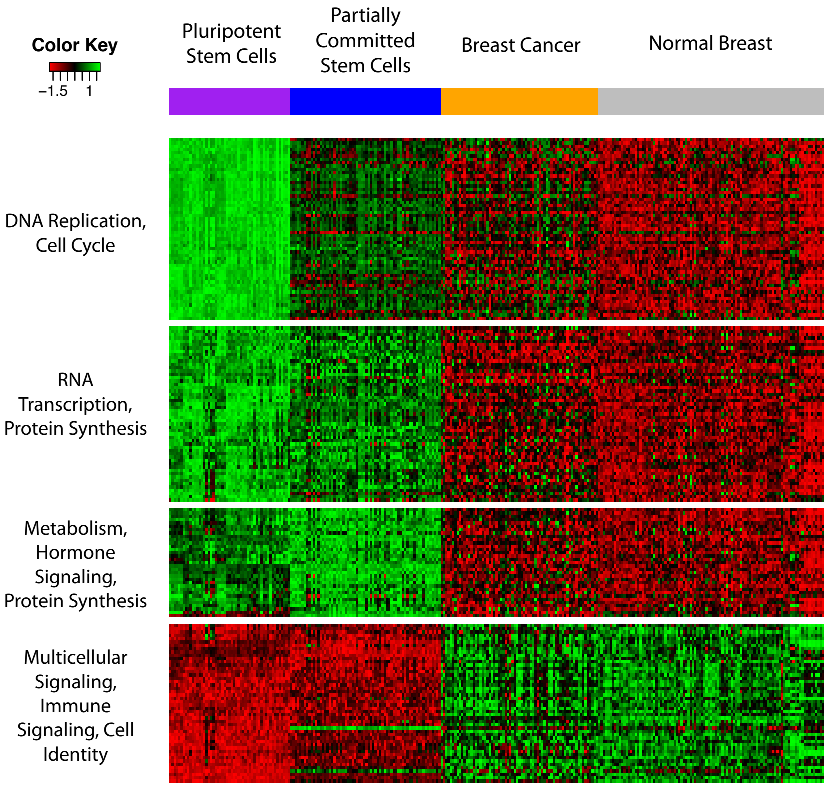


Figure 4-8: Four distinct expression modules (row clusters) are apparent within the stem cell genes. To demonstrate the transcriptome-wide implications of these profiles, this figure displays a series of cell types, ranging from fully differentiated (normal breast), through the associated malignancy, partially committed stem cells, and pluripotent stem cells. Each gene (row) has been independently z-score normalized to improve readability and highlight cluster-specific trends. Biological significance of each cluster was determined by GO analysis (see Tables D.7 - D.10 in Appendix D). The individual genes represented in each cluster can be found in Tables D.3 - D.6 in Appendix D.

To illustrate the gene expression trends apparent within each gene cluster, Figure 4-8 shows a heatmap of their profiles across pluripotent and partially committed stem cells, as well as malignant and normal breast samples. Genes active in DNA replication, cell cycle regulation and RNA transcription (see Tables D.7, D.8 in Appendix D) are most highly expressed in the pluripotent stem cells, and less so, respectively, through increasing levels of cellular differentiation / decreasing pluripotentiality, consistent with prior studies of the dynamics of stem cell cycling and regeneration [101, 130]. Genes related to metabolism and hormone signaling (Table D.9) show peak expression intensity among the partially committed stem cells, while exhibiting low intensity among the fully differentiated tissue and tumor samples. Correspondingly, genes responsible for multicellular signaling and cellular identity (Table D.10) are most highly expressed in the fully differentiated tissue and malignant samples. Within each functional module, the tumor samples trend away from the respective normal tissue, echoing stem cell-like transcriptional activity.

4.4.4 Grading of tumors

We used the stemness index that we derived from the stem cell gene set to evaluate the transcriptional profiles of several graded tumor data sets. Our goal was to evaluate whether our molecular marker for tissue-agnostic stem cell-like transcriptional activity was representative of poor clinical prognosis. We included four publicly-available data sets in this analysis. For each data set, we computed the samples stemness index (via PCA over the stem cell gene set) to identify the dominant differences between the samples within the context of the stem cell genes.

We identified four independent data series containing expression profiles for graded tumors of various tissue types in GEO (GSE4290, GSE23593, GSE17537, GSE18842) on Affymetrix HG-U133 Plus 2.0. Each series was pre-processed (MAS5.0 normalized, summarized) as previously described. Within each series, the stem cell gene set summary values were computed, again, via PCA over this gene set, allowing us to associate a value with each sample indicating its relative stem-like expression activity.

This analysis revealed that our stemeness index correlates with tumor grade for a variety of primary tissues. Figure 4-9 shows the distribution of stemness index values for the four tissue types graded tumor samples. In each case, the transcriptional activity of the stem cell gene set defines a clear separation between the high- and low-graded tumors, while also providing a molecular foundation based on stem-like

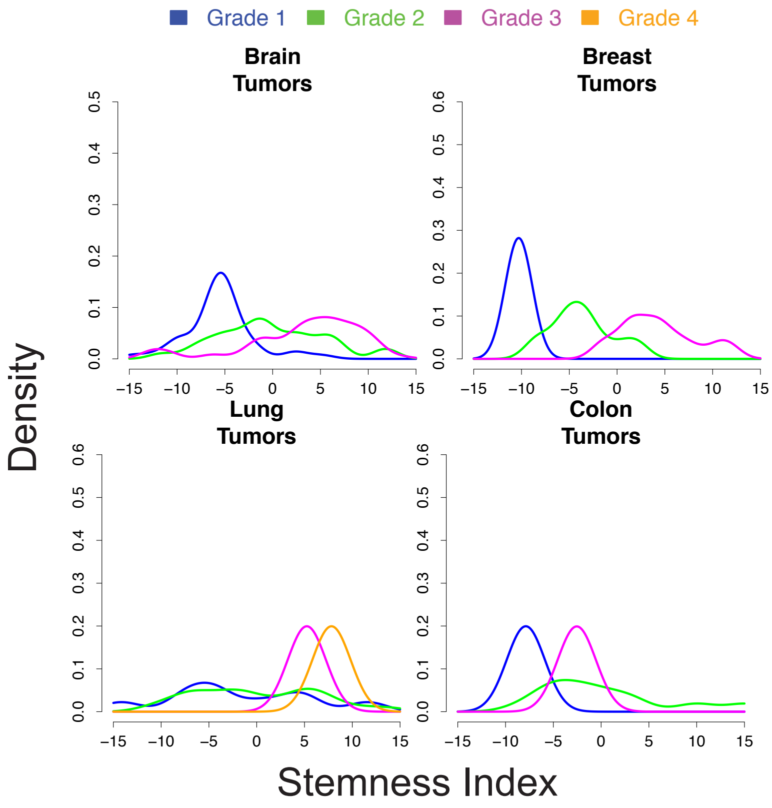


Figure 4-9: Stem cell-like activity correlates with tumor grade in various solid malignancies. Each panel displays the distribution, within the space of the stem cell genes, of graded tumor samples for one particular tissue type. Our stemness index consistently separates high-grade tumors from low grade ones. Based on this transcriptional index, the mid-grade tumors are less well defined.

expression for the clinical difficulty in classifying mid-grade tumors. [40, 134]

4.4.5 Biological implications

The increasing volume of evidence supporting a pervasive connection between cancer and stem cells suggests significant therapeutic implications. Current therapies are evaluated based on their ability to reduce the overall size of a tumor. Regimens that target cancer stem cells, however, may have more success in preventing long-term recurrence [138]. Molecular signatures that are capable of grading pluripotentiality and proliferative potential represent an important step in designing such regimens and guiding therapeutic procedures.

Indeed, gene expression signatures derived from breast cancer stem cells have been shown to separate patients with early-stage breast cancer into high-risk and low-risk groups [76]. Similar methods with broad applicability will pave the way for individually tailored treatment strategies. Diverse malignant tissue samples have been shown to exhibit a broadly similar trend within a large gene expression database, but no specific connection has been made in this context to stem cell-like activity [77]. Identifying an unbiased transcriptional measure of stemness conserved across embryonic and adult stem cells, and relating that signature to malignancy, has remained a challenge [38, 101, 140].

While a large volume of evidence indicates that only a small number of tumor cells are capable of self-renewal, controversy remains as to the exact origin of these cells. The hierarchical cancer stem cell hypothesis suggests that these cells arise from normal pluripotent or multipotent stem cells that have lost the ability to regulate their proliferative activity. Under this model, the phenotypic diversity observed in many tumors is viewed as the result of this defective stem cell population mismanaging the process of normal organogenesis. Alternatively, the stochastic model of tumorigenesis suggests that proliferative tumor cells arise from normal fully differentiated or committed progenitor cells that acquire the ability to self renew [90], and that tumor cell phenotype variation is the result of these mutated cells differentiating in a random fashion [50].

Regardless of the origin of proliferative tumor cells, our results indicate that there is a high degree of stem cell-specific gene expression programming observable in heterogeneous tumor samples. Our data indicates the need for more detailed transcriptional assays comparing proliferative tumor cells to both ES / iPS cells and bulk heterogeneous tumor cells, as well as normal tissue cells. Our data suggests the hypothesis that the gene expression patterns observed in heterogeneous tumor samples may be due to the effect of a small population of cancer stem cells in combination with a large number of partially differentiated cells. It is plausible that, while the partially differentiated mass of the tumor behaves transcriptionally similar to healthy tissue, the small population of proliferative tumor cells push the observation of the aggregate mRNA back along the spectrum of stem cell-like activity.

Chapter 5

Data begets data: Efficiently expanding an existing curated expression database

Building large, highly curated databases of gene expression data is a task that has been attempted by many. Although some have the financial resources to generate new data [2], the repurposing of existing gene expression data sets is the more common (and economical) route. In general, this curation effort takes two forms: highly specific sample curation for a specific analysis or classification task (e.g. [53, 113, 115, 119, 124, 133]), or the accumulation of a vast array of samples for the purpose of having a diverse dataset to perform various data mining procedures (e.g. [3, 19, 39, 56, 63, 77, 81, 116]). Indeed, our work here falls into the latter category, but has the potential to be used for the former.

An ideal expression database is one that can be searched for a phenotype (or a set of phenotypes) to obtain the expression vectors for all samples relating to that phenotype. The National Library of Medicine's Gene Expression Omnibus (GEO) [13] does allow for searching by phenotype[1], but it simply returns webpages of dataset as opposed to pre-processed data that can be used as-is in an analysis. While it is a

[1] Searching GEO works like any other search – one types in some keywords and the user is given a page with all of the results that are deemed to be related to the search query.

simple matter to write a program that obtains the datasets that are deemed relevant by GEO's search logic (GEO provides FTP access to the raw expression data), one of the largest stumbling blocks to the repurposing of existing datasets is the lack of standardized nomenclature when defining the content of the data series and samples [19]. This means that multiple queries have to be made for a given phenotype in order to obtain complete coverage. Althought the use of the standardized concepts in the Unified Medical Language System (UMLS) [17] does circumvent this issue to a large degree, the automated labeling procedure using the MetaMap [7] program (or any other NLP based program) is unfortunately plagued with both false positives and false negatives (refer to Chapter 2 as to how we used MetaMap and the UMLS ontology to build our Concordiafied version of GEO). Thus, as also described by Butte et al. [19], manual intervention is necessary to verify the labelings. Regrettably, this process is both time consuming and error prone.

The beauty of working with expression data, however, is that not only do we have the text that describes the experiment that was performed, but also the actual expression intensity values for each of the samples. As we showed in Chapter 3, we can take new expression samples, and with a high degree of accuracy, label it with its UMLS concepts by just using the expression data. Taking this into account, could we not combine both textual and expression information when querying for samples of interest? Although previous work shows that it is possible to query for expression samples using an expression vector as input [39], here we focus on the task of efficiently expanding a curated database. By combining the contextual information provided by the text and hints to the underlying process provided by the gene expression data we can expand an existing curated database very quickly.

5.1 Seeing is believing: Active learning

The fundamental problem is that building a curated database is a tiresome endeavor. Furthermore, it is often the case that when building such a database not all phenotypes are equally important. For example, if a research group is currently studying autism, it would much rather add a lot of autism samples to the database than a random selection of samples that cover a wide range of phenotypes. Taking this into account, we want to be able to increase the size of a curated database as quickly and as efficiently as possible. Assuming that every entry into the database has to be manually verified for accuracy by an "expert"[2], the most efficient way to add all samples for a given phenotype is to return all of the samples that have yet to be curated for that phenotype. Clearly, if we had a highly trained classifier (whether it uses just text, just expression, or a combination of both) that can do this for us, that would be great. Unfortunately, creating such a classifier requires training data – both positive and negative examples that can be used to teach the classifier as to what truly is an autism sample.

This is where the supervised learning paradigm of *active learning* can be applied. The active learning framework makes use of samples that have previously been correctly labeled to guide the labeling of future samples. Rather than having an expert annotate all samples (since we assume that to annotate all samples for a given phenotype, this expert must have also had to annotate, or skip over, many samples that were not for this phenotype) to create a highly accurate classifier, we seed the classifier with a handful of annotated samples, and then have it improve its understanding of the phenotype as more samples are annotated. Therefore, at first the classifier may return poor results, but as the expert curates the database, (s)he will be provided with better and better samples (i.e. at some point every sample being looked at should be of the phenotype in question). Not only does this improve annotation speed, it also ensures a higher degree of accuracy.

Traditionally, active learning is used to reduce the amount of data that needs to be labeled to train an accurate classifier. Active learning is comprised of repeatedly training a classifier, selecting the next set of samples to annotate based on the classification results, annotating those samples, and the retraining the classifier with the updated training set to repeat the procedure again (see Algorithm 3). At each round,

[2] It probably wouldn't be too far from the truth to state that in most cases this expert is a new graduate student.

the set of samples that are chosen to be labeled are removed from the unlabeled set ($T_{U,i}$) and then added to the labeled set to be used for the next round of learning ($T_{K,i+1}$). For example, Singh et al. [120] used active learning to reduce the number of microarray experiments that need to be performed for time-series experiments. As our task is to build a *large*, curated database, unlike these previous works, however, we employ active learning not to reduce the number of samples to be labeled to create a great classifier, but rather to reorganize the samples to be presented to the expert labeler in an order that is conducive for fast and accurate labeling. For a full review of active learning and its applications see the *Active Learning Literature Survey* [117].

Algorithm 3: Active learning

```
Input: $T_U$: The set of unlabeled gene expression samples
Input: $T_{K,0}$: A small set of known labeled samples
Input: $T_T$: A set of testing data used to evaluate the trained classifier
Input: $C$: A classifier
Output: $T_K$: The set of labeled samples
for $i$ in 0... do
    $train(C, T_{K,i})$
    $evaluate(C, T_{K,i})$
    $scores_i \leftarrow score(T_{U,i})$
    $T_{C,i} \leftarrow choose(scores_i)$
    $T_{C*,i} \leftarrow label(T_{C,i})$
    $T_{U,i+1} \leftarrow T_{U,i} \setminus T_{C*,i}$
    $T_{K,i+1} \leftarrow T_{K,i} \cup T_{C*,i}$
```

5.2 The baseline: What do we have to beat?

Before we continue, an important question to ask is whether or not we need to perform active learning at all. Since we are already using MMTx [7] to annotate the free text for each sample with its corresponding Unified Medical Language System (UMLS) [17] concepts, if a user is interested in concept c, can we just assume that MMTx is correct and return all samples annotated with c to the user? Although there is noise associated with the labelings produced by MMTx, they are not all incorrect (if they were, why would anyone use MMTx?).

When we annotated 3030 gene expression samples from NCBI's Gene Expression Omnibus (GEO) [13] with MMTx (as explained in Chapter 2), there were a total of 2267 unique UMLS concepts. The annotations associated with these samples were then manually verified and about a third of them (633 unique concepts) were kept. We also added 39 new concepts that MMTx did not return for any of the 3030 samples. Although these concepts were then "mapped back up" the ontology[3] in the Concordiafied version of GEO that we used in the analyses of the previous chapters, here we just use the raw results returned by MMTx without any additional processing.

Using just the 633 concepts that were kept after manual verification, we can then compute the accuracy of MMTx in regards to its labeling of the text. Four common statistics that are measured are *sensitivity*, *specificity*, *precision*, and *recall*. Sensitivity is defined as the fraction of actual positives which are correctly identified as positives (e.g. the fraction of all breast samples that were annotated as being from breast tissue). We compute this value as follows:

$$\mathit{sensitivity} = \frac{TP}{TP + FN} \tag{5.1}$$

where TP is the number of *true positives* (the number of samples that were correctly labeled) and FN is the number of *false negatives* samples (the samples that were incorrectly not labeled with the concept but should have been). Sensitivity is also know as the *true positive rate*. Specificity, on the other hand, is the fraction of negatives which are identified as negatives (e.g. the fraction of samples that are not breast samples, that were identified as not being breast tissue samples) and is computed as

[3] Recall, that if a sample was annotated with "breast cancer" we then also annotated that sample with all ancestor concepts in the UMLS ontology. Continuing the example, the sample would then also be annotated with concepts such as "malignant neoplasm" and "disease." See Chapter 2 for details.

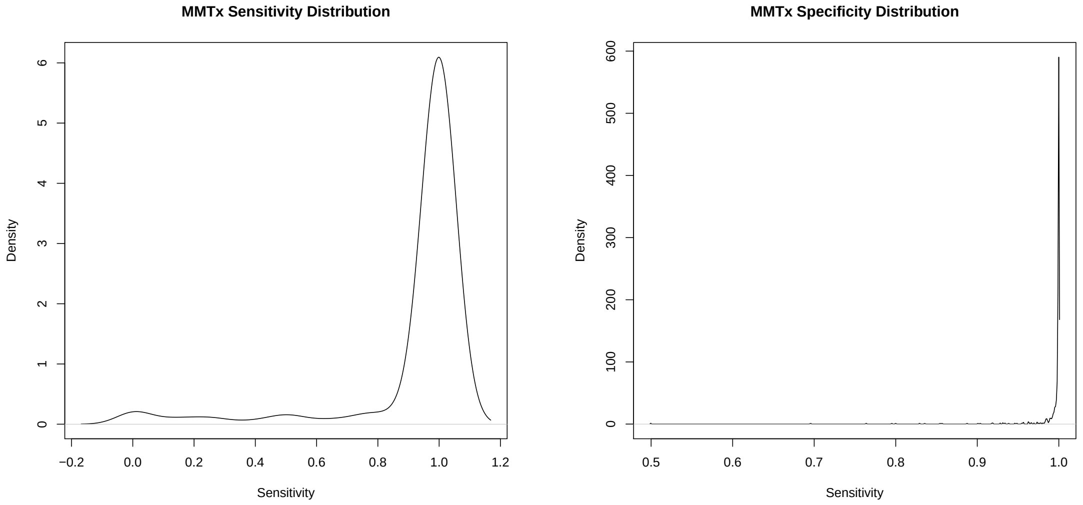


Figure 5-1: The distribution of sensitivity and specificity values of the 633 UMLS concepts that were kept. We see that both of these values are very hight.

follows:

$$specificity = \frac{TN}{TN + FP} \tag{5.2}$$

where TN is the number of *true negatives* (the number of samples that were correctly labeled as *not* having the phenotype) and FP is the number of *false positives* samples (the samples that were incorrectly not labeled with the concept with they should not have been). Sensitivity and specificity are also referred to as *type I* and *type II* errors.

Closely related to sensitivity and specificity are precision and recall. Precision is the fraction of results that are actually correct positives. It is computed as follows:

$$precision = \frac{TP}{TP + FP} \tag{5.3}$$

where TP is the number of true positives and FP is the number of false positives. Recall, on the other hand is the fraction of results that are actually correct and have not been missed.

$$recall = \frac{TP}{TP + FN} \tag{5.4}$$

where TP is the number of true positives and FN is the number of false negatives.

As confirmed by Butte, et al [19], we see that MMTx generally labels samples

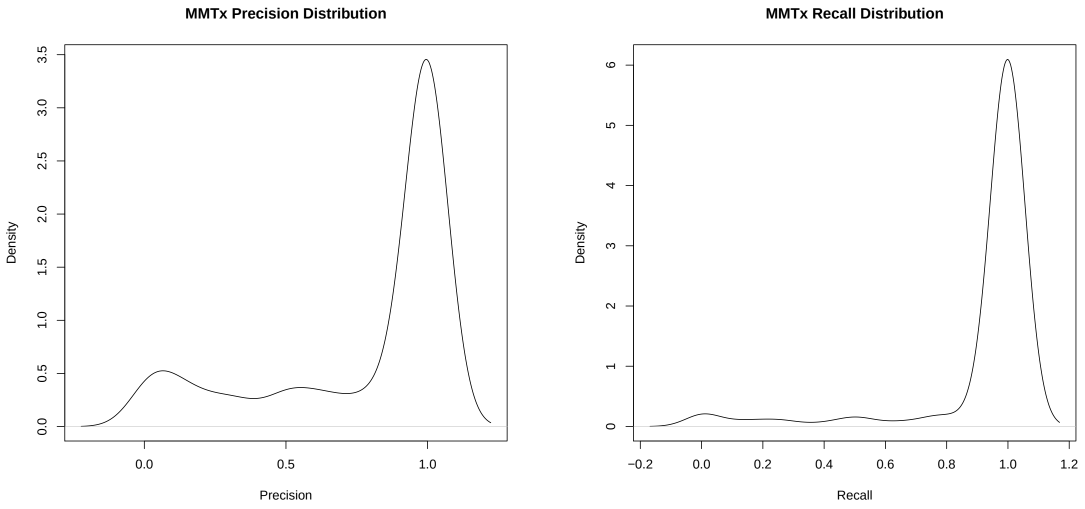


Figure 5-2: The distribution of precision and recall values of the 633 UMLS concepts that were kept.

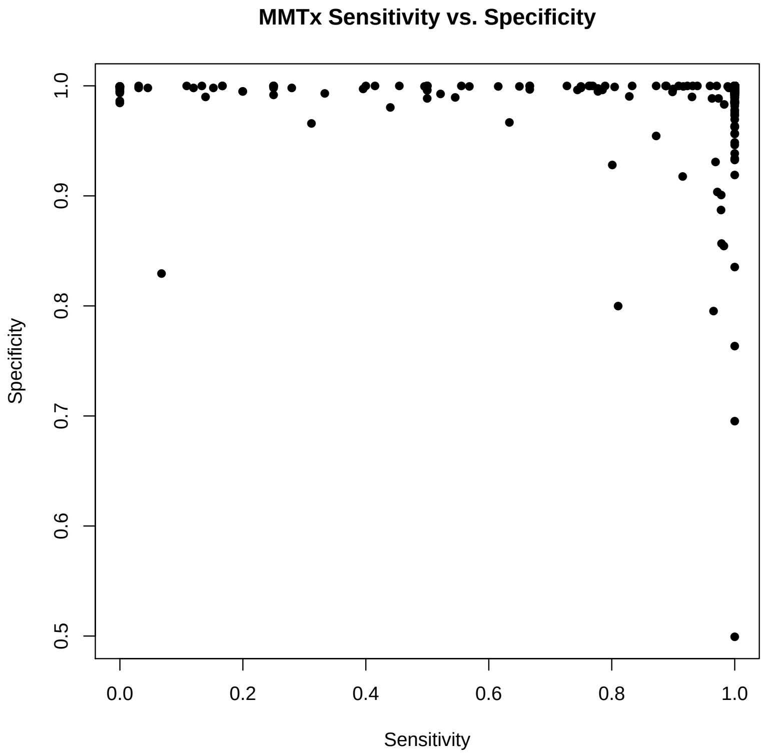


Figure 5-3: The MMTx sensitivity and specificity values for the 633 UMLS concepts that were kept. Table 5.1 contains the concepts with low sensitivity values.

correctly assuming that the text was there to begin with. As shown in the density estimates of sensitivity, specificity, precision, and recall in Figures 5-1 and 5-2, we see that a significant number of the concepts annotations were accurate. This tells us that by just parsing the text available to us using MMTx, a large portion of the hard work has been completed for us. Of course, we must not forget that we are only measuring the results for the 633 UMLS concepts that we deemed to be relevant after having removed 1634 irrelevant concepts. When we plot sensitivity vs. specificity for each of the 633 concepts (Figure 5-3), we again see that the majority of the concepts perform well. What we see from the concepts that had a low sensitivity (Table 5.1) is that many of the are concepts that can be easily obtained incorrectly. For example, "Adult" has a sensitivity of 0.25. In several of the texts that contain the word "Adult," the words "child" (which maps to the concept "Childhood" with sensitivity of 0.1) or "Adolescent" (sensitivity of 0.17) are also used. For instance, an excerpt from the series description for GSE2842 is (bold face added for clarity):

> Glucocorticoids (GC) are in most chemotherapy protocols for lymphoid malignancies, particularly **childhood** acute lymphoblastic leukemia (ALL) for their ability to induce apoptosis in malignant blast. The underlying mechanism, however, has so far only been investigated in model systems. ... For comparison, expression profiles were generated from an **adult** ALL patient, peripheral blood lymphocytes from GC-exposed healthy donors, GC-sensitive and -resistant ALL cell lines and mouse thymocytes treated with GC in vivo and in vitro.

Not only do we have to contend with contradictory concepts (such as "adult" and "childhood") as indicated above, shorthand or uncommon abbreviations cause other erroneous mappings. For instance, a nightmare scenario would be a sentence like, "This experiment compared breast cancer (BC), lung cancer (LC), and prostate cancer (PC) samples to matched normals." Running this through MMTx results in the concepts in Table 5.2.

As one can see, all of the correct concepts are identified by MetaMap ("Normal," "Malignant neoplasm of breast," "Malignant neoplasm of lung," "Malignant neo-

Concept	Sensitivity
B-Cell Lymphomas	0.4
Leukocytes	0.395833333333333
Glucocorticoids	0.333333333333333
Human cells	0.311688311688312
leukemia	0.279569892473118
Adult	0.25
Brain Neoplasms	0.25
Glioma	0.25
Lymphoblastic Leukemia	0.25
Lymphoma, Diffuse	0.25
Diffuse Large B-Cell Lymphoma	0.25
Head	0.2
Neck	0.2
Adolescent	0.166666666666667
Precursor Cell Lymphoblastic Leukemia Lymphoma	0.166666666666667
Acute leukemia	0.152173913043478
Blood specimen	0.13953488372093
Malignant neoplasm of stomach	0.133333333333333
ovarian neoplasm	0.12
Childhood	0.108695652173913
Homo sapiens	0.0679287305122494
Myeloid Leukemia	0.0454545454545455
Leukemia, Myelocytic, Acute	0.0307692307692308
Malignant neoplasm of skin	0.0307692307692308
Human tissue	0
Cultured Cells	0
Depletion	0
prednisolone	0
Whole blood sample	0
Infection	0
Injury	0
Cancer of Neck	0
Cancer of Head	0
Pediatric	0
chronic	0
Myeloid	0
Pregnant - adjective	0
Monozygotic twins	0
Hypertensive disease	0

Table 5.1: The concepts and sensitivity values for the UMLS concept that had a specificity of 0.4 or less.

To - dosing instruction fragment	To
Matches	Normal
Specimen from breast	Malignant neoplasm of breast
Breast	Entire breast
Malignant neoplasm of lung	Malignant Neoplasms
Primary malignant neoplasm	Bicarbonates
Malignant neoplasm of lung	Lung
Entire lung	Specimen from prostate
Malignant neoplasm of prostate	Prostate
Entire prostate	Palmitoylcarnitine
Lecithin	Phosphocreatine
Phosphatidylcholines	

Table 5.2: The UMLS concepts as a result of running MMTx on the sentence, "This experiment compared breast cancer (BC), lung cancer (LC), and prostate cancer (PC) samples to matched normals." Although there are many correct concepts, there are also many spurious ones.

plasm of prostate," "Breast," "Lung," and "Prostate"). However, unless each of the individual samples in the series has the correct phenotypic label describing which of the labels is correct, which it usually does not, we have no way to tell which sample corresponds to which phenotype. This sort of scenario is, unfortunately, the norm and not the exception.

5.3 Expanding the database using only text

Having examined the accuracy of the MMTx [7] as a labeling engine, let us turn our attention to using those UMLS [17] concept annotations produced by MMTx for the task of building and expanding a curated database. As this section only deals with the text itself, these results are applicable to a broad audience (not just those building an expression database). As is common with text based classification (e.g. spam filtering), we will assess the performance of a naïve Bayes classifier when applied to this task.

5.3.1 Brief introduction to naïve Bayes classifiers

Although not necessary for the understanding of the results below, here we present a (very) brief (and non mathematical) introduction to naïve Bayes classifiers. Interested readers should refer to the book *Introduction to Information Retrieval* [80] for a detailed (and mathematical) discussion of the topic.

A naïve Bayes classifier is a probabilistic classifier that is based on Bayes theorem with a strong assumption of independence. Unlike other probabilistic models such as a Bayesian network, each of the features in the model are completely independent. For example, if one aims to classify whether a given vehicle is an Aston Martin DB5 (James Bond's car), the features that may be included in the model may include: number of wheels, size of wheels, number of doors, size of engine, etc. Thus, when creating a naïve Bayes classifier, we assume that all of these features are completely independent. In other words, the number of wheels that a vehicle has has no relationship with the number of doors. While this strong assumption of independence is usually not accurate[4], it is a good approximation that works well in many areas (especially text classification, such as spam filtering [25]).

To train the classifier, we simply provide it with positive and negative examples to learn from. Each of the examples is presented to the classifier as a *feature vector*. A feature vector is a list (a vector) of values for each of the features that we have data for (the number of doors, size of wheels, etc.). The parameters of the model (the probability that the vehicle is a DB5 given the number of doors, etc.) are then estimated by *maximum likelihood estimation* (MLE). Simply stated, MLE tries to compute the probability of a vehicle being the DB5 given the current data. Thus, if we had 5 examples for the "number of doors" parameter as shown in Table 5.3 the probability of being a Aston Martin DB5 is 50% if the vehicle as 2 doors (it's either the DB5 or the M3) and 0% if it has any other number of doors. Once we have trained the classifier and all of its parameters, we show it a new feature vector and

[4] If the vehicle has two wheels it most likely has no doors (it's a motorcycle), if it has four wheels it most likely has two or four doors (a regular car), and if it has 18 wheels it probably has 2 doors (a truck). Of course, in this example, any vehicle that doesn't have four wheels can immediately be assumed not to be the DB5.

it computes the probability that that car as described by that feature vector is an Aston Martin DB5.

Car	No. of Doors
Aston Martin DB5	2
BMW M3	2
VW GTI	3
BMW 7 Series	4
Toyota Prius	4

Table 5.3: Example input for a naïve Bayes classifier parameter. After seeing these 5 examples, the classifier would say that the probability of being a Aston Martin DB5 is 50% if the vehicle as 2 doors (it's either the DB5 or the M3) and 0% if it has any other number of doors.

5.3.2 Learning from text

Now that we have a basic understanding of how the classifier works, we show how it is applied to the task of classifing whether a sample is indeed related to the phenotype of interest (e.g. is it really a lung tissue sample) or not. We define the feature vector of a sample by the indicator vector of UMLS concepts that the free text was associated with. Each entry in the vector corresponds to whether or not the given sample was annotated with that concept by MMTx, and is set to 1 if it was and 0 if it was not. For example, Table 5.4 depicts two samples and their corresponding indicator values for six concepts. Sample s_0's free text was annotated with concepts c_0, c_1, and c_3, while sample s_1 was annotated with c_0, c_3, c_4, and c_5.

We tested the efficacy of performing active learning on merely the concepts that each sample was associated with by employing a naïve Bayes classifier. For example,

	c_0	c_1	c_2	c_3	c_4	c_5
s_0	1	1	0	1	0	0
s_1	1	0	0	1	1	1

Table 5.4: An example of two samples and their indicator vectors corresponding to which concepts they have been labeled with. Sample s_0's free text was annotated with concepts c_0, c_1, and c_3, while sample s_1 was annotated with c_0, c_3, c_4, and c_5.

can we look at all of the concepts and predict whether or not a sample is a breast cancer sample. Thus, using the nomenclature of the active learning algorithm described in Algorithm 3, each round of learning consists of training a naïve Bayes classifier with the current set of labeled concept indicator feature vectors, evaluating the classifier on a held out testing set for classification performance, and then choosing a set of new samples to label. These labeled samples are then removed from the unlabeled set and placed in the labeled set for the next iteration of learning.

5.3.3 Scoring strategies

The hardest part of performing effective active learning is in the step where the set of samples to be labeled is chosen. In order to have active learning be useful, we need to pick the samples that will help us the most in future rounds of active learning. In order to understand the efficacy of active learning, we implemented several scoring functions. Although a typical use-case is to return one sample at a time at each iteration, we chose to return five at a time.

Maximum (minimum) entropy

Given a classifier, the maximum entropy scoring method scores the samples closest to the decision boundary the highest. For instance, this scoring method has previously been used for statistical natural language parsing [131]. For example, assume that we have a two-class classification problem such that the output is either "positive" or "negative." The sample that should be chosen next is the one that the classifier is the most unsure of; the one closest to the classifier's decision boundary. This "closeness" to the decision boundary can be summarized by the following entropy (H) function.

$$H(s) = -p(pos) \times \log(p(pos)) - p(neg) \times \log(p(neg)) \tag{5.5}$$

If the probability of being in either the "positive" or "negative" case is high, then entropy will be low. On the other hand, if the probability of being in either one is 50%, then the entropy will be high. Thus, by picking samples that have a high entropy, we are guaranteeing that at each iteration the sample that causes the most confusion for the classifier is labeled to be added to the next iteration's training set.

We also implemented the minimum entropy function. Unlike the maximum entropy function, the samples that the classifier is most confident about at each iteration are chosen to be added to the next iteration's training set. Although this function may be the incorrect choice when attempting to minimize the number of labeling steps required to maximize the classifier accuracy, it is more in line with the underlying requirements of a large-scale annotation effort. In other words, the minimum entropy function will return the samples that the classifier is most confident about, and thus provide us with the set of most-likely-to-be-correct samples that should be annotated.

MMTx labeling

Instead of using a naïve Bayes classifier to compute which samples should be labeled based on an entropic measure, we can trust the labeling of MMTx to guide us in the process of making the choice. Intuitively, if a sample has a given concept, then it is much more likely to actually being whatever that concept says it is than if it were not labeled with it. Thus, by simply returning high scores for samples that have been labeled with the concept for which the classifier is being turned for, we are returning MMTx's best guess as to how to train the classifier.

For this strategy we also implemented two variations. The first was to simply return all MMTx labeled samples (subsequently denoted as *allsamples*) and the second was to return three samples corresponding to the MMTx label along with two random ones (subsequently denoted as *mostwithconcept*). Although it might have been a bit obscure as to why we chose to return five samples at each iteration, it was to allow for a bit of randomness in this step. In other words this *mostwithconcept* strategy is akin to letting the method explore random parts of the text space.

Random

Finally, in order to measure the performance of the aforementioned scoring strategies, we also implemented a random scoring function that arbitrarily scores each sample. This amounts to ignoring the active learning framework completely and provides a baseline to compare the other metrics against.

5.3.4 Quantifying performance

To measure the performance of the active learning procedure using the various scoring schemes the following was performed: for a given concept c, 10 samples, five that were known to be positively associated with the concept of interest and five that were not, were randomly chosen and used for the initial starting set ($T_{K,0}$). At each of the 500 iterations, the classifier was trained on the set of samples with known labels ($T_{K,i}$). The samples with unknown labels ($T_{U,i}$) were then each scored and the top five highest scoring samples were then chosen for subsequent labeling. We ran this for each UMLS concept that we had data for.

In order to ensure reproducibility, we performed cross validation on this active learning approach. We randomly split the data into 75% training and 25% testing sets 5 times. All of the active learning was performed on only the training set. The performance of at each iteration for all of the scoring schemes was measured using only the testing set. We also ensured that for each split of the data, the starting set of 10 samples was identical for each scoring strategy.

5.3.5 Performance

Classifier tuning performance

In an attempt to minimize confusion, first let us first examine the actual performance of the naïve Bayes classifier. In this scenario, the aim is to maximize the predictive performance of the classifier itself. In other words, we are attempting to answer the question, “which samples should I annotate next to build the best classifier?” Note, that this does not answer the original aim of "which sample is most likely one that I am looking for?" However, it is clear that this can be viewed as a dual problem. If we have a good classifier that can predict whether or not something is, for example a lung sample, then it will be easy to return the next lung sample to be labeled. However, we don't actually want to spend the time to build this classifier (since we do not want to waste time training it), and would rather just have a black box that gives us the best one to annotate next, regardless of how it affects the performance of the underlying classifier.

Figure 5-4 depicts the performance of a naïve Bayes classifier across 500 iterations of active learning. We use the *F-measure* to report a single value for the performance. The F-measure is the harmonic mean of precision and recall:

$$F = 2 \times \frac{precision \times recall}{precision + recall} \tag{5.6}$$

As to be expected, the scoring strategy that performs the best (i.e. shows the most improvement in the performance of the classifier during the active learning) is the maximum entropy method. Since this strategy continually picks the samples that are most confusing the the classifier, it becomes easier and easier as more iterations are performed. On the other hand minimum entropy and *allwithconcept* perform the worst since they continually return concepts that the doesn't help the classifier learn something new. Interestingly the *mostwithconcept* does a lot better than the previous two; much more on par with random.

Not only is the performance of the text based classification method underwhelming, it does not really address the underlying problem of sorting the samples in the most favorable way such that labeling becomes easier. Since we are building a classifier at each iteration of learning, we are actually tuning a classifier instead of tuning the sorting method. In other words, this active learning procedure attempts to make the best naïve Bayes classifier at predicting whether, given just the concept indicator vector, a sample is associated with a certain concept with the fewest number of samples. Although choosing samples closest to the decision boundary is typically what

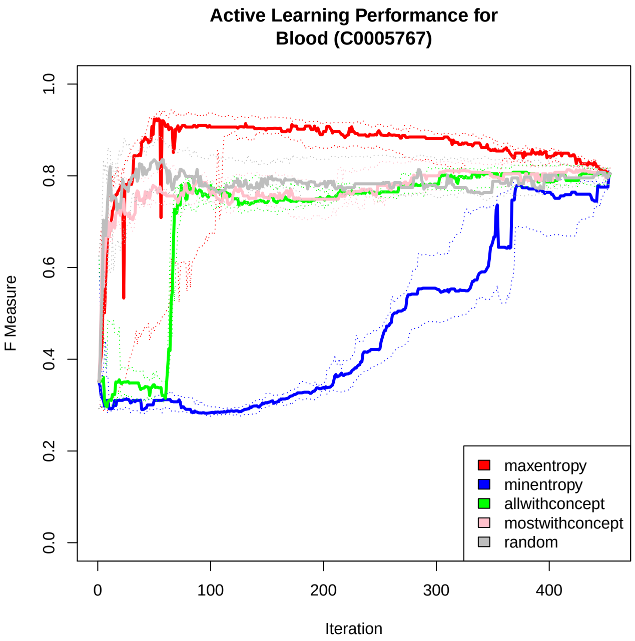


Figure 5-4: The naïve Bayes classifier performance for various sample choosing metrics for blood samples. Here we see that the only sample scoring method that performs better than random is maximum entropy. Although this is only one example, the maximum entropy strategy was generally the best method for all other concepts as well.

one wants to do when trying to improve the classification accuracy of classifier, in our application of active learning, we *do not* want to tune a classifier as we are only looking to increase the size of the database as efficiently and with as few errors as possible. In other words, we want to do exactly the opposite of the optimal solution of finding the samples near the decision boundary (maximum entropy) and return the samples that we are most sure about.

Database labeling performance

Unlike the previous section which attempted to show the improvement of the classifier performance, we now turn to examining the performance of how well each of the scoring strategies fares when we attempt to minimize the number of iterations needed to label the samples of interest. Again, here we do not care about how well the underlying classifier performs, but rather want to minimized the amount of work

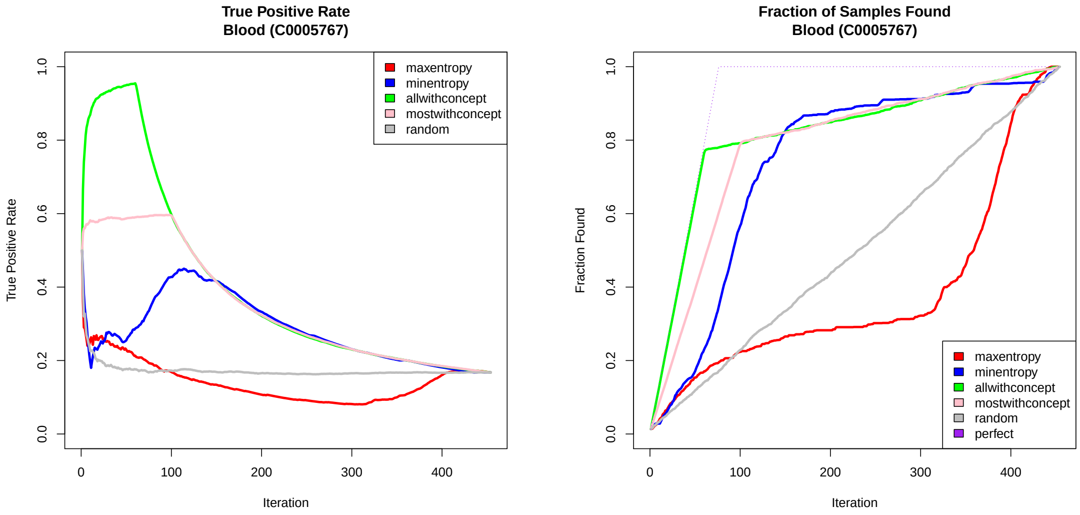


Figure 5-5: The performance of the various scoring schemes at the task of returning samples to be labeled under the assumption that the minimum number of iterations should be performed. (a) shows the true positive rate across 500 iterations while (b) shows the fraction of blood samples that were find by the i-th iteration.

the user must perform. What we see is that the methods that performed poorly in training the classifier, performed well under this metric, and vice versa (Figure 5-5).

For example, using just the results from MMTx and returning those samples, outperforms both methods that attempt to learn what it means to be a blood sample. In other words, attempting to use the other UMLS concepts associated with a sample to try to predict whether or not a new sample is a blood sample performs worse than just assuming that a sample is a blood sample if MMTx labeled it as such. Furthermore, the method that works best at building the best classifier the quickest (*maxentropy*) performs the worst in this scenario. This makes intuitive sense as returning the samples that the classifier is least sure about (the ones with the maximal entropy) are the least likely ones to be blood samples.

As can be seen in Figure 5-5(b), the largest drawback to using only the text based data is that any sample that is not labeled by MMTx to be of a given type, will most likely not be picked up. For example, by about the 80th iteration the method that simply returns the blood samples labeled by MMTx (*allwithconcept*) goes from being nearly perfect to random.

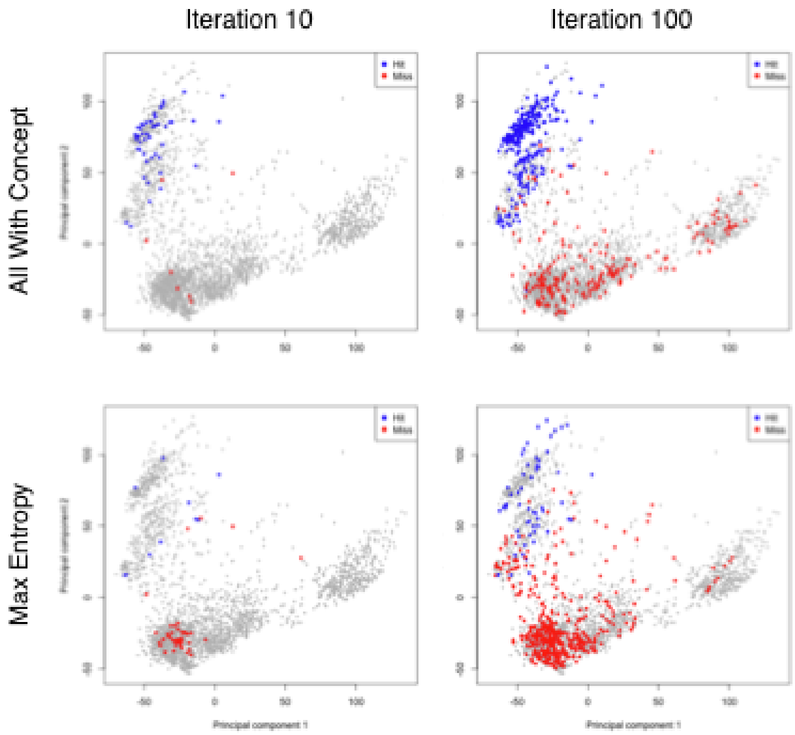


Figure 5-6: We can visualize the performance of the labeling performance of the various scoring schemes by viewing the labeling progress on the transcriptomic landscape. Here we show the labeling progress at iteration 10 and 100 for *allwithconcept* and *maxentropy*. We clearly see that *allwithconcept* does a much better job of picking the blood samples (see Section 3.2.1).

We can visualize the performance of the labeling performance of the various scoring schemes by viewing the labeling progress on the transcriptomic landscape. Recall from Section 3.2.1 that the upper left hand cluster of the transcriptomic landscape corresponds to blood tissue samples. Figure 5-6 shows the labeling progress at iteration 10 and 100 for *allwithconcept* and *maxentropy*. We clearly see that *allwithconcept* does a much better job of picking the blood samples. Although we only use the expression information here as a way to visualize the performance of the scoring methods, the following section will cover a method to include the gene expression information to enhance the labeling procedure.

5.4 Expanding the database using text & expression data

Although the training of a classifier to improve its classification accuracy, such as the naïve Bayes classifier in the previous section, is the general use-case for active learning, it is not exactly what we are after. As aforementioned, we are looking for an optimal sorting such that an expert labeler has the minimum amount of work to do. While this sorting can be achieved with a traditional classifier by ordering all unlabeled samples by they classification confidence, as in the use of the previously detailed minimum entropy function, we will be examining non-classifier based methods. In addition to the sorting rather than classifying paradigm shift, gene expression data will be incorporated to aid in the sorting process.

The two main ways that one could use multiple sources of information to generate a sorting (or a classification) is either by performing the sorting (classification) independently for each source of information and then combining them or by first combining the sources of information together into one rich feature vector and sorting (classifying) the data based on that single feature vector. Both methodologies have their relative merits. The former "feature poor" method, allows one to independently use each source of information and use a domain specific sorting (classification) engine for the particular source of information. However, this strength is also its weakness as it requires the tuning of the weights used to combine the independently generated results and does not allow any possible correlations between the sources of information to "boost" the signal during the sorting (classification) process. On the other hand, the "feature rich" method uses a single sorting (classification) methodology on one large feature vector allowing the algorithm to "decide" what the relative merits of each feature are. This method, however, requires the combination of all data sources into one feature vector as input to a single sorting (classification) method, and is difficult to apply when the sources of information come from vastly different domains (such as text and gene expression, for example).

5.4.1 Scoring strategies

To include both textual and expression data in labeling data, we score each data type (text and expression) independently and then compute a weighted score based on a linear combination of the two scores. The score for each sample is computed as follows:

$$Score = \alpha \times Score_{expression} + (1 - \alpha) \times Score_{text} \tag{5.7}$$

where α is the weight used to tune the importance of the expression data. As we shall see, each of the two score components ($Score_{expression}$ and $Score_{text}$) range between 0 and 1, and thus the final combined score also ranges between 0 and 1.

Text based scoring

As we saw in the previous section, a classifier based text scoring method (e.g. using a naïve Bayes classifier) did not yield favorable results in minimizing the number of labeling rounds to perform. Thus, we limited our scoring metrics here to two that were based on the MMTx derived concept annotations. The first scoring method is identical to the one used in the previous section. Namely, if a sample was annotated by MMTx to be associated with a concept, it was given a score of 1; if not labeled with the concept, it was given a score of 0 (scores ranged between 0 and 1 where 1 is the highest possible score). If there were multiple samples with the same score, a sample was simply chosen at random among them. We call this scoring metric is *binary*.

The second variation that was tested was a weighted version (aptly called *weighted*) where the weights were based on all of the concepts associated with the samples that had been previously annotated. The score of a new sample was 0 if it was not labeled with the concept of interest (e.g. if we were looking for brain samples and the current sample was not annotated with the concept for brain by MMTx). If it was labeled with the concept of interest, then the score was a value between 0 and 1 depending on how similar all of the concepts that it was labeled with were to concepts of those samples that had already been labeled in previous iterations. Formally, the distance between a new sample and a previously labeled sample was computed as the Manhattan distance[5] of the of the samples' concept indicator vectors (see Table 5.4). This distance was computed to all previously labeled samples, and the score of the new sample was the mean of all of the Manhattan distances. This method, while still only yielding non-zero scores for samples that have actually been annotated by MMTx with the concept of interest, allows for a ranking of samples depending on the other concepts it was labeled with.

Expression based scoring

The expression based score for each sample is based on its correlation to the samples that have already been labeled. The correlation was computed using all 20252 genes measured on the Affymetrix HG-U133 Plus 2.0 array. Thus, samples having a high correlation to the previously labeled samples with the phenotype of interest have higher scores than those that have a lower correlation. We shall call this group of samples that have previously been labeled to be associated with phenotype p of interest as S_p.

As there is always more than one sample in S_p (we start with five positively labeled samples), we must summarize the distance of each unlabeled sample s, to S_p. Two summarization methods were chosen: *mean* and *centroid*. When using the *mean* method, we simply take the mean distance of s to all samples in S_p. Alternatively, when using the *centroid* method we first find the sample in S_p that is closest to the *centroid* of S_p and then compute the distance of s to that "centroid" sample[6].

[5] The Manhattan distance a version of Euclidean distance and is defined as $\sqrt{\sum_{i=1}^{n} x_i - y_i}$ where x and y are the input vectors. Unlike Euclidean distance, the $(x_i - y_i)$ is not raised to the power of two. However, since we are applying this Manhattan to binary indicator vectors, the Manhattan distance is equivalent to the Euclidean distance.

[6] The true centroid of a cluster (in this case S_p) is the point that minimizes the distance to all points in the cluster. Since this is an artificial point and may not actually exist, we find and use the point in the cluster that is closest to the centroid as the centroid.

5.4.2 Quantifying performance

To measure the performance of the active learning procedure using the various scoring schemes the following was performed: for a given concept c, 10 samples, five that were known to be positively associated with the concept of interest and five that were not, were randomly chosen and used for the initial starting set ($T_{K,0}$). At each of the 2000 iterations, the "classifier" was trained on the set of samples with known labels ($T_{K,i}$). The samples with unknown labels ($T_{U,i}$) were then each scored and the highest scoring sample was then chosen for labeling.

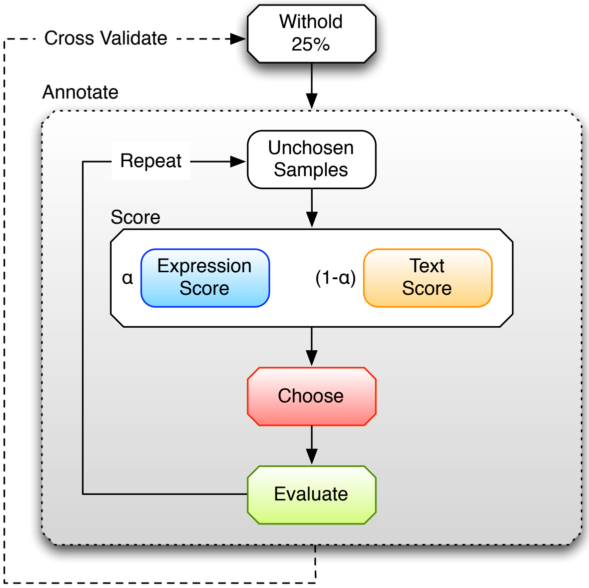


Figure 5-7: To ensure reproducibility of the results we performed cross validation. In each of the runs, 25% of the data was withheld and used as a testing set; the remaining 75% of the data was used to perform the active learning. At each iteration in the learning process, we scored the unchosen samples using a weighted score that was based on both the expression data's signal and the signal from the text. The highest scoring samples were then chosen and labeled before repeating the learning step again.

In order to ensure reproducibility, we performed cross validation on this active learning approach. We randomly split the data into 75% training and 25% testing sets 10 times. All of the active learning was performed on only the training set. The performance of at each iteration for all of the scoring schemes was measured using only the testing set. We also ensured that for each split of the data, the starting set of 10 samples was identical for each scoring strategy.

5.4.3 Performance

As the goal of the labeling task is to be able to label all samples in the database with a given phenotype as quickly and efficiently as possible, we compute the true positive rate (sensitivity) and the fraction of positive samples found at each iteration. The true positive rate at iteration i shows how correct we have been at picking out the samples of the phenotype of interest p in the last i iterations. For example, if we have picked 10 samples related to p of interest in 10 iterations, then our true positive rate at iteration 10 is 1. The fraction of samples found shows how quickly we find all the samples associated with p. If we have found half of the samples that are truly associated with p at iteration 20, then the value for the fraction of samples found at iteration 20 is 0.5.

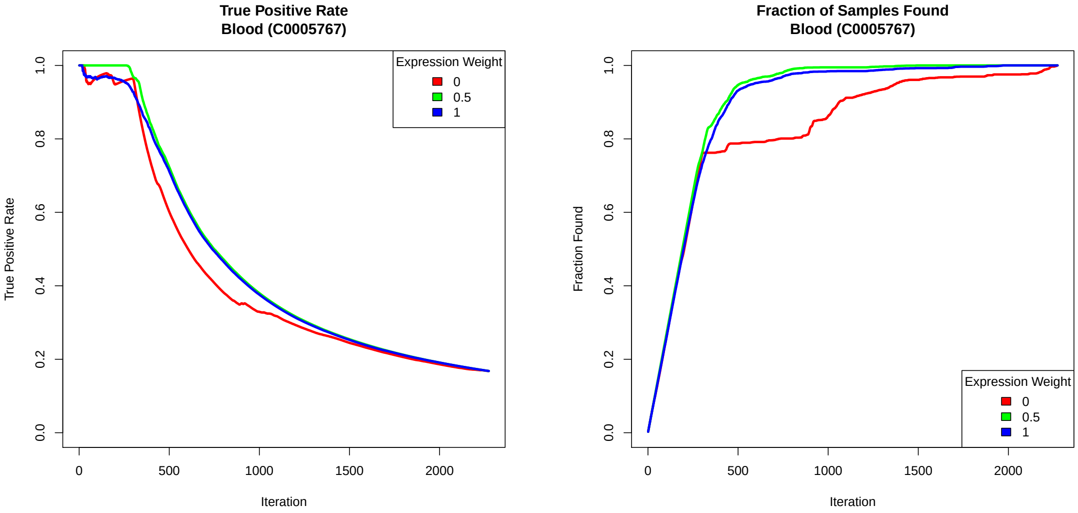


Figure 5-8: Database labeling performance for blood samples when using text and gene expression information. Figure (a) shows the true positive rate, while (b) shows the fraction of samples found. The three curves are the average across 10 cross validation runs when we set α, the weight of the expression signal to 0 (red), 0.5 (green), and 1(blue). When the expression data is included (when α is 0.5 or 1), we are able to label all of the blood samples in the database much faster.

Figures 5-8 - 5-10 depict the sensitivity and the fraction of samples found for three phenotypes, blood, liver, and lung, when using the *binary* method for the textual

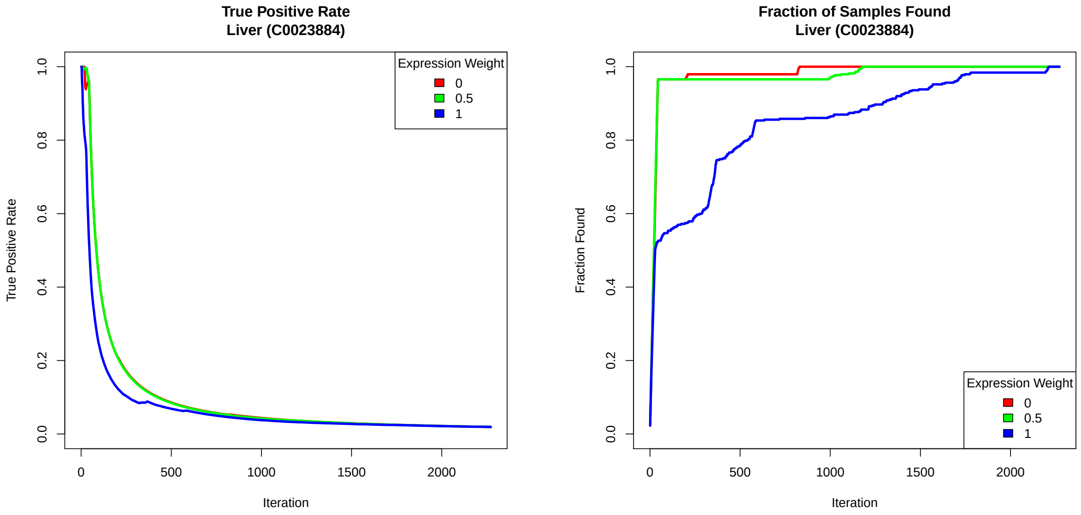


Figure 5-9: Database labeling performance for liver samples when using text and gene expression information. Figure (a) shows the true positive rate, while (b) shows the fraction of samples found. The three curves are the average across 10 cross validation runs when we set α, the weight of the expression signal to 0 (red), 0.5 (green), and 1(blue). Unlike the case with the blood samples (Figure 5-8), just using the expression data (α set to 1) results in poor labeling performance.

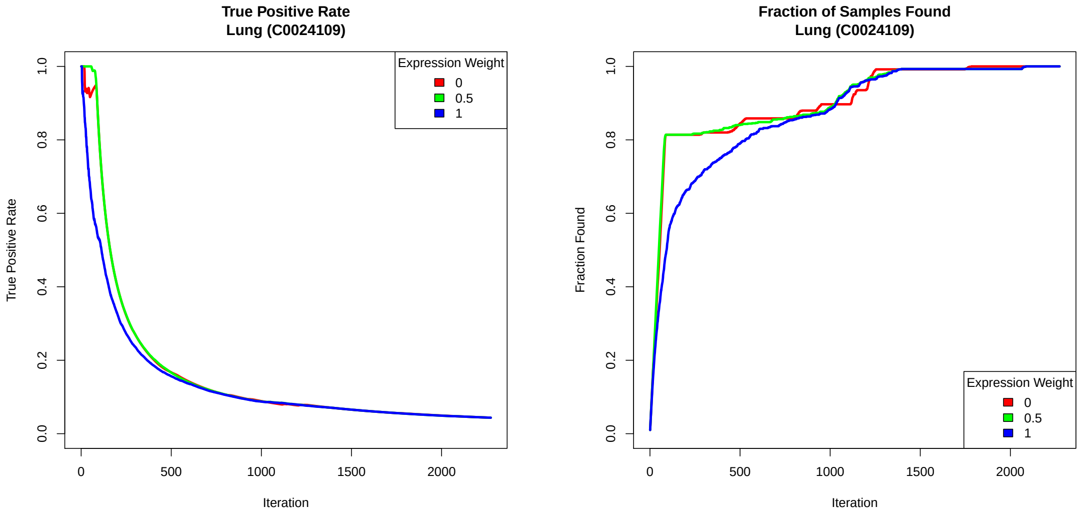


Figure 5-10: Database labeling performance for lung samples when using text and gene expression information. Figure (a) shows the true positive rate, while (b) shows the fraction of samples found. The three curves are the average across 10 cross validation runs when we set α, the weight of the expression signal to 0 (red), 0.5 (green), and 1(blue). Here is an example were all there methods perform comparatively.

data and the *centroid* method for the expression data. Each of the lines in the figures represents the average true positive rate (sensitivity) or average fraction of samples found across the 10 cross validation runs for three different expression weights (α set to 0 (red), 0.5 (green), and 1(blue)). We find that depending on the phenotype, the utility of the expression data signal varies widely. For example, when labeling all of the blood samples in the database (Figure 5-8), the expression signal provides the majority of the signal such that we find almost all blood samples in about 500 labeling iterations when we set α to 0.5 or 1. On the other hand, when labeling liver samples (Figure 5-9), the expression signal alone does not perform well. Labeling the lung samples (Figure 5-10) provides us with an example where the textual information is initially much better than the expression data, but after about the 700th labeling iteration, they behave almost identically.

Interestingly, we obtain the best of both worlds when we set α to 0.5. This suggests, that by combining the textual information with the biological signal from the expression data, we can quickly and efficiently label new samples to grow an existing curated expression database. Neither source of information results in consistently good results when used independently, but when combined we are able to leverage the strengths of each one. While we only show the results for one of the scoring method combinations, *binary* concept vector and *centroid* based expression distance, the results for the other combinations of scoring schemes behaved similarly.

Chapter 6

Drug similarity: A transcriptomic view

High costs in drug design and development have resulted in the use of computational methods to help reduce both production time and cost [132]. For instance, Pfizer recently created a tool that combines multiple sources of data to "visualize" the drug target landscape in order to generate therapeutical hypotheses about chemical compounds [20]. Other researchers have taken multiple sources of drug data to predict drug-drug relationships [49] and provide possible new indications for existing drugs [122].

The use of gene expression data in providing targeted therapeutics or showing the effect a drug has on a certain disease has also become common place. For example, gene expression based high-throughput screening revealed that all-trans retinoic acid (ATRA) showed clinical promise for a rare subtype of leukemia known as promyelocytic leukemia (APL) [127]. Similarly, gefitinib was shown to induce myeloid differentiation of acute myeloid leukemia in a cell-line based study [126]. To provide broad access to such gene expression based drug data, the Connectivity Map (CMAP) [69] was introduced in 2006, and enables researchers to obtain relevant drugs based on up- and down-regulatad genes (see Section 6.1.1).

After analyzing various of sources of data independently (Section 6.1), we create a drug-drug similarity network by combing disparate sources of data (Section 6.2).

Leveraging the tissue specific marker genes introduced earlier (Chapter 4), we show how we can create tissue specific drug-similarity rankings using a large curated expression database (as introduced in Chapter 2). As one of the goals of computational drug analysis methods is to enable targeted therapeutics, we anticipate that methods such as these can provide the foundation for future researchers.

6.1 Types of drug data

6.1.1 Connectivity Map

The goal of the Connectivity Map (CMAP) is to "provide a generic solution to this problem by attempting to describe all biological statesphysiological, disease, or induced with a chemical or genetic constructin terms of genomic signatures" [69]. Using a Kolmogorov-Smirnov statistic based pattern matching strategy, the CMAP database will return the list of drugs by how closely it resembles the user submitted pattern of input gene signatures. Ultimately, this is useful if, for instance, we know a set of genes that are differentially regulated in a certain disease condition that we hope to find a counteracting drug for. For example, if the we submit the genes g_1, g_2, and g_3 as up-regulated genes and g_4 and g_5 as the set of down-regulated genes, CMAP will return an ordered list of drugs based on how closely they match the given input pattern of genes. Thus, drugs that very closely match the up- and down-regulation patterns of the input genes are drugs that mimic the expression response of the disease, while drugs that have an opposite response could potentially be used to counter the effects of the disease by repressing the up-regulated genes and over-expressing the down-regulated ones.

The current version of CMAP is comprised of 7056 microarray samples divided into 956 controls and 6100 drug treatment samples. These 6100 samples correspond to a total of 1309 unique drugs performed across five difference cell lines (Table E.1). The control samples are used by CMAP to generate *difference profiles* for each treatment sample by subtracting the baseline expression of an untreated sample from

	HL60	MCF7	PC3	SKMEL5	ssMCF7	Total
HG-U133A	396	218	148	22	23	807
High Throughput HG-U133A	1010	3149	1870	0	0	6029
High Throughput HG-U133A EA	0	220	0	0	0	220
Total	1406	3587	2018	22	23	7056

Table 6.1: Cross-tab of the number of CMAP samples that were performed on the various gene expression platforms and the corresponding cell lines.

the expression of samples that were treated with a drug compound. Unlike many expression studies that try to minimize the technical variables, not only where these samples performed on three different platforms and two different mediums (Tables 6.1, 6.2 and Appendix E), there is also a wide discrepancy in the number of times a particular drug's expression response was measured (Table 6.2). For example, there are a total of 182 trichostatin A samples while only a single sample for drugs such as cantharidin and gefitinib.

Although there are some discrepancies in the CMAP data, researchers have used the Connectivity Map for such things as to find potential therapeutic agents for colon cancer [42], for cancers that have become resistant to chemotherapy [106], and to discover possible treatments for diseases that affect particular pathways [110]. More recently Sirota et al. [122] generated disease profiles from public expression data and validated the drug cimetidine for use in treatment for lung adenocarcinoma in a mouse model.

Batch effect in CMAP

Before using the data provided by CMAP, we performed some data validation. Using the transcriptomic landscape detailed in Chapter 3, we plotted all of the raw CEL file data used by the Connectivity Map onto this landscape (Figure 6-1). Unsurprisingly, what becomes immediately apparent is the cell line effect. Regardless of the drugs that the samples were treated with, the dominant signal is the tissue of origin (Figure 6-1(a)). A second "batch effect" is related to the platform that the experiment was performed on. As can be seen by the HL60 cell line samples (in red in Figure 6-

	HL60	MCF7	PC3	SKMEL5	ssMCF7	Total
trichostatin A	34	92	55	0	1	182
tanespimycin	12	36	12	1	1	62
LY-294002	13	34	12	1	1	61
valproic acid	14	31	10	1	1	57
sirolimus	10	25	8	0	1	44
fulvestrant	6	21	12	0	1	40
estradiol	8	19	8	0	2	37
haloperidol	7	19	6	0	0	32
monorden	4	12	5	1	0	22
tretinoin	5	13	4	0	0	22
thioridazine	4	11	5	0	0	20
chlorpromazine	4	11	4	0	0	19
fluphenazine	4	10	3	1	0	18
wortmannin	4	10	2	1	1	18
clozapine	4	10	3	0	0	17
genistein	3	11	3	0	0	17
alpha-estradiol	3	9	3	0	1	16
prochlorperazine	4	9	3	0	0	16
trifluoperazine	4	9	3	0	0	16
troglitazone	4	7	4	1	0	16
15-delta prostaglandin J2	3	8	3	1	0	15
geldanamycin	3	10	2	0	0	15
nordihydroguaiaretic acid	3	8	2	0	2	15
rosiglitazone	4	7	3	0	0	14
acetylsalicylic acid	3	8	2	0	0	13
alvespimycin	3	7	2	0	0	12
vorinostat	3	7	2	0	0	12
pioglitazone	0	6	5	0	0	11
metformin	1	7	2	0	0	10
naproxen	2	4	3	0	0	9

Table 6.2: Cross-tab of the number of CMAP samples that were performed for the top 30 most frequent treatments and their corresponding cell lines.

1(a)), they are distinctly separated by the array type in Figure 6-1(b). Since this clear separation is seen when the CMAP samples are viewed in the context of the transcriptomic landscape, it is not surprising that we see the same clustering by cell line type and array technology when we perform the PCA analysis solely on the 7056 CAMP samples (Figures 6-1(c) and 6-1(d)).

When we take this analysis one step farther and limit the view of the samples

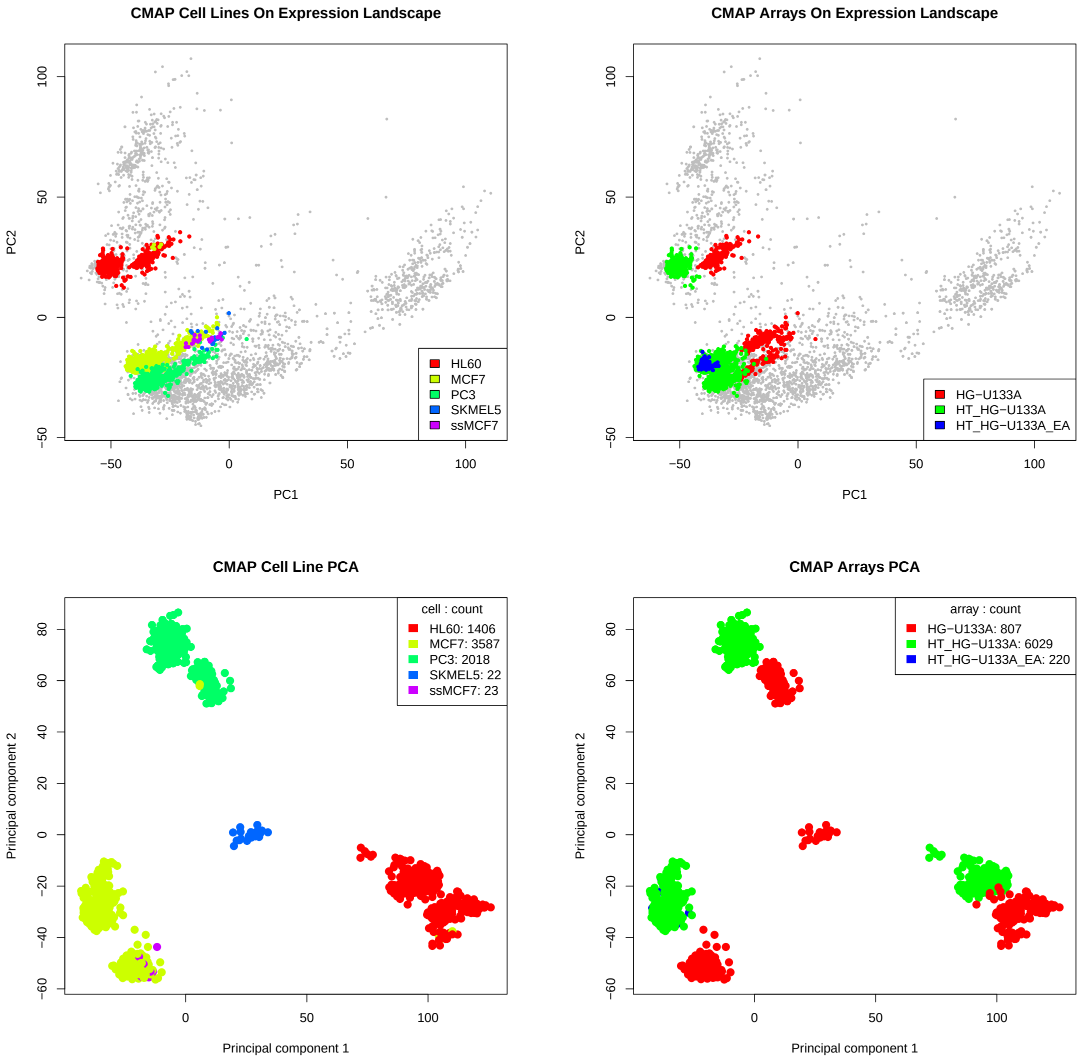


Figure 6-1: The 7056 samples (processed directly from the raw CEL files) from CMAP plotted on the transcriptomic landscape and (b) colored by the array technology and (a) cell line types that were used. Figures (c) and (d) depict the principal component analysis of just the 7056 CMAP samples without mapping it in to the transcriptomic landscape. We see the same pattern of clear separation by cell line and array technology.

to just those performed on the MCF7 breast cancer cell line and using the high throughput HG-U133A array, we see that the samples group together by the batch in which they were performed. Although these sorts of batch effects are not uncommon in expression studies, it points as to why the method used by the Connectivity Map uses differential profiles based on a control for each batch. Furthermore, these strong data artifacts mean that any subsequent analysis of this data must also use some sort of differential expression based analysis.

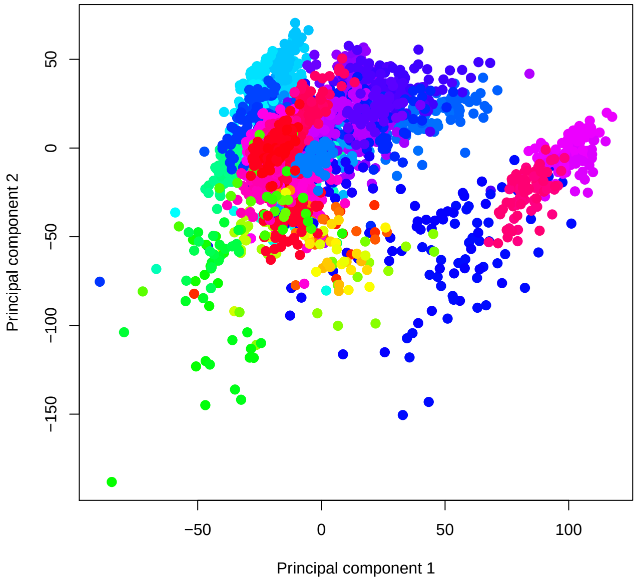


Figure 6-2: The first two principal components of the ranked raw expression data performed using the MCF7 breast cancer cell line on the high throughput HG-U133A array show such that each point in the figure is a sample and the color depicts the batch it originated from. There is a distinct batch effect even when we restrict the expression samples to only one cell line and one platform.

Thus, in order to contend with the multitude of experimental parameters (various expression platforms, cell-line types, etc) we limited our subsequent analyses to only the difference profiles for the 3149 samples performed on the high throughput HG-U133A platform using the MCF7 cell line. This subset of data only used DMSO as the medium treating the cell-lines with the small molecule compounds. Although this is less than half of the data that is available in CMAP, it represents the largest subset of data that was performed on the most similar of experimental conditions.

After the initial trimming of samples, we performed a first pass analysis on the difference profiles that CMAP uses to perform its analysis. Interestingly, the first two principal components of a PCA analysis of the samples across all genes generates two distinct clusters (Figure 6-3(a)). At first glance it appears that the samples

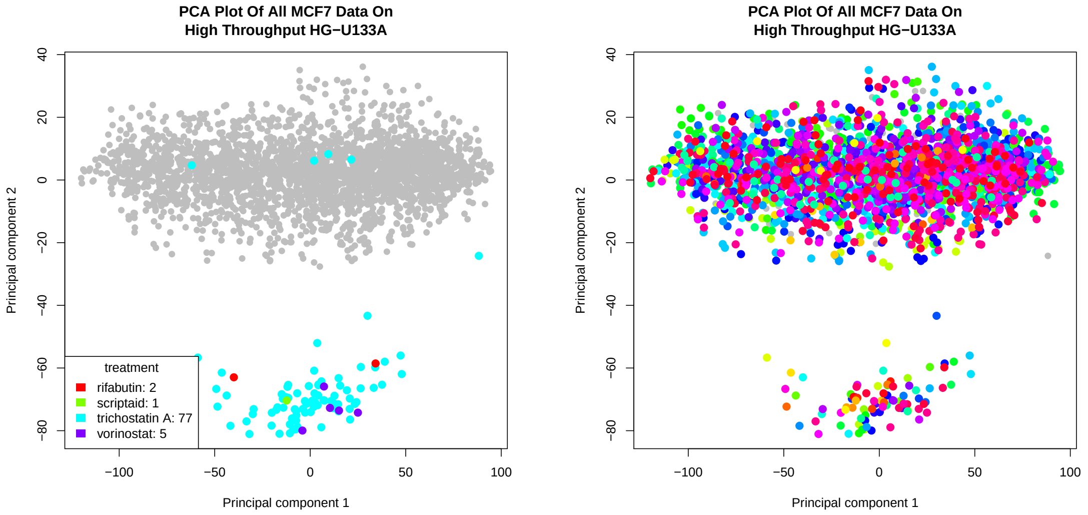


Figure 6-3: The first two principal components of the difference profiles of the MCF7 samples reveals two distinct clusters. (a) The bottom cluster is mainly comprised of samples treated with trichostatin A and at first glance appears to indicate a "trichostatin A signal." Unfortunately, we see four other trichostatin A samples in the main cluster and thus indicates that this may be an artifact of the data rather than a novel finding. (b). Shows the the clusters colored by batches. There does not appear to be a single bad batch that is causing this clustering.

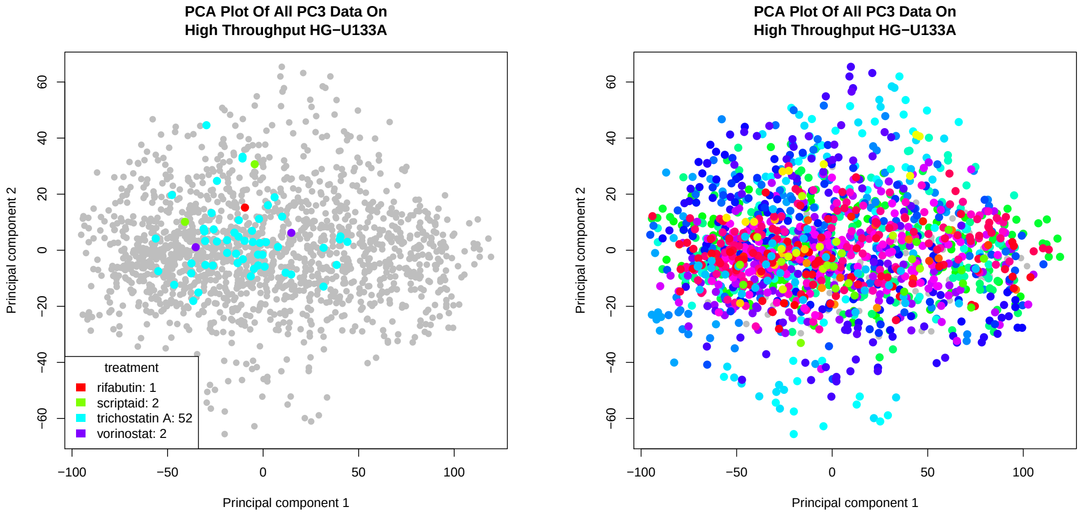


Figure 6-4: Unlike the two clusters that resulted from the PCA analysis of the MCF7 data (Figure 6-3), the first two principal components of the PC3 cell line data reveals one single cluster. (a) Coloring the samples that were in the bottom cluster in the MCF7 PC plot shows that they can be found in the middle of the cluster of PC3 samples. (b). Coloring the plot by batch does not reveal and clear "batch effect."

treated with trichostatin A have a very different expression profile compared to all other treated samples. Unfortunately, not only do we find four other trichostatin A samples in the midst of all other treated MCF7 samples, this same finding not recapitulated in the PC3 cell line data (Figure 6-4(a)). If we examine the batches that these samples originated from (Figures 6-3(b) and 6-4(b)) it does not seem to indicate a single "bad apple." Although a previous study found a relationship between the drugs found in the bottom cluster [58], as we could not verify the cause of this clustering, the 86 samples corresponding to the bottom cluster were also removed from subsequent analyses.

6.1.2 DrugBank

Unlike the Connectivity Map that contains the expression response of various cell lines to different small molecule compounds, DrugBank [63] is a database that contains various information such as the chemical formula, indication, toxicity, manufactures, and target genes for 6707 small molecule and drug compounds. While it does not contain any biological data like CMAP, it can be an integral source of information. For example, PREDICT [49], a method for identifying drug relationships, makes heavy use of the data available in DrugBank.

Drug target gene relationship network

One interesting piece of information provided by DrugBank is the set of genes that a given drug targets. These target genes are the genes corresponding to the protein(s) that a drug targets. Using this information one can perform various genomic studies to quantify the relationships between various drugs. Without any additional sources of data, we can create a drug relationship network based on the number of target genes that they share. The intuition here is that drugs that have similar gene targets should have a similar purpose. Such a target gene relationship network is depicted in Figure 6-5 in which each node represents a drug and an edge exists if the two drugs it connects have five or more target genes in common. The nodes of the network are

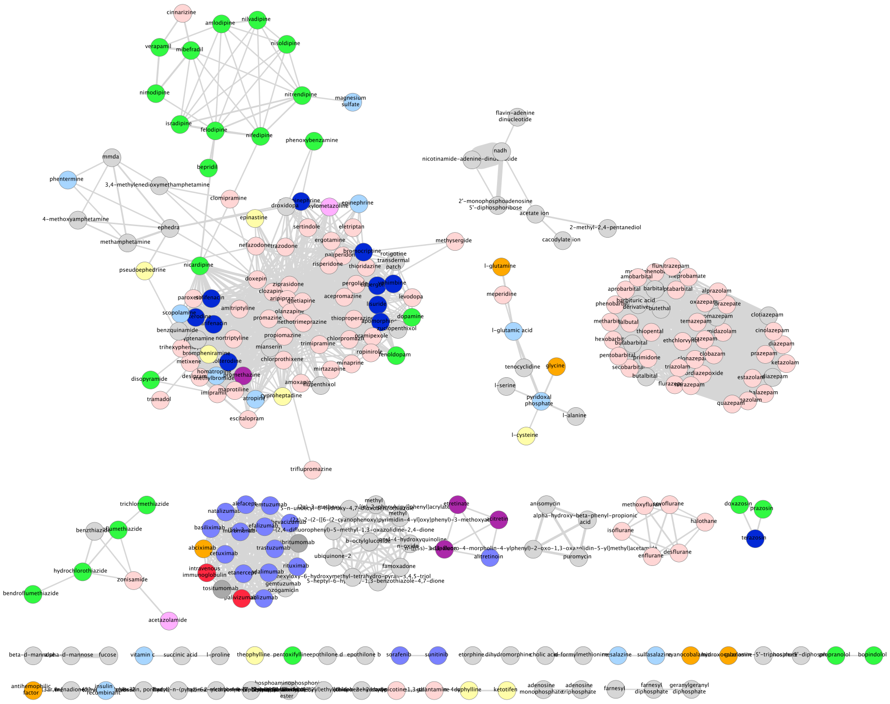


Figure 6-5: A drug relationship network based on common target genes of drugs. Each node depicts a single drug and an edge is drawn between two drugs if they have at least 5 target genes in common. The weight of the edge indicates a relative measure of how many target genes two drugs have in common. The nodes are colored by the Anatomical Therapeutic Chemical Classification System (ATC) code for that drug. There is distinct clustering by the ATC code of what the drug is indicated to be used for.

colored by the Anatomical Therapeutic Chemical Classification System[1] (ATC) code for that drug. Interestingly, we see clear groupings of many cardiovascular drugs. represented as green nodes, and of drugs that are related to the "nervous system" (anesthetics, anti-parkinson drugs, psycholeptics, etc.), which are represented as pink nodes. Furthermore, we see that many anti-neoplastic drugs cluster together (circular cluster of light purple nodes near the bottom of the figure). We will expand on this

[1] The ATC codes are a system of alphanumeric codes developed by the World Health Organization to classify drugs and other medical products that has 14 main groups: "Alimentary tractandmetabolism", "Bloodandblood forming organs", "Cardiovascular system", "Dermatologicals", "Genito-urinary systemandsex hormones", "Systemichormonalpreparations, excluding sex hormones and insulins", "Antiinfectivesfor systemic use", "Antineoplasticandimmunomodulatingagents", "Musculo-skeletal system", "Nervous system", "Antiparasiticproducts,insecticidesandrepellents", "Respiratory system", "Sensory organs", and "Various".

introductory analysis in Section 6.2.

Drug target genes and CMAP expression

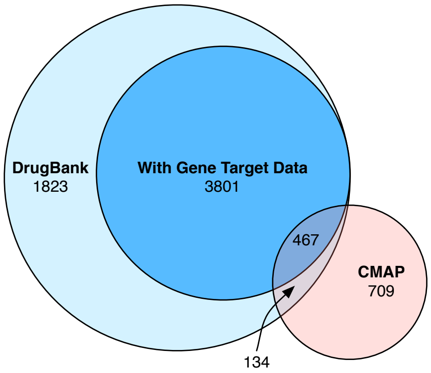


Figure 6-6: The overlap of drugs in the DrugBank database and those in CMAP.

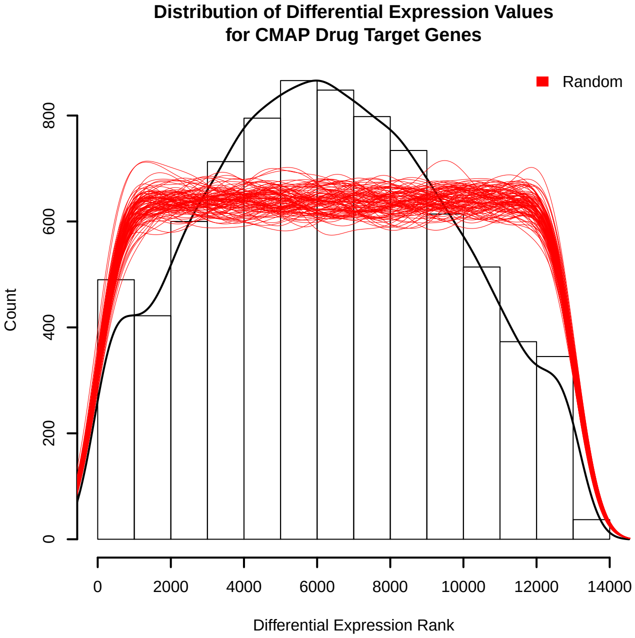


Figure 6-7: The distribution of differential expression ranks in CMAP for the genes deemed to be the target genes according to DrugBank.

In an effort to ascertain the effect a drug has on the gene expression of the genes corresponding to the protein that it targets, we analyzed the target genes in DrugBank [63] and their expression patterns in the Connectivity Map. Of the 6826 drugs present in DrugBank, 4268 of them had gene target information that corresponded to a human NCBI Entrez gene identifier. Of these, 467[2] are represented in CMAP (6-6). The mapping between DrugBank and CMAP was done by simple string matching; if the drug had the same name (ignoring capitalization) in both DrugBank and CMAP, it was deemed to be the same drug. Using the differential expression profiles provided by CMAP, Figure 6-7 depicts the distribution of the ranks of gene expression differences in CMAP for the genes that are deemed to be the target genes for the corresponding treatments. A low rank indicates that the difference between the expression value in the treatment and control samples was the small, while a high rank indicates that the differential expression was substantial. In other words, if a particular drug's effect on expression is high, then it should have a high rank. As a baseline, we compare these ranks with 100 distribution of ranks obtained by randomly selecting target genes for each drug (in red). What we find is that there is no clear evidence for a statistically significant effect on expression when examining the target genes in combination with the CMAP difference profiles. Interestingly, the random sampling indicates (unsurprisingly) a uniform distribution of differential expression ranks, while the actual distribution of target gene expression difference ranks looks approximately normal.

6.2 Drug similarity networks

As we were not able to draw any broad conclusions from the transcriptomic differences caused by the drugs' effects on the target genes using the CMAP data, we turned to a more fine-grained approach. As we had done when generating the target gene relationship network (Figure 6-5), we created similarity networks for drugs based on various sources of information. This type of network analysis was previously employed in such endeavors as generating a human disease network [46] and finding relationships between phenotypes and genotypes [19]. The underlying assumption is

[2] Recall that although CMAP contains expression information for 1309 different small compounds, we are only using the data corresponds to the subset of samples performed using the MCF7 cell line on the high throughput HG-U133A array.

that drugs have various properties, and each of these properties can provide a source of information about how similar two drugs are. For example, if two drugs target the same genes, they are probably more similar than two drugs that do not.

6.2.1 Similarity measures

Although there are a plethora of possible similarity measures for drugs, we used five metrics that were based on chemical structure, drug target genes, and the drug's effect on expression.

Atom pair distance

The first measure of similarity that we used was *atom pair* distance, which is "defined in terms of the atomic environments of, and shortest path separations between, all pairs of atoms in the topological representation of a chemical structure." [21] Atom pair distances have become widely used in searching for chemical compounds in large databases and virtual screening efforts [62] as they provide a metric for computing the similarity between chemical compounds based on it's atomic structure. For the purpose of this work we used the R interface of the publicly available ChemMine toolkit [10] called ChemminR to compute the atom pair distances between the chemical compounds found in DrugBank. We shall refer to this distance as the *atom pair* distance or *structure* distance.

Target gene based distances

We employed three distance metrics based on the target genes for the drugs. First, we computed the similarity between two drugs based on the number of target genes they have in common. To account for varying numbers of target genes for different drugs, we used the Jaccard index. The Jaccard index of two drugs X and Y with respective set of target genes x and y is the number of intersecting target genes divided by the number of total target genes. Mathematically, that is:

$$JI_{X,Y} = \frac{x \cup y}{x \cap y} \tag{6.1}$$

In other words, drugs that target all of the same genes have a Jaccard index of 1, while those that target completely disjoint sets of genes have a Jaccard index of 0. We shall refer to this distance as the *target gene overlap* distance.

Another source of information provided by DrugBank is the target protein family (Pfam) domain [123] that the drug targets. Briefly, Pfam is a database that contains evolutionarily related proteins (protein families) and provides the multiple sequence alignment and hidden Markov model (HMM) for the the proteins in each of the families. Using the Pfam domains that each of the drugs target, we can compute the Jaccard index (in the same fashion as with the target genes), to compute the similarity between two drugs.[3] This distance metric will be referred to as the *Pfam overlap* distance.

The final target gene based distance uses the target genes for each drug in conjunction with a protein-protein interaction (PPI) network to see how close the target genes are. For this, we employed the curated PPI network data that was used for IsoBase [96]. The distance between two drugs X and Y with respective set of target genes x and y is the minimum number of hops in the PPI network between any of the target genes. Mathematically, that is:

$$PPI_{X,Y} = \min_{x,y} nhops(x_i, y_j) \tag{6.2}$$

Here we try to capture a more abstract sort of similarity between the target genes that takes into account how these genes (well, the proteins that theses genes encode for) interact. Even if two drugs have differing target genes, if those genes are very close together in PPI space (it could be the case the drug X's target genes directly

[3] Although we did not perform this in this work, a future step would be to not just use the direct Pfam domain hits by the drug, but also include similar Pfam domains. Expanding the data as such, we could hope to find other genes that this drug may bind to due to non-specific binding.

interact with one or more of drug Y's target genes) then these two drugs should be considered similar. We shall refer to this distance metric as *PPI* distance.

Expression effect distance

The final source of similarity metric is the effect the drug has on expression. Here we made use of the CMAP [69] data and computed the correlation of each drug's expression difference profile (see Section 6.1.1) to all others. Thus, drugs that have highly correlated difference profiles should be more similar as they have a similar effect on gene expression. Again, it is important to note, that due to the various dataset effects in the CMAP data, we only used the data corresponding to the samples performed using the MCF7 breasts cancer cell line and performed on the high throughput HG-U133A platform. This final distance metric will be referred to as *expression* distance.

Restricting the similarities

As each of the above similarity metrics produces a similarity measure for all pairs of drugs, the network representation would yield a fully connected graph. As this is completely uninformative, we discarded any similarity score that was below a 95% cutoff. Although there are other methods of defining a cutoff as to the amount of similarity is to be deemed as significant, we chose to use a simple percentile cutoff to ensure that each of the networks had roughly the same number of edges.

6.2.2 Comparing the similarity networks

Given a matrix of similarity scores (in our case, we have five matrices, each corresponding to one of the aforementioned scoring metrics), we convert it to a network representation by drawing an connection (edge) between two drugs (nodes) such that the weight of the edge is the similarity score. Note, since we restricted the similarity matrices to only contain the top 5% of the scores, we are not left with a complete graph. After having converted each similarity matrix into a graph, we then examined

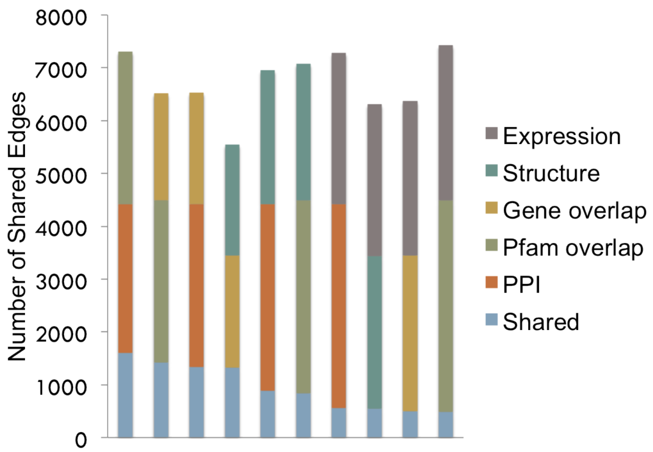


Figure 6-8: A pairwise comparison between all five types of similarity networks. Each bar represents the number of edges in the respective network such that light blue part of the bar corresponds to the number of edges shared by the two networks being compared and the other two colors are the remaining edges in the network.

how much overlap there is between the five similarity measures by counting the number of edges that were shared among the various networks. Figure 6-8 depicts the pairwise similarities between all five networks where we count the number of edges that are common in both networks. Each bar in the graph represents a comparison between two networks such that the light blue segment at the bottom is the number of edges that they had in common. As to be expected, the three networks that had the most in common were those that were derived from the same source information: target gene overlap, Pfam overlap, and PPI. Interestingly however, all of these distance metrics are at least 50% different from each other. The distance metric that provided the most differing set of similarities was the expression based distance. Thus, it appears that the structural similarity of a drug is more closely aligned with the genes that it targets.

6.2.3 Consensus similarity network

While examining each similarity network in isolation can provide insights as to how similar drugs are in the particular domain the similarity score was computed for, by aggregating the results across multiple networks we can see which drugs are similar in more than one domain. We call this "stacking" of similarity networks a *consensus* network. To generate a consensus network we take each of the five aforementioned networks and convert the similarity score matrices into binary indicator matrices such that there is a 1 if the edge exists in that network, and 0 if it does not. Recall, that we previously limited each of the similarity networks to only contain the top 5% of the edges and thus we will not obtain a complete graph. By adding these binary indicator matrices we are left with a consensus matrix that contains numbers between 0 (no edge between the two drugs) and 5 (all similarity networks said these two drugs were similar).

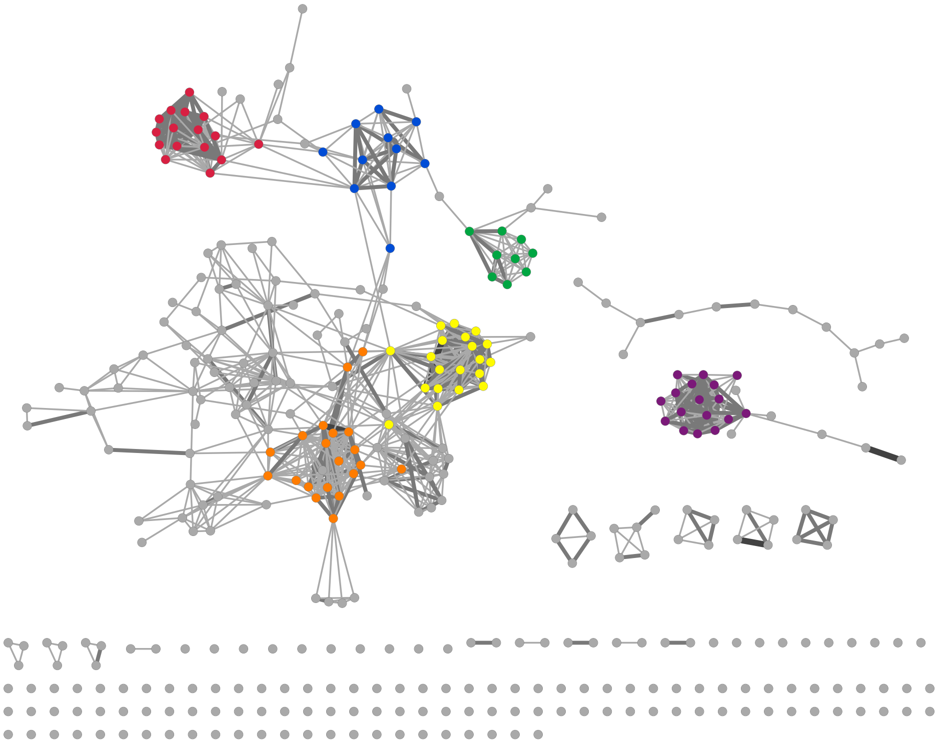


Figure 6-9: A consensus similarity network such that each node represents a drug and an edge exists between two drugs if they were found to be similar (i.e. had an edge) in at least 3 of the 5 independent similarity networks. The weight of the edge is relative to the number of networks that contained the edge; it is thickest when all five networks deemed the two drugs connected by the edge to be similar. The coloring of the nodes indicate interesting highly connected clusters. For example, the drugs represented by the purple nodes (bottom right) are mostly anti-inflammatory drugs while many of the drugs in yellow cluster are adrenergic beta-antagonists.

Further restricting our view of drug similarity to only keep the edges that are present in three or more of the similarity networks (we shall refer to this as the *3,5 concensus network* as it contains at least three out of the five independent networks), we are left with a network that looks like Figure 6-9. As before, each node in the network represents a drug and there is an edge between two drugs if it was deemed similar (in this case in at least 3 of the 5 independent similarity networks). The weight of the edge is relative to the number of networks that contained the edge; it is thickest when all five networks deemed the two drugs connected by the edge to be similar. The coloring of the nodes indicate interesting highly connected subgraphs clusters such that the subgraph has an edge connectivity of more than half of the number of nodes in the subgraph [52]. For example, the drugs represented by the purple nodes (bottom right) are mostly anti-inflammatory drugs while many of the drugs in yellow cluster are adrenergic beta-antagonists. Also of note, the uncolored subgraph that is above the purple component consists of many hypoglycemic agents.

Drugs	Similarities
rimexolone, budesonide, and fluocinonide	Anti-inflammatory glucocorticoid steroids.
atropine, mepenzolate bromide, and difenidol	Atropine and mepenzolate bromide are both parasympatholytics (anti-muscarinic) while mepenzolate Bromide and difenidol are both diphenylmethanes.
carteolol, nadolol, and levobunolol	All beta blockers. carteolol and nadolol are used for treatment of angina, arrhythmia, and hypertension.
testosterone and cyproterone	Cyproterone suprresses testosterone
daunorubicin and dauxorubicin	Chemotherapy drugs; Anthracycline antibiotics
irinotecan and camptothecin	Anticancer; irinotecan is an analog of camptothecin
selegiline and pargyline	MAO-B inhibitors

Table 6.3: The set of related drugs when we enforce that the drugs be similar using all five similarity metrics.

When we require that there be an edge between two drugs in all five networks (i.e. complete consensus, or a *5,5 consensus network*) we are left with three subgraphs containing three drugs each along with several drug pairs (Table 6.3). It is worth noting that these drugs are indeed very similar to each other and that there are examples of these drugs used together. For instance, daunorubicin and dauxorubicin are used in chemotherapy [9] and were part of a Phase III clinical trial for AIDS-related Kaposi's sarcoma [44].

6.2.4 Potential applications

It is promising to see that the use of a method based on the consensus of similarity across a wide range of drug properties (structure, target genes, and effect on expression) can provide us with a way to glean new insight into the relationship between various drugs. By expanding this work to include such things as side-effects we could not only infer the similarity of drugs, but possibly provide potential alternatives to existing drug therapies. Furthermore, it could also point to novel applications of existing drugs to treat conditions that they were not initially indicated for. Unlike other methods that are based on predicting drug class [49], the use of a network allows researchers to effortlessly explore the connections between drugs.

As additional example, in the original CMAP paper [69], put forth two interesting drug associations as "test cases." First, 17B-estradiol (ER ligand) was analyzed and found to be similar to estradiol, fluvestrant and genistein. Indeed, using our *3,5 consensus network*, we see that estradiol, fluvestrant, and dienstrol (also an ER agonist) are in a clique. Furthermore, we find that genistein is connected to dienestrol, thus recapitulating their findings. Their second example made use of phenothiazine for which they had strong similarity findings for trifluoperazine, thioridazine, and fluphenazine. They, however, were only able to find a weak association to another anti-psychotic drug, haloperidol. When we searched for these drugs in our consensus network, we were able to not only recapitulate their findings, but also show a clear connection between fluphenazine and haloperidol.

Although clinical applications of this method is still distant, an expansion of this type of analysis that allows researchers to layer various similarity measures upon each other to find interesting connections between drugs may well aid in finding new uses for existing drugs or finding better alternatives to therapies with detrimental side effects.

6.3 Leveraging Concordia stem cell marker genes

We showed in Section 4.4 how we can make use of 189 stem cell marker genes to not only stratify pluripotentiality and malignancy, but also to provide clinical gradings for various types of tumors. Naturally, one would inquire as to how these genes fair in the context of the Connectivity Map. As the stem cell marker genes were derived from data performed on the HG-U133 Plus 2.0 array, there unfortunately is not a complete overlap with the set of genes for which data is available in CMAP (which was performed using the HG-U133A array). As such, the following analysis is performed using only the 140 genes out of the 189 genes that were common to both platforms.

6.3.1 Stem cell genes as a CMAP query

The first, and most naïve, analysis that we can perform with these genes is to perform a traditional CMAP query with these genes. As CMAP requires a list of up- and down-regulated genes for its input query signatures[4], we computed mean difference of expression for the each of the 140 genes as compared to the background expression intensity. For example, one of the marker genes in the list is FGF2 fibroblast growth factor. To compute whether FGF2 is up- or down-regulated in stem cells, we took all samples associated with stem cells (the same ones used to derive the stem cell marker gene set) and computed the mean expression for FGF2. Similarly, using all samples not associated with stem cells we computed the mean background expression for FGF2. The set of up-regulated stem cell genes was thus the ones that had a mean expression level greater than the background, and conversely, the set of down-regulated genes were those that had a mean expression that was lower than the

[4] The default CMAP tool (`http://www.broadinstitute.org/cmap/`) requires probe level identifiers for its input query signature. To perform this query based on the stem cell marker genes, we set up a local instance of CMAP and summarized all of the probe level data from CMAP to gene level data (each gene's value is the mean of all probe values for that gene). Other than this gene level summarization, our local instance is identical to the original online tool.

background.[5] Table E.5 contains the set of 140 genes along with their respective mean differences from the background distributions.

The result of a CMAP query is a sorted list of the 6100 treated samples in the CMAP database such that those with the highest "similarity" to the input gene signature are those that have all of the up-regulated input genes up-regulated and all of the down-regulated input genes down-regulated. As such, the treatments corresponding to those samples can be viewed as having the same effect as the input gene signatures. Conversely, if we are searching for a treatment that has the opposite effect of the input gene signature (i.e. if we are looking for a drug that undoes the transcriptomic effect of a particular disease) then the treatments that are most "dissimilar," having the up-regulated input genes down-regulated and the down-regulated input genes up-regulated, are of interest. Since we saw that the stem cell marker gene set is related to malignancy, we would thus want to find the most "dissimilar" treatments that undo what the cancer does. Note, as we are using the traditional use-case of CMAP, we are including all 6100 treatment samples regardless of cell line and array technology that was used.s

When we perform the CMAP query with the 109 up-regulated and 31 down-regulated stem cell maker genes we find that the treatments that occur most frequently in the top 50 "dissimilar" (i.e. undoes what the stem cell signature does) samples are trichostatin A, LY-294002, trifluoperazine, and sirolimus. For instance, trichostatin A inhibits histone deacetylase (HDAC) enzymes and inhibits cell cycle during the beginning of the growth stage. It has also recently been shown to have a potential for the regulation of hematopoietic progenitor/stem cell frequencies [92]. LY-294002 is a derivative of quercetin, which, according to the American Cancer Society, "has been promoted as being effective against a wide variety of diseases, including cancer."[6] However, they also state that there is no clinical evidence that shows that it can prevent or treat cancer. Similarly, trifluoperazine has been found to inhibit DNA repair to induce cell death in non small cell lung carcinoma [98] while sirolimus has

[5] For those who are curious, FGF2 is an up-regulated gene in the stem cell population.

[6] `http://www.cancer.org/Treatment/TreatmentsandSideEffects/ComplementaryandAlternativeMedicine/DietandNutrition/quercetin`

been shown to block cell cycle in keratinocyte stem cells [60]. Although this analysis is far from conclusive, it is promising to see that stem cell marker genes found using a large heterogeneous database of gene expression data indeed are related to drugs that have previously been implicated to affect the cell cycle process and in the treatment cancer treatment.

6.3.2 Stem cell marker genes as a lens into cell-cycle and cancer drug space

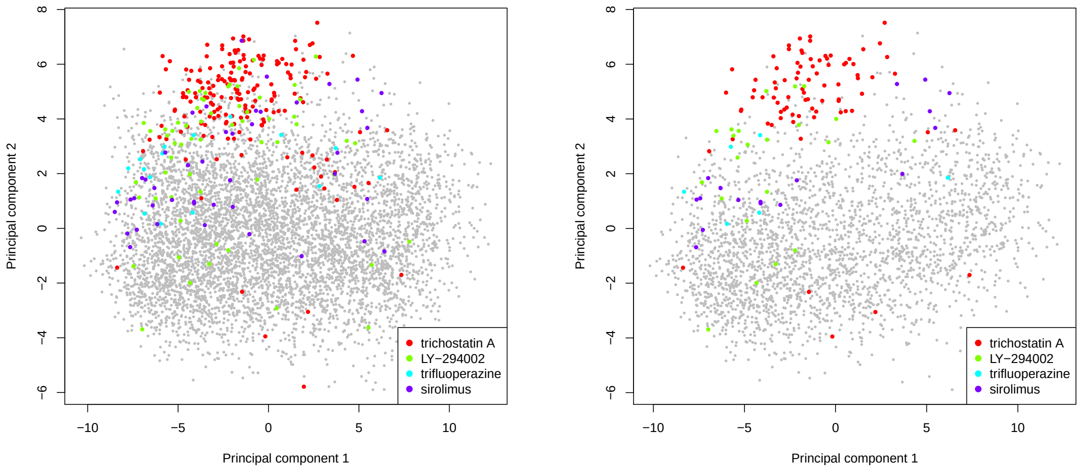


Figure 6-10: To examine the four drugs (trichostatin A, LY-294002, trifluoperazine, and sirolimus) from the previous section, we took the (a) 6100 difference profiles and the (b) 2740 MCF7 samples performed on the high throughput HG-U133A array and computed their first two principal components using only the 140 stem cell marker genes. Interestingly wee see that the majority of the samples for the four drugs appear in one localized neighborhood of the plot.

Instead of using the the stem cell marker genes to form an input query to CMAP, we can use them as a lens to view the CMAP difference profiles (just like we used the breast gene set to view the metastasized breast cancer samples in Section 4.3). To further examine the results from the previous section, we took the 6100 difference profiles and computed their first two principal components using only the 140 stem cell marker genes[7] and overlaid the location of the samples that were treated with

[7] Note, we are only using 140 because this is the set of genes of the 189 stem cell marker genes

trichostatin A, LY-294002, trifluoperazine, and sirolimus (Figure 6-10). Regardless of whether we use all of the 6100 profile samples available in CMAP (Figure 6-10(a)) or restrict our view to just the 2740 that were performed using the MCF7 breast cancer cell line on the high throughput HG-U133A array (Figure 6-10(b)), the samples for the four drugs appear to be localized in one neighborhood of the plot.

If we then identify the treatments of the samples in the local neighborhood of the upper left corner ($PC_1 < 2$ and $PC_2 > 2$) of the PCA plot in Figure 6-10(b) we find HDAC inhibitors such as scriptaid and vorinstat(Table 6.4). Furthermore thioridazine, trifluoperazine, and chlorcyclizine were found to reverse chemotherapy resistance in KB carcinoma[8] cells [4]. Although it is hard to draw conclusions about the treatments that have only a few samples, it is interesting to see HDAC inhibitors and cancer therapy drugs such as gefinitib in the local neighborhood of drugs that we previously found to be related to cell-cycle and cancer therapy using the CMAP query signature in the previous section.

that are present on both the HG-U133A and HG-U133 Plus 2.0 arrays.

[8] KB carcinoma cells are a cell line derived from a human carcinoma of the nasopharynx.

Drug	In Neighborhood	Total	Percentage
scriptaid	3	3	1
MS-275	2	2	1
quinostatin	2	2	1
cantharidin	1	1	1
dexverapamil	1	1	1
gefitinib	1	1	1
HC toxin	1	1	1
oxamic acid	1	1	1
PHA-00665752	1	1	1
tyrphostin AG-1478	1	1	1
vorinostat	11	12	0.92
trichostatin A	154	182	0.85
amantadine	3	4	0.75
daunorubicin	3	4	0.75
emetine	3	4	0.75
etoposide	3	4	0.75
ouabain	3	4	0.75
perhexiline	3	4	0.75
suloctidil	3	4	0.75
fendiline	2	3	0.67
latamoxef	2	3	0.67
nystatin	2	3	0.67
reserpine	2	3	0.67
rifabutin	2	3	0.67
STOCK1N-35215	2	3	0.67
terfenadine	2	3	0.67
LY-294002	38	61	0.62
thioridazine	12	20	0.6
trifluoperazine	8	16	0.5
chlorcyclizine	3	6	0.5

Table 6.4: The top 30 treatments (by percentage) of the treatments found in the upper left neighborhood of Figure 6-10(b). The first column indicates the number of samples that were found in the neighborhood while the second indicates the total number of samples in all of CMAP for that drug. Although it is hard to draw conclusions about the treatments that have only a few samples, it is interesting to see HDAC inhibitors such as scriptaid and vorinstat in a neighborhood of drugs that we previously deemed to have an effect on cell-cycle and cancer.

6.4 Drug target genes and Concordia expression

While the Connectivity Map [69] provides a laudable piece of work that allows for finding potential therapeutic agents for various conditions, its greatest asset (the difference profile) is also it's greatest weakness. Although they attempt to treat various cell lines with different chemical compounds, it is not possible for them to perform a gene expression experiment for each drug on every type of tissue. For example, just because a drug caused the up-regulation of a certain gene in the MCF7 breast cancer cell line, that does not mean it will have the same effect on lung tumor tissue sample. However, if one hopes to repurpose existing drugs (or even screen novel compounds that have yet gone to market), it is imperative to be able to see their potential effect on expression in the context of different tissues.

As such, what we need is a way to quantify the importance of a drug in the context of various tissues. In Chapter 3 we detailed a method of curating the publicly available gene expression samples in the NCBI Gene Expression Omnibus (GEO) [13] which we then used in Chapter 4 to generate tissue specific marker gene scores. Using this framework, we can provide a tissue specific view of the effects of a drug.

6.4.1 Concordia marker gene scores to examine drug target genes

In two earlier sections of this chapter (Sections 6.1.2 and 6.2) we showed how we can use the target gene information of a drug provided by DrugBank [63] to generate drug relationship networks. Here we want to identify whether the target genes of a drug are genes that are implicated as being marker genes for the tissue in which the disease that the drug addresses occurs. Stated another way, we want to see whether the drug is targeting genes that are deemed "important" in the tissue in which the disease occurs.

To test whether the target genes of drugs are indeed related to tissue specific marker genes, we took 121 drugs across five drug classes (hypoglycemic agents, anti-inflammatory agents, antipsychotic agents, HMG-CoA reductase inhibitors, and antineoplastic agents) as indicated by DrugBank [63] and examined the marker gene scores (see Section 4.1) of their target genes for various types of tissues. More specifically, for each drug, we computed its "target gene score" as the z-score of the marker

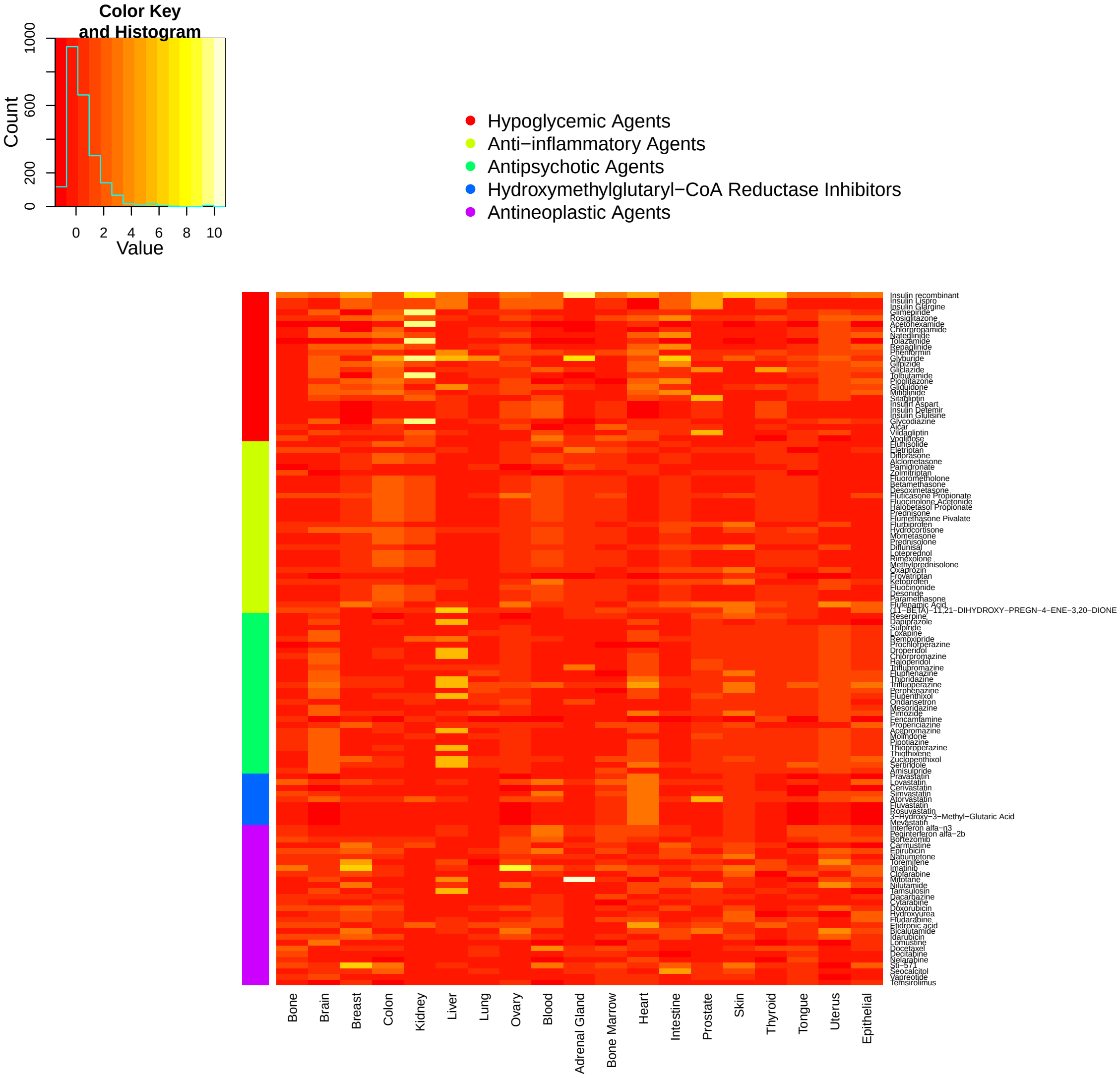


Figure 6-11: A heatmap of the drug target gene marker gene scores. Each row corresponds to one of the 121 and each column a tissue type. The color of each i, j entry corresponds to the marker gene score for that target genes for drug i in tissue type j such that red is low and yellow/white is high. Interestingly, many of the hypoglycemic agents' target genes have high marker gene scores in kidney tissue while anti-psychotics have high marker gene scores in liver tissue.

gene scores for each of it's target genes for each tissue. For example, the drug Temsirolimus used to treat advanced renal cell carcinoma has one target gene, MTOR. We then compute the marker gene score for MTOR for each tissue type and compute the z-score of MTOR's marker gene score as compared to the marker gene scores for all other genes for each tissue type. If a drug has more than one target gene (many of them do), we take the the maximum value as this indicates the drug's maximum possible affect on a marker gene for that tissue (Figure 6-11).

Several insights can to be gleaned from this. First, many of the hypoglycemic agents have high target gene scores in the kidney, one of the two organs in which insulin is removed from the body (the liver is the other organ). When we examine these hypoglycemic drugs more closely, we see, for example, that glyburide's target genes have a high marker gene score for adrenal gland tissue. Interestingly, it is known that glyburide may cause adrenal insufficiency. Secondly, many antipsychotic drugs appear to have an affect on genes with high marker gene scores for the liver. It has been shown that a significant number of patients treated with antipsychotic drugs have alterations of liver function tests [41]. Furthermore, Imatinib, a drug used treat certain types of leukemia, is currently in phase II clinical trials for treating ovarian cancer [1]; in Figure 6-11 we see a clear indication that Imatinib targets genes related to expression activity in the ovary. Interestingly, the drug that had the highest target gene score for the adrenal gland (and the highest overall scoring target gene score) was Mitotane, a drug that is used to treat cancer in the adrenal gland. Although these results only pertain to a small fraction of the available drugs and possible tissue types, we see that by combining the marker gene score data from Chapter 4 we can provide new insight into the workings of drugs on a molecular level without performing any new gene expression study.

6.4.2 Drug target gene expression correlation

In the previous section we explored the "marker geneness" of the target genes of drugs and showed that in several instances the target genes of drugs indeed are related to marker genes for various tissues. While that analysis provides useful insight as to how "important" the target genes of a drug are for various tissues, it does not give us a way to easily compare different drugs under different tissue conditions. In other words, it doesn't address whether the target genes of drugs act in a more coordinated fashion in one tissue than another. Furthermore, it doesn't allow us to provide a ranking of how similar other drugs are to a given drug in different tissue contexts.

Computing the tissue specific drug target gene similarities

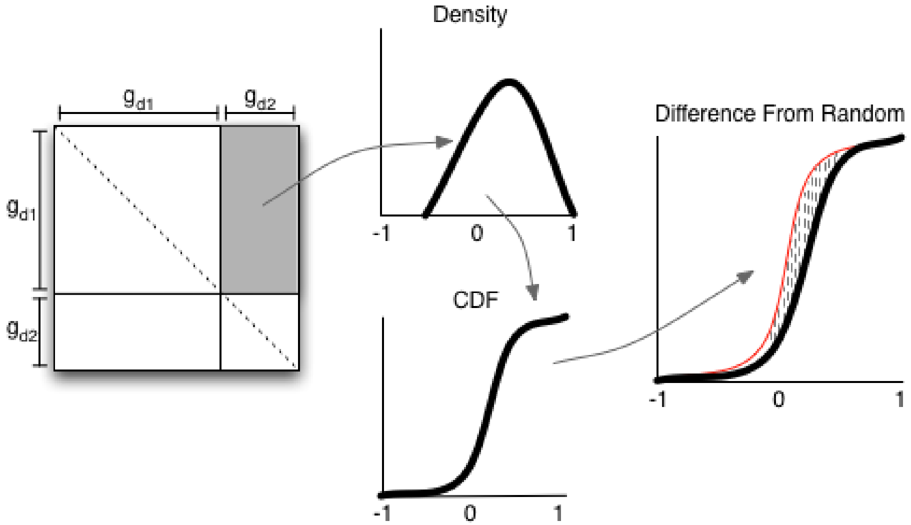


Figure 6-12: We compute the similarity between two drugs by computing the "difference" in the observed distribution of expression correlations for the pair's target genes as compared to the correlation distribution of random samplings of genes.

To compute tissue specific drug similarity scores for drugs we take each drug pair and compute the expression correlation of their target genes and compare how the distribution of these correlations differs from random (Figure 6-12). Since we are only interested in the "cross-talk" between the target genes for the two drugs, we only examine the expression correlations in the upper right hand corner of the correlation matrix (i.e. the intersection of the correlations of the target genes for drug pair of drugs). We then take these gene-gene correlation scores and convert their distribution into a cumulative density function (CDF) which is then compared to the expression correlation CDFs of random samplings of genes equivalent in number to the observed set. We ensure that if the drug pair being tested has any target genes in common, that the number of overlapping genes in the random set is equivalent to the number in the drug pair's target genes. We then compute the z-score (standard score) of the observed CDF as compared to the mean and standard deviation of the randomly generated CDFs at fixed intervals. We set the fixed intervals to start at a correlation of 0 and increment the interval by 0.01 each time yielding a total of 100 z-scores between a correlation of 0 and 1. This method is very similar to computing the Kolmogorov-Smirnov (KS) statistic but can be done more efficiently. If the sum or the z-scores is negative (i.e. the observed CDF is shifted to the right of the random CDFs) then the target genes of the two drugs are more correlated than by random chance. The more negative, the more significant the result. If, on the other hand the sum of the z-score is positive, it means that the target genes are less correlated than by random.

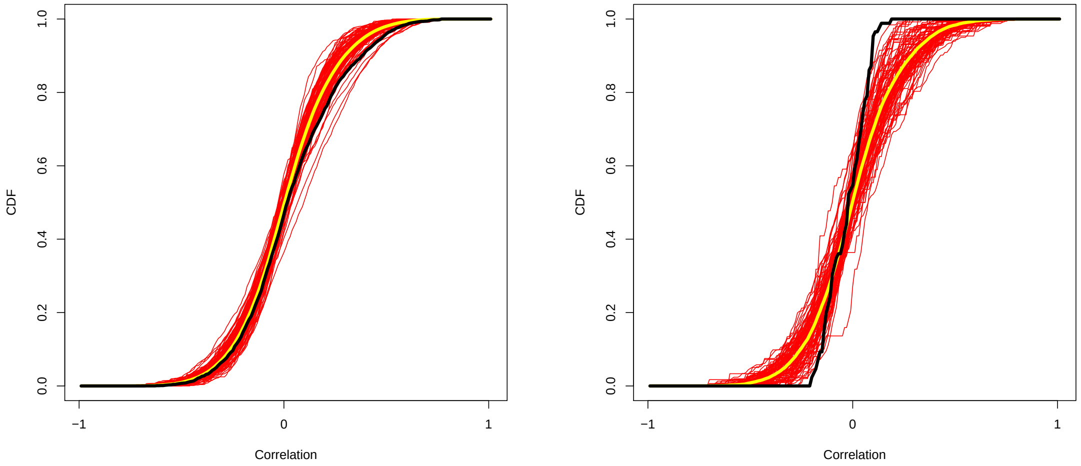


Figure 6-13: The CDFs of the correlations distribution of (a) haloperidol vs. molindone's and (b) haloperidol vs. voglibose's target genes and their immediate PPI neighbors. The black line in each of the plots depicts the actual CDF of the distribution of correlation while the yellow line depicts the mean of 100 randomly sampled correlation distribution CDFs. The red lines depict each of the 100 random CDFs that were generated. In both cases we used the gene-gene correlations as computed using only the brain tissue samples. (a) shows an example of two drugs that have target genes who's expression is more correlated than random while (b) depicts the opposite.

We also extended the above method to not just incorporate the target genes of each of the drugs, but also their immediate neighbors in protein-protein (PPI) interaction space. In other words, if a drug is indicated as having a single target gene, we find all neighbors of that target genes in the PPI network and add them to the list of the drug's target genes. Again, we used the same PPI network data that was used for for IsoBase [96]. The intuition for this is that it is not just the coordinated expression of only the exact targets of a drug that is important, but also the coordinated expression of the genes that are interaction partners.[9]

Thus far we have not talked about how we get tissue specific drug similarities. To account of different tissues, the gene-gene correlations that were used to compute the drug similarities were generated by using different subsets of the manually curated gene expression data (see Chapter 3) we obtained from GEO [13]. For example, to see how related the effect of a drug is in brain tissue, we computed the gene-gene correlations using just the gene expression samples corresponding to brain tissue.

As an example, Figure 6-13 depicts the CDFs of the correlations distribution of haloperidol vs. molindone's (Figure 6-13(a)) and haloperidol vs. voglibose's (Figure 6-13(b)) target genes and their immediate PPI neighbors. The black line in each of the plots depicts the actual CDF of the distribution of correlation while the yellow line depicts the mean of 100 randomly sampled correlation distribution CDFs. The red lines depict each of the 100 random CDFs that were generated. In both cases we used the gene-gene correlations as computed using only the brain tissue samples. Here we see that the expression correlation for haloperidol and molindone's target genes are more correlated than chance while haloperidol and voglibose's are less correlated. This makes sense as the former are both used to treat psychotic disorders while voglibose is a hypoglycemic agent.

To compare the efficacy of using the target gene's correlations to one another, we also computed the Jaccard index between the overlap target genes of the drugs (as in Section 6.2) as a baseline. This was done for both the cases: 1) when just the target genes and 2) when the target gene list was expanded by the target gene's PPI neighbors. Furthermore, instead of using the raw Jaccard index, we performed 1000 random samplings and computed the empirical p-value. The random sampling was done in the same manner as previously described. The use of these p-values provides us with a baseline of what just the target genes tell us about the drug similarity

[9] Again, it is not that these genes are interacting in the cell, it is the proteins that are encoded by these genes that interact. However, we are assuming that if the proteins that the genes encode for interact, then the genes that encode for these proteins should also have some sort of coordinated expression.

without including any expression information.

Analyzing tissue specific drug target gene similarities

The drug-drug similarity scores were computed for 121 drugs form five drug classes (hypoglycemic agents, anti-inflammatory agents, antipsychotic agents, HMG-CoA reductase inhibitors, and antineoplastic agents) as indicated by DrugBank [63] across eight distinct tissue types along with one case where all tissue data was used. Given a drug and a tissue type, we can now examine what other drugs are deemed to be similar in the context of the target genes and the correlated expression patterns of those target genes. We find that when we use a specific tissue for the gene-gene correlations and include the neighbors of the target genes that we do a better job of finding drugs of similar class to a given drug than if we do not.

For example, Figure 6-14(a) depicts the 120 drugs sorted in order form most similar to most different to haloperidol (an anti-psychotic drug) when just the target genes and all tissue samples (brain, blood, lung, colon, etc.) are used to compute the gene-gene correlations. Figure 6-14(b) depicts the result when we used just the brain tissue samples to compute the gene-gene correlations. The black dots in the figures represent the actual sum of z-scores for each drug as computed in the aforementioned manner. The horizontal blue line indicates the 5% significance cutoff. Any values below this line are scores that are in the top 5% of all scores. The colored squares indicate the drug type category for each of the drugs. For example, the leftmost entry in Figure 6-14(a) is voglibose, a hypoglycemic agent. The colors of these squares are a gradient from red to yellow such that red indicates the lowest score and yellow is the highest. This just provides another visual indication as to how good the score for that drug is.

Two things are immediately apparent. First, in both cases, regardless of whether or not we use all of the tissue samples or just the brain tissue samples, hypoglycemic agents appear to be the most similar to the anti-psychotic drug haloperidol. Second, although we see hypoglycemic agents as being the most similar to haloperidol when using the brain tissue samples, we see that only one of the results is significant.

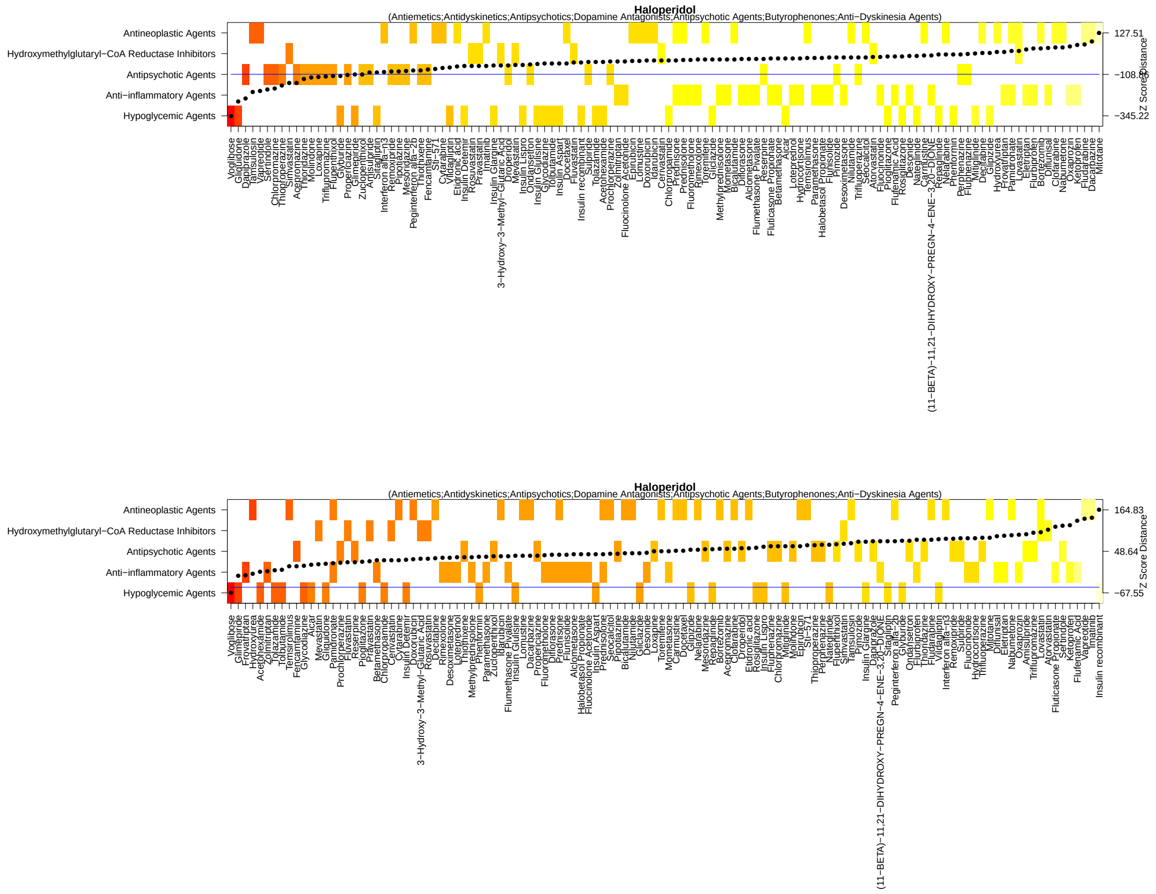
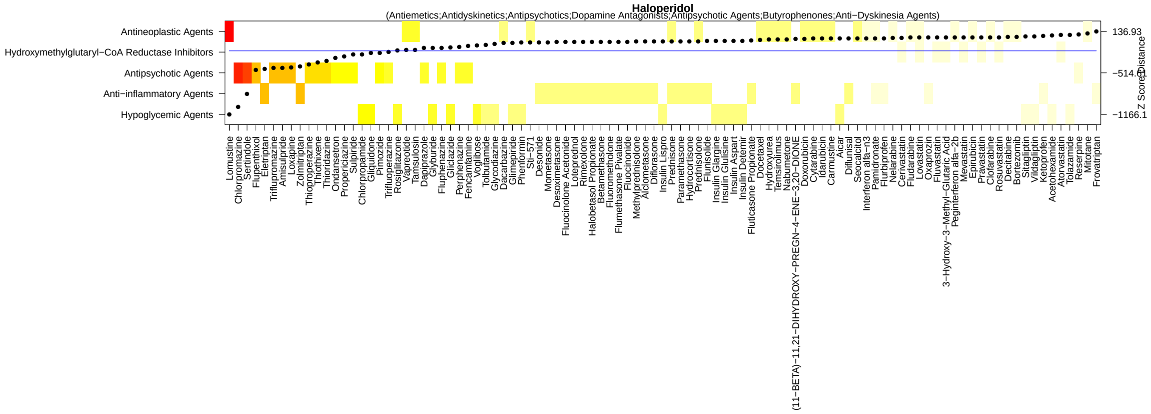


Figure 6-14: These figures depict the 120 drugs sorted in order form most similar to most different to haloperidol when just the target genes are used to compute the similarity to haloperidol. (a) shows the result when all tissue (brain, blood, lung, colon, etc.) is used to compute the similarities while (b) shows the result when only the brain tissue is used. The black dots in the figures represent the actual sum of z-scores for each drug as computed in the aforementioned manner. The horizontal blue line indicates the 5% significance cutoff. Any values below this line are scores that are in the top 5% of all scores across all 7260 drug-drug similarities (121 drugs results in $\frac{121 \times 120}{2}$ total distances). The colored squares indicate the drug type category for each of the drugs. For example, the leftmost entry in (a) is voglibose, a hypoglycemic agent. The colors of these squares are a gradient from red to yellow such that red indicates the lowest score and yellow is the highest. This just provides another visual indication as to how good the score for that drug is.

If instead of looking at just the expression correlation of the target genes of the drugs, but rather expand the set of genes to include the immediate PPI network neighbors, we see than most of the drugs that are similar to haloperidol are indeed anti-psychotic drugs (Figure 6-15). Furthermore, when we restrict the gene-gene correlations to be computed on only the brain tissue samples, we see that there is only one drug (etidronic acid) that is significantly similar to haloperidol.



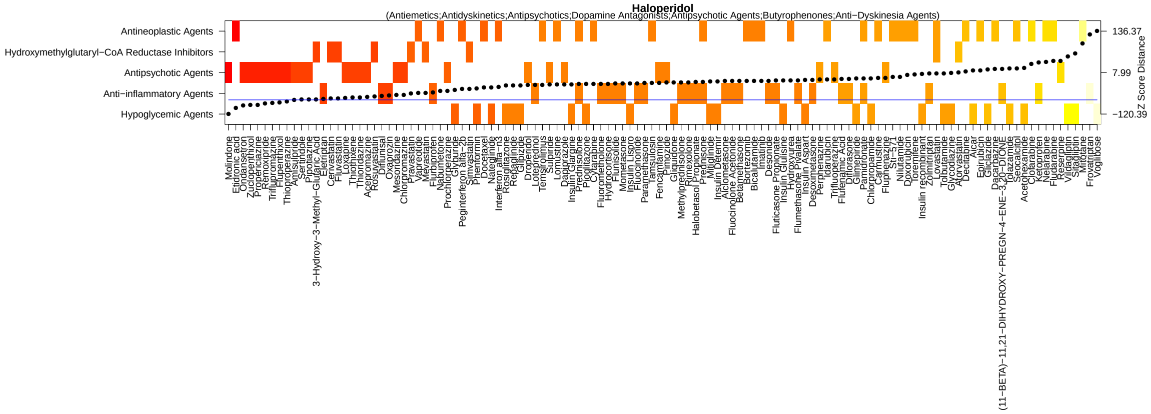


Figure 6-15: These figures depict the 120 drugs sorted in order form most similar to most different to haloperidol when target genes along with their immediate PPI neighbors are used to compute the similarity to haloperidol. (a) shows the result when all tissue (brain, blood, lung, colon, etc.) is used to compute the similarities while (b) shows the result when only the brain tissue is used. The meaning of the dots, 5% threshold line, and colored boxes is identical to Figure 6-14.

We then compared these results to when just the empirical p-value of the Jaccard indices of the target gene overlaps were used to see if the correlation structure of the expression data is providing us with any additional useful information. Figure 6-16 shows the same 120 drugs and how similar they are to haloperidol when when use just to target genes (Figure 6-16(a)) and when we expanded the set to include the

PPI neighbors (Figure 6-16(b)). Unlike in the previous plots where points below the horizontal blue line corresponded to the top 5% most similar values, here the blue line indicates a hard 5% p-value cutoff. Any value below the line had an empirical p-value of less than 0.05. Again, just as in the case when we used just the correlation of the target genes of the drugs, the most similar drugs to haloperidol appear to be hypoglycemic agents. However, when we use the PPI neighbors we do see that many of the anti-psychotic drugs cluster together (Figure 6-16(b)).

Now that we have looked at a particular example (haloperidol's most similar drugs under various conditions), let us take an aggregate look at the results. To see how well a particular variation (tissue type, whether or not to include the PPI neighbors) does, for each drug we computed the number of other drugs that were of the same drug category below and below the 5 % threshold. In other words, of the results that were significant, we want to see how many of them were similar drugs. Figures 6-17 and 6-18 respectively show these results for when only the target genes are used and when the PPI neighbors are included. The values in both of these figure range from 0 (none of the drugs below the 5% threshold were of the same drug category) and 1 (all of the drugs under the 5% threshold were of the same category). For instance, in the row for "Insulin recombinant" in Figure 6-17 we see that in the all tissue data and blood tissue data conditions many of the most similar drugs were also hypoglycemic agents.

Interestingly we see that both the hypoglycemic agents and anti-inflammatory agents appear to be performing better when the PPI neighbors are not included. In the case of the hypoglycemic agents we see that they perform very well even when all of the tissue data was used. An explanation for this may be that a large portion of the signal from all of the data is recapitulated in the blood only gene correlations. Furthermore, since the expression signal of blood tissue is very distinct (Section 3.2.1), it makes intuitive sense that the expression signature of the target genes of the hypoglycemic agents is also very similar.

On the other hand, the results for anti-psychotic and anti-neoplastic agents appears to be better when the PPI neighbors are included. For example, we see that

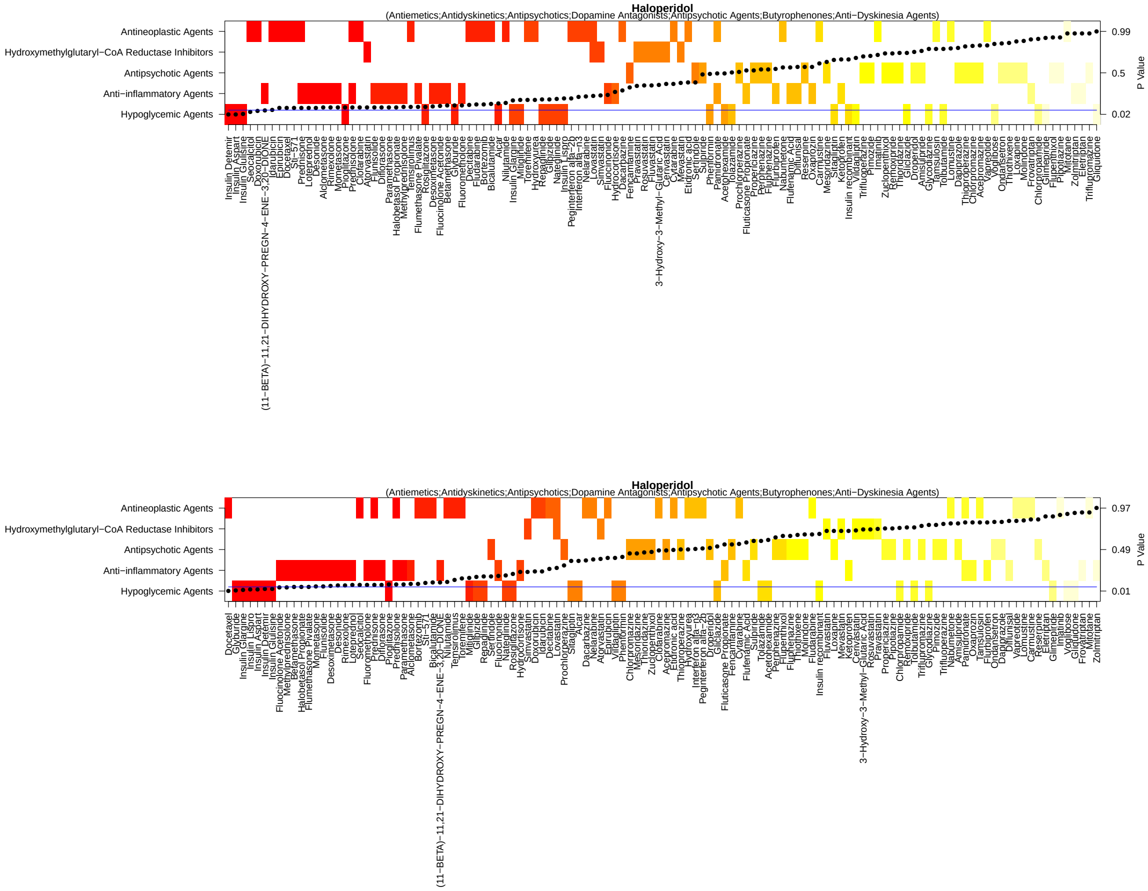


Figure 6-16: These figures depict the 120 drugs sorted in order form most similar to most different to haloperidol when the empirical p-value of the Jaccard index of the target gene overlap is used to compute the similarity to haloperidol. (a) shows the result when only the target genes are used while (b) shows the result when this set is expanded to include the PPI neighbors of the target genes. The black dots in the figures represent the empirical p-value of the Jaccard index for the overlap between haloperidol and the drug noted on the x-axis. The horizontal blue line indicates the 5% significance cutoff. Any values below this line are p-values that are below 0.05. The colored squares indicate the drug type category for each of the drugs. For example, the leftmost entry in (a) is voglibose, a hypoglycemic agent. The colors of these squares are a gradient from red to yellow such that red indicates the lowest score and yellow is the highest. This just provides another visual indication as to how good the score for that drug is.

a large portion of the anti-neoplastic drugs have other anti-neoplastic drugs as close neighbors when either color or kidney tissue gene correlations are used. Furthermore, we also see that unlike when the PPI neighbors are not included, the anti-psychotic

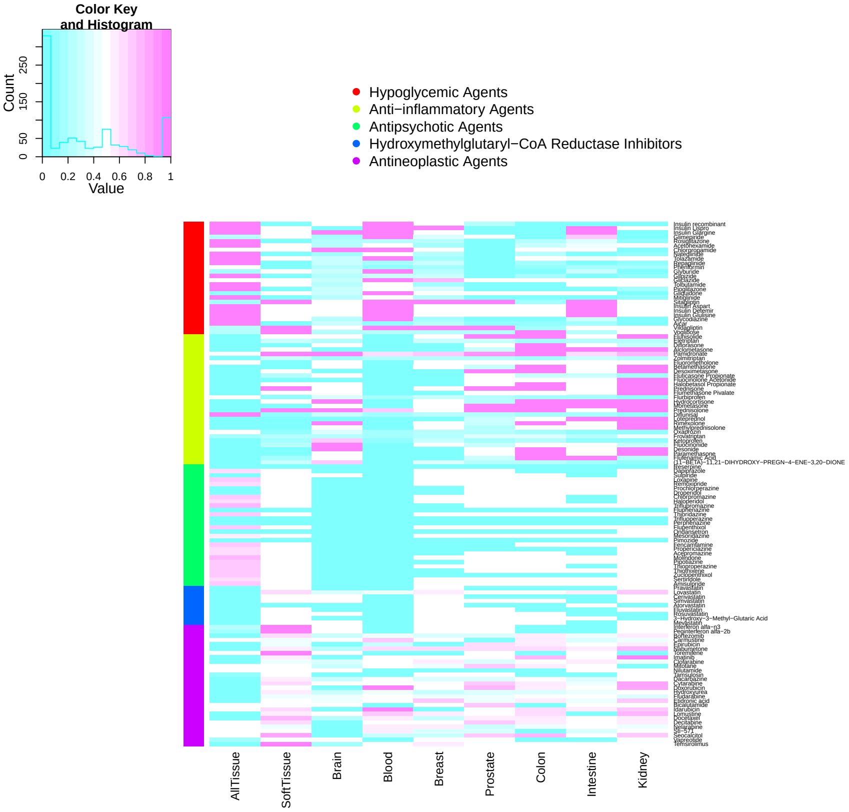


Figure 6-17: A heatmap that depicts the "performance" of the finding the most similar drugs when only the target genes for each drug are used. Each row is the result for one drug across the 9 different tissue conditions. The color in each cell is related to the fraction of the similar samples for the the drug in the row were below the 5% threshold and of the same drug category. The values range from 0 (none of the drugs below the 5% threshold were of the same drug category) and 1 (all of the drugs under the 5% threshold were of the same category) For instance, in the row for "Insulin recombinant" we see that in the all tissue data and blood tissue data conditions many of the most similar drugs were also hypoglycemic agents.

drugs perform relatively well when the brain tissue gene correlations are used.

While these results are preliminary, and a more in-depth study is required, it provides the potential for tissue specific analysis of the effects of drugs. For instance, if we find two drugs that target similar target genes in different tissues, it may be a good idea to investigate the possibility of an adverse interaction between the two drugs. In future work it will be necessary to ensure that we include more tissue types, and include tissue specific PPI networks. Tissue specific PPI networks are imperative because not all proteins are present in all tissues. For example, if a given protein is never expressed in a particular tissue, we should not include the gene corresponding to that protein when looking at the target gene neighbors. Thus, in combination with the similarity network results from Section 6.1.2 and marker gene based results from Section 6.4.1 this sort of analysis can provide useful insights to drug repurposing, may highlight potential adverse drug interactions, and can lay the foundations for future drug analyses.

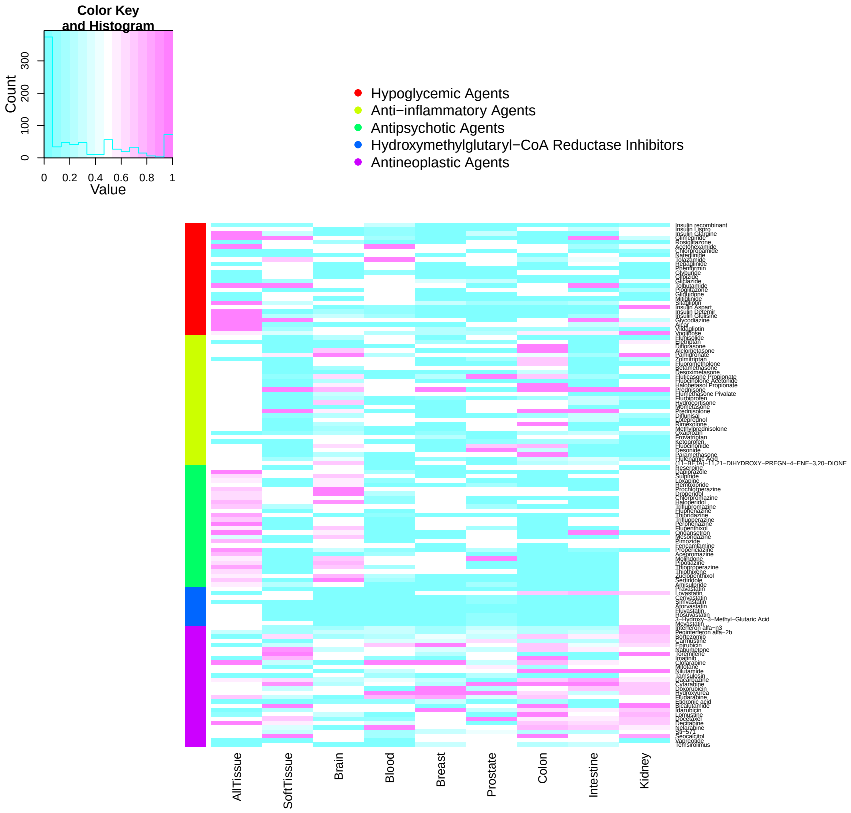


Figure 6-18: A heatmap that depicts the "performance" of the finding the most similar drugs when both the target genes and their immediate PPI neighbors are used. Each row is the result for one drug across the 9 different tissue conditions. The color in each cell is related to the fraction of the similar samples for the the drug in the row were below the 5% threshold and of the same drug category. The values range from 0 (none of the drugs below the 5% threshold were of the same drug category) and 1 (all of the drugs under the 5% threshold were of the same category).

Chapter 7

Concluding remarks

In this work, we have detailed an ontological database framework called Concordia and showed how it can be applied to create a curated gene expression database (Chapter 2). Employing this large set of expression samples labeled with specific phenotypic identifiers, we showed how we can label new expression samples (Chapter 3), how we can elucidate phenotype specific marker genes (Chapter 4), and how it can be applied to creating tissue specific drug similarity scores (Chapter 6). As we will want to expand this database in the future, we also covered active learning based techniques that can be used to efficiently grow the expression database (Chapter 5).

The impact of this enabling technology is extremely broad and significant, finding application in both clinical and biomedical research communities. Our open, standards-based framework for making medical text machine-intelligible may eventually allow our system to lie at the core of large electronic medical record systems, improving the ability of medical professionals and researchers alike to answer questions about both individual patients and patient populations as a whole. In addition, it has been noted that the lack of reproducibility in scientific research has been detrimental to the biomedical community [88]. By making a highly curated phenotype-genotype expression database like ours available to the public, we can make it possible for researchers to perform analyses of their results against a large corpus of data, limiting the effect of the “incidentalome” [66].

7.1 Future work

7.1.1 Applying the Concordia framework to other domains

Having shown the preliminary feasibility of this methodology using the microarray data from NCBI's Gene Expression Omnibus (GEO) [13], we intend to further develop the GEO prototype and then create an application for electronic health records using the same framework. Current medical record software appear to index their records on some basic patient identifier(s) (name, date of birth, social security number, etc.) and are optimized for searching based on those identifiers. This clearly is not ideal for finding the set of all patients with a given phenotypic condition for a clinical study. By using MetaMap [7] to annotate the various fields in each record (doctor's notes, diagnoses, etc.) we can map each record to the set of UMLS concepts that describe its contents. A user of the application derived from the Concordia framework can then search for records based on any medical concept found in UMLS. For example, one can query for skin rashes and cancer to find patients for a study on paraneoplastic processes. Since the UMLS concepts reside in an ontology, even if a patient record never explicitly mentions the general term "cancer," but rather a more specific type of cancer, the system would correctly include this patient's record in the search results.

Although these two individual systems (expression and EHR) can provide useful insights on their own, they could be combined for an even greater impact. If gene expression experiments were to be introduced as a diagnostic tool, then a clinician could find where in the space of diseases this patient's expression profile lies by simply finding the other gene expression experiments that are most similar. We have shown that, indeed, experiments pertaining to the same phenotypic condition cluster together even if they originated from different data sets. Using this information, the clinician could then find all patients that have had the same diagnosis and quickly identify a treatment option. Furthermore, if the gene expression experiments being searched were not the publicly available ones from GEO, but were the actual experiments performed on other patients, one could efficiently locate the other patients that not only had the most similar expression profiles but also the most similar background (gender, ethnicity, weight, etc.). This type of cross data searching and matching is infeasible with current clinical software but would represent only the beginning of what could become possible with the proposed framework.

7.1.2 Expanding the expression database

As it currently stands, we have a highly curated expression database containing 3030 gene expression samples performed on the Affymetrix HG-U133 Plus 2.0 platform. While we were able to perform various studies and extract meaningful biological results using this database, a larger database with more samples for each phenotype and more overall phenotypes, would undoubtedly yield other novel findings. As we demonstrated in Chapter 5, we can use active learning techniques to efficiently expand an existing database for samples relating to particular phenotypes. In addition, we could expand upon the approach we detailed, and employ the concept enrichment statistics described in Chapter 3 in place of (or in addition to) the expression correlation based method.

Furthermore, it would behoove us not to expand the database to include data performed on other organisms and microarray platforms. While this new expression data may not be directly comparable to the data that currently resides in the database, it would enable us to map the results from a greater range of biological studies on to the transcriptomic landscape of tissue and disease.

7.1.3 Concept enrichment using marker genes

The current implementation of the concept enrichment utilizes the sample correlations based on all 20252 genes on the Affymetrix HG-U133 Plus 2.0 platform. However, in Chapter 4 we detailed a method that allows us to compute the relative importance of each gene to a phenotype via the marker gene score. As there are genes that provide no additional signal (and may actually be adding noise) in the expression space of a particular phenotype, it would be worthwhile to explore the effect of using only the most significant genes (highest marker gene scores) when computing the sample correlations. This, however, adds additional complexity when labeling new samples as the correlations need to be computed not just once to all samples in the database, but once for every phenotype to all samples (since we are using a different set of genes to compute the correlations).

7.1.4 Targeted drug therapeutics

As was mentioned in Chapter 6, targeted therapeutics are the future of drug design and administration. By leveraging expression data, we can provide tissue (or any other phenotype) specific similarity measures for drugs. As such, it would be feasible to expand this work to find novel repurposing of existing drugs so that it can be administered for the treatment of a different disease. Unlike the Connectivity Map (CMAP) [69], the use of untreated tissue samples to examine the possible transcriptomic effect of drugs will enable much larger scale high-throughput studies.

As an alternative to using existing expression profiles to generate hypotheses about the potential effects of drugs, it may be possible to use gene expression data of patients to provide them with highly targeted therapeutic regiments. For example, we could create a database of expression profiles for all patients (or a significant subset) that visit a hospital. Using this expression data, in conjunction with their medical records, it could potentially be plausible to find the "most similar patients" and administer the treatment that was seen to be most efficacious.

Appendix A

Data in Concordia

A.1 GEO data in Concordia

A.1.1 GEO series

These are the 192 distinct GEO Series (GSEs) for which there is data in Concordia:

GSE15431, GSE15578, GSE5040, GSE13314, GSE13313, GSE12172, GSE3077, GSE7896, GSE11045, GSE5764, GSE14302, GSE13067, GSE5460, GSE12583, GSE8121, GSE7224, GSE15389, GSE7821, GSE13309, GSE2435, GSE6364, GSE13732, GSE14434, GSE7757, GSE7553, GSE13300, GSE13307, GSE8023, GSE6575, GSE6891, GSE3284, GSE14054, GSE6791, GSE15583, GSE8545, GSE11151, GSE15392, GSE7753, GSE11348, GSE15396, GSE15636, GSE10780, GSE15395, GSE16059, GSE16054, GSE5109, GSE14103, GSE5372, GSE15773, GSE3202, GSE13828, GSE7117, GSE14429, GSE10317, GSE6764, GSE15455, GSE3744, GSE14886, GSE4183, GSE15459, GSE4182, GSE14746, GSE11375, GSE4250, GSE15645, GSE14618, GSE9576, GSE6969, GSE5060, GSE14519, GSE5264, GSE8052, GSE2125, GSE12891, GSE7127, GSE5058, GSE14380, GSE9891, GSE3061, GSE14615, GSE3062, GSE7637, GSE15132, GSE9899, GSE3526, GSE4107, GSE16032, GSE7832, GSE10715, GSE10714, GSE6004, GSE3325, GSE15658, GSE13887, GSE12417, GSE14801, GSE13059, GSE9440, GSE14386, GSE2817, GSE2634, GSE9593, GSE10927, GSE14491, GSE15918, GSE14020, GSE16028, GSE5081, GSE10334, GSE8514, GSE15148, GSE2677, GSE5281, GSE11083, GSE14825, GSE14905, GSE13975, GSE13351, GSE6532, GSE6465, GSE11100, GSE4567, GSE13355, GSE10327, GSE6460, GSE14711, GSE13471, GSE14017, GSE2842, GSE13670, GSE13141, GSE13671, GSE14479, GSE8507, GSE13205, GSE15499, GSE12195, GSE15329, GSE4086, GSE2109, GSE2555, GSE12453, GSE13136, GSE10846, GSE6257, GSE14844, GSE4498, GSE15602, GSE12452, GSE14468, GSE14841, GSE14842, GSE4737, GSE15460, GSE7152, GSE12390, GSE7153, GSE4218, GSE15368, GSE4219, GSE4217, GSE7888, GSE7305, GSE12667, GSE7011, GSE15090, GSE7307, GSE15091, GSE12662, GSE3678, GSE11882, GSE13294, GSE15176, GSE5110, GSE15175, GSE16130, GSE13911, GSE4488, GSE13506, GSE6872, GSE15615, GSE13987, GSE15477, GSE6351, GSE13985, GSE15083, GSE3292, GSE13904, GSE15709, GSE11135, GSE15209, GSE15372

A.1.2 GEO samples

These are the 3030 distinct GEO samples (GSMs) for which there is data in Concordia:

GSM175794, GSM170979, GSM175795, GSM46884, GSM175796, GSM175797, GSM170978, GSM175790, GSM175791, GSM46888, GSM175792, GSM117730, GSM203686, GSM402327, GSM175793, GSM175798, GSM353935, GSM175799, GSM159011, GSM352110, GSM353933, GSM203696, GSM318104, GSM402317, GSM117720, GSM203699, GSM46878, GSM159001, GSM117710, GSM402307, GSM353915, GSM159031, GSM152689, GSM318124, GSM117700, GSM152681, GSM379868, GSM117701, GSM46898, GSM352123, GSM353925, GSM159021, GSM152699, GSM318114, GSM379858, GSM363401, GSM260997, GSM194307, GSM363406, GSM363403, GSM117770, GSM117772, GSM187610, GSM261007, GSM187611, GSM350298, GSM318144, GSM187616, GSM194309, GSM187617, GSM194308, GSM187618, GSM187619, GSM187612, GSM187613, GSM187614, GSM152669, GSM187615, GSM194313, GSM194314, GSM194311, GSM353905, GSM194312, GSM199397, GSM117763, GSM194310, GSM76489, GSM117761, GSM261017, GSM117756, GSM187621, GSM67186, GSM187622, GSM117755, GSM152670, GSM187620, GSM318134, GSM350288, GSM187629, GSM152679, GSM187627, GSM187628, GSM187625, GSM187626, GSM187623, GSM187624, GSM175777, GSM175776, GSM260977, GSM175779, GSM175778, GSM76499, GSM117751, GSM175775, GSM187630, GSM337197, GSM152649, GSM337199, GSM337198, GSM385721, GSM363411, GSM175789, GSM363412, GSM175788, GSM260987, GSM175787, GSM325807, GSM175782, GSM175781, GSM117741, GSM175780, GSM175786, GSM363415, GSM175785, GSM175784, GSM175783, GSM280370, GSM152659, GSM361954, GSM391367, GSM211122, GSM280847, GSM371106, GSM148611, GSM148610, GSM211132, GSM325817, GSM85486, GSM325812, GSM361964, GSM391357, GSM280837, GSM325827, GSM148605, GSM211142, GSM148606, GSM148607, GSM148608, GSM148609, GSM85496, GSM260967, GSM279060, GSM279061, GSM279062, GSM279063, GSM279064, GSM279065, GSM211102, GSM46824, GSM348321, GSM325837, GSM46828, GSM211112, GSM151998, GSM151999, GSM151996, GSM151997, GSM151994, GSM151995, GSM151992, GSM151993, GSM151990, GSM46818, GSM151991, GSM46817, GSM85476, GSM238798, GSM201248, GSM238799, GSM201249, GSM201246, GSM201247, GSM201244, GSM201245, GSM270842, GSM270843, GSM270844, GSM270840, GSM261088, GSM231885, GSM270841, GSM231886, GSM46848, GSM151980, GSM261092, GSM151982, GSM261091, GSM151981, GSM151984, GSM201254, GSM151983, GSM201253, GSM151986, GSM201252, GSM151985, GSM201251, GSM151988, GSM201250, GSM151987, GSM151989, GSM201259, GSM231899, GSM201255, GSM201256, GSM201257, GSM201258, GSM270834, GSM261096, GSM261099, GSM231896, GSM231897, GSM46838, GSM270839, GSM270838, GSM151971, GSM270837, GSM151970, GSM270836, GSM270835, GSM151975, GSM201263, GSM151974, GSM201262, GSM151973, GSM201265, GSM151972, GSM201264, GSM301697, GSM151979, GSM151978, GSM151977, GSM201261, GSM46833, GSM151976, GSM201260, GSM151969, GSM151966, GSM151965, GSM151968, GSM46868, GSM151967, GSM151962, GSM201232, GSM201231, GSM151964, GSM201230, GSM151963, GSM201233, GSM201234, GSM201235, GSM201236, GSM201237, GSM385383, GSM201238, GSM201239, GSM231876, GSM231874, GSM46858, GSM238795, GSM238794, GSM238797, GSM238796, GSM238791, GSM201241, GSM238790, GSM201240, GSM46850, GSM238793, GSM201243, GSM238792, GSM279753, GSM173679, GSM325787, GSM53033, GSM386413, GSM60985, GSM173684, GSM317736, GSM279743, GSM173685, GSM173682, GSM173683, GSM306190, GSM173680, GSM173681, GSM211092, GSM317739, GSM80602, GSM80601, GSM80600, GSM173688, GSM270809, GSM173689, GSM173686, GSM173687, GSM60972, GSM386403, GSM316693, GSM238875, GSM238877, GSM238870, GSM211082, GSM238873, GSM280897, GSM279774, GSM238874, GSM238871, GSM238872, GSM351404, GSM238867, GSM238865, GSM238864, GSM316683, GSM238868, GSM211072, GSM238860, GSM238861, GSM199307, GSM238862, GSM279763, GSM238863, GSM66937, GSM325797, GSM360316, GSM238854, GSM238856, GSM238855, GSM238858, GSM238857, GSM316673, GSM80632, GSM80633, GSM80634, GSM80635, GSM80630, GSM80631, GSM340514, GSM372286, GSM238851, GSM280877, GSM372289, GSM372288, GSM372287, GSM238848, GSM401152, GSM238846, GSM238847, GSM372292, GSM238844, GSM401156, GSM372293, GSM238845, GSM372290, GSM238842, GSM372291, GSM238843, GSM80629, GSM386453, GSM80626, GSM80625, GSM360329, GSM80628, GSM80627, GSM80645, GSM80646, GSM80643, GSM75017, GSM80644, GSM80641, GSM340504, GSM80642, GSM80640, GSM372295, GSM372294, GSM280887, GSM372297, GSM238841, GSM372296, GSM279784, GSM238840, GSM372299, GSM372298, GSM401162, GSM238835, GSM238837, GSM238838, GSM401165, GSM279794, GSM238834, GSM386443, GSM80639, GSM238839, GSM80638, GSM80637, GSM80636, GSM80610, GSM176306, GSM80611, GSM203716, GSM80612, GSM176304, GSM80613, GSM176305, GSM176302, GSM176303, GSM352580, GSM176300, GSM176301, GSM238822, GSM280857, GSM238823, GSM238820, GSM401132, GSM238821, GSM238826, GSM238827, GSM238824, GSM238825, GSM80604, GSM80603, GSM60960, GSM80606, GSM80605, GSM386433, GSM80608, GSM80607, GSM80609, GSM176319, GSM179951, GSM80620, GSM179950, GSM80623, GSM176315, GSM80624, GSM176316, GSM80621, GSM176317, GSM203706, GSM80622, GSM176318, GSM176312, GSM176313, GSM176310, GSM238810, GSM280867, GSM238811, GSM238812, GSM238813, GSM401142, GSM238815, GSM238816, GSM80617, GSM386423, GSM238817, GSM80616, GSM238818, GSM80615, GSM238819, GSM80614, GSM80619, GSM80618, GSM152759, GSM152757, GSM187702, GSM350248, GSM238807, GSM152755, GSM238806, GSM80669, GSM238809, GSM238808, GSM238803, GSM238802, GSM238805, GSM238804, GSM401112, GSM238801, GSM238800, GSM80671, GSM203732, GSM80670, GSM176321, GSM176320, GSM117680, GSM176323, GSM203736, GSM176322, GSM175840, GSM176325, GSM175841, GSM176324, GSM80679, GSM175842, GSM176327, GSM80678, GSM175843, GSM176326, GSM80677, GSM175844, GSM176329, GSM80676, GSM175845, GSM176328, GSM80675, GSM175846, GSM80674, GSM175847, GSM179940, GSM80673, GSM175848, GSM199357, GSM80672, GSM175849, GSM175839, GSM152749, GSM350258, GSM345187, GSM401122, GSM80680, GSM176332, GSM176331, GSM80682, GSM176330, GSM80681, GSM176336, GSM175830, GSM176335, GSM176334, GSM176333, GSM203726, GSM80688, GSM175833, GSM179930, GSM80687, GSM301707, GSM175834, GSM117690, GSM176339, GSM175831, GSM176338, GSM80689, GSM175832, GSM176337, GSM80684, GSM175837, GSM80683, GSM175838, GSM199367, GSM80686, GSM175835, GSM80685, GSM175836, GSM80649, GSM80647, GSM80648, GSM187722, GSM281019, GSM350268, GSM175860, GSM176345, GSM175861, GSM176344, GSM175862, GSM117660, GSM176347, GSM203756, GSM175863, GSM176346, GSM176341, GSM176340, GSM176343, GSM176342, GSM80653, GSM175868, GSM80652, GSM175869, GSM80651, GSM340534, GSM80650, GSM152739, GSM80657, GSM53093, GSM175864, GSM199377, GSM80656, GSM175865, GSM80655, GSM175866, GSM80654, GSM175867, GSM179920, GSM80658, GSM80659, GSM281009, GSM187712, GSM176360, GSM401102, GSM176361, GSM350278, GSM175851, GSM176358, GSM175852, GSM176357, GSM203746, GSM176356, GSM175850, GSM117670, GSM176355, GSM176354, GSM176353, GSM80660, GSM176352, GSM179918, GSM80662, GSM368398, GSM175859, GSM152729, GSM80661, GSM53083, GSM340524, GSM80664, GSM175857, GSM80663, GSM175858, GSM80666, GSM175855, GSM80665, GSM175856, GSM80668, GSM175853, GSM179910, GSM80667, GSM175854, GSM176359, GSM199387, GSM317794, GSM316663, GSM176370, GSM176372, GSM176371, GSM351424, GSM175806, GSM350208, GSM175807, GSM175808, GSM175809, GSM179900, GSM175801, GSM389778, GSM175800, GSM175803, GSM122548, GSM152719, GSM175802, GSM175805, GSM53073, GSM175804, GSM176362, GSM176363, GSM203776, GSM176364, GSM345147, GSM176365, GSM199317, GSM176366, GSM176367, GSM306160, GSM176368, GSM176369, GSM176383, GSM176382, GSM176381, GSM316653, GSM350218, GSM351414, GSM95519, GSM389788, GSM95522, GSM95523, GSM95524, GSM53063, GSM95525, GSM152709, GSM176375, GSM199327, GSM176376, GSM95520, GSM345137, GSM176373, GSM203766, GSM95521, GSM176374, GSM176392, GSM345177, GSM170983, GSM176391, GSM170980, GSM176390, GSM95509, GSM95508, GSM350228, GSM175828, GSM175829, GSM95513, GSM80696, GSM175825, GSM95514, GSM80697, GSM53053, GSM175824, GSM170597, GSM199337, GSM95511, GSM80694, GSM175827, GSM170596, GSM122528, GSM95512, GSM80695, GSM175826, GSM170595, GSM95517, GSM175821, GSM95518, GSM175820, GSM95515, GSM80698, GSM175823, GSM95516, GSM80699, GSM175822, GSM306180, GSM170590, GSM176388, GSM176389, GSM80692, GSM170594, GSM176384, GSM95510, GSM80693, GSM170593, GSM176385, GSM80690, GSM170592, GSM176386, GSM80691, GSM170591, GSM176387, GSM203796, GSM170992, GSM345167, GSM350238, GSM175819, GSM53043, GSM53046, GSM175817, GSM175818, GSM95500, GSM175816, GSM95501, GSM175815, GSM95502, GSM175814, GSM199347, GSM95503, GSM175813, GSM95504, GSM175812, GSM170589, GSM95505, GSM175811, GSM170588, GSM95506, GSM175810, GSM95507, GSM306170, GSM345157, GSM203786, GSM176396, GSM385060, GSM73686, GSM76579, GSM345117, GSM337033, GSM158711, GSM385070, GSM345127, GSM76587, GSM76585, GSM340494, GSM96276, GSM337023, GSM76559, GSM361371, GSM60588, GSM176297, GSM176296, GSM337013, GSM361381, GSM158731, GSM114096, GSM76569, GSM335834, GSM345107, GSM176287, GSM155701, GSM176294, GSM176295, GSM176292, GSM176293, GSM176290, GSM176291, GSM337003, GSM158721, GSM175890, GSM175892, GSM175891, GSM175894, GSM175893, GSM175896, GSM175895, GSM89091, GSM60562, GSM175898, GSM175897, GSM175899, GSM385020, GSM306210, GSM155711, GSM361351, GSM385010, GSM152769, GSM390943, GSM270789, GSM337073, GSM89081, GSM155721, GSM361361, GSM385030, GSM306220, GSM387979, GSM152779, GSM337063, GSM175872, GSM76595, GSM175871, GSM89071, GSM175874, GSM89072, GSM175873, GSM60548, GSM175870, GSM101100, GSM175879, GSM101101, GSM385040, GSM101102, GSM101103, GSM175876, GSM101104, GSM389824, GSM361331, GSM175875, GSM101105, GSM175878, GSM101106, GSM175877, GSM152789, GSM390158, GSM337053, GSM281029, GSM387969, GSM76590, GSM89060, GSM175885, GSM89061, GSM175884, GSM175883, GSM175882, GSM175881, GSM175880, GSM60538, GSM361341, GSM385050, GSM306200, GSM175889, GSM175888, GSM175887, GSM389813, GSM175886, GSM270799, GSM387959, GSM152799, GSM337043, GSM281039, GSM143900, GSM378170, GSM387949, GSM88971, GSM51690, GSM261312, GSM46948, GSM46941, GSM395790, GSM387939, GSM361321, GSM88981, GSM46938, GSM261302, GSM51680, GSM46936, GSM395780, GSM387929, GSM88991, GSM88997, GSM46928, GSM310839, GSM310838, GSM261332, GSM280009, GSM38103, GSM38104, GSM38100, GSM387919, GSM94603, GSM94604, GSM46918, GSM94605, GSM261322, GSM134589, GSM134588, GSM134587, GSM134586, GSM134584, GSM187595, GSM187596, GSM187593, GSM93568, GSM187594, GSM187599, GSM187597, GSM187598, GSM287293, GSM387909, GSM134591, GSM403597, GSM401092, GSM73656, GSM88949, GSM46975, GSM46976, GSM280028, GSM46973, GSM173691, GSM173690, GSM328997, GSM46960, GSM46961, GSM88955, GSM73666, GSM46968, GSM88951, GSM187586, GSM187587, GSM187588, GSM187589, GSM187584, GSM187585, GSM187590, GSM187592, GSM187591, GSM73676, GSM88961, GSM46958, GSM88962, GSM175903, GSM175904, GSM175901, GSM175902, GSM372348, GSM175900, GSM199417, GSM175909, GSM175908, GSM350308, GSM175907, GSM175906, GSM175905, GSM372358, GSM184639, GSM199427, GSM401062, GSM184636, GSM184637, GSM101095, GSM184638, GSM350318, GSM101096, GSM101097, GSM101098, GSM101099, GSM336033, GSM336983, GSM401076, GSM184640, GSM184641, GSM184644, GSM184645, GSM184642, GSM184643, GSM184648, GSM401072, GSM184649, GSM184646, GSM184647, GSM101998, GSM199407, GSM336043, GSM250001, GSM143898, GSM184650, GSM184651, GSM184652, GSM184653, GSM184654, GSM184655, GSM184656, GSM184657, GSM184658, GSM401082, GSM184659, GSM80900, GSM365142, GSM310849, GSM176409, GSM80901, GSM365143, GSM80902, GSM365140, GSM176407, GSM80903, GSM365141, GSM176408, GSM80904, GSM310845, GSM238951, GSM189790, GSM310846, GSM176406, GSM310847, GSM310848, GSM310844, GSM339558, GSM339559, GSM339566, GSM277701, GSM339565, GSM339568, GSM238949, GSM339567, GSM339562, GSM339561, GSM339564, GSM184665, GSM339563, GSM184664, GSM238943, GSM184663, GSM189782, GSM365139, GSM238944, GSM184662, GSM189783, GSM365138, GSM339560, GSM238941, GSM184661, GSM189784, GSM365137, GSM238942, GSM184660, GSM189785, GSM365136, GSM238947, GSM189786, GSM365135, GSM238948, GSM189787, GSM365134, GSM238945, GSM189788, GSM365133, GSM238946, GSM189789, GSM80913, GSM365151, GSM336993, GSM176418, GSM365152, GSM176419, GSM80911, GSM365153, GSM80912, GSM365154, GSM310858, GSM176414, GSM189781, GSM310859, GSM176415, GSM189780, GSM176416, GSM365150, GSM310857, GSM176417, GSM176410, GSM176411, GSM310852, GSM176412, GSM310853, GSM176413, GSM46908, GSM310850, GSM310851, GSM339569, GSM387575, GSM189779, GSM277711, GSM365149, GSM189773, GSM365148, GSM189774, GSM189771, GSM189772, GSM365145, GSM189777, GSM365144, GSM189778, GSM365147, GSM189775, GSM365146, GSM189776, GSM365160, GSM176427, GSM365161, GSM176428, GSM176425, GSM189770, GSM176426, GSM365162, GSM176429, GSM387565, GSM310860, GSM176420, GSM310861, GSM310862, GSM176423, GSM176424, GSM176421, GSM176422, GSM189768, GSM189769, GSM365158, GSM189764, GSM365157, GSM189765, GSM365156, GSM189766, GSM365155, GSM189767, GSM189760, GSM189761, GSM238963, GSM189762, GSM365159, GSM189763, GSM176436, GSM176437, GSM176438, GSM176439, GSM176430, GSM176431, GSM94599, GSM176432, GSM94598, GSM176433, GSM176434, GSM176435, GSM339557, GSM189759, GSM189757, GSM189758, GSM189755, GSM189756, GSM189753, GSM189754, GSM238952, GSM189751, GSM238953, GSM189752, GSM238955, GSM187600, GSM345097, GSM125006, GSM187606, GSM187605, GSM187608, GSM187607, GSM187602, GSM187601, GSM187604, GSM187603, GSM242672, GSM175989, GSM242673, GSM158791, GSM176446, GSM100898, GSM175985, GSM150220, GSM176228, GSM176440, GSM187609, GSM176227, GSM242674, GSM175987, GSM150222, GSM76509, GSM242675, GSM175988, GSM169531, GSM150221, GSM176229, GSM176441, GSM175981, GSM150224, GSM176224, GSM175982, GSM150223, GSM176223, GSM175983, GSM150226, GSM176226, GSM175984, GSM150225, GSM176225, GSM176220, GSM176448, GSM150227, GSM176447, GSM176222, GSM175980, GSM176221, GSM176449, GSM345087, GSM176240, GSM176456, GSM175978, GSM176455, GSM175979, GSM176454, GSM175976, GSM176453, GSM175977, GSM176452, GSM175974, GSM176239, GSM176451, GSM175975, GSM176238, GSM176450, GSM176237, GSM175973, GSM176236, GSM176235, GSM176234, GSM176233, GSM176232, GSM100888, GSM176231, GSM176230, GSM391616, GSM365113, GSM365114, GSM125026, GSM365115, GSM365116, GSM365117, GSM365118, GSM345077, GSM365119, GSM277721, GSM176206, GSM176205, GSM175965, GSM176208, GSM363399, GSM175966, GSM176207, GSM363398, GSM175967, GSM176466, GSM176209, GSM363396, GSM363395, GSM306240, GSM365121, GSM365120, GSM365124, GSM365125, GSM365122, GSM125016, GSM391626, GSM365123, GSM67153, GSM365128, GSM365129, GSM365126, GSM365127, GSM351339, GSM277731, GSM169530, GSM80567, GSM277094, GSM175954, GSM176219, GSM80566, GSM277095, GSM175955, GSM176218, GSM80569, GSM277092, GSM175952, GSM176217, GSM80568, GSM277093, GSM175953, GSM176216, GSM80563, GSM277098, GSM175958, GSM169525, GSM80562, GSM277099, GSM175959, GSM169524, GSM80565, GSM277096, GSM175956, GSM169527, GSM80564, GSM277097, GSM175957, GSM169526, GSM169529, GSM176211, GSM306230, GSM169528, GSM176210, GSM80561, GSM365132, GSM277090, GSM175950, GSM176215, GSM365131, GSM277091, GSM175951, GSM176214, GSM365130, GSM176213, GSM176212, GSM350348, GSM151324, GSM363383, GSM175949, GSM158741, GSM176271, GSM176270, GSM176273, GSM176272, GSM176267, GSM176268, GSM372301, GSM175940, GSM176269, GSM372300, GSM336013, GSM80571, GSM176263, GSM80572, GSM176264, GSM176265, GSM80570, GSM176266, GSM80575, GSM175946, GSM80576, GSM372306, GSM175945, GSM80573, GSM76549, GSM175948, GSM80574, GSM372308, GSM175947, GSM80579, GSM372303, GSM363379, GSM175942, GSM372302, GSM175941, GSM80577, GSM372305, GSM363377, GSM175944, GSM80578, GSM372304, GSM175943, GSM388709, GSM363390, GSM151314, GSM350358, GSM363392, GSM363394, GSM175938, GSM175939, GSM158751, GSM391606, GSM176280, GSM336023, GSM176278, GSM176279, GSM80580, GSM60601, GSM176276, GSM80581, GSM176277, GSM80582, GSM176274, GSM80583, GSM176275, GSM80584, GSM175937, GSM80585, GSM76539, GSM363385, GSM175936, GSM158761, GSM80586, GSM372318, GSM175935, GSM80587, GSM363387, GSM175934, GSM80588, GSM175933, GSM80589, GSM363389, GSM175932, GSM175931, GSM175930, GSM350328, GSM175927, GSM175928, GSM175929, GSM151344, GSM176251, GSM89101, GSM176250, GSM80593, GSM176241, GSM80594, GSM176242, GSM80591, GSM176243, GSM80592, GSM176244, GSM176245, GSM80590, GSM176246, GSM176247, GSM176248, GSM76529, GSM175920, GSM176249, GSM80599, GSM242653, GSM175922, GSM242652, GSM175921, GSM80597, GSM242651, GSM175924, GSM80598, GSM372328, GSM242650, GSM175923, GSM80595, GSM175926, GSM158771, GSM80596, GSM175925, GSM175918, GSM175919, GSM175916, GSM175917, GSM151334, GSM350338, GSM96266, GSM176262, GSM176261, GSM176260, GSM176254, GSM176255, GSM176252, GSM176253, GSM242668, GSM176258, GSM242667, GSM176259, GSM176256, GSM242669, GSM176257, GSM372338, GSM175911, GSM175910, GSM242666, GSM76519, GSM175915, GSM175914, GSM175913, GSM175912, GSM158781, GSM377475, GSM113822, GSM158811, GSM85219, GSM85217, GSM85218, GSM371383, GSM85215, GSM85216, GSM199167, GSM350139, GSM125066, GSM148493, GSM113812, GSM148491, GSM148495, GSM148496, GSM158801, GSM357635, GSM371373, GSM199157, GSM125076, GSM148488, GSM335978, GSM148485, GSM125036, GSM148487, GSM199197, GSM350155, GSM350156, GSM199187, GSM350158, GSM102578, GSM350151, GSM350152, GSM350153, GSM350154, GSM125046, GSM335988, GSM159162, GSM371393, GSM350150, GSM350146, GSM102568, GSM350147, GSM199177, GSM350144, GSM350145, GSM350142, GSM249991, GSM350143, GSM350140, GSM350141, GSM350148, GSM125056, GSM350149, GSM277695, GSM158851, GSM277696, GSM114526, GSM176182, GSM176183, GSM176184, GSM114525, GSM176185, GSM176180, GSM176181, GSM176179, GSM51710, GSM176176, GSM176175, GSM176178, GSM176177, GSM249981, GSM151304, GSM158841, GSM114535, GSM176173, GSM176174, GSM176171, GSM176172, GSM261292, GSM176170, GSM387809, GSM114534, GSM261282, GSM176169, GSM51700, GSM176168, GSM176167, GSM176166, GSM176165, GSM176164, GSM277691, GSM249971, GSM113802, GSM114506, GSM158831, GSM114504, GSM114505, GSM125086, GSM261272, GSM387819, GSM249961, GSM85227, GSM85226, GSM85228, GSM158821, GSM85221, GSM85220, GSM85223, GSM85222, GSM85225, GSM114515, GSM85224, GSM114516, GSM125096, GSM176186, GSM387829, GSM261262, GSM249950, GSM402152, GSM335522, GSM150209, GSM386291, GSM249940, GSM312934, GSM161820, GSM102512, GSM80800, GSM287323, GSM261252, GSM387839, GSM361610, GSM102518, GSM371309, GSM371306, GSM371305, GSM371308, GSM371307, GSM371302, GSM327292, GSM371301, GSM371304, GSM371303, GSM249930, GSM150201, GSM150208, GSM161810, GSM335512, GSM161811, GSM287333, GSM161812, GSM161813, GSM361620, GSM312924, GSM102508, GSM387849, GSM102507, GSM261242, GSM327282, GSM150210, GSM161819, GSM249920, GSM161818, GSM161815, GSM161814, GSM161817, GSM161816, GSM312911, GSM312912, GSM155672, GSM312910, GSM155671, GSM287343, GSM387859, GSM261232, GSM312913, GSM312914, GSM361242, GSM161806, GSM161805, GSM161804, GSM161803, GSM249910, GSM161809, GSM155681, GSM161808, GSM161807, GSM312900, GSM312901, GSM287353, GSM312906, GSM312907, GSM312908, GSM387869, GSM312909, GSM261222, GSM312902, GSM312903, GSM312904, GSM312905, GSM155691, GSM249900, GSM183234, GSM261212, GSM387879, GSM102553, GSM102555, GSM102556, GSM155651, GSM102558, GSM183230, GSM386245, GSM335572, GSM387889, GSM155668, GSM155669, GSM261202, GSM155665, GSM155666, GSM155667, GSM183240, GSM102548, GSM155661, GSM155670, GSM391596, GSM386255, GSM335562, GSM152009, GSM102538, GSM152006, GSM152005, GSM152008, GSM152007, GSM287303, GSM152002, GSM152001, GSM152004, GSM152003, GSM387899, GSM152000, GSM335552, GSM386225, GSM335938, GSM171597, GSM199027, GSM286700, GSM152017, GSM102528, GSM152016, GSM152015, GSM287313, GSM152014, GSM183220, GSM260703, GSM152013, GSM312944, GSM260702, GSM152012, GSM152011, GSM152010, GSM335532, GSM335542, GSM386235, GSM377465, GSM335942, GSM335941, GSM335940, GSM199037, GSM327202, GSM80868, GSM80867, GSM80869, GSM80874, GSM80870, GSM80871, GSM80872, GSM80873, GSM333446, GSM199047, GSM151294, GSM327212, GSM198042, GSM80887, GSM80888, GSM80885, GSM80886, GSM80883, GSM80884, GSM80881, GSM80882, GSM333436, GSM317934, GSM317933, GSM151284, GSM199057, GSM198052, GSM80845, GSM198053, GSM198050, GSM327222, GSM198051, GSM198049, GSM198048, GSM80851, GSM198047, GSM198046, GSM80853, GSM198045, GSM198044, GSM198043, GSM151274, GSM199067, GSM80861, GSM80865, GSM80866, GSM80864, GSM333456, GSM287383, GSM93939, GSM80823, GSM93938, GSM80824, GSM80825, GSM80826, GSM199077, GSM337202, GSM199087, GSM337203, GSM279998, GSM337200, GSM337201, GSM80831, GSM93944, GSM93943, GSM93941, GSM287373, GSM93946, GSM350413, GSM93948, GSM337205, GSM337204, GSM337207, GSM74882, GSM337206, GSM337209, GSM337208, GSM337210, GSM337211, GSM337212, GSM337213, GSM337214, GSM199097, GSM93954, GSM80844, GSM80843, GSM80842, GSM80841, GSM93950, GSM287363, GSM93952, GSM80801, GSM80802, GSM80803, GSM80804, GSM350423, GSM80805, GSM80806, GSM80807, GSM80808, GSM80809, GSM337219, GSM337218, GSM337217, GSM337216, GSM337215, GSM337224, GSM337225, GSM337222, GSM337223, GSM337220, GSM337221, GSM80811, GSM286660, GSM80810, GSM80814, GSM80815, GSM80812, GSM93927, GSM80813, GSM80818, GSM287393, GSM80819, GSM80816, GSM80817, GSM337227, GSM371403, GSM337226, GSM350433, GSM337229, GSM337228, GSM337233, GSM337234, GSM337235, GSM337236, GSM337230, GSM337231, GSM337232, GSM80822, GSM80821, GSM80820, GSM286650, GSM176128, GSM176129, GSM38094, GSM158891, GSM337241, GSM176120, GSM337240, GSM176121, GSM337243, GSM176122, GSM337242, GSM176123, GSM337245, GSM176124, GSM337244, GSM176125, GSM76640, GSM337247, GSM272315, GSM176126, GSM337246, GSM176127, GSM337237, GSM337238, GSM350443, GSM337239, GSM176130, GSM125106, GSM286690, GSM286670, GSM176139, GSM337250, GSM75563, GSM337254, GSM176133, GSM337253, GSM176134, GSM337252, GSM176131, GSM337251, GSM176132, GSM378160, GSM337258, GSM176137, GSM76630, GSM337257, GSM176138, GSM337256, GSM176135, GSM337255, GSM176136, GSM337248, GSM48672, GSM350453, GSM337249, GSM176141, GSM176140, GSM286680, GSM337260, GSM158871, GSM75553, GSM119369, GSM176146, GSM176147, GSM337269, GSM176148, GSM176149, GSM176142, GSM89001, GSM176143, GSM176144, GSM176145, GSM176150, GSM74892, GSM242033, GSM176152, GSM242032, GSM176151, GSM350463, GSM337259, GSM158861, GSM277681, GSM158881, GSM119379, GSM176159, GSM337279, GSM176157, GSM176158, GSM176155, GSM199107, GSM176156, GSM89011, GSM176153, GSM176154, GSM176163, GSM350473, GSM176162, GSM176161, GSM176160, GSM175998, GSM175999, GSM175996, GSM175994, GSM277678, GSM175995, GSM175992, GSM175993, GSM175990, GSM175991, GSM38054, GSM89021, GSM76600, GSM179780, GSM337289, GSM350168, GSM359509, GSM199117, GSM50703, GSM139018, GSM139017, GSM139019, GSM151264, GSM179790, GSM89031, GSM242031, GSM38064, GSM337299, GSM38068, GSM350178, GSM119359, GSM119354, GSM199127, GSM179784, GSM179786, GSM89041, GSM139002, GSM176103, GSM139003, GSM176102, GSM139004, GSM176105, GSM139005, GSM176104, GSM80891, GSM80890, GSM76620, GSM176101, GSM176100, GSM38074, GSM199137, GSM80899, GSM176107, GSM80898, GSM350188, GSM176106, GSM80897, GSM176109, GSM176108, GSM80889, GSM103559, GSM89046, GSM150196, GSM150197, GSM150198, GSM150199, GSM139015, GSM176116, GSM139016, GSM176115, GSM139013, GSM176114, GSM89051, GSM139014, GSM176113, GSM139011, GSM176112, GSM139012, GSM176111, GSM76610, GSM176110, GSM139010, GSM350198, GSM38084, GSM199147, GSM176119, GSM176118, GSM176117, GSM139009, GSM139008, GSM139007, GSM125116, GSM139006, GSM194087, GSM194088, GSM194089, GSM203643, GSM194083, GSM194084, GSM96897, GSM194085, GSM203646, GSM96898, GSM158911, GSM194086, GSM343815, GSM159051, GSM187752, GSM281300, GSM231907, GSM231906, GSM194091, GSM194090, GSM102458, GSM194093, GSM194092, GSM102455, GSM387029, GSM312875, GSM102450, GSM102451, GSM203656, GSM158901, GSM194096, GSM194097, GSM194094, GSM194095, GSM261192, GSM343825, GSM231916, GSM159041, GSM187762, GSM261184, GSM249890, GSM281310, GSM102447, GSM199297, GSM102449, GSM102448, GSM387019, GSM312862, GSM158931, GSM203666, GSM159071, GSM211450, GSM158463, GSM158464, GSM187732, GSM377358, GSM231926, GSM349749, GSM211449, GSM249880, GSM387009, GSM176098, GSM176099, GSM312894, GSM102478, GSM312896, GSM312897, GSM312898, GSM312899, GSM211446, GSM281320, GSM211447, GSM199287, GSM211448, GSM194075, GSM158921, GSM159061, GSM194078, GSM194079, GSM203676, GSM402247, GSM194076, GSM194077, GSM176097, GSM187742, GSM176096, GSM176095, GSM343805, GSM176094, GSM176093, GSM176092, GSM231936, GSM176091, GSM349739, GSM176090, GSM249870, GSM176089, GSM176087, GSM318094, GSM176088, GSM402257, GSM194082, GSM281330, GSM102468, GSM194081, GSM194080, GSM199277, GSM170833, GSM187792, GSM176080, GSM176081, GSM176082, GSM231946, GSM176083, GSM176084, GSM176085, GSM176086, GSM159091, GSM158951, GSM152569, GSM281340, GSM402267, GSM102498, GSM272305, GSM249860, GSM176077, GSM318084, GSM176076, GSM176079, GSM176078, GSM261151, GSM261152, GSM85506, GSM170835, GSM176070, GSM176071, GSM176074, GSM176075, GSM176072, GSM231956, GSM176073, GSM231950, GSM388192, GSM158941, GSM231952, GSM159081, GSM152579, GSM102488, GSM402277, GSM176068, GSM85513, GSM261146, GSM176067, GSM85514, GSM261143, GSM176066, GSM85515, GSM249850, GSM176065, GSM85516, GSM318074, GSM170823, GSM85517, GSM261142, GSM85518, GSM85519, GSM176069, GSM176061, GSM170850, GSM176062, GSM231966, GSM176063, GSM359583, GSM176064, GSM170855, GSM353428, GSM261182, GSM170853, GSM187772, GSM343837, GSM176060, GSM203626, GSM152589, GSM158971, GSM388182, GSM402287, GSM158981, GSM335602, GSM261172, GSM170858, GSM176059, GSM176058, GSM261174, GSM170857, GSM176055, GSM176054, GSM249840, GSM176057, GSM176056, GSM176052, GSM231976, GSM176053, GSM359593, GSM176050, GSM249820, GSM152594, GSM176051, GSM343847, GSM170841, GSM187782, GSM170844, GSM170843, GSM152599, GSM203636, GSM158961, GSM203641, GSM323169, GSM402297, GSM323168, GSM176049, GSM176048, GSM261162, GSM170848, GSM176047, GSM171011, GSM170849, GSM176046, GSM249830, GSM171012, GSM176045, GSM176044, GSM176043, GSM261113, GSM211032, GSM261112, GSM329007, GSM261117, GSM261116, GSM137954, GSM287463, GSM387731, GSM386393, GSM335622, GSM155968, GSM367219, GSM155969, GSM315621, GSM280907, GSM231986, GSM249810, GSM211042, GSM261102, GSM315622, GSM183301, GSM315623, GSM183300, GSM315624, GSM315625, GSM183302, GSM329017, GSM137964, GSM387741, GSM117629, GSM261109, GSM335612, GSM117632, GSM249800, GSM312816, GSM277128, GSM277129, GSM277126, GSM277127, GSM277125, GSM261134, GSM211052, GSM261132, GSM287443, GSM335642, GSM261138, GSM261137, GSM137934, GSM137931, GSM38376, GSM155989, GSM335652, GSM155988, GSM277132, GSM277131, GSM277130, GSM280927, GSM277137, GSM277138, GSM277139, GSM211062, GSM277133, GSM261122, GSM277134, GSM277135, GSM277136, GSM387721, GSM137945, GSM335632, GSM137944, GSM287453, GSM261127, GSM117649, GSM38386, GSM373559, GSM280917, GSM137994, GSM277109, GSM287423, GSM277108, GSM277103, GSM277102, GSM277101, GSM277100, GSM277107, GSM277106, GSM277105, GSM277104, GSM201302, GSM377338, GSM201301, GSM201300, GSM155920, GSM277110, GSM280947, GSM201304, GSM201303, GSM155923, GSM155922, GSM155921, GSM38356, GSM155928, GSM155927, GSM287433, GSM155919, GSM387789, GSM158465, GSM158466, GSM158467, GSM158468, GSM312826, GSM158469, GSM353885, GSM377348, GSM158471, GSM280937, GSM158470, GSM158473, GSM158472, GSM158475, GSM158474, GSM335662, GSM38366, GSM287403, GSM102438, GSM353895, GSM280967, GSM155948, GSM155947, GSM287413, GSM137984, GSM102428, GSM312849, GSM211022, GSM211012, GSM280957, GSM101301, GSM38346, GSM117610, GSM80725, GSM272192, GSM80724, GSM272193, GSM80727, GSM327342, GSM272190, GSM80726, GSM335582, GSM272191, GSM80729, GSM386311, GSM80728, GSM280979, GSM138034, GSM272295, GSM183260, GSM80730, GSM239824, GSM80731, GSM239825, GSM80732, GSM272185, GSM239826, GSM80733, GSM80734, GSM272183, GSM80738, GSM335592, GSM80737, GSM386301, GSM272180, GSM80736, GSM272181, GSM80735, GSM327352, GSM272182, GSM117587, GSM80739, GSM337309, GSM280989, GSM138044, GSM80740, GSM272177, GSM80741, GSM286730, GSM272176, GSM183250, GSM272172, GSM80742, GSM272175, GSM80743, GSM272174, GSM327322, GSM183290, GSM386331, GSM272170, GSM53113, GSM272171, GSM80749, GSM80748, GSM280999, GSM138054, GSM272169, GSM134694, GSM272164, GSM272163, GSM272162, GSM272275, GSM272161, GSM286720, GSM272168, GSM80750, GSM80751, GSM272165, GSM386321, GSM183280, GSM80759, GSM327332, GSM80758, GSM53103, GSM80757, GSM272160, GSM134690, GSM134691, GSM134692, GSM134693, GSM272159, GSM134688, GSM272158, GSM134687, GSM134689, GSM272151, GSM272150, GSM272152, GSM272155, GSM272154, GSM183270, GSM272285, GSM272157, GSM80761, GSM387799, GSM286710, GSM272156, GSM337339, GSM201279, GSM401293, GSM201278, GSM201277, GSM316703, GSM53133, GSM137924, GSM201286, GSM201287, GSM201284, GSM201285, GSM201282, GSM201283, GSM201280, GSM201281, GSM119685, GSM119684, GSM119683, GSM119682, GSM179801, GSM201267, GSM119688, GSM179800, GSM201266, GSM119687, GSM201269, GSM337349, GSM119686, GSM201268, GSM119681, GSM53123, GSM119680, GSM316713, GSM137912, GSM137910, GSM80701, GSM80700, GSM138004, GSM201273, GSM138003, GSM201274, GSM119679, GSM138002, GSM201275, GSM201276, GSM137916, GSM201270, GSM137914, GSM201271, GSM201272, GSM179810, GSM201299, GSM337319, GSM80706, GSM53153, GSM117577, GSM80707, GSM80708, GSM316723, GSM80709, GSM80702, GSM80703, GSM80704, GSM80705, GSM80710, GSM80712, GSM80711, GSM347925, GSM347924, GSM137904, GSM347923, GSM347922, GSM347921, GSM138014, GSM201289, GSM201288, GSM124996, GSM179820, GSM337329, GSM80719, GSM80717, GSM80718, GSM53143, GSM80715, GSM352629, GSM179827, GSM80716, GSM80713, GSM80714, GSM80723, GSM272194, GSM80722, GSM272195, GSM80721, GSM272196, GSM80720, GSM272197, GSM347916, GSM272198, GSM272199, GSM347918, GSM347917, GSM162960, GSM201290, GSM162961, GSM201291, GSM162962, GSM201292, GSM201293, GSM201294, GSM201295, GSM201296, GSM138024, GSM201297, GSM201298, GSM119649, GSM176025, GSM162954, GSM119648, GSM176026, GSM359603, GSM162957, GSM119647, GSM176027, GSM272215, GSM170867, GSM162956, GSM119646, GSM176028, GSM176021, GSM176022, GSM176023, GSM199217, GSM176024, GSM53173, GSM158991, GSM176029, GSM53170, GSM378838, GSM378837, GSM378836, GSM378831, GSM119651, GSM378830, GSM170862, GSM119652, GSM179830, GSM176031, GSM119650, GSM176030, GSM378835, GSM170865, GSM162958, GSM119655, GSM378834, GSM170866, GSM162959, GSM119656, GSM378833, GSM119653, GSM378832, GSM119654, GSM119636, GSM176038, GSM119635, GSM176039, GSM272225, GSM119638, GSM176036, GSM162943, GSM119637, GSM176037, GSM162942, GSM176034, GSM162941, GSM119639, GSM176035, GSM162940, GSM176032, GSM176033, GSM53163, GSM199227, GSM378826, GSM378825, GSM95473, GSM378828, GSM378827, GSM95475, GSM95474, GSM378829, GSM95477, GSM53167, GSM95476, GSM95479, GSM370399, GSM176042, GSM95478, GSM176041, GSM378820, GSM119640, GSM176040, GSM179840, GSM119641, GSM378822, GSM119642, GSM378821, GSM119643, GSM378824, GSM119644, GSM378823, GSM119645, GSM176000, GSM176001, GSM162931, GSM176002, GSM162930, GSM176003, GSM162933, GSM176004, GSM162932, GSM176005, GSM162935, GSM119669, GSM176006, GSM162934, GSM119668, GSM176007, GSM95480, GSM176008, GSM176009, GSM95488, GSM95487, GSM119670, GSM95486, GSM378819, GSM95485, GSM378818, GSM95484, GSM378817, GSM95483, GSM378816, GSM95482, GSM378815, GSM95481, GSM378814, GSM378813, GSM162936, GSM119677, GSM378812, GSM337359, GSM162937, GSM119678, GSM378811, GSM162938, GSM119675, GSM162939, GSM159101, GSM119673, GSM119674, GSM119671, GSM95489, GSM119672, GSM179850, GSM176012, GSM176013, GSM199207, GSM176010, GSM179870, GSM176011, GSM272205, GSM119658, GSM176016, GSM272204, GSM119657, GSM176017, GSM176014, GSM272202, GSM119659, GSM176015, GSM272201, GSM95490, GSM176018, GSM95491, GSM176019, GSM53183, GSM281280, GSM95497, GSM95496, GSM281290, GSM95499, GSM95498, GSM95493, GSM95492, GSM95495, GSM45796, GSM95494, GSM119664, GSM162928, GSM119665, GSM337369, GSM159111, GSM119666, GSM119667, GSM119660, GSM176020, GSM179860, GSM119661, GSM162929, GSM119662, GSM119663, GSM272143, GSM301693, GSM272144, GSM272145, GSM152619, GSM80771, GSM272146, GSM199257, GSM80778, GSM80777, GSM272140, GSM80776, GSM272255, GSM272141, GSM272142, GSM272147, GSM179880, GSM272148, GSM272149, GSM159122, GSM327302, GSM301687, GSM80783, GSM272134, GSM80782, GSM272135, GSM80785, GSM80784, GSM152609, GSM80787, GSM80786, GSM301680, GSM80789, GSM199267, GSM80788, GSM350078, GSM272265, GSM162902, GSM272138, GSM272139, GSM179890, GSM80781, GSM272136, GSM80780, GSM272137, GSM162906, GSM162905, GSM162904, GSM159132, GSM399579, GSM80779, GSM327312, GSM301677, GSM80799, GSM80798, GSM80797, GSM80796, GSM80795, GSM199237, GSM80794, GSM80793, GSM80792, GSM80791, GSM80790, GSM119628, GSM119629, GSM272235, GSM249790, GSM119626, GSM119627, GSM119624, GSM119625, GSM119634, GSM119633, GSM119632, GSM119631, GSM119630, GSM159142, GSM152639, GSM238763, GSM301667, GSM272245, GSM199247, GSM152629, GSM119617, GSM119618, GSM119619, GSM119615, GSM119616, GSM119621, GSM119620, GSM119623, GSM119622, GSM159152, GSM301657, GSM152624

A.2 UMLS concepts in Concordia

A.2.1 Direct concept hits for the text associated with the 3030 GEO samples

These are the 1489 Unified Medical Language System [17] (UMLS) that we used to annotate the 3030 GEO samples in the database:

3,4-Methylenedioxyamphetamine; Abdominal mass; Abdominal Pain; Acromegaly; Adenocarcinoma; Papillary adenocarcinoma; adenoma; Adipose tissue; Adrenal Cortex; Adrenal Glands; Adult; Alcohol consumption; Ethanol; Aldosterone; Alzheimer's Disease; American Indians; Amniocentesis; Amniotic Fluid; Amygdaloid structure; Androgens; Refractory anaemia with excess blasts; Aorta; Appendix; Arsenic; Arthritis; Rheumatoid Arthritis; Asthma; Ataxia; Autistic Disorder; Autophagy; B-Lymphocytes; Benign prostatic hypertrophy; Bifidobacterium; Biopsy; Black race; Bladder; Blood; In Blood; Blood Cells; Bone Marrow; Bone Marrow Cells; Brain; Brain Neoplasms; Branchioma; Breast; Malignant neoplasm of breast; Breast Diseases; Bronchi; Bronchoscopy; Burkitt Lymphoma; Malignant Neoplasms; Carcinoid Tumor; Carcinoma; Malignant tumor of colon; Rectal Carcinoma; Malignant neoplasm of skin; Malignant neoplasm of thyroid; Basal cell carcinoma; Bronchioloalveolar Carcinoma; Adenocarcinoma, Mucinous; Carcinoma, Non-Small-Cell Lung; Carcinoma, Papillary; Renal Cell Carcinoma; Squamous cell carcinoma; Carcinoma, Transitional Cell; Cartilage; Caucasoid Race; Cecum; Cell Line; Cell Line, Transformed; Cells; Cultured Cells; Cerebellum; Cerebral cortex; Cervix Uteri; Oral Tobacco; Child; Chondrosarcoma; Chronic Disease; Cisplatin; Colon; Colonic Neoplasms; colonoscopy; Carcinoma of the Large Intestine; Constipation; Contraceptive Agents; Contraceptives, Oral; Corpus Callosum; Coughing; Cystadenocarcinoma; Cyst; Diarrhea; Dimethyl Sulfoxide; Disease; Duodenum; Embryo; Emetine; Endometriosis, site unspecified; Endometrium; Endothelium; Epithelial Cells; Epithelium; Herpesvirus 4, Human; Escherichia coli; Esophagogastric Junction; Esophagus; Limb structure; Mammalian Oviducts; Fatigue; Female; Fetus; Fibroblasts; Nonproliferative fibrocystic disease; fibrosarcoma; Ficoll; frontal lobe; Gallbladder; Ganglia; Ganglia, Spinal; gastric fundus; Gastritis; Gingiva; Glioblastoma; Glioma; Globus Pallidus; Glucocorticoids; Growth Factor; Head; Headache; Heart; Heart Atrium; Heart Ventricle; Hela Cells; Hematopoietic stem cells; Hemoptysis; Primary carcinoma of the liver cells; Hippocampus (Brain); Hispanic Americans; Hodgkin Disease; hypercholesterolemia; Hypercholesterolemia, Familial; Hypertensive disease; Hypothalamic structure; ileum; Bone structure of ilium; Infection; Human Papillomavirus; Inflammatory Bowel Diseases; Intestinal Mucosa; Large Intestine; Intestines, Small; Intestines; Irritable Bowel Syndrome; jejunum; Job's Syndrome; Joints; Kidney; Structure of cortex of kidney; Structure of medulla of kidney; Kidney Neoplasms; Knee; leiomyosarcoma; leukemia; Chronic Lymphocytic Leukemia; Acute Erythroblastic Leukemia; Lymphoblastic Leukemia; Acute lymphocytic leukemia; Leukemia, Lymphocytic, Acute, L1; Acute monocytic leukemia; Leukemia, Myelocytic, Acute; Myeloid Leukemia; Leukemia, Myelomonocytic, Acute; Acute Promyelocytic Leukemia;

Leukemia, T-Cell; Leukocytes; liposarcoma; Liver; Lung; Chronic Obstructive Airway Disease; Lymph; lymph nodes; Lymphocyte; Lymphoma; Lymphoma, Follicular; Reticulosarcoma; Lymphoma, Non-Hodgkin's; Macaca mulatta; macrophage; Male gender; Malignant neoplasm of stomach; Mammography; Mediastinum; medulloblastoma; melanoma; Melena; Tissue membrane; Mental Retardation; Midbrain structure; Cercopithecus aethiops; Monkeys; monocyte; Oral mucous membrane structure; Mucous Membrane; Multiple Myeloma; Multiple Sclerosis; Muscle; Muscular Atrophy; Spinal Muscular Atrophy; Muscular Dystrophies; Mutation; myometrium; Nasopharynx; Neck; African race; Neoplasm Metastasis; Neoplasms; Neuroblastoma; neutrophil; Nipples; Nodule; Nose; Nucleus Accumbens; Obesity; Occipital lobe; Omentum; osteosarcoma; Ovarian Carcinoma; Ovary; Pain; Pancreas; Papillomavirus; Parathyroid gland; Parathyroid Neoplasms; Parietal Lobe; Parkinson Disease; Parotid Gland; Pectoralis Muscles; Pelvis; penis; Pericardial sac structure; Periodontitis; Peritoneum; Pharyngeal structure; Phycoerythrin; Pituitary Adenoma; Pituitary Gland; Placenta; Plasma Cells; Plasmids; Pontine structure; prednisolone; Pregnant Women; Primates; Prostate; Psoriasis; Pulmonary artery structure; Structure of putamen; Pylorus; Radiation therapy; Androgen Receptor; Rectum; Rhabdomyosarcoma; Rheumatism; Rhinovirus; Riboflavin; Salivary Glands; Saphenous Vein; Metastatic to; Sezary Syndrome; Septic Shock; Sigmoidoscopy (procedure); skin disorder; Skin Neoplasms; Smoking; sperm cell; Sphingosine; Spinal Cord; Spleen; Starvation; Stem cells; Steroid 11-beta-Monooxygenase; Stomach; Streptococcus; Substantia nigra structure; Synovial Fluid; Synovial Membrane; T-Lymphocyte; Tamoxifen; Temporal Lobe; Testis; Thalamic structure; Thymus Gland; Thyroid Gland; thyroid neoplasm; Body tissue; Tobacco; Encounter due to tobacco use; Tongue; Palatine Tonsil; Trachea; Structure of trigeminal ganglion; Twin Multiple Birth; Monozygotic twins; Umbilical vein; Ureter; Urethra; Urinary tract; Uterine Fibroids; Uterus; Vagina; Veins; Vena caval structure; Vestibular nucleus structure; Visual Cortex; Vomiting; Vulva; Body Weight decreased; Caucasians; Woman; Wounds and Injuries; arsenic trioxide; matrigel; Rolipram; sphingosine 1-phosphate; Aldosterone Synthase; Asians; Branchial Clefts-Congenital disorder; B-Cell Lymphomas; Lymphoma, Diffuse; Diffuse Large B-Cell Lymphoma; Lymphoma, T-Cell, Cutaneous; Macrophages, Alveolar; alpha-beta T-Cell Receptor; Helicobacter; Acute leukemia; African American; Hispanics; Homo sapiens; Synovial biopsy; Bleeding of vagina; Structure of superior frontal gyrus; Structure of middle temporal gyrus; Structure of subthalamic nucleus; Malignant neoplasm of tongue; Malignant neoplasm of gallbladder; Uterine Cancer; Malignant neoplasm of ureter; Malignant neoplasm of brain; Umbilical Cord Blood; Stromal Cells; Prefrontal Cortex; Structure of entorhinal cortex; Ventral Tegmental Area; Injury; Blood specimen; Muscle biopsy; Fiberoptic bronchoscopy; Bronchial; Coronary artery; Cervical; Dorsal; Peripheral; Basal; chronic; Induced; Invasive; Malignant - descriptor; Nodular; Normal; Papillary; Lobular; Uninvolved; Undifferentiated; Adenocarcinoma, Oxyphilic; Adolescent; Undifferentiated carcinoma; Posterior root of spinal nerve; Adenosquamous carcinoma; Malignant Mixed Tumor; Endometrial Stromal Sarcoma; Adrenal Cortical Adenoma; Adenoma, Villous; Adenomatous Polyps; Adenocarcinoma, Clear Cell; Adrenocortical carcinoma; Carcinoma, Endometrioid; Carcinoma, Lobular; Mucoepidermoid Carcinoma; Carcinoma, Neuroendocrine; Cystadenocarcinoma, Papillary; Cystadenocarcinoma, Serous; Carcinoma, Large Cell; Rhabdoid Tumor; Lesion; Epithelial; Squamous epithelial cell; Skin structure of nipple; Subcutaneous Fat; Collecting duct; Lactiferous duct; Iliac crest structure; Structure of deltoid muscle; Entire biceps brachii; Structure of vastus lateralis muscle; Structure of synovial tissue of joint; Synovial fluid mononuclear cell; soft tissue; Endothelial Cells; Transitional epithelial cell; Oral cavity; Papilla of tongue; Fundus of abomasum; Colonic epithelium; Transverse colon; Renal pelvis; Body of uterus; Endometrio-; Foreskin of penis; Frontal lobe gyrus; Temporal lobe gyrus; Cerebellar hemisphere structure; Cerebellar vermis structure; Hematopoietic; lymphoblast; peripheral blood; Pelvic peritoneum; Childhood; Autistic thinking; Memory impairment; Gallbladder Carcinoma; Endometrium normal; Uterus normal; Nonhuman Primates; Gastrointestinal Stromal Tumors; Muscular Dystrophy, Facioscapulohumeral; Carcinoma of Nasopharynx; Papillary thyroid carcinoma; Epidermal Growth Factor; Malignant neoplasm of lung; mucosa-associated lymphoid tissue lymphoma; Skeletal muscle structure; Systemic Inflammatory Response Syndrome; Systemic infection; Cyclin-Dependent Kinases; control; Interferon beta-1a; Multiple tumors; Skeletal bone; Dermatomyositis, Childhood Type; Childhood asthma; Acute gastric mucosal erosion; Retroperitoneal mass; Rhinovirus infection; Red stools; Subcutaneous Tissue; Cancer of Head and Neck; Malignant Bone Neoplasm; Refractory anemia with excess blasts in transformation (clinical); Mixed Oligodendroglioma-Astrocytoma; Caco-2 Cells; atorvastatin; Cervix carcinoma; Organic arsenic; Human rhinovirus; Acinar; Normal tissue morphology; Septic; Depletion; Myeloma cell; Metaplastic polyp; Metaplastic; Secretory endometrium; [M]Squamous cell carcinoma, metastatic NOS; Squamous cell carcinoma, keratinizing; Papillary transitional cell carcinoma; [M]Adenocarcinoma, metastatic, NOS; Papillary serous cystadenocarcinoma; Serous surface papillary carcinoma; Mucin-producing adenocarcinoma; Blast (physical force); Smoker; Non-smoker; Endometriosis of uterus; Malignant neoplasm of liver; Cancer of Intestines; Malignant neoplasm of pancreas; Pelvic mass; Blast Cell; whole blood; Marrow; Malignant neoplasm of prostate; Escherichia coli O157; Fibroblast Growth Factor 2; CBFbeta-MYH11 fusion protein; Bone Tissue; Chemotherapy Regimen; Lupus Erythematosus; Indian ethnic group; Developmental delay (disorder); Human cells; Bone marrow specimen; Primary operation; Myeloid; Follicular; Cirrhotic; Tobacco smoke; Fetal brain; Serous; Human tissue; Oral; Salivary; Paravertebral; subcutaneous; Whole blood sample; Synovial fluid cells; Whole; Non-small cell; Tobacco smoking behavior; Chewed tobacco consumption; Entire temporal lobe gyrus; Airway structure; Cigarette consumption; Tumor tissue sample; Pulmonary lymphoma; Spinal; Oropharyngeal; Rectum and sigmoid colon; Adrenal; Fetal; Probiotics; Small Intestine - Duodenum; Balanced salt solution; Tobacco use; Malignant neoplasm of esophagus; Thyroid carcinoma; Ewings sarcoma; Pregnant - adjective; Chronic Childhood Arthritis; Exocrine pancreas; Malignant Glioma; Colorectal; Tongue Carcinoma; Persistent cough; Placenta healthy; Vulva normal; Vagina normal; Normal ovary; Breast normal; Nipple normal; Fallopian tube normal; Joint normal; Skeletal muscle normal; Biopsy of jejunum; Tongue normal; Stomach normal; Liver normal; Penis normal; Gastric biopsy sample; Colonic biopsy sample; Basal Cell; Embryonic Stem Cells; Steroid biosynthesis; Nasal Epithelium; Ureter Carcinoma; Prostate carcinoma; epoxomicin; Cigarette; Epithelial ovarian cancer; Breast Carcinoma; Skin tissue; Carcinoma of lung; Regression - mental defense mechanism; Duct (organ) structure; Umbilical Blood; Colon Carcinoma; Stomach Carcinoma; Skin carcinoma; Bone carcinoma; Small; Fundus; Malignant neoplasm of kidney; Cancer of Neck; Sarcoma, metastatic; Brain Tumor, Primary; Cancer of Head; Tonsil; Uterine carcinoma; Radiation; Pluripotent Stem Cells; R-1881; Stromal Neoplasm; ovarian neoplasm; Microsatellite Instability;

Mammary gland; Inorganic arsenic; Cholecystolithiasis; Hurthle Cells; adalimumab; Skin; Invasive Ductal Breast Carcinoma; Malignant neoplasm of ovary; ezetimibe; kinase inhibitor; Ezetrol; Ductal Carcinoma; Lymphatic Endothelial Cells; Mesenchymal Stem Cells; HCT116 Cells; sarcoma; Bifidobacterium lactis; Chromophobe Renal Cell Carcinoma; Collecting Duct Carcinoma (Kidney); Metaplastic carcinoma; Superior mediastinal lymph node; Entire pulmonary artery; Entire substantia nigra; Entire thalamus; Skin fibroblast; Cutaneous lymphoma; Entire fallopian tube; Entire limb; Entire skeletal muscle (organ); Branchial Clefts; Entire oropharynx; Entire hypothalamus; Entire pons; Entire superior frontal gyrus; Entire middle temporal gyrus; Entire subthalamic nucleus; Entire putamen; Entire rib; Entire entorhinal cortex; Inflammatory disorder; Abdominal bloating; Amniotic fluid specimen; Precursor B-cell lymphoblastic leukemia; Entire synovial tissue of joint; Medulla; Primary malignant neoplasm; Papillary Renal Cell Carcinoma; Peripheral blood mononuclear cell; Entire vastus lateralis muscle; Acute myelomonocytic leukemia with abnormal eosinophils; Classical Hodgkin's Lymphoma; Stromal sarcoma; Adrenal carcinoma; Renal carcinoma; Metastatic Carcinoma; Gastric erosion; Systemic onset juvenile chronic arthritis; torcetrapib; Tumor Necrosis Factor-alpha; Mammary Neoplasms; Ductal; Pediatric; Pectoral; Coronary; metastatic qualifier; Mediastinal; urinary; Colorectal Cancer; Ductal Breast Carcinoma; Adenocarcinoma, Endometrioid; Glioblastoma Multiforme; Cirrhosis; Gastric; Ventral; Fetal Stem Cells; Acute myeloid leukemia without maturation; Acute Myeloid Leukemia (AML-M2); Embryonic Cell; Precursor Cell Lymphoblastic Leukemia Lymphoma

Appendix B

Transcriptomic landscape: Differentially expressed genes in brain, blood, and soft tissue

B.1 Over-expressed genes in soft tissue

Table B.1:

GO ID	GO Term	P Value
GO:0005584	collagen type I	0.017
GO:0005583	fibrillar collagen	0
GO:0032964	collagen biosynthetic process	0
GO:0001527	microfibril	0
GO:0043205	fibril	0.005
GO:0030057	desmosome	0
GO:0048407	platelet-derived growth factor binding	0
GO:0030199	collagen fibril organization	0
GO:0005520	insulin-like growth factor binding	0
GO:0005581	collagen	0
GO:0032963	collagen metabolic process	0
GO:0044259	multicellular organismal macromolecule metabolic process	0
GO:0044236	multicellular organismal metabolic process	0.001
GO:0044420	extracellular matrix part	0
GO:0005201	extracellular matrix structural constituent	0
GO:0030198	extracellular matrix organization	0
GO:0005604	basement membrane	0
GO:0043588	skin development	0.001
GO:0005200	structural constituent of cytoskeleton	0.001
GO:0010035	response to inorganic substance	0.033
GO:0001649	osteoblast differentiation	0.039
GO:0009612	response to mechanical stimulus	0
GO:0043062	extracellular structure organization	0

Continued on Next Page...

Table B.1 – Continued

GO ID	GO Term	P Value
GO:0006956	complement activation	0.001
GO:0070161	anchoring junction	0.018
GO:0002541	activation of plasma proteins involved in acute inflammatory response	0.002
GO:0009987	cellular process	0.013
GO:0005911	cell-cell junction	0.036
GO:0016043	cellular component organization	0.048
GO:0031960	response to corticosteroid stimulus	0
GO:0031012	extracellular matrix	0
GO:0005578	proteinaceous extracellular matrix	0
GO:0016337	cell-cell adhesion	0.008
GO:0019838	growth factor binding	0
GO:0030154	cell differentiation	0
GO:0008201	heparin binding	0
GO:0051384	response to glucocorticoid stimulus	0
GO:0001525	angiogenesis	0.017
GO:0008544	epidermis development	0
GO:0005539	glycosaminoglycan binding	0
GO:0005198	structural molecule activity	0
GO:0006959	humoral immune response	0.041
GO:0001871	pattern binding	0
GO:0030247	polysaccharide binding	0
GO:0030855	epithelial cell differentiation	0.004
GO:0048869	cellular developmental process	0.017
GO:0044421	extracellular region part	0
GO:0009628	response to abiotic stimulus	0.049
GO:0005576	extracellular region	0
GO:0005615	extracellular space	0
GO:0048545	response to steroid hormone stimulus	0
GO:0050896	response to stimulus	0.05
GO:0007584	response to nutrient	0.028
GO:0009888	tissue development	0
GO:0007155	cell adhesion	0
GO:0022610	biological adhesion	0
GO:0009725	response to hormone stimulus	0
GO:0009719	response to endogenous stimulus	0.008
GO:0010033	response to organic substance	0
GO:0009605	response to external stimulus	0.02
GO:0048856	anatomical structure development	0
GO:0042221	response to chemical stimulus	0
GO:0032502	developmental process	0
GO:0006950	response to stress	0.023

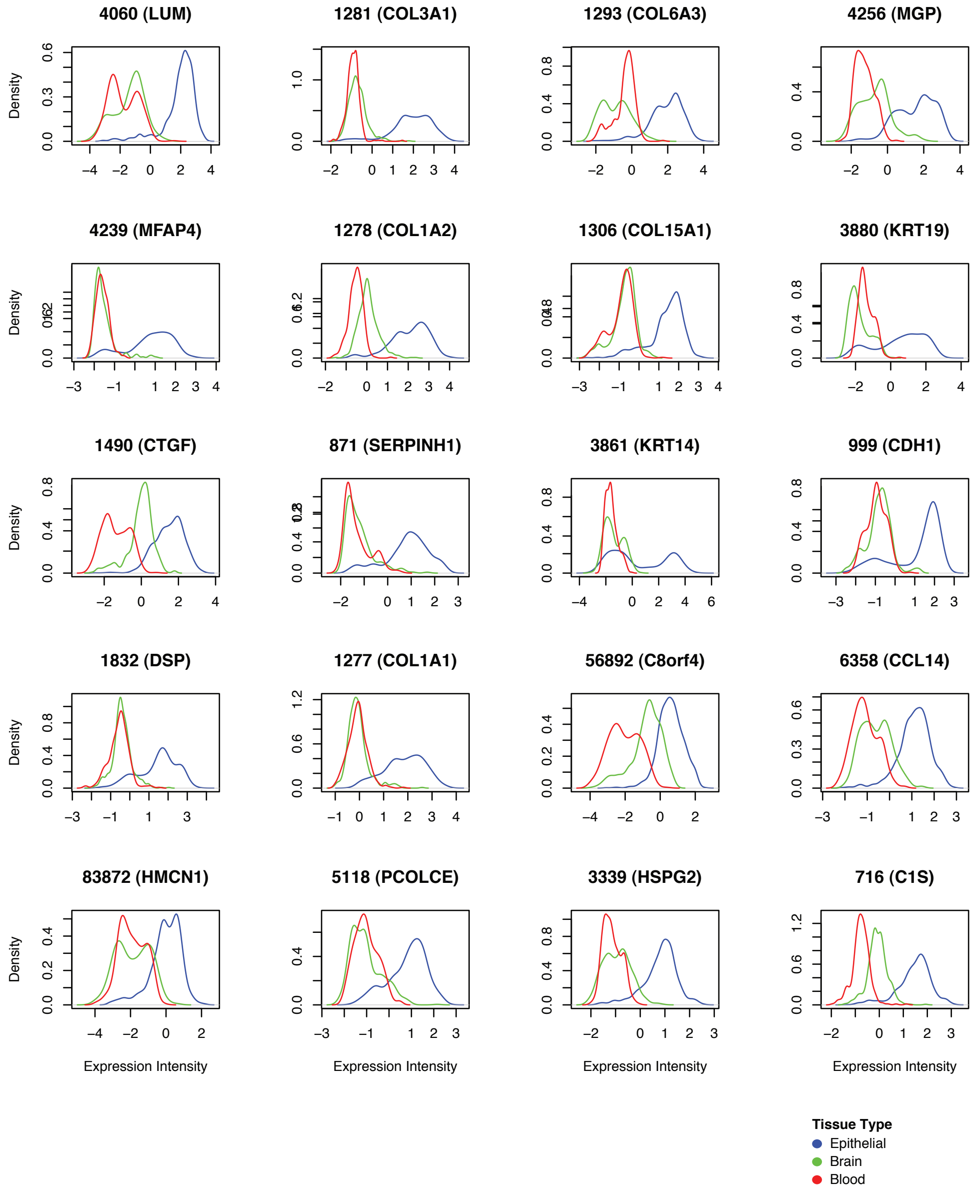


Figure B-1: Expression intensity distribution of the top 20 over-expressed soft tissue genes. Each plot corresponds to the kernel density estimate of expression values for the gene named above each plot for the three broad tissue types, blood, brain, and soft tissue. We see that the expression values of soft tissue specific genes such as COL3A1, COL6A3, KRT19, KRT14, and CADH1 are markedly higher in samples corresponding to soft tissues than in samples of the other two types.

B.2 Over-expressed genes in blood

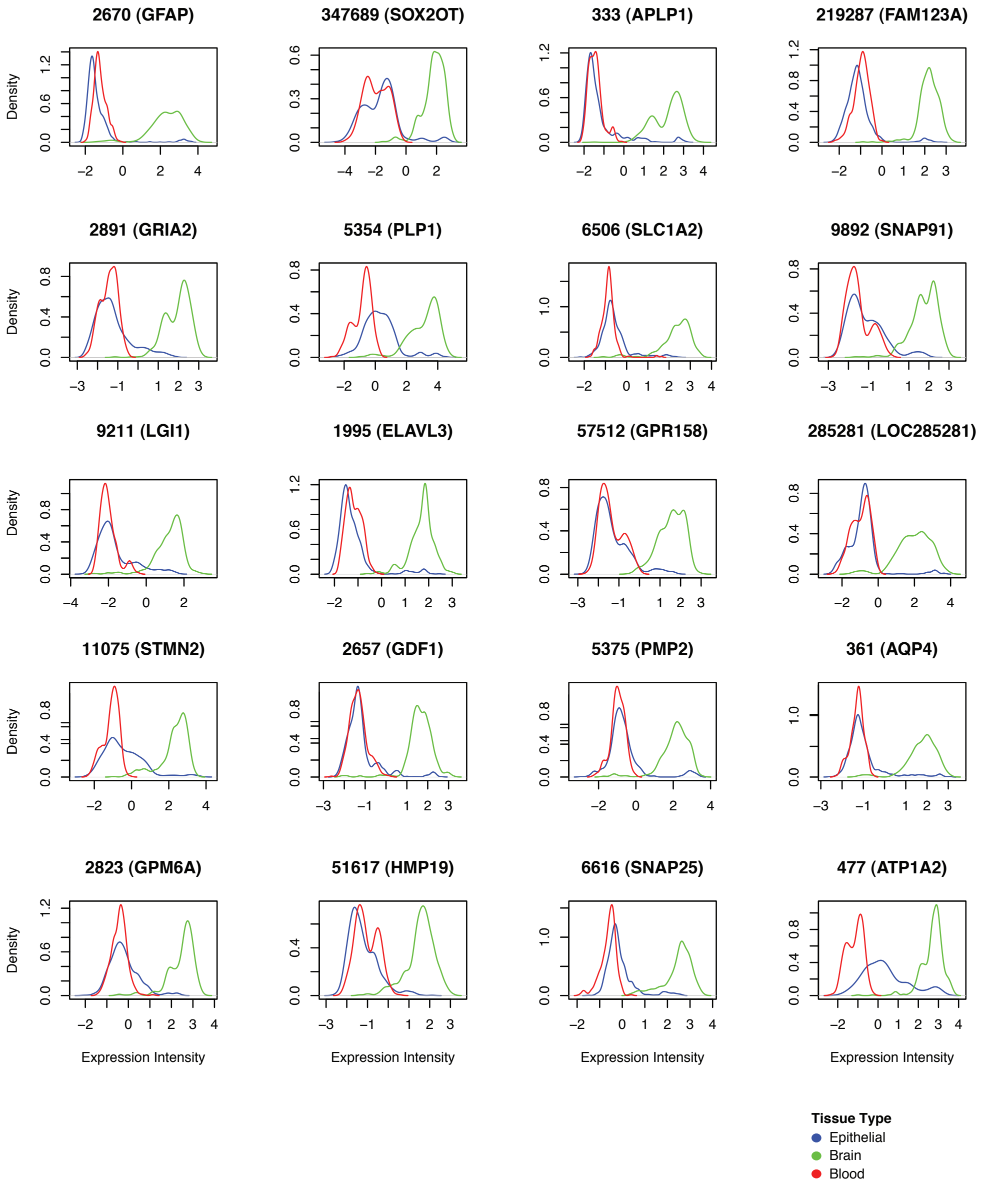


Figure B-2: Expression intensity distribution of the top 20 over-expressed brain tissue genes. Each plot corresponds to the kernel density estimate of expression values for the gene named above each plot for the three broad tissue types, blood, brain, and soft tissue. We see that the expression values of brain specific genes such as GFAP, APLP1, GRIA2, PLP1, and SLC1A2 are markedly higher in samples corresponding to brain tissue than in samples of the other two types.

Table B.2:

GO ID	GO Term	P Value
GO:0042105	alpha-beta T cell receptor complex	0
GO:0045730	respiratory burst	0.008
GO:0050857	positive regulation of antigen receptor-mediated signaling pathway	0.041
GO:0005833	hemoglobin complex	0
GO:0005344	oxygen transporter activity	0.001
GO:0042101	T cell receptor complex	0.002
GO:0050854	regulation of antigen receptor-mediated signaling pathway	0.005
GO:0031640	killing of cells of another organism	0.004
GO:0045058	T cell selection	0.035
GO:0003823	antigen binding	0
GO:0001906	cell killing	0.036
GO:0050830	defense response to Gram-positive bacterium	0
GO:0009620	response to fungus	0.009
GO:0006968	cellular defense response	0
GO:0001608	nucleotide receptor activity, G-protein coupled	0.045
GO:0045028	purinergic nucleotide receptor activity, G-protein coupled	0.045
GO:0004715	non-membrane spanning protein tyrosine kinase activity	0.036
GO:0042742	defense response to bacterium	0
GO:0031225	anchored to membrane	0.014
GO:0006935	chemotaxis	0
GO:0042330	taxis	0
GO:0050870	positive regulation of T cell activation	0.015
GO:0009617	response to bacterium	0
GO:0042110	T cell activation	0
GO:0006955	immune response	0
GO:0002376	immune system process	0
GO:0050863	regulation of T cell activation	0.004
GO:0040011	locomotion	0
GO:0046649	lymphocyte activation	0
GO:0007626	locomotory behavior	0
GO:0006952	defense response	0
GO:0050867	positive regulation of cell activation	0.014
GO:0045321	leukocyte activation	0
GO:0051707	response to other organism	0
GO:0009897	external side of plasma membrane	0.044
GO:0002684	positive regulation of immune system process	0
GO:0001775	cell activation	0
GO:0051249	regulation of lymphocyte activation	0.01
GO:0050865	regulation of cell activation	0.002
GO:0002694	regulation of leukocyte activation	0.008
GO:0006954	inflammatory response	0
GO:0002682	regulation of immune system process	0
GO:0007610	behavior	0.002
GO:0009607	response to biotic stimulus	0
GO:0030246	carbohydrate binding	0.038
GO:0009611	response to wounding	0
GO:0009605	response to external stimulus	0.001
GO:0005887	integral to plasma membrane	0
GO:0031226	intrinsic to plasma membrane	0
GO:0051704	multi-organism process	0.003
GO:0004872	receptor activity	0
GO:0004871	signal transducer activity	0
GO:0060089	molecular transducer activity	0
GO:0006950	response to stress	0
GO:0050896	response to stimulus	0
GO:0005886	plasma membrane	0
GO:0044459	plasma membrane part	0

Continued on Next Page. . .

Table B.2 – Continued

GO ID	GO Term	P Value
GO:0007166	cell surface receptor linked signaling pathway	0
GO:0004888	transmembrane receptor activity	0.012
GO:0023033	signaling pathway	0
GO:0023052	signaling	0.003
GO:0016020	membrane	0
GO:0044425	membrane part	0
GO:0031224	intrinsic to membrane	0.002
GO:0016021	integral to membrane	0.012

B.3 Over-expressed genes in brain

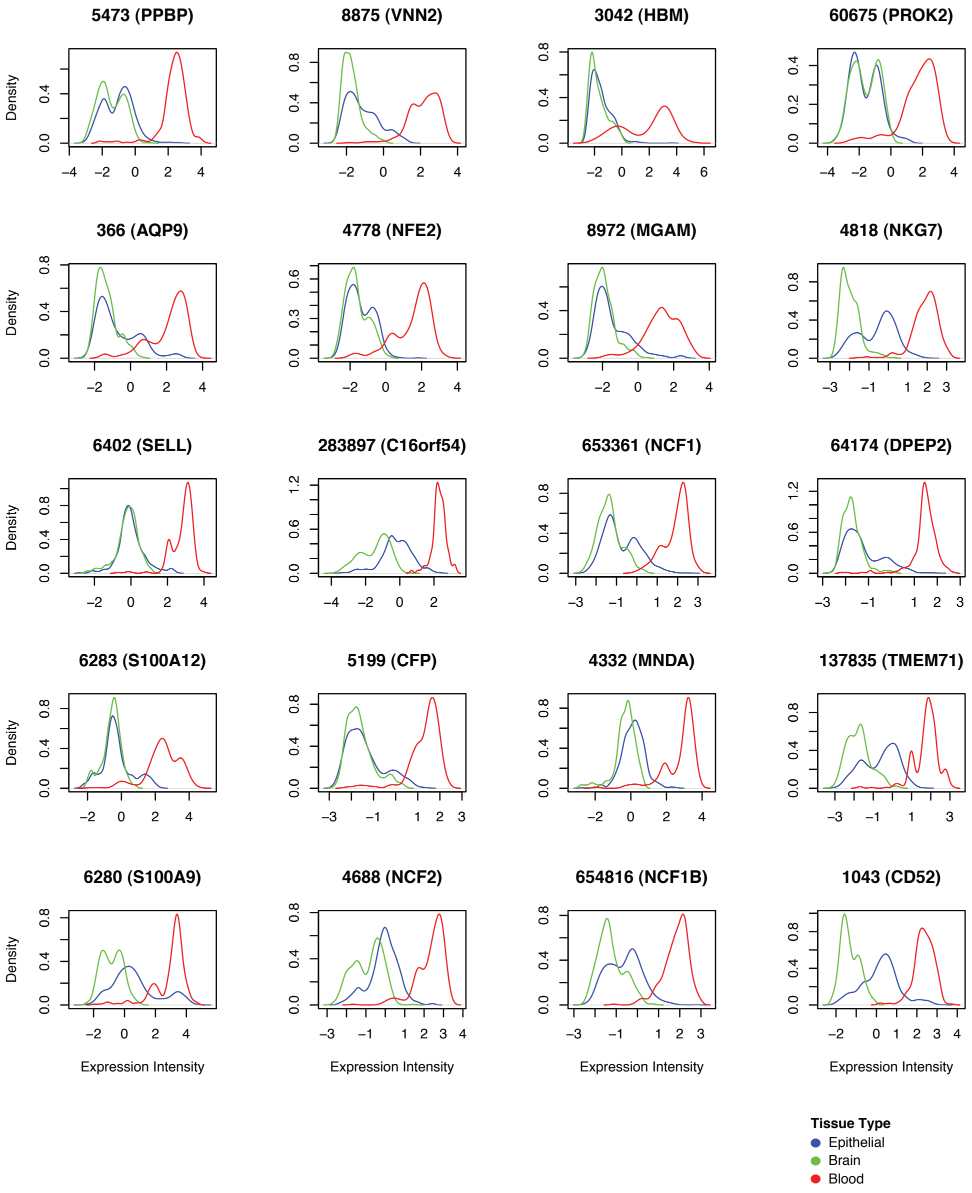


Figure B-3: Expression intensity distribution of the top 20 over-expressed blood genes. Each plot corresponds to the kernel density estimate of expression values for the gene named above each plot for the three broad tissue types, blood, brain, and soft tissue. We see that that the expression value of brain specific genes such as HBM, PPBP, VNN2, SELL, and NFE2 are markedly higher in samples corresponding to blood than in samples of the other two types.

Table B.3:

GO ID	GO Term	P Value
GO:0045110	intermediate filament bundle assembly	0.044
GO:0005883	neurofilament	0.001
GO:0060052	neurofilament cytoskeleton organization	0.013
GO:0007269	neurotransmitter secretion	0.02
GO:0001505	regulation of neurotransmitter levels	0
GO:0006836	neurotransmitter transport	0
GO:0008021	synaptic vesicle	0.013
GO:0043197	dendritic spine	0.032
GO:0044309	neuron spine	0.032
GO:0033267	axon part	0
GO:0030424	axon	0
GO:0007409	axonogenesis	0
GO:0043005	neuron projection	0
GO:0008509	anion transmembrane transporter activity	0.035
GO:0048812	neuron projection morphogenesis	0
GO:0007417	central nervous system development	0
GO:0048858	cell projection morphogenesis	0
GO:0044456	synapse part	0
GO:0045202	synapse	0
GO:0044463	cell projection part	0
GO:0032990	cell part morphogenesis	0.003
GO:0007268	synaptic transmission	0
GO:0022891	substrate-specific transmembrane transporter activity	0.018
GO:0022857	transmembrane transporter activity	0.04
GO:0005215	transporter activity	0.007
GO:0045211	postsynaptic membrane	0.019
GO:0042995	cell projection	0
GO:0030054	cell junction	0
GO:0007399	nervous system development	0
GO:0048731	system development	0
GO:0022838	substrate-specific channel activity	0.036
GO:0051234	establishment of localization	0.02
GO:0007267	cell-cell signaling	0.021
GO:0006810	transport	0.04
GO:0015075	ion transmembrane transporter activity	0.013
GO:0007154	cell communication	0.02
GO:0006811	ion transport	0.017
GO:0044459	plasma membrane part	0.003
GO:0048856	anatomical structure development	0.033

Appendix C

Concordia performance

C.1 Cross-validation performance of Concordia

The leave-one-out cross validation performance of the 1489 disease and anatomy concepts as computed by the method outlined in Chapter 4

Table C.1: Cross-validation performance of Concordia

Concept	AUC	Num Samples	Num Series
Anatomic structures	0.860795353	2954	154
Body Regions	0.860795353	2954	154
Physical anatomical entity	0.860795353	2954	154
body system	0.852956551	2603	131
Body tissue	0.837848153	2474	112
Body organ structure	0.744149574	2433	118
Body part	0.906311914	1914	83
Body substance	0.835581871	1916	85
Entire subdivision of organ system	0.742509803	1595	100
Musculoskeletal System	0.742132317	1594	100
Skeletal system	0.742132317	1594	100
SKELETAL SYSTEM: GENERAL TERMS	0.742132317	1594	100
Skeletal System (Bones of Head, Rib Cage and Vertebral Column)	0.742132317	1594	100
SOFT TISSUES, SMOOTH MUSCLE AND CARTILAGINOUS TISSUES	0.736962657	1571	98
Soft Tissue, Bone and Cartilage	0.736962657	1571	98
soft tissue	0.685021181	1513	98
Disorder by body site	0.741622966	1194	100
Neck, chest, abdomen, and pelvis	0.897100766	1322	73
Disorder of body system	0.741326811	1141	97
Body space structure	0.848075943	1551	62
Body material	0.694239734	1665	70
Body cavities	0.850792747	1537	60
Neck, chest and abdomen	0.897120386	1269	66
Trunk structure	0.859012735	1234	68
Structure of subregion of trunk	0.858528777	1232	66
Chest, abdomen, and pelvis	0.868753539	1193	66
Chest and abdomen	0.871699619	1140	59

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Cells	0.913694148	710	90
Neoplasms	0.79627514	910	78
Neoplasm and/or hamartoma	0.79627514	910	78
sex	0.756465161	1457	51
Organ part	0.858393737	1199	53
Unspecified Neoplasms and Tumor Cells	0.79737436	902	75
Malignant Neoplasms	0.815926867	855	74
Malignant tumor of unknown origin or ill-defined site	0.815926867	855	74
Malignant neoplasm of other and unspecified site otherwise specified	0.815926867	855	74
Malignant neoplasm of other and unspecified sites	0.815926867	855	74
Malignant Neoplasm (Morphology)	0.812252124	835	67
Primary malignant neoplasm	0.811627883	833	66
Upper body structure	0.896343213	1117	44
Upper body part structure	0.896343213	1117	44
Connective Tissue	0.691873179	917	68
Body tissue material	0.691873179	917	68
Skeletal material	0.691873179	917	68
Neoplasm by body site	0.770455134	755	68
Hemic and Immune Systems	0.968525364	681	56
Neoplasms by Site	0.764899866	717	62
Neoplasms by Histologic Type	0.796900372	773	55
Cellular Structures	0.879070575	575	67
Lower body structure	0.799826114	827	51
Lower body part structure	0.799826114	827	51
Abdomen and pelvis	0.80237391	820	50
Entire cell	0.886090444	566	65
Structure of viscus	0.802896481	834	44
Musculoskeletal Diseases	0.647263327	671	62
Connective Tissue Diseases	0.647263327	671	62
Musculoskeletal and connective tissue disorders	0.647263327	671	62
Abdominal Cavity	0.812552248	767	43
ABDOMEN INCLUDING PERITONEUM AND RETROPERITONEUM	0.812552248	767	43
Abdomen	0.812552248	767	43
Disorder of body cavity	0.725818039	661	49
Fluids and Secretions	0.98873893	526	45
Body Fluids	0.989560932	524	44
Male gender	0.768290336	694	40
Blood	0.99298379	508	41
Female	0.700241563	764	36
Disorder of trunk	0.826006589	546	42
Bone and/or joint structure	0.853841241	487	45
Hematological system	0.907304659	467	44
Hematopoietic System	0.907304659	467	44
Skeletal bone	0.863690184	475	45
Integumentary system	0.827015849	530	42
SKIN AND SKIN APPENDAGES	0.827015849	530	42
INTEGUMENTARY SYSTEM: GENERAL TERMS	0.827015849	530	42
Immune system	0.906197905	437	42
Structure of lymphoreticular system	0.906197905	437	42
Neoplasms, Glandular and Epithelial	0.906278199	503	32
Bone Marrow	0.912430297	407	37
Bone Marrow and Erythropoietic Tissues	0.912430297	407	37
Neck and chest	0.860055289	505	31
Gastrointestinal system	0.918517706	466	31
Gastrointestinal tract structure	0.918517706	466	31
DIGESTIVE SYSTEM: GENERAL TERMS	0.918517706	466	31
Head and neck structure	0.978785357	680	19
Neoplasm of trunk	0.86726834	440	33

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Digestive organ structure	0.928308168	442	30
DIGESTIVE ORGANS: GENERAL TERMS	0.928308168	442	30
DISORDERS OF THE MUSCLES, LIGAMENTS, FASCIAE AND OTHER SOFT TISSUES	0.724249863	429	39
Skeleton	0.782528241	495	31
Epithelioma	0.925849089	486	26
Entire body organ	0.843848235	554	25
Entire anatomical structure	0.843848235	554	25
Bone and Bones	0.789295752	480	30
Carcinoma	0.937877567	459	25
Malignant epithelial neoplasm - category	0.937877567	459	25
Pelvis and lower extremities	0.869981436	455	27
Pelvis	0.875675176	448	26
Lower trunk structure	0.875675176	448	26
Structure of abdominal viscus	0.889200394	395	29
Head	0.984123118	634	16
Structure of breast and/or endocrine system	0.882905169	431	25
Head part	0.984555368	621	15
Other diseases of blood or blood-forming organs	0.939489687	274	35
Disorder of cellular component of blood	0.939489687	274	35
Disorder of hematopoietic structure	0.939489687	274	35
Hematological Disease	0.939489687	274	35
Genitourinary system	0.876727537	427	23
Urinary tract	0.876727537	427	23
Urinary system	0.876727537	427	23
URINARY TRACT: GENERAL TERMS	0.876727537	427	23
Structure of thorax, including mediastinum and diaphragm	0.884483878	376	23
Upper trunk structure	0.884483878	376	23
Chest	0.884483878	376	23
Digestive System Disorders	0.814596577	333	27
Intra-abdominal digestive structure	0.950259537	319	24
Blood Cells	0.922545955	287	27
Structure of product of conception	0.973690213	486	15
Disorder of abdomen	0.807550896	322	26
Bone marrow part	0.925127094	257	25
Structure of myelopoietic tissue	0.925127094	257	25
Leukocytes	0.925127094	257	25
Urogenital organ	0.884348608	365	18
Structure of anatomical reproductive system	0.884348608	365	18
Genitalia	0.884348608	365	18
Genital system	0.884348608	365	18
Immune System Diseases	0.917118014	207	30
Disorder of pelvis	0.826190109	291	23
Reticuloendothelial System	0.908491689	224	27
Abdominal mass	0.80899537	274	23
Abdominal Neoplasms	0.809412205	273	23
Developmental body structure	0.9814019	435	11
Embryonic Structures	0.9814019	435	11
Gland	0.864800244	301	18
Complex structure derived from epithelium	0.864800244	301	18
Cardiovascular Diseases	0.924081132	222	22
Cultured Cells	0.987418163	152	30
Adenoma AND/OR adenocarcinoma	0.900045077	309	16
ADENOMAS AND ADENOCARCINOMAS	0.898977891	305	16
Cell Line	0.988847222	150	29
Nervous system structure	0.997338868	530	8
Other part of nervous system	0.997338868	530	8
Mononuclear cell (histiocyte, lymphocyte, plasma cell)	0.923247787	207	22

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Reticuloendothelial cell	0.923247787	207	22
Central nervous system part	0.997433424	521	8
Neuraxis	0.997433424	521	8
Brain and spinal cord structure	0.997433424	521	8
Structure of body cavity subdivision	0.879796195	335	14
Pelvic cavity structure	0.879796195	335	14
CLINICAL CLASSIFICATION OF NEOPLASMS OF THE MUSCULOSKELETAL SYSTEM AND SOFT TISSUES	0.807428644	232	22
Disorder of soft tissue	0.768224999	233	23
[X]Other soft tissue disorders	0.768224999	233	23
Cranial cavity structure	0.996742397	497	8
Peripheral blood mononuclear cell	0.923089866	204	21
Malignant neoplasm of abdomen	0.852137791	210	22
Brain	0.997067302	488	8
Intracranial structure	0.997067302	488	8
Pelvic genital structure	0.878519688	315	14
Neoplasm, uncertain whether benign or malignant	0.95582066	193	21
[X]Malignant neoplasms of lymphoid, hematopoietic and related tissue	0.967070052	190	21
Hematopoietic Neoplasms	0.967070052	190	21
Neoplasm of hematopoietic cell type	0.967070052	190	21
Disorder of the genitourinary system	0.85773095	241	18
Disorder of hematopoietic morphology	0.966877674	189	20
Malignant adenomatous neoplasm - category	0.921305759	282	14
Adenocarcinoma	0.91945118	277	14
peripheral blood	0.948089669	180	19
Structure of respiratory system and/or intrathoracic structure	0.785398275	227	18
Intestines	0.954716029	223	15
Lower Gastrointestinal Tract	0.954716029	223	15
Stem cells	0.930699608	179	19
Endocrine system	0.865576316	238	15
Endocrine Glands	0.870520061	236	15
Structure of endocrine system	0.870520061	236	15
Thoracic Diseases	0.883076805	203	17
Respiration Disorders	0.883076805	203	17
DISEASES OF THE SINUSES, NOSE, PHARYNX AND LARYNX	0.883076805	203	17
skin disorder	0.815144635	185	19
Skin and subcutaneous tissue disorders	0.815144635	185	19
Disorder of integument	0.815144635	185	19
RESPIRATORY SYSTEM: GENERAL TERMS	0.90312674	189	16
Respiratory System	0.90312674	189	16
Other female genital tract	0.869372565	277	11
Female genitalia	0.869372565	277	11
Female genitourinary system	0.869372565	277	11
Disorder of digestive organ	0.925218679	165	17
Pelvic organ	0.872254428	270	11
Female internal genitalia structure	0.872649335	268	11
Pelvic cavity female genital structure	0.872649335	268	11
Human material	0.935804891	899	3
Human surgical material	0.935804891	899	3
Human tissue	0.935804891	899	3
Upper female genital structure	0.875670647	260	11
Neuromuscular Diseases	0.853808859	191	15
nervous system disorder	0.853808859	191	15
Neuropathy	0.853808859	191	15
Nervous system and sense organ diseases	0.853808859	191	15
Myopathy	0.853808859	191	15
[X]Other disorders of the nervous system	0.853808859	191	15
Neuromuscular Junction Diseases	0.853808859	191	15

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
DISORDERS OF PERIPHERAL NERVOUS SYSTEM: GENERAL TERMS	0.853808859	191	15
Disorder of skeletal muscle	0.853808859	191	15
Nerve, plexus and root disorders	0.853808859	191	15
Peripheral Neuropathy	0.853808859	191	15
Breast	0.928091168	195	13
Disorder of digestive tract	0.921074977	157	16
Disorder of immune structure	0.96008056	133	18
Non-infectious disorder of lymphatics	0.96008056	133	18
Lymphatic Diseases	0.96008056	133	18
Lymphatic Vessel Diseases	0.96008056	133	18
Respiratory tract structure	0.914390521	179	14
Telencephalon	0.985696581	422	5
Prosencephalon	0.985696581	422	5
Brain tissue	0.985696581	422	5
Supratentorial brain part	0.985696581	422	5
Embryonic nervous system structure	0.985696581	422	5
Brain part	0.985696581	422	5
Nervous structure of head	0.985696581	422	5
Regional nervous structure	0.985696581	422	5
Nervous structure of head and neck	0.985696581	422	5
Cerebrum	0.985696581	422	5
Neoplasm of intra-abdominal organs	0.820813464	155	16
Skin AND subcutaneous tissue structure	0.936626656	194	11
Integumentary system part	0.936626656	194	11
Soft tissue lesion	0.791637671	156	16
Hematologic Neoplasms	0.992818515	151	13
Leukemia (category)	0.992818515	151	13
leukemia	0.992818515	151	13
CLINICAL CLASSIFICATION OF NEOPLASMS OF THE HEMATOPOIETIC AND IMMUNE SYSTEMS	0.992818515	151	13
Thoracic Neoplasms	0.948782839	167	12
Mediastinal Diseases	0.948782839	167	12
[D]Chest mass	0.948782839	167	12
DISEASES OF THE PLEURA, MEDIASTINUM AND DIAPHRAGM	0.948782839	167	12
Thoracic cavity structure	0.806660862	181	13
Immunoproliferative neoplasm	0.960789907	123	16
Lymphoreticular tumor	0.960789907	123	16
[M]Miscellaneous myeloproliferative and lymphoproliferative disorders	0.960789907	123	16
Hematopoietic neoplasm of uncertain behavior	0.960789907	123	16
Immunoproliferative Disorders	0.960789907	123	16
Malignant immunoproliferative neoplasm	0.960789907	123	16
Immunoproliferative morphology	0.960789907	123	16
Lymphoid neoplasm	0.960789907	123	16
Lymphoproliferative Disorders	0.960789907	123	16
Cerebral hemisphere structure (body structure)	0.984203701	379	5
Structure of skin and/or surface epithelium	0.929849639	196	10
Primary malignant neoplasm of bone marrow	0.993372685	150	12
Upper digestive tract structure	0.914264264	169	11
Structure of lung and/or mediastinum	0.831865487	170	12
Structure of thoracic viscus	0.833626675	169	12
Large Intestine	0.955636743	156	11
Traumatic abnormality	0.865541486	133	14
GENERAL AND COMPRESSION INJURIES	0.865541486	133	14
GENERAL INJURIES	0.865541486	133	14
Injury	0.865541486	133	14
Integumentary system subdivision	0.938580862	188	9
Entire skin	0.938580862	188	9
Skin	0.938580862	188	9

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
CLINICAL CLASSIFICATION OF NEOPLASMS OF THE DIGESTIVE SYSTEM	0.849413338	124	15
Endocrine System Diseases	0.910750297	133	13
GENERAL AND POLYGLANDULAR ENDOCRINE DISORDERS	0.910750297	133	13
GENERAL AND GROWTH RELATED DISORDERS	0.910750297	133	13
Acute leukemia	0.992114711	144	11
Acute leukemia (category)	0.992114711	144	11
Female Reproductive System Disorder	0.830014077	172	11
reproductive system disorder	0.830014077	172	11
Female Genital Diseases	0.830014077	172	11
Primary malignant neoplasm of trunk	0.853404172	126	14
Disorder of skeletal system	0.891154898	104	16
Disorder of immune function	0.893311447	97	16
Lower respiratory tract structure	0.94260856	142	10
Lower respiratory system structure	0.94260856	142	10
Large intestine part	0.954574889	140	10
CNS disorder	0.950732979	126	11
Neck	0.926122775	129	11
Scalp and/or neck structure	0.926122775	129	11
Face and/or neck structure	0.926122775	129	11
Complication	0.863676241	117	13
Poisoning / injury	0.863676241	117	13
POISONINGS: GENERAL TYPES	0.863676241	117	13
Poisoning	0.863676241	117	13
Sequela of disorder	0.863676241	117	13
Bone Marrow Cells	0.919730967	129	11
Colon	0.957816236	134	10
Disorder of head	0.940922395	136	10
Digestive System Neoplasms	0.91056671	108	13
Neoplasm of digestive organ	0.91056671	108	13
[X]Malignant neoplasm of digestive organs	0.91056671	108	13
Soft Tissue Neoplasms	0.862589352	122	12
Pulmonary structure including vessels and lymphoid tissue	0.947662442	132	10
Noninflammatory disorder of the female genital organs	0.856026394	142	10
Pelvic mass	0.856026394	142	10
Genitourinary Neoplasms	0.854429494	141	10
Pelvic Neoplasms	0.854429494	141	10
Gastrointestinal Diseases	0.9385517	116	11
Cerebral hemisphere part	0.987087843	290	4
Skin lesion	0.874163434	108	12
Lung	0.952565902	131	9
Uterus	0.916445334	175	7
UTERUS: GENERAL TERMS	0.916445334	175	7
Endocrine Gland Neoplasms	0.918531693	122	10
Female genital organ part	0.917254244	173	7
Immunologic cell	0.953745764	89	13
Uterus part	0.916187934	172	7
[X]Other specified respiratory disorders	0.900153068	134	9
Other disorders of lung	0.900153068	134	9
Other respiratory system diseases NOS	0.900153068	134	9
Disorder of lower respiratory system	0.900153068	134	9
DISEASES OF THE LUNG: GENERAL TERMS	0.900153068	134	9
Lung diseases	0.900153068	134	9
Gonadal structure	0.945116752	126	9
Stomatognathic System	0.942421013	123	9
Mouth and/or pharynx structures	0.942421013	123	9
CLINICAL CLASSIFICATION OF NEOPLASMS OF THE GENITOURINARY SYSTEM	0.831083631	124	10

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Soft tissue tumor AND/OR sarcoma	0.894420302	96	12
Sarcoma - category	0.894420302	96	12
Connective and Soft Tissue Neoplasm	0.894420302	96	12
sarcoma	0.894420302	96	12
Leukocytes, Mononuclear	0.959415318	89	12
Brain Diseases	0.96444927	116	9
Tissue membrane	0.858421184	104	11
Upper aerodigestive tract	0.955354786	113	9
Inflammation of specific body systems	0.789088761	91	13
Inflammation of specific body structures or tissue	0.789088761	91	13
Inflammation of specific body organs	0.789088761	91	13
Inflammatory disorder	0.789088761	91	13
Cerebral cortex	0.988689443	236	4
Layer of cerebrum	0.988689443	236	4
Malignant neoplasm of pelvis	0.886186182	104	10
Malignant neoplasm of genitourinary organ NOS	0.886186182	104	10
Marrow lymphoid tissue	0.982844922	82	11
Lymphocyte	0.982844922	82	11
Face	0.952046729	103	9
Male Genital Organs	0.908565602	88	11
Male genitourinary tract	0.908565602	88	11
Neoplasms, Nerve Tissue	0.942557141	93	10
Neoplasms, Germ Cell and Embryonal	0.942557141	93	10
Neuroectodermal Tumors	0.942557141	93	10
Hematopoietic precursor cell	0.949869526	131	7
Lobe of brain	0.987886124	218	4
Cerebral lobe	0.987886124	218	4
Animal Structures	0.941326491	109	8
Mammalian Oviducts	0.951279092	107	8
animal Oviduct	0.951279092	107	8
Uterine adnexae structure	0.951279092	107	8
Ovary and/or broad ligament structures	0.947738	103	8
Ovary	0.947738	103	8
Digestive organ part	0.85285045	91	10
Regional musculoskeletal structure	0.871264407	79	11
Skin Neoplasms	0.901783328	93	9
Neoplasm of integumentary system	0.901783328	93	9
Limb structure	0.910886673	72	11
Primary malignant neoplasm of soft tissues	0.904030962	88	9
Gastrointestinal Neoplasms	0.923123104	86	9
Oral region	0.949576787	94	8
Oral cavity	0.949576787	94	8
Body of uterus	0.941591198	151	5
CLINICAL CLASSIFICATION OF NEOPLASMS OF THE SKIN	0.904218465	87	9
Malignant neoplasm of skin	0.904218465	87	9
Primary malignant neoplasm of skin	0.904218465	87	9
Malignant neoplasm of thorax	0.953098002	72	10
Extremity part	0.929292929	66	11
Lymphoid leukemia (category)	0.995833737	84	8
Lymphoblastic Leukemia	0.995833737	84	8
Adult Stem Cells	0.935222713	101	7
Disorder of lower gastrointestinal tract	0.94470788	85	8
Intestinal Diseases	0.94470788	85	8
Propensity to adverse reactions	0.842927076	68	11
Hypersensitivity	0.842927076	68	11
Immune hypersensitivity disorder by mechanism	0.842927076	68	11
Hypersensitivity disorder	0.842927076	68	11
Adverse reactions	0.842927076	68	11

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Bone Diseases	0.972210425	70	9
Liver and/or biliary structure	0.932231914	59	11
Mammary Neoplasms	0.965571757	69	9
Malignant neoplasm of breast	0.965571757	69	9
Breast Diseases	0.965571757	69	9
Vascular Diseases	0.924512637	108	6
GENERAL VASCULAR DISORDERS	0.924512637	108	6
Carcinoma of the Large Intestine	0.963114694	69	9
Liver	0.931544995	58	11
Mouth region part	0.962988999	88	7
SKELETAL MUSCULAR SYSTEM: GENERAL TERMS	0.985010901	86	7
Muscle	0.985010901	86	7
Muscle structure	0.985010901	86	7
Types and Parts of Skeletal Muscles	0.985010901	86	7
Skeletal Muscular System (Muscles of Head, Neck, Mouth and Upper Extremity)	0.985010901	86	7
Skeletal muscle system structure	0.985010901	86	7
Neuroendocrine Tumors	0.95962756	87	7
Regional bone structure	0.872130383	71	9
Skeletal System (Bones of Shoulder Girdle, Pelvis and Extremities)	0.927519696	59	10
Epithelial Cells	0.858768407	42	15
System disorder of the nervous system	0.973479132	111	5
Lower genitourinary tract structure	0.885750122	67	9
Structure of soft tissues of head and neck	0.972995275	78	7
Oral soft tissues	0.972995275	78	7
Structure of soft tissues of head	0.972995275	78	7
Lower male genitourinary tract structure	0.890040213	65	9
Back	0.767798133	94	7
Back structure, including back of neck	0.767798133	94	7
Regional back structure	0.767798133	94	7
Genital Neoplasms, Female	0.906491402	92	6
Anogenital region	0.921632318	54	10
Disorder of upper digestive tract	0.932437729	76	7
Mucous Membrane	0.963991525	80	6
Upper extremity part	0.923096036	61	8
Upper Extremity	0.923096036	61	8
Endometrium	0.95604468	93	5
Kidney and/or ureter structures	0.888950617	60	8
Intra-abdominal urinary structure	0.888950617	60	8
Neurodegenerative Disorders	0.978100363	108	4
Kidney	0.892252223	59	8
Retroperitoneal Space	0.929105471	75	6
T-Lymphocyte	0.99042471	70	6
B-cell neoplasm	0.911092675	44	10
Myeloproliferative disease	0.996339917	66	6
Bone Marrow Diseases	0.996339917	66	6
Myeloid Leukemia	0.996339917	66	6
Leukemia, Myelocytic, Acute	0.996339917	66	6
Skeletal tissue	0.934237452	70	6
Bone Tissue	0.934237452	70	6
Structure of shoulder and/or upper arm	0.923367003	60	7
Shoulder	0.923367003	60	7
Colonic Diseases	0.95768938	57	7
Disorder of large intestine	0.95768938	57	7
Diseases and Syndromes of Colon, Appendix and Rectum	0.95768938	57	7
Degenerative disorder	0.971820447	98	4
CLINICAL CLASSIFICATION OF NEOPLASMS OF THE RESPIRATORY SYSTEM	0.970887741	98	4

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Respiratory Tract Neoplasms	0.970887741	98	4
Neoplasm of lower respiratory tract	0.970887741	98	4
Lung Neoplasms	0.970887741	98	4
Malignant neoplasm of lung	0.970887741	98	4
Anterior perineum	0.928114094	50	8
External genitalia	0.928114094	50	8
Blast Cell	0.913205095	96	4
Pectoral girdle structure	0.923206641	54	7
Exocrine Glands	0.879361785	65	6
Allergic disorder by body site affected	0.879049525	43	9
Temporal Lobe	0.96914216	117	3
Autoimmune Diseases	0.928837793	40	9
Infectious and parasitic diseases NOS	0.97870768	85	4
INFECTIOUS AND PARASITIC DISEASES: GENERAL TERMS	0.97870768	85	4
Communicable Diseases	0.97870768	85	4
Epithelium	0.881669888	47	8
Lymphoma, Diffuse	0.91590301	40	9
Malignant lymphoma, diffuse	0.91590301	40	9
Diffuse low grade B-cell lymphoma morphology	0.91590301	40	9
Low grade B-cell lymphoma morphology	0.91590301	40	9
B-cell lymphoma morphology	0.91590301	40	9
Unspecified and Diffuse Lymphomas	0.91590301	40	9
Lymphoma, Non-Hodgkin's	0.91590301	40	9
Lymphoma	0.91590301	40	9
Head and Neck Neoplasms	0.804197324	40	10
Disorder of upper gastrointestinal tract	0.949484821	56	6
Chronic Disease	0.868639252	52	7
[X]Diseases of esophagus, stomach and duodenum	0.951535523	55	6
Intestinal Neoplasms	0.939129106	55	6
Cancer of Intestines	0.939129106	55	6
Limbic System	0.945704783	109	3
Stomach Diseases	0.95120221	54	6
Diseases and Syndromes of Stomach and Duodenum	0.95120221	54	6
Primary malignant neoplasm of intra-abdominal organs	0.909111106	56	6
Lower urinary tract	0.884958781	57	6
Bladder and outflow structure	0.884958781	57	6
Pelvic cavity urinary structure	0.884958781	57	6
Malignant squamous tumor	0.961447259	78	4
Squamous Cell Neoplasms	0.961447259	78	4
Squamous cell carcinoma - category	0.961447259	78	4
[M]Papillary and squamous cell neoplasms	0.961447259	78	4
Urinary outflow structure	0.899006875	55	6
Breast part	0.982384659	75	4
Colorectal Neoplasms	0.962403491	51	6
Colonic Neoplasms	0.962403491	51	6
Malignant tumor of colon	0.962403491	51	6
Mass of colon	0.962403491	51	6
Rectal Diseases	0.962403491	51	6
Malignant neoplasm of large intestine	0.962403491	51	6
Anorectal disorder	0.962403491	51	6
Neoplasms, Ductal, Lobular, and Medullary	0.979846382	58	5
Ductal, lobular AND/OR medullary neoplasm	0.979846382	58	5
ovarian neoplasm	0.985859073	70	4
Ovarian Diseases	0.985859073	70	4
Gonadal Disorders	0.985859073	70	4
Neoplasm of uterine adnexa	0.985859073	70	4
Adnexal Diseases	0.985859073	70	4
Stomach and Omentum	0.880863512	52	6

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Thyroid and/or parathyroid structures	0.99480563	46	6
Malignant neoplasm of female genital organ	0.979987271	56	5
Malignant neoplasm of other and unspecified female genital organs	0.979987271	56	5
Male external genitalia structure	0.933642869	41	7
Bone structure of head and/or neck	0.90272065	59	5
Bone structure of cranium	0.90272065	59	5
Bones of cranium and face	0.90272065	59	5
Musculoskeletal structure of head	0.90272065	59	5
Musculoskeletal structure of head and neck	0.90272065	59	5
Bone structure of head	0.90272065	59	5
Pluripotent Stem Cells	0.996533169	53	5
Mesenchymal Stem Cells	0.998486893	66	4
Inflammatory disorder of musculoskeletal system	0.89003006	36	8
Prostatic and/or seminal vesicle structures	0.891170534	47	6
Minor pelvis	0.891170534	47	6
Male urinary outflow structure	0.891170534	47	6
Prostate and vas deferens structures	0.891170534	47	6
Prostate	0.891170534	47	6
Male internal genital organ	0.891170534	47	6
Pelvic cavity male genital structure	0.891170534	47	6
GENERAL CONVENIENCE TERMS	0.957050207	131	2
Other mental disorders	0.957050207	131	2
Schizophrenia and Disorders with Psychotic Features	0.957050207	131	2
Mental disorders	0.957050207	131	2
9-72 PSYCHOTIC DISORDERS NEC in SNMI98	0.957050207	131	2
Psychotic Disorders	0.957050207	131	2
Adrenal Glands	0.990503356	50	5
General Cytologic Alterations	0.890608599	45	6
Abnormal cell	0.890608599	45	6
Urologic Diseases	0.810384134	49	6
Urologic Neoplasms	0.810384134	49	6
URINARY TRACT DISEASES: GENERAL TERMS	0.810384134	49	6
Malignant tumor of urinary system	0.82529203	48	6
Arthritis	0.894450006	33	8
Other and unspecified arthropathies	0.894450006	33	8
Arthropathies NOS	0.894450006	33	8
DERANGEMENTS OF THE JOINTS OTHER THAN VERTEBRAL COLUMN	0.894450006	33	8
Mechanical joint disorder	0.894450006	33	8
Thyroid Gland	0.999657493	44	5
Rheumatism	0.91674193	34	7
Malignant melanoma - category	0.977681674	74	3
Nevi and Melanomas	0.977681674	74	3
Melanocytic neoplasm	0.977681674	74	3
melanoma	0.977681674	74	3
Nevus AND/OR melanoma	0.977681674	74	3
Esophageal and/or gastric structures	0.952011911	45	5
Mouth, esophagus and stomach structures	0.952011911	45	5
Leukocyte Disorders	0.939452575	38	6
Structure of digestive system mucous membrane	0.971257448	55	4
Mediastinum	0.89565277	39	6
Breast Carcinoma	0.969737078	43	5
Primary malignant neoplasm of breast	0.969737078	43	5
frontal lobe	0.962626941	54	4
Extrapyramidal system	0.959566	108	2
Infratentorial brain part	0.949878385	71	3
Brain Stem	0.949878385	71	3
Infratentorial brain structure	0.949878385	71	3

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Rheumatoid Arthritis	0.92804053	31	7
Delayed hypersensitivity disorder	0.92804053	31	7
Secondary inflammatory arthritis	0.92804053	31	7
Arthropathy associated with a hypersensitivity reaction	0.92804053	31	7
Arthropathy associated with another disorder	0.92804053	31	7
Cancer of ovary and other female genital organs	0.98684912	51	4
Malignant neoplasm of ovary	0.98684912	51	4
CLINICAL CLASSIFICATION OF NEOPLASMS OF THE ENDOCRINE SYSTEM	0.942647269	35	6
Chronic inflammatory disorder	0.865021403	45	5
Degenerative Diseases, Central Nervous System	0.982250516	95	2
Hereditary AND/OR degenerative disease of central nervous system	0.982250516	95	2
Embryonic Stem Cells	0.994912283	31	6
B-Cell Lymphomas	0.910409174	29	7
Hippocampus (Brain)	0.970426009	63	3
Hippocampal Formation	0.970426009	63	3
Structure of archicortex	0.970426009	63	3
Cancer; other primary	0.923661035	33	6
Cancer of Head and Neck	0.923661035	33	6
Stomach and/or duodenal structures	0.969442212	37	5
Structure of soft tissues of trunk	0.611251251	32	9
Ductal Carcinoma	0.971613624	45	4
Lymphoid system structure	0.924162257	27	7
Lymphoid organ structure	0.924162257	27	7
Lymphatic System	0.924162257	27	7
Lymphoid Tissue	0.924162257	27	7
Stomach	0.967796705	36	5
myometrium	0.995057317	58	3
Smooth muscle (tissue)	0.995057317	58	3
HEART: GENERAL TERMS	0.94233838	36	5
CARDIOVASCULAR SYSTEM: GENERAL TERMS	0.94233838	36	5
Cardiovascular system	0.94233838	36	5
Cardiovascular structure of trunk	0.94233838	36	5
HEART AND PERICARDIUM	0.94233838	36	5
Heart	0.94233838	36	5
Heart AND pericardium structure	0.94233838	36	5
Regional cardiovascular structure	0.94233838	36	5
Intrathoracic cardiovascular structure	0.94233838	36	5
Neurologic Manifestations	0.769576204	44	5
Cerebral cortex part	0.981008689	85	2
Cerebral gyrus	0.981008689	85	2
Gyrus of brain	0.981008689	85	2
Squamous cell carcinoma	0.973136228	56	3
Upper respiratory tract	0.945898453	34	5
Pharynx and/or larynx structures	0.945898453	34	5
Ear, nose and throat	0.945898453	34	5
PHARYNX - OROPHARYNX AND HYPOPHARYNX	0.945898453	34	5
Pharyngeal structure	0.945898453	34	5
Organ dysfunction syndrome	0.988512532	81	2
Bacterial Infections	0.988512532	81	2
Shock	0.988512532	81	2
Bacterial infections - causative organisms	0.988512532	81	2
Systemic Inflammatory Response Syndrome	0.988512532	81	2
Systemic infection	0.988512532	81	2
Acute Disease	0.988512532	81	2
Acute disease of cardiovascular system	0.988512532	81	2
Infection by site	0.988512532	81	2
Connective Tissue Cells	0.952872034	24	7

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Muscle, Striated	0.961394427	23	7
Skeletal muscle structure	0.961394427	23	7
Mature (peripheral) B-cell neoplasm	0.918352051	28	6
Disorder of basophils	0.9886844	31	5
Basophilic leukemia	0.9886844	31	5
Disorder involving basophils and mast cells	0.9886844	31	5
Malignant white blood cell disorder	0.9886844	31	5
Acute Basophilic Leukemia	0.9886844	31	5
DISEASES OF THE LIVER AND BILIARY SYSTEM	0.910833254	28	6
Bone structure of face	0.992281879	50	3
Dentition	0.992281879	50	3
Jaw	0.992281879	50	3
Structure of gum and supporting structure of tooth	0.992281879	50	3
Oral hard tissue structure	0.992281879	50	3
Teeth and Tooth Structures	0.992281879	50	3
Gingiva	0.992281879	50	3
Maxillofacial bone structure	0.992281879	50	3
TEETH, GUMS AND SUPPORTING STRUCTURES: GENERAL TERMS	0.992281879	50	3
Periodontium	0.992281879	50	3
Structure of teeth, gums, and supporting structures	0.992281879	50	3
Tooth structure	0.992281879	50	3
Brain stem part	0.968215416	51	3
Midbrain and pons	0.968215416	51	3
Benign Neoplasm	0.580777863	51	5
Tracheobronchial tree part	0.989084098	29	5
Tracheobronchial structure	0.989084098	29	5
Acute infectious disease	0.989208494	70	2
Cardiovascular Infections	0.989208494	70	2
Septic Shock	0.989208494	70	2
Disorder of neck	0.987936614	28	5
Acute lymphoblastic leukemia - category	0.981072142	28	5
Acute lymphocytic leukemia	0.981072142	28	5
Basal Ganglia	0.957292471	70	2
Basal ganglia and capsules	0.957292471	70	2
Neck Neoplasms	0.991724325	27	5
Layer of adrenal gland	0.993690251	44	3
Endocrine gland part	0.993690251	44	3
Adrenal part	0.993690251	44	3
Adrenal Cortex	0.993690251	44	3
Diseases and Syndromes of Peritoneum, Omentum and Mesentery	0.922423086	35	4
Peritoneal Diseases	0.922423086	35	4
Primary malignant neoplasm of pelvis	0.757257067	28	6
Uterine Diseases	0.852336909	49	3
Myeloid Cells	0.958760115	26	5
Cell content alteration	0.958760115	26	5
Phagocytes	0.958760115	26	5
Inflammatory disorder of digestive system	0.948365397	43	3
Inflammatory disorder of digestive tract	0.948365397	43	3
Midbrain structure	0.966070632	42	3
Precursor Cell Lymphoblastic Leukemia Lymphoma	0.99952862	60	2
Other gastrointestinal cancer	0.91795572	32	4
Tumor of esophagus, stomach and duodenum	0.935483871	31	4
Nervous system tumor morphology	0.940473634	17	7
Central nervous system tumor morphology	0.940473634	17	7
Neoplasms, Neuroepithelial	0.940473634	17	7
Glioma	0.940473634	17	7
Stomach Neoplasms	0.932511111	30	4
Malignant neoplasm of stomach	0.932511111	30	4

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Musculoskeletal structure of limb	0.955653831	19	6
Bronchial	0.987865691	22	5
Organ cavity	0.832953498	26	5
Lower Extremity	0.976372289	18	6
Hematopoietic stem cells	0.994123539	35	3
Cancer of Neck	0.995826078	26	4
[X]Inflammatory polyarthropathies	0.980794838	26	4
Chronic arthritis of juvenile onset	0.980794838	26	4
Chronic polyarticular juvenile rheumatoid arthritis	0.980794838	26	4
Chronic arthropathy	0.980794838	26	4
Chronic arthritis	0.980794838	26	4
Chronic Childhood Arthritis	0.980794838	26	4
Chronic disease of musculoskeletal system	0.980794838	26	4
Polyarthropathy	0.980794838	26	4
Adipose tissue	0.974546758	26	4
Primary malignant neoplasm of urinary system	0.749115082	27	5
Diencephalon part	0.925054759	54	2
Structure of diencephalon	0.925054759	54	2
Airway structure	0.980614618	20	5
Body conduit	0.980614618	20	5
Cervix Uteri	0.907309817	21	5
Normal pregnancy and/or delivery	0.967987732	49	2
Twin Multiple Birth	0.967987732	49	2
Maternal AND/OR fetal condition affecting labor AND/OR delivery	0.967987732	49	2
Abnormal products of conception	0.967987732	49	2
MATERNAL AND FETAL CONDITIONS AFFECTING LABOR AND DELIVERY	0.967987732	49	2
Hemorrhagic complication of pregnancy	0.967987732	49	2
Complications of pregnancy, childbirth and the puerperium	0.967987732	49	2
Disorder of labor / delivery	0.967987732	49	2
Disorder of pregnancy	0.967987732	49	2
Pregnancy, Multiple	0.967987732	49	2
Pregnancy Complications	0.967987732	49	2
Disorder of product of conception	0.967987732	49	2
Delivery AND/OR maternal condition affecting management	0.967987732	49	2
Umbilical Cord Blood	0.98659523	32	3
Cancer of Urinary Tract	0.823079481	23	5
Intestinal Mucosa	0.987446595	47	2
Layers of gastrointestinal wall	0.987446595	47	2
Intestinal wall structure	0.987446595	47	2
Structure of gastrointestinal mucous membrane	0.987446595	47	2
Retroperitoneal mass	0.934945046	33	3
Uterine Neoplasms	0.788099341	39	3
Testis	0.993652492	23	4
Scrotal and testis structures	0.993652492	23	4
Retroperitoneal Neoplasms	0.944369163	32	3
Blood Vessels	0.902209302	20	5
Prostate mass	0.817365812	22	5
Disorder of male reproductive system	0.817365812	22	5
Malignant neoplasm of prostate	0.817365812	22	5
Disorder of the lower urinary tract	0.817365812	22	5
DISEASES OF THE LOWER URINARY TRACT: GENERAL CONDITIONS	0.817365812	22	5
malignant tumor of male genital organ	0.817365812	22	5
Prostatic Diseases	0.817365812	22	5
Prostatic Neoplasms	0.817365812	22	5
Genital Neoplasms, Male	0.817365812	22	5
Genital Diseases, Male	0.817365812	22	5
Small Intestine - Duodenum	0.847446994	26	4

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Intestines, Small	0.847446994	26	4
SMALL INTESTINE: GENERAL TERMS	0.847446994	26	4
Benign epithelial neoplasm - category	0.734011111	30	4
Benign adenomatous neoplasm - category	0.734011111	30	4
adenoma	0.734011111	30	4
Lymphoid precursor cell	0.999611825	44	2
lymphoblast	0.999611825	44	2
[X]Malignant neoplasm of thyroid and other endocrine glands	0.954295051	23	4
Malignant neoplasm of endocrine gland	0.954295051	23	4
Cerebral degeneration presenting primarily with dementia	0.995016423	87	1
Alzheimer's Disease	0.995016423	87	1
[X]Dementia in other diseases classified elsewhere	0.995016423	87	1
DEMENTIAS IN THE SENIUM AND PRESENIUM	0.995016423	87	1
Other cerebral degeneration NOS	0.995016423	87	1
Degenerative brain disorder	0.995016423	87	1
Delirium, Dementia, Amnestic, Cognitive Disorders	0.995016423	87	1
Tauopathies	0.995016423	87	1
Dementia	0.995016423	87	1
Dementing Neurological Diseases and Syndromes	0.995016423	87	1
Disease of liver and bile duct	0.904857627	19	5
Malignant neoplasm of liver	0.904857627	19	5
Liver neoplasms	0.904857627	19	5
Liver diseases	0.904857627	19	5
Acute Myeloid Leukemia (AML-M2)	0.986247606	21	4
Structure of soft tissues of abdomen	0.561973276	24	6
EMBRYO AND FETUS	0.868641799	13	7
penis	0.875996016	18	5
Malignant Glioma	0.932550208	14	6
Bronchial Diseases	0.973688928	26	3
Lung Diseases, Obstructive	0.973688928	26	3
Pharyngeal part	0.971269077	26	3
Antibody-Producing Cells	0.982384293	11	7
B-Lymphocytes	0.982384293	11	7
Tongue	0.94076412	20	4
Skin tissue	0.999688663	75	1
Hereditary and degenerative nervous system conditions	0.924063591	27	3
Occipital lobe	0.968327822	38	2
Serous sac	0.798288328	18	5
Serous Membrane	0.798288328	18	5
Bronchi	0.989929172	18	4
Corpus striatum structure	0.981235392	35	2
Lentiform nucleus structure	0.981235392	35	2
Neoplasm Metastasis	0.914861897	25	3
Neoplastic Processes	0.914861897	25	3
Hemorrhage	0.877586295	26	3
Hemorrhage of blood vessel	0.877586295	26	3
Myomatous neoplasm	0.975130493	23	3
Peritoneal sac	0.787801878	17	5
Peritoneal Cavity	0.787801878	17	5
Structure of cavity of serous sac	0.787801878	17	5
Structure of serous cavity	0.787801878	17	5
Peritoneum	0.787801878	17	5
Frontal lobe gyrus	0.982731878	34	2
Lactiferous duct	0.999361019	66	1
Mammary lobe	0.999361019	66	1
Glandular structure of breast	0.999361019	66	1
Duct (organ) structure	0.999361019	66	1
Thyroid lump	0.997053312	22	3

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Malignant neoplasm of thyroid	0.997053312	22	3
thyroid neoplasm	0.997053312	22	3
Thyroid Diseases	0.997053312	22	3
Spinal Cord	0.995085995	33	2
Vertebral column	0.995085995	33	2
BONES OF VERTEBRAL COLUMN	0.995085995	33	2
Structure of vertebral region of back	0.995085995	33	2
Spinal cord, roots and ganglia structure	0.995085995	33	2
Primary malignant neoplasm of gastrointestinal tract	0.960565067	34	2
Primary malignant neoplasm of large intestine	0.978705979	33	2
Colon Carcinoma	0.978705979	33	2
Primary malignant neoplasm of colon	0.978705979	33	2
Primary malignant neoplasm of intestinal tract	0.978705979	33	2
Nerve	0.976026532	33	2
Spinal nerve structure	0.976026532	33	2
Nerve part	0.976026532	33	2
Peripheral Nervous System	0.976026532	33	2
Non-Autonomic Spinal Nerves	0.976026532	33	2
Peripheral Nerves	0.976026532	33	2
Extrapyramidal Disorders	0.949830754	22	3
Movement Disorders	0.949830754	22	3
Other and unspecified extrapyramidal diseases and abnormal movement disorders	0.949830754	22	3
Motion and Coordination Diseases and Syndromes	0.949830754	22	3
Liver tumor morphology	0.944840743	16	4
Adenocarcinoma of liver	0.944840743	16	4
Primary carcinoma of the liver cells	0.944840743	16	4
Primary malignant neoplasm of liver	0.944840743	16	4
Neoplasm of body of uterus	0.793317267	38	2
Nose and nasopharynx structure	0.988811111	30	2
Endometriosis, site unspecified	0.980633333	30	2
Disorder characterized by pain	0.980633333	30	2
Hypothalamic structure	0.971711111	30	2
Benign neoplasm of trunk	0.936441179	31	2
Benign neoplasm of abdomen	0.936441179	31	2
Metencephalon	0.982821818	29	2
hindbrain	0.982821818	29	2
Regional skeletal muscle structure	0.936436934	12	5
Kidney part	0.885644653	21	3
Pain	0.927266667	30	2
Sensory and Pain Diseases and Syndromes	0.927266667	30	2
Pain Disorder	0.927266667	30	2
Adrenal mass	0.998125331	27	2
Tumors of Adrenal Cortex	0.998125331	27	2
Adrenal Cortex Diseases	0.998125331	27	2
Adrenal Gland Diseases	0.998125331	27	2
Adrenal Gland Neoplasms	0.998125331	27	2
lymph nodes	0.893914292	12	5
Regional vascular structure	0.95645197	18	3
Serous membrane part	0.80121931	16	4
Omentum	0.80121931	16	4
Ganglia, Sensory	1	25	2
Structure of nervous system ganglion	1	25	2
Ganglia	1	25	2
Leukemia, T-Cell	0.999483221	50	1
Structure of putamen	0.994276206	25	2
Neostriatum	0.994276206	25	2
Temporal lobe gyrus	0.974468337	51	1

Continued on Next Page…

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Immediate hypersensitivity	0.992372712	25	2
Asthma	0.992372712	25	2
Obstruction of lower respiratory tract	0.992372712	25	2
Respiratory Hypersensitivity	0.992372712	25	2
Respiratory Insufficiency	0.992372712	25	2
Hypersensitivity disease	0.992372712	25	2
Airway Obstruction	0.992372712	25	2
Stomach part	0.972491751	17	3
Region of stomach	0.972491751	17	3
Lower female genital structure	0.971261787	17	3
Fetus	0.895269355	11	5
GENERAL CONDITIONS OF THE KIDNEY AND URETER	0.944074567	26	2
Kidney Neoplasms	0.944074567	26	2
Kidney Diseases	0.944074567	26	2
Malignant neoplasm of kidney	0.975720466	25	2
Tumor Cells, Cultured	0.99094323	12	4
Cell Line, Tumor	0.99094323	12	4
Disorder of small intestine	0.989756598	24	2
Inflammatory Bowel Diseases	0.989756598	24	2
Gastritis	0.989756598	24	2
Gastroenteritis	0.989756598	24	2
Thalamic structure	0.967356953	24	2
Musculoskeletal structure of lower limb	0.957256461	12	4
Bone of limb	0.957256461	12	4
Bone structure of lower limb	0.957256461	12	4
Bone and/or joint structure of limb	0.957256461	12	4
Musculoskeletal structure of trunk	0.925833886	12	4
Small intestine part	0.919334771	16	3
Nutrition Disorders	0.912193506	12	4
Developmental Disabilities	0.994223041	44	1
Mental disorder of infancy, childhood or adolescence	0.994223041	44	1
Mental disorder usually first evident in infancy, childhood AND/OR adolescence	0.994223041	44	1
Mental Disorders Diagnosed in Childhood	0.994223041	44	1
Developmental mental disorder	0.994223041	44	1
Leiomyomatous neoplasm - category	0.992247945	22	2
Tegmentum Mesencephali	0.983302103	22	2
Midbrain part	0.983302103	22	2
Cerebral Peduncle	0.983302103	22	2
Dermatitis	0.930127142	15	3
Small Intestine - Jejunum and Ileum	0.926390271	15	3
Carcinoma, Papillary	0.941972921	22	2
Cerebellum	1	20	2
Cardiovascular organ part	0.997873754	20	2
Heart part	0.997873754	20	2
Disorder of soft tissue of body cavity	0.996528239	20	2
Disorder of soft tissue of head	0.996528239	20	2
Mouth Diseases	0.996528239	20	2
DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY	0.996528239	20	2
Disorder of oral soft tissues	0.996528239	20	2
Circulatory system disease NOS	0.996478405	20	2
Malignant neoplasm of soft tissues of thorax	0.989571913	13	3
Disorder of soft tissue of trunk	0.989571913	13	3
Skin disorder of breast	0.989571913	13	3
Primary malignant neoplasm of skin of chest	0.989571913	13	3
Primary malignant neoplasm of soft tissues of trunk	0.989571913	13	3
Primary malignant neoplasm of soft tissues of thorax	0.989571913	13	3
Primary malignant neoplasm of skin of trunk	0.989571913	13	3
Primary malignant neoplasm of chest wall	0.989571913	13	3

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Carcinoma, Lobular	0.989571913	13	3
Malignant neoplasm of skin of trunk	0.989571913	13	3
Neoplasm of skin region	0.989571913	13	3
Disorder of body wall	0.989571913	13	3
Primary malignant neoplasm of skin of breast	0.989571913	13	3
Neoplasm of soft tissues of thorax	0.989571913	13	3
Neoplasm of skin of chest	0.989571913	13	3
Neoplasm of skin of breast	0.989571913	13	3
Neoplasm of skin of trunk	0.989571913	13	3
Disorder of skin AND/OR subcutaneous tissue of trunk	0.989571913	13	3
Neoplasm of soft tissues of trunk	0.989571913	13	3
Neoplasm of chest wall	0.989571913	13	3
Hereditary Diseases	0.855280195	11	4
Parkinson Disease	0.976367355	19	2
Basal Ganglia Diseases	0.976367355	19	2
Parkinsonian Disorders	0.976367355	19	2
Carcinoma of genital organs NOS	0.88089712	14	3
Carcinoma of genitourinary organ	0.88089712	14	3
Endocrine tumor morphology	0.947808572	13	3
Noninfectious, erythematous, papular AND/OR squamous disease	0.929017618	13	3
Cerebral white matter structure	0.996753726	18	2
Corpus Callosum	0.996753726	18	2
White matter structure of brain and spinal cord	0.996753726	18	2
Child Development Disorders, Pervasive	0.992549487	35	1
Psychoses with origin in childhood	0.992549487	35	1
Autistic Disorder	0.992549487	35	1
[X]Unspecified disorder of psychological development	0.992549487	35	1
Pervasive Development Disorder	0.992549487	35	1
Endothelial Cells	0.980341194	7	5
MULTIPLE SYSTEM MALFORMATIONS AND CHROMOSOMAL DISEASES	0.84513245	10	4
Congenital Disorders	0.84513245	10	4
Vascular structure of trunk	0.929865253	12	3
Malignant neuroendocrine neoplasm, neural	0.988888554	11	3
Embryonal neuroepithelial tumor	0.988888554	11	3
Neuronal and mixed neuronal-glial tumor	0.988888554	11	3
Neuroepitheliomatous neoplasm	0.988888554	11	3
Neuroectodermal Tumor, Primitive	0.988888554	11	3
Bacteria	0.903744201	12	3
Prokaryote	0.903744201	12	3
Musculoskeletal structure of pelvis	0.984070583	11	3
Structure of superior frontal gyrus	0.98377165	33	1
Disorder of lipoprotein AND/OR lipid metabolism	0.977235087	11	3
Other disorders of metabolism	0.977235087	11	3
Metabolic Diseases	0.977235087	11	3
HYPERALIMENTATION AND OBESITY	0.973561384	11	3
Overnutrition	0.973561384	11	3
Obesity	0.973561384	11	3
Other endocrine/nutritional/metabolic disorder	0.973561384	11	3
Cranial nerve part	0.943050702	17	2
Cranial Nerves	0.943050702	17	2
Structure of layer of kidney	0.999046118	16	2
Nerve Tissue	0.999004645	16	2
Spinal nerve root structure	0.999004645	16	2
Peripheral nerve part	0.999004645	16	2
Nerve root structure	0.999004645	16	2
Ganglia, Spinal	0.999004645	16	2
Adrenocortical carcinoma	0.997573822	16	2

Continued on Next Page…

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Non-Occupational Pulmonary Diseases and Syndromes	0.952362311	11	3
Amygdaloid structure	0.981668879	16	2
Veins	0.864724705	9	4
Venous system	0.864724705	9	4
VEINS - TYPE AND STRUCTURE	0.864724705	9	4
ARTERIES: TYPE AND STRUCTURE	0.928091782	11	3
Systemic vascular structure	0.928091782	11	3
Systemic arterial structure	0.928091782	11	3
Artery of trunk	0.928091782	11	3
Arteries	0.928091782	11	3
Arterial system	0.928091782	11	3
Head Neoplasms	0.762574454	8	5
Extracellular Fluid	0.918034268	11	3
Extracellular Space	0.918034268	11	3
Posterior root of spinal nerve	0.998629077	15	2
Skin part	0.93563865	8	4
SKIN REGION: GENERAL TERM	0.93563865	8	4
Skin region	0.93563865	8	4
Skin of trunk, NOS	0.93563865	8	4
Skin of part of trunk	0.93563865	8	4
Skin AND subcutaneous tissue structure of trunk	0.93563865	8	4
Nervous System Neoplasms	0.854023912	7	5
Central Nervous System Neoplasms	0.854023912	7	5
Intracranial mass	0.854023912	7	5
Brain Neoplasms	0.854023912	7	5
Neoplasms, Intracranial	0.854023912	7	5
Visual Cortex	0.978788889	30	1
Body wall structure	0.807017544	9	4
Salivary Glands	0.933609272	10	3
Cardiac internal structure	0.999668435	14	2
Cardiac chamber structure	0.999668435	14	2
neutrophil	0.998531641	14	2
granulocyte	0.998531641	14	2
Neurosecretory Systems	0.993416067	14	2
Hypothalamus, Middle	0.993416067	14	2
Hypothalamo-Hypophyseal System	0.993416067	14	2
Hypothalamus part	0.993416067	14	2
Pituitary and/or pineal structures	0.993416067	14	2
Pituitary Gland	0.993416067	14	2
Systemic circulatory system	0.86618961	8	4
Afterbirth	0.852043349	8	4
Structure of middle temporal gyrus	0.969936709	28	1
Layer of temporal lobe	0.969936709	28	1
Cerebral dorsum structure	0.969936709	28	1
Gray matter of temporal lobe	0.969936709	28	1
Acute myeloid leukemia without maturation	0.984111221	9	3
Female perineal structure	0.981205635	9	3
Vulva	0.981205635	9	3
Vulval and/or female perineal structures	0.981205635	9	3
Female external genitalia structure	0.981205635	9	3
Esophagus	0.875529801	10	3
Nipples	0.949280959	9	3
Proximal stomach	0.946816727	9	3
Synovial Membrane	0.841680486	15	2
ARTICULAR SYSTEM - JOINTS	0.841680486	15	2
ARTICULAR SYSTEM: GENERAL TERMS	0.841680486	15	2
Joint part	0.841680486	15	2
Membrane organ structure	0.841680486	15	2

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Joints	0.841680486	15	2
Soft tissue joint component	0.841680486	15	2
Joint Capsule	0.841680486	15	2
Types and Parts of Joints	0.841680486	15	2
Articular system	0.841680486	15	2
Cecum	0.907315458	9	3
Primary malignant neoplasm of male genital organ	0.935468244	13	2
Prostate carcinoma	0.935468244	13	2
Primary malignant neoplasm of prostate	0.935468244	13	2
Childhood asthma	0.992958527	24	1
Exanthema	0.990004418	12	2
Disorder of keratinization	0.990004418	12	2
Cell-mediated cytotoxic disorder	0.990004418	12	2
Cutaneous hypersensitivity	0.990004418	12	2
Acquired disorder of keratinization	0.990004418	12	2
Histologic type of inflammatory skin disorder	0.990004418	12	2
Psoriasis	0.990004418	12	2
Other psoriasis	0.990004418	12	2
Skin Diseases, Papulosquamous	0.990004418	12	2
Inflammatory hyperkeratotic dermatosis	0.990004418	12	2
Pain finding at anatomical site	0.945544093	25	1
Ventral Tegmental Area	0.975811796	12	2
Abdominal Pain	0.96174318	24	1
Pain of truncal structure	0.96174318	24	1
Benign neoplasm of other endocrine glands and related structures	0.949856417	12	2
Benign tumor of endocrine gland	0.949856417	12	2
Region of cerebral cortex	0.986104886	23	1
Surface of brain	0.986104886	23	1
Structure of entorhinal cortex	0.986104886	23	1
Cerebral medial surface structure	0.986104886	23	1
Parahippocampal Gyrus	0.986104886	23	1
Region of temporal cortex	0.986104886	23	1
Congenital abnormal shape	0.83184376	9	3
CONGENITAL ANOMALIES: GENERAL TERMS	0.83184376	9	3
Congenital growth alteration	0.83184376	9	3
Deformity	0.83184376	9	3
Other and unspecified congenital anomalies	0.83184376	9	3
Congenital Abnormality	0.83184376	9	3
jejunum	0.929285399	12	2
Nasopharynx	0.998148412	21	1
Parameningeal structure in the context of malignancy	0.998148412	21	1
Reticuloendotheliosis	0.995397351	10	2
Malignant histiocytic neoplasm	0.995397351	10	2
Histiocytosis	0.995397351	10	2
Histiocytic Disorders, Malignant	0.995397351	10	2
Histiocytosis, Langerhans-Cell	0.995397351	10	2
Lung Diseases, Interstitial	0.995397351	10	2
Histiocytic neoplasm (morphology)	0.995397351	10	2
Monocytic leukemia	0.995397351	10	2
Histiocytic syndrome	0.995397351	10	2
Dendritic cell neoplasm	0.995397351	10	2
Acute monocytic/monoblastic leukemia	0.995397351	10	2
Langerhans cell histiocytosis - category	0.995397351	10	2
Acute monocytic leukemia	0.995397351	10	2
Entire viscus	0.99423588	20	1
Hollow viscus	0.99423588	20	1
Abdominal organ	0.99423588	20	1
Entire fallopian tube	0.99423588	20	1

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Entire pelvic organ	0.99423588	20	1
Entire female internal genital organ	0.99423588	20	1
Entire pelvic viscus	0.99423588	20	1
Entire female genital organ	0.99423588	20	1
Intra-abdominal genital structure	0.99423588	20	1
Uterine Fibroids	0.99154485	20	1
Benign myomatous tumor	0.99154485	20	1
Benign neoplasm of female genital organ, site unspecified	0.99154485	20	1
Benign neoplasm of body of uterus	0.99154485	20	1
Benign neoplasm of uterus NOS	0.99154485	20	1
Benign leiomyomatous neoplasm - category	0.99154485	20	1
Benign genital neoplasm	0.99154485	20	1
Benign neoplasm corpus uteri NEC	0.99154485	20	1
Thoracic Arteries	0.940551014	7	3
Structure of brachiocephalic artery	0.940551014	7	3
Artery of mediastinum	0.940551014	7	3
Supraaortic branch of thoracic aorta	0.940551014	7	3
Structure of artery of thorax AND/OR abdomen	0.940551014	7	3
Branch of thoracic aorta	0.940551014	7	3
Substantia nigra structure	0.977218543	10	2
Midbrain nucleus	0.977218543	10	2
Diffuse high grade B-cell lymphoma	0.97031405	5	4
High grade B-cell lymphoma	0.97031405	5	4
Peripheral and visceral atherosclerosis	1	19	1
Peripheral Vascular Diseases	1	19	1
Multiple Myeloma	1	19	1
Paraproteinemias	1	19	1
Skin Manifestations	1	19	1
Vascular Hemostatic Disorders	1	19	1
Purpura and other hemorrhagic conditions	1	19	1
Other paraproteinemias	1	19	1
[X]Diseases of arteries, arterioles and capillaries	1	19	1
Blood Protein Disorders	1	19	1
Blood Coagulation Disorders	1	19	1
Gammopathy	1	19	1
Monoclonal Gammapathies	1	19	1
Plasma Cell Neoplasm	1	19	1
White blood cell abnormality	1	19	1
Purpura	1	19	1
Hemorrhagic Disorders	1	19	1
Plasmacytoma - category	1	19	1
Plasma cell myeloma - category	1	19	1
Immunosecretory disorder	1	19	1
Plasma cell myeloma/plasmacytoma	1	19	1
Myeloma cell	1	19	1
Abnormal hematopoietic cell	1	19	1
Abnormal cellular component of blood	1	19	1
Clotting or bleeding disorder NOS	1	19	1
[X]Coagulation defects, purpura and other hemorrhagic conditions	1	19	1
Plasmacytoma	1	19	1
Malignant immunoproliferative disease (clinical)	1	19	1
Coagulation and hemorrhagic disorders	1	19	1
Other peripheral vascular disease	1	19	1
Purpura, Nonthrombocytopenic	1	19	1
Gingival and periodontal disease NOS	0.999160971	19	1
Jaw Diseases	0.999160971	19	1
Inflammatory disorder of jaw	0.999160971	19	1
Inflammatory disorder of head	0.999160971	19	1

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Disorder of teeth AND/OR supporting structures	0.999160971	19	1
Chronic disease of teeth AND/OR supporting structures	0.999160971	19	1
Chronic digestive system disorder	0.999160971	19	1
Disorder of face	0.999160971	19	1
Periodontal Diseases	0.999160971	19	1
Periodontitis	0.999160971	19	1
Neoplasms, Cystic, Mucinous, and Serous	0.920797342	20	1
Cystic, mucinous AND/OR serous neoplasm	0.920797342	20	1
Spleen	1	9	2
Base of skull structure	0.997940344	9	2
Structure of organ cavity subdivision	0.997940344	9	2
Intracranial ganglion	0.997940344	9	2
Structure of fossa of cranial cavity	0.997940344	9	2
Structure of middle fossa of cranial cavity	0.997940344	9	2
Structure of cranial nerve ganglion	0.997940344	9	2
Trigeminal nerve structure	0.997940344	9	2
Structure of trigeminal ganglion	0.997940344	9	2
Parietal Lobe	0.987016808	9	2
Functional disorder of intestine	0.980911398	9	2
DISEASES OF THE GALLBLADDER AND BILE DUCTS	0.975725477	9	2
Biliary Tract Diseases	0.975725477	9	2
Gall Bladder Diseases	0.975725477	9	2
Endometrial Neoplasms	0.972185333	18	1
Endometrial disorder	0.972185333	18	1
Pontine structure	0.958108058	9	2
Still's disease with juvenile onset and/or adult onset	0.990628844	17	1
Systemic onset juvenile chronic arthritis	0.990628844	17	1
Endometrioid tumor	0.974073134	17	1
Malignant endometrioid tumor	0.974073134	17	1
Carcinoma, Endometrioid	0.974073134	17	1
ATRIA: GENERAL TERMS	1	8	2
Urethra	1	8	2
Heart Atrium	1	8	2
macrophage	0.999834547	8	2
Structure of medulla of kidney	0.99975182	8	2
Acute Promyelocytic Leukemia	0.999586367	8	2
Acute myeloid leukemia with recurrent genetic abnormality	0.999586367	8	2
Structure of cortex of kidney	0.9987591	8	2
Vagina	0.998676373	8	2
Structure of pyloric portion of stomach	0.99851092	8	2
Part of pyloric region	0.99851092	8	2
Pylorus	0.99851092	8	2
Oral mucous membrane structure	0.998180013	8	2
Body orifice mucosa	0.998180013	8	2
gastric fundus	0.9975182	8	2
Lymph	0.991727333	8	2
Proteobacteria	0.971376572	8	2
Gram-Negative Bacteria	0.971376572	8	2
Structure of bone (organ)	0.963807081	8	2
Type of bone	0.963807081	8	2
Vestibular nucleus structure	0.963352085	8	2
Pons part	0.963352085	8	2
Structure of vestibular system	0.963352085	8	2
Intracranial nerve structure	0.963352085	8	2
Structure of cranial nerve nucleus	0.963352085	8	2
pontine nuclei	0.963352085	8	2
Special sensory system	0.963352085	8	2
Pontine cranial nerve nucleus	0.963352085	8	2

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Fibroblasts	0.960261071	4	4
Coughing	0.959045289	16	1
T-Cell Lymphoma	0.914160608	4	4
T-cell lymphoma morphology	0.914160608	4	4
T-cell AND/OR NK-cell neoplasm	0.914160608	4	4
Persistent cough	0.953300166	15	1
Congenital chromosomal disease	0.893530774	8	2
Other condition due to autosomal anomaly	0.893530774	8	2
Autosomal hereditary disorder	0.893530774	8	2
Neoplasms, Complex and Mixed	0.893034414	8	2
Larynx and/or tracheal structures	0.998865838	7	2
Trachea	0.998865838	7	2
Musculoskeletal structure of upper limb	0.997259109	7	2
Skeletal muscle structure of upper limb	0.997259109	7	2
monocyte	0.995274325	7	2
Marrow Monocytes and Plasma Cells	0.995274325	7	2
Systemic venous structure	0.988752894	7	2
Type of vein	0.988752894	7	2
Lower extremity part	0.89877686	5	3
sperm cell	1	13	1
Meiotic cell	1	13	1
Germ Cells	1	13	1
Skeletal Muscular System (Muscles of Trunk, Perineum and Lower Extremity)	0.825520661	5	3
Skeletal muscle structure of trunk	0.825520661	5	3
Primary malignant neoplasm of endocrine gland	0.871603421	7	2
Structure of region of lymphatic system	0.808595041	5	3
Structure of peripheral vein	1	6	2
Peripheral vascular system	1	6	2
Venous structure of limb	1	6	2
Saphenous Vein	1	6	2
Vascular structure of lower limb	1	6	2
Structure of pelvic and leg veins	1	6	2
Stromal Cells	1	12	1
Structure of vein of lower extremity	1	6	2
Structure of superficial vein of lower extremity	1	6	2
Vascular structure of limb	1	6	2
Structure of superficial vein	1	6	2
Heart Ventricle	0.999889771	6	2
White Adipose Tissue	0.999669312	6	2
Subcutaneous Fat	0.999669312	6	2
Subcutaneous Tissue	0.999669312	6	2
Chronic Lymphocytic Leukemia	0.999503968	6	2
Coronary artery	0.998567019	6	2
Mammary gland	0.992504409	6	2
Skin and subcutaneous tissue structure of genitalia	0.991407799	4	3
Male perineal structure	0.991407799	4	3
Skin and subcutaneous tissue structure of pelvis	0.991407799	4	3
Glans penis and/or preputial structures	0.991407799	4	3
Skin structure of anogenital region	0.991407799	4	3
Skin and subcutaneous tissue structure of perineum	0.991407799	4	3
Male genital organ part	0.991407799	4	3
Penis part	0.991407799	4	3
Structure of soft tissues of perineum	0.991407799	4	3
Soft tissues of pelvis	0.991407799	4	3
Skin structure of lower trunk	0.991407799	4	3
Skin of penis	0.991407799	4	3
Skin structure of male genitalia	0.991407799	4	3
SKIN OF PERINEUM AND GENITALIA	0.991407799	4	3

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Skin structure of perineum	0.991407799	4	3
Skin structure of external genitalia	0.991407799	4	3
Skin of pelvis	0.991407799	4	3
Spermatic cord and/or male perineal structures	0.991407799	4	3
Skin structure of male perineum	0.991407799	4	3
Structure of skin and/or mucosa of anogenital area	0.991407799	4	3
Foreskin of penis	0.991407799	4	3
Skin of part of pelvic region	0.991407799	4	3
Skin of part of anogenital region	0.991407799	4	3
Skin of part of male external genitalia	0.991407799	4	3
Skin of part of genitalia	0.991407799	4	3
Skin of part of penis	0.991407799	4	3
PLACENTA AND MEMBRANES	0.980434303	6	2
Diffuse non-Hodgkin's lymphoma	0.948777264	4	3
Diffuse Large B-Cell Lymphoma	0.948777264	4	3
Diffuse large B-cell lymphoma - category	0.948777264	4	3
Benign neoplasm of intra-abdominal organs	1	11	1
Benign neoplasm of adrenal gland	1	11	1
Benign neoplasm of adrenal cortex	1	11	1
Adrenal Cortical Adenoma	1	11	1
Benign neoplasm of retroperitoneum	1	11	1
Structure of subthalamic nucleus	0.971152398	11	1
Subthalamic structure	0.971152398	11	1
Neoplasms, Connective Tissue	0.706710744	5	3
Adenocarcinoma, Mucinous	0.954500286	11	1
Large blood vessel structure	0.85587522	6	2
Structure of great blood vessel (organ)	0.85587522	6	2
Type of vessel	0.85587522	6	2
SPECIFIC ENDOMETRIOSES	0.998609272	10	1
Endometriosis of uterus	0.998609272	10	1
Endometriosis of pelvis	0.998609272	10	1
Cervical	0.997019868	10	1
Globus Pallidus	0.995397351	10	1
Malignant retroperitoneal tumor	0.824018959	6	2
Entire putamen	0.987636364	5	2
Neuroblastoma	0.976066116	5	2
Ewings sarcoma-primitive neuroectodermal tumor (PNET)	0.976066116	5	2
[M]Miscellaneous tumor NOS	0.976066116	5	2
Skin tumor of neural origin	0.976066116	5	2
Oropharyngeal	0.966214876	5	2
Papillary adenocarcinoma	0.964635762	10	1
Anorectal structure	0.954834437	10	1
Lower bowel structures	0.954834437	10	1
Rectum	0.954834437	10	1
Pelvic alimentary structure	0.954834437	10	1
Complex mixed AND/OR stromal neoplasm	0.944330579	5	2
Body surface region	0.922049587	5	2
Sense Organs	1	9	1
Nose	1	9	1
Entire skeletal muscle (organ)	0.997136879	3	3
Monozygotic twins	0.994409504	9	1
Neurobehavioral Manifestations	0.990474089	9	1
Mental Retardation	0.990474089	9	1
Chest wall structure	0.888859504	5	2
Part of chest wall	0.888859504	5	2
Entire nucleus of brain	0.926661518	9	1
Structure of large artery	0.828033058	5	2
Type of artery	0.828033058	5	2

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
High grade T-cell lymphoma morphology	0.89362405	3	3
Reticulosarcoma	0.89362405	3	3
Thigh structure	1	4	2
Structure of quadriceps femoris muscle	1	4	2
Structure of vastus lateralis muscle	1	4	2
Skeletal muscle structure of thigh	1	4	2
Skeletal muscle structure of hip	1	4	2
Muscle of hip AND thigh	1	4	2
Skeletal muscle structure of perineum	1	4	2
Thigh part	1	4	2
Skeletal muscle structure of lower limb	1	4	2
Entire quadriceps femoris muscle	1	4	2
Hip region structure	1	4	2
Entire vastus lateralis muscle	1	4	2
Skeletal muscle structure of pelvis	1	4	2
Cholelithiasis	0.999669093	8	1
Cholecystolithiasis	0.999669093	8	1
Calculi	0.999669093	8	1
Biliary calculi	0.999669093	8	1
Multiple Sclerosis	0.999214097	8	1
Autoimmune Diseases of the Nervous System	0.999214097	8	1
Demyelinating Autoimmune Diseases, CNS	0.999214097	8	1
Demyelinating Diseases	0.999214097	8	1
Demyelinating disease of central nervous system	0.999214097	8	1
Deficiency anemias NOS	0.989920687	4	2
Anemia	0.989920687	4	2
Refractory anemias	0.989920687	4	2
Refractory anaemia with excess blasts	0.989920687	4	2
Dysmyelopoietic Syndromes	0.989920687	4	2
Other deficiency anemias NOS	0.989920687	4	2
Other anemias NOS	0.989920687	4	2
Red blood cell disorder	0.989920687	4	2
Anemia due to decreased red cell production	0.989920687	4	2
Developmental delay (disorder)	0.989741893	8	1
Leukemia, Myelomonocytic, Acute	0.988873263	8	1
Nucleus Accumbens	0.988087359	8	1
[M]Complex mixed and stromal neoplasms	0.983559154	4	2
Primary malignant neoplasm of retroperitoneum	0.781884298	5	2
Waldeyer's ring	0.972488434	4	2
Body region wall	0.972488434	4	2
Structure of lymphatic system of head and neck	0.972488434	4	2
Lymphatic vessel	0.972488434	4	2
Wall of oropharynx	0.972488434	4	2
Structure of lymphatic vessel of head and neck	0.972488434	4	2
Tonsil and adenoid structure	0.972488434	4	2
lymphatic system of head	0.972488434	4	2
lateral wall of oropharynx	0.972488434	4	2
Palatine Tonsil	0.972488434	4	2
Low grade B-cell lymphoma	0.964144085	4	2
Uterine Cancer	0.769586777	5	2
Cancer of uterus and cervix	0.769586777	5	2
Virus Diseases	0.95935228	4	2
Specific viral infections	0.95935228	4	2
Firmicutes	0.925644415	4	2
Bacilli class	0.925644415	4	2
Gram-Positive Bacteria	0.925644415	4	2
Extra-embryonic structure	0.781081379	3	3
Bone structure of spine and/or pelvis	1	7	1

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
hip bone	1	7	1
Bone structure of ilium	1	7	1
Bone part	1	7	1
Ilium part	1	7	1
Iliac crest structure	1	7	1
Structure of flat bone	1	7	1
Bone structure of pelvic region and/or thigh	1	7	1
Bony pelvis	1	7	1
Campylobacterales	0.999621946	7	1
Helicobacter	0.999621946	7	1
HCT116 Cells	0.999621946	7	1
Helicobacteraceae	0.999621946	7	1
Epsilonproteobacteria	0.999621946	7	1
Colonic epithelium	0.999621946	7	1
Colonic mucous membrane	0.999621946	7	1
Structure of intestinal epithelium	0.999621946	7	1
Subclass Aerobic-Microaerophilic, Motile Curved Gram-Negative Bacteria	0.999621946	7	1
[M]Adenocarcinoma, metastatic, NOS	0.872559563	8	1
Pancreas	0.865251157	4	2
Congenital hypergammaglobulinemia	0.982987571	7	1
Job's Syndrome	0.982987571	7	1
Congenital immunodeficiency disease	0.982987571	7	1
Qualitative abnormality of granulocyte	0.982987571	7	1
Disorder of neutrophils	0.982987571	7	1
Immunologic Deficiency Syndromes	0.982987571	7	1
Non-malignant white cell disorder	0.982987571	7	1
Chemotactic disorder	0.982987571	7	1
Autosomal recessive hereditary disorder	0.982987571	7	1
Phagocyte Bactericidal Dysfunction	0.982987571	7	1
Abdominal bloating	0.973772506	7	1
Flatulence, eructation, and gas pain	0.973772506	7	1
[D]Gas pain (abdominal)	0.973772506	7	1
Pain of digestive structure	0.973772506	7	1
Metastatic Carcinoma	0.951420065	7	1
Hela Cells	1	3	2
medulloblastoma	1	6	1
Primary malignant neoplasm of thyroid gland	0.999614198	6	1
Papillary thyroid carcinoma	0.999614198	6	1
Primary malignant neoplasm of neck	0.999614198	6	1
Structure of deltoid muscle	0.99928351	6	1
Structure of skeletal muscle of shoulder	0.99928351	6	1
Carcinoma, Transitional Cell	0.998126102	6	1
Transitional Cell Neoplasm	0.998126102	6	1
[M]Transitional cell papilloma or carcinoma NOS	0.998126102	6	1
Upper urinary tract structure	0.998126102	6	1
Upper genitourinary tract structure	0.998126102	6	1
Papillary serous cystadenocarcinoma	0.992504409	6	1
Entire substantia nigra	0.984032596	3	2
Gastrointestinal Hemorrhage	0.968065191	3	2
Maintenance chemotherapy; radiotherapy	0.964891975	6	1
Chemotherapy Regimen	0.964891975	6	1
Upper gastrointestinal disorders	0.944169144	3	2
Neoplasms, Muscle Tissue	0.916969497	3	2
Malignant myomatous tumor	0.916969497	3	2
Superior mediastinum	0.909150975	3	2
Osseous AND/OR chondromatous neoplasm	0.843078956	3	2
Amniotic Fluid	1	5	1
Pneumocyte	0.999867769	5	1

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Macrophages, Alveolar	0.999867769	5	1
Mononuclear phagocyte system	0.999867769	5	1
Colonic Diseases, Functional	0.998942149	5	1
Irritable Bowel Syndrome	0.998942149	5	1
Renal collecting system structure	0.997487603	5	1
Renal pelvis	0.997487603	5	1
Complex epithelial neoplasm	0.926479339	5	1
Hereditary disorder by system	0.685827552	3	2
Cancer of Head	1	4	1
Skin and subcutaneous tissue structure of chest	0.998182419	4	1
Skin structure of breast	0.998182419	4	1
Anterior chest wall structure	0.998182419	4	1
Structure of soft tissues of thorax	0.998182419	4	1
Skin of chest	0.998182419	4	1
Skin structure of nipple	0.998182419	4	1
Skin structure of upper trunk	0.998182419	4	1
Structure of surface region of thorax	0.998182419	4	1
Skin of anterior surface of thorax	0.998182419	4	1
Skin of anterolateral surface of thorax	0.998182419	4	1
Nipple part	0.998182419	4	1
Skin of part of front of thorax	0.998182419	4	1
Skin of part of breast	0.998182419	4	1
Skin of part of thorax	0.998182419	4	1
Skin of part of anterolateral surface of thorax	0.998182419	4	1
Precursor B-cell neoplasm	0.997769332	4	1
Precursor B-cell lymphoblastic leukemia	0.997769332	4	1
Precursor B-lymphoblastic leukemia/lymphoblastic lymphoma	0.997769332	4	1
Other and unspecified gastrointestinal disorders	0.991242564	4	1
Constipation	0.991242564	4	1
Squamous epithelial cell	0.983972241	4	1
Adenocarcinoma of pelvis	0.982650364	4	1
Primary malignant neoplasm of kidney	0.982650364	4	1
Renal glomerular disease	0.982650364	4	1
RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES	0.982650364	4	1
Renal Cell Carcinoma	0.982650364	4	1
Malignant tumor of kidney parenchyma	0.982650364	4	1
Adenosquamous carcinoma	0.943985459	4	1
Neoplasm of cerebrum	0.419337077	3	3
Transitional epithelial cell	0.894828156	4	1
Primary malignant neoplasm of intrathoracic organs	0.593436846	3	2
Primary malignant neoplasm of lung	0.593436846	3	2
Primary malignant neoplasm of respiratory tract	0.593436846	3	2
Tongue part	0.883179114	4	1
Tongue surface region	0.883179114	4	1
Papilla of tongue	0.883179114	4	1
Dorsum of tongue	0.883179114	4	1
Systemic artery of trunk	0.882187707	4	1
Aorta	0.882187707	4	1
Synovial Fluid	1	3	1
Lactobacillales	1	3	1
Streptococcaceae	1	3	1
Streptococcus	1	3	1
Synovial fluid mononuclear cell	1	3	1
ileum	1	3	1
Diffuse low grade B-cell lymphoma	1	3	1
Marginal Zone B-Cell Lymphoma	1	3	1
Catalase-negative Gram-positive coccus	1	3	1
Facultative anaerobic bacteria	1	3	1

Continued on Next Page...

Table C.1 – Continued

Concept	AUC	Num Samples	Num Series
Fastidious bacteria	1	3	1
Gram-Positive Cocci	1	3	1
Fastidious bacterium	1	3	1
Cocci	1	3	1
mucosa-associated lymphoid tissue lymphoma	1	3	1
Rhinovirus infection	0.9994494	3	1
Abnormal coordination	0.9994494	3	1
Dyskinetic syndrome	0.9994494	3	1
Ataxia	0.9994494	3	1
RNA Virus Infections	0.9994494	3	1
Picornaviridae Infections	0.9994494	3	1
Joint and/or tendon synovial structure	0.99867856	3	1
Synovial joint structure	0.99867856	3	1
Structure of synovial tissue of joint	0.99867856	3	1
Rectum and sigmoid colon	0.997467239	3	1
Entire entorhinal cortex	0.996476159	3	1
Hemoptysis	0.980839115	3	1
Respiratory tract hemorrhage	0.980839115	3	1
Myositis	0.971038432	3	1
Polymyositis	0.971038432	3	1
Dermatomyositis	0.971038432	3	1
Rheumatic and Collagen Muscle Diseases and Syndromes	0.971038432	3	1
Dermatomyositis, Childhood Type	0.971038432	3	1
Primary malignant neoplasm of head	0.360973461	3	2

C.2 Concept enrichment of metastasis samples

The top 50 enriched concepts for samples belonging to series containing metastasis samples. For each sample its series ID (GSE), the tissue type, the primary tumor (if applicable), and metastasis site (if applicable) are included. For details on how these concept labels were obtained, see Chapter 4.

Table C.2: Concept enrichment of metastasis samples

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM242028	GSE9576	Liver	Midgut	Liver	Other and unspecified gastrointestinal disorders; Constipation; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Papillary adenocarcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Gastrointestinal Hemorrhage; Metastatic Carcinoma; Hemoptysis; Respiratory tract hemorrhage; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; [M]Adenocarcinoma, metastatic, NOS; Rectum and sigmoid colon; Complex epithelial neoplasm; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Upper gastrointestinal disorders; Adrenal mass; Tumors of Adrenal Cortex; Adrenal Cortex Diseases; Adrenal Gland Diseases; Adrenal Gland Neoplasms; Adrenocortical carcinoma; Urethra; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Endocrine tumor morphology; Transitional epithelial cell; ileum; Stromal Cells; Colonic Diseases, Functional
GSM242029	GSE9576	Liver	Midgut	Liver	Urethra; Endocrine tumor morphology; Benign neoplasm of intra-abdominal organs; Primary malignant neoplasm of male genital organ; Benign neoplasm of adrenal gland; Prostate carcinoma; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Primary malignant neoplasm of prostate; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; ileum; Adenosquamous carcinoma; Colonic Diseases, Functional; Irritable Bowel Syndrome; Structure of medulla of kidney; gastric fundus; Proximal stomach; Gastrointestinal Hemorrhage; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Other and unspecified gastrointestinal disorders; Constipation; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Functional disorder of intestine; Layer of adrenal gland; Endocrine gland part; Adrenal part; Adrenal Cortex; Adrenal mass; Tumors of Adrenal Cortex; Adrenal Cortex Diseases; Adrenal Gland Diseases; Adrenal Gland Neoplasms; Rectum and sigmoid colon; Urinary outflow structure
GSM242030	GSE9576	Liver	Midgut	Liver	Urethra; Endocrine tumor morphology; Benign neoplasm of intra-abdominal organs; Primary malignant neoplasm of male genital organ; Benign neoplasm of adrenal gland; Prostate carcinoma; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; gastric fundus; Primary malignant neoplasm of prostate; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Adenosquamous carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Proximal stomach; Colonic Diseases, Functional; Irritable Bowel Syndrome; Structure of medulla of kidney; Gastrointestinal Hemorrhage; ileum; Complex epithelial neoplasm; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Layer of adrenal gland; Endocrine gland part; Adrenal part; Adrenal Cortex; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Rectum and sigmoid colon; Papillary serous cystadenocarcinoma; Mammary gland; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Adrenal Glands; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Urinary outflow structure

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM242031	GSE9576	Ileum Mucosa	NA	NA	Other and unspecified gastrointestinal disorders; Upper gastrointestinal disorders; Campylobacterales; Urethra; Helicobacter; Colonic Diseases, Functional; Constipation; Helicobacteraceae; ileum; Epsilonproteobacteria; Irritable Bowel Syndrome; Rectum and sigmoid colon; Gastrointestinal Hemorrhage; Functional disorder of intestine; Subclass Aerobic-Microaerophilic, Motile Curved Gram-Negative Bacteria; Adenosquamous carcinoma; Papillary serous cystadenocarcinoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Metastatic Carcinoma; Proteobacteria; Gram-Negative Bacteria; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Complex epithelial neoplasm; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Primary malignant neoplasm of gastrointestinal tract; Disorder of small intestine; Inflammatory Bowel Diseases; Gastritis; Gastroenteritis; Skin tissue; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face
GSM242032	GSE9576	Ileum Mucosa	NA	NA	Other and unspecified gastrointestinal disorders; Upper gastrointestinal disorders; Campylobacterales; Urethra; Helicobacter; Colonic Diseases, Functional; Constipation; Helicobacteraceae; ileum; Epsilonproteobacteria; Irritable Bowel Syndrome; Rectum and sigmoid colon; Gastrointestinal Hemorrhage; Functional disorder of intestine; Subclass Aerobic-Microaerophilic, Motile Curved Gram-Negative Bacteria; Adenosquamous carcinoma; Papillary serous cystadenocarcinoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Metastatic Carcinoma; Complex epithelial neoplasm; Primary malignant neoplasm of gastrointestinal tract; Anorectal structure; Lower bowel structures; Rectum; Pelvic alimentary structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Skin tissue; Disorder of small intestine; Inflammatory Bowel Diseases; Gastritis; Gastroenteritis; Proteobacteria; Gram-Negative Bacteria; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head
GSM242033	GSE9576	Ileum Mucosa	NA	NA	Other and unspecified gastrointestinal disorders; Upper gastrointestinal disorders; Campylobacterales; Urethra; Helicobacter; Colonic Diseases, Functional; Constipation; Helicobacteraceae; ileum; Epsilonproteobacteria; Irritable Bowel Syndrome; Adenosquamous carcinoma; Rectum and sigmoid colon; Gastrointestinal Hemorrhage; Functional disorder of intestine; Subclass Aerobic-Microaerophilic, Motile Curved Gram-Negative Bacteria; Papillary serous cystadenocarcinoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Complex epithelial neoplasm; Metastatic Carcinoma; Anorectal structure; Lower bowel structures; Rectum; Pelvic alimentary structure; Primary malignant neoplasm of gastrointestinal tract; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Skin tissue; Proteobacteria; Gram-Negative Bacteria; Disorder of small intestine; Inflammatory Bowel Diseases; Gastritis; Gastroenteritis; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM242034	GSE9576	Midgut	Midgut	NA	Urethra; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Adenosquamous carcinoma; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Gastrointestinal Hemorrhage; gastric fundus; Proximal stomach; Endocrine tumor morphology; ileum; Urinary outflow structure; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Lower urinary tract; Bladder and outflow structure; Pelvic cavity urinary structure; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Complex epithelial neoplasm; Rectum and sigmoid colon; Papillary serous cystadenocarcinoma; Colonic Diseases, Functional; Irritable Bowel Syndrome; Prostatic and/or seminal vesicle structures; Minor pelvis; Male urinary outflow structure; Prostate and vas deferens structures; Prostate; Male internal genital organ; Pelvic cavity male genital structure; Layer of adrenal gland; Endocrine gland part; Adrenal part
GSM242035	GSE9576	Midgut	Midgut	NA	Urethra; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Adenosquamous carcinoma; Benign neoplasm of retroperitoneum; gastric fundus; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Proximal stomach; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Endocrine tumor morphology; Complex epithelial neoplasm; ileum; Papillary serous cystadenocarcinoma; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Gastrointestinal Hemorrhage; Rectum and sigmoid colon; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Mammary gland; Urinary outflow structure; Layer of adrenal gland; Endocrine gland part; Adrenal part; Adrenal Cortex; Colonic Diseases, Functional; Irritable Bowel Syndrome; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Lower urinary tract
GSM242036	GSE9576	Midgut	Midgut	NA	Urethra; Colonic Diseases, Functional; Benign neoplasm of intra-abdominal organs; ileum; Primary malignant neoplasm of male genital organ; Benign neoplasm of adrenal gland; Prostate carcinoma; Benign neoplasm of adrenal cortex; Irritable Bowel Syndrome; Structure of medulla of kidney; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; gastric fundus; Primary malignant neoplasm of prostate; Gastrointestinal Hemorrhage; Adenosquamous carcinoma; Proximal stomach; Other and unspecified gastrointestinal disorders; Constipation; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Endocrine tumor morphology; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Rectum and sigmoid colon; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Papillary serous cystadenocarcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Functional disorder of intestine; Stomach part; Region of stomach; Urinary outflow structure; Layer of adrenal gland; Endocrine gland part; Adrenal part; Adrenal Cortex; Abdominal bloating

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM242037	GSE9576	Mucosal Layer	NA	NA	Other and unspecified gastrointestinal disorders; Upper gastrointestinal disorders; Urethra; Colonic Diseases, Functional; Constipation; ileum; Irritable Bowel Syndrome; Adenosquamous carcinoma; Rectum and sigmoid colon; Gastrointestinal Hemorrhage; Functional disorder of intestine; Papillary serous cystadenocarcinoma; Campylobacterales; Helicobacter; Helicobacteraceae; Epsilonproteobacteria; Subclass Aerobic-Microaerophilic, Motile Curved Gram-Negative Bacteria; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Skin tissue; Complex epithelial neoplasm; Anorectal structure; Lower bowel structures; Rectum; Pelvic alimentary structure; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Metastatic Carcinoma; Primary malignant neoplasm of gastrointestinal tract; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; gastric fundus; Proximal stomach; Disorder of small intestine; Inflammatory Bowel Diseases; Gastritis; Gastroenteritis; Stomach part; Region of stomach; [M]Adenocarcinoma, metastatic, NOS; Proteobacteria
GSM242038	GSE9576	Mucosal Layer	NA	NA	Other and unspecified gastrointestinal disorders; Upper gastrointestinal disorders; Campylobacterales; Urethra; Helicobacter; Colonic Diseases, Functional; Constipation; Helicobacteraceae; ileum; Epsilonproteobacteria; Irritable Bowel Syndrome; Adenosquamous carcinoma; Rectum and sigmoid colon; Gastrointestinal Hemorrhage; Functional disorder of intestine; Subclass Aerobic-Microaerophilic, Motile Curved Gram-Negative Bacteria; Papillary serous cystadenocarcinoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Anorectal structure; Lower bowel structures; Rectum; Pelvic alimentary structure; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Complex epithelial neoplasm; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Primary malignant neoplasm of gastrointestinal tract; Skin tissue; Stomach part; Region of stomach; Proximal stomach; gastric fundus; Metastatic Carcinoma; Disorder of small intestine; Inflammatory Bowel Diseases; Gastritis; Gastroenteritis; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Proteobacteria; Gram-Negative Bacteria
GSM242039	GSE9576	Mucosal Layer	NA	NA	Other and unspecified gastrointestinal disorders; Upper gastrointestinal disorders; Urethra; Colonic Diseases, Functional; Constipation; ileum; Irritable Bowel Syndrome; Adenosquamous carcinoma; Rectum and sigmoid colon; Gastrointestinal Hemorrhage; Functional disorder of intestine; Campylobacterales; Helicobacter; Helicobacteraceae; Epsilonproteobacteria; Subclass Aerobic-Microaerophilic, Motile Curved Gram-Negative Bacteria; Papillary serous cystadenocarcinoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Anorectal structure; Lower bowel structures; Rectum; Pelvic alimentary structure; Complex epithelial neoplasm; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Skin tissue; Primary malignant neoplasm of gastrointestinal tract; Metastatic Carcinoma; Disorder of small intestine; Inflammatory Bowel Diseases; Gastritis; Gastroenteritis; Proteobacteria; Gram-Negative Bacteria; [M]Adenocarcinoma, metastatic, NOS; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM277231	GSE10961	Liver	Colon	Liver	Other and unspecified gastrointestinal disorders; Urethra; Constipation; Papillary serous cystadenocarcinoma; Hemoptysis; Adenosquamous carcinoma; Rectum and sigmoid colon; Respiratory tract hemorrhage; Metastatic Carcinoma; Functional disorder of intestine; Colonic Diseases, Functional; Complex epithelial neoplasm; Irritable Bowel Syndrome; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Adenocarcinoma, Mucinous; Primary malignant neoplasm of gastrointestinal tract; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Gastrointestinal Hemorrhage; Superior mediastinum; Carcinoma, Transitional Cell; Transitional Cell Neoplasm; [M]Transitional cell papilloma or carcinoma NOS; Upper urinary tract structure
GSM277236	GSE10961	Liver	Colon	Liver	Cholelithiasis; Other and unspecified gastrointestinal disorders; White Adipose Tissue; Colonic Diseases, Functional; Constipation; Cholecystolithiasis; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Calculi; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of adrenal cortex; Irritable Bowel Syndrome; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Biliary calculi; Adenosquamous carcinoma; DISEASES OF THE GALLBLADDER AND BILE DUCTS; Biliary Tract Diseases; Gall Bladder Diseases; Rectum and sigmoid colon; Urethra; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; ileum; Complex epithelial neoplasm; Functional disorder of intestine; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Papillary serous cystadenocarcinoma; Adenocarcinoma, Mucinous; Endocrine tumor morphology; Gastrointestinal Hemorrhage; Anorectal structure; Lower bowel structures; Rectum; Pelvic alimentary structure; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; Metastatic Carcinoma; HYPERALIMENTATION AND OBESITY; Overnutrition; Obesity; Other endocrine/nutritional/metabolic disorder; gastric fundus
GSM277238	GSE10961	Liver	Colon	Liver	Urethra; White Adipose Tissue; Colonic Diseases, Functional; Subcutaneous Fat; Subcutaneous Tissue; Irritable Bowel Syndrome; Adenosquamous carcinoma; Cholelithiasis; Cholecystolithiasis; Calculi; Biliary calculi; Complex epithelial neoplasm; Rectum and sigmoid colon; Proximal stomach; gastric fundus; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; DISEASES OF THE GALLBLADDER AND BILE DUCTS; Biliary Tract Diseases; Gall Bladder Diseases; Papillary serous cystadenocarcinoma; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Metastatic Carcinoma; Adenocarcinoma, Mucinous; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; HYPERALIMENTATION AND OBESITY; Overnutrition; Obesity; Other endocrine/nutritional/metabolic disorder; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Other and unspecified gastrointestinal disorders; Constipation

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM277239	GSE10961	Liver	Colon	Liver	Cholelithiasis; Other and unspecified gastrointestinal disorders; White Adipose Tissue; Colonic Diseases, Functional; Constipation; Cholecystolithiasis; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Calculi; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of adrenal cortex; Irritable Bowel Syndrome; Structure of medulla of kidney; Structure of cortex of kidney; Adrenal Cortical Adenoma; Adenosquamous carcinoma; Benign neoplasm of retroperitoneum; Structure of layer of kidney; Biliary calculi; Rectum and sigmoid colon; DISEASES OF THE GALLBLADDER AND BILE DUCTS; Biliary Tract Diseases; Gall Bladder Diseases; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Urethra; Gastrointestinal Hemorrhage; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; Complex epithelial neoplasm; Anorectal structure; Lower bowel structures; Rectum; Pelvic alimentary structure; Functional disorder of intestine; Papillary serous cystadenocarcinoma; ileum; Adenocarcinoma, Mucinous; Endocrine tumor morphology; Proximal stomach; gastric fundus; Upper gastrointestinal disorders
GSM277246	GSE10961	Liver	Colon	Liver	Other and unspecified gastrointestinal disorders; Constipation; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Adenosquamous carcinoma; Rectum and sigmoid colon; [M]Adenocarcinoma, metastatic, NOS; Papillary adenocarcinoma; Hemoptysis; Respiratory tract hemorrhage; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Functional disorder of intestine; Colonic Diseases, Functional; Complex epithelial neoplasm; Irritable Bowel Syndrome; Urethra; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Transitional epithelial cell; Carcinoma, Transitional Cell; Transitional Cell Neoplasm; [M]Transitional cell papilloma or carcinoma NOS; Upper urinary tract structure; Upper genitourinary tract structure; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Renal collecting system structure; Renal pelvis; ileum; Primary malignant neoplasm of gastrointestinal tract; Neoplasm Metastasis; Neoplastic Processes; Adenocarcinoma, Mucinous; Gastrointestinal Hemorrhage; Benign neoplasm of intra-abdominal organs
GSM277248	GSE10961	Liver	Colon	Liver	Other and unspecified gastrointestinal disorders; Urethra; Constipation; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Rectum and sigmoid colon; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Transitional epithelial cell; Functional disorder of intestine; Colonic Diseases, Functional; Irritable Bowel Syndrome; Papillary adenocarcinoma; Complex epithelial neoplasm; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Hemoptysis; Respiratory tract hemorrhage; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Carcinoma, Transitional Cell; Transitional Cell Neoplasm; [M]Transitional cell papilloma or carcinoma NOS; Upper urinary tract structure; Upper genitourinary tract structure; Renal collecting system structure; Renal pelvis; Anorectal structure; Lower bowel structures; Rectum; Pelvic alimentary structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Primary malignant neoplasm of gastrointestinal tract; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal)

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM277253	GSE10961	Liver	Colon	Liver	Transitional epithelial cell; Other and unspecified gastrointestinal disorders; Constipation; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Rectum and sigmoid colon; Papillary adenocarcinoma; Colonic Diseases, Functional; Irritable Bowel Syndrome; Functional disorder of intestine; Hemoptysis; Respiratory tract hemorrhage; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Gastrointestinal Hemorrhage; ileum; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Metastatic Carcinoma; Primary malignant neoplasm of gastrointestinal tract; Upper gastrointestinal disorders; Complex epithelial neoplasm; [M]Adenocarcinoma, metastatic, NOS; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Urethra; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Structure of pyloric portion of stomach; Part of pyloric region
GSM277256	GSE10961	Liver	Colon	Liver	Other and unspecified gastrointestinal disorders; Constipation; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Adenosquamous carcinoma; Rectum and sigmoid colon; Hemoptysis; Respiratory tract hemorrhage; Complex epithelial neoplasm; [M]Adenocarcinoma, metastatic, NOS; Functional disorder of intestine; Colonic Diseases, Functional; Irritable Bowel Syndrome; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Urethra; Papillary adenocarcinoma; Neoplasm Metastasis; Neoplastic Processes; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Transitional epithelial cell; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Carcinoma, Transitional Cell; Transitional Cell Neoplasm; [M]Transitional cell papilloma or carcinoma NOS; Upper urinary tract structure; Upper genitourinary tract structure; Renal collecting system structure; Renal pelvis; Adenocarcinoma, Mucinous; Primary malignant neoplasm of gastrointestinal tract; ileum; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal)
GSM277466	GSE10961	Liver	Colon	Liver	Other and unspecified gastrointestinal disorders; Anorectal structure; Urethra; Constipation; Lower bowel structures; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Rectum; Pelvic alimentary structure; Rectum and sigmoid colon; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Adenocarcinoma, Mucinous; Hemoptysis; Respiratory tract hemorrhage; Papillary adenocarcinoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Functional disorder of intestine; Colonic Diseases, Functional; Irritable Bowel Syndrome; Skin tissue; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Primary malignant neoplasm of gastrointestinal tract; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Abdominal bloating

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM277469	GSE10961	Liver	Colon	Liver	Other and unspecified gastrointestinal disorders; Colonic Diseases, Functional; Constipation; Papillary serous cystadenocarcinoma; Irritable Bowel Syndrome; Adenosquamous carcinoma; Rectum and sigmoid colon; Functional disorder of intestine; Metastatic Carcinoma; Hemoptysis; Respiratory tract hemorrhage; [M]Adenocarcinoma, metastatic, NOS; Complex epithelial neoplasm; Gastrointestinal Hemorrhage; Papillary adenocarcinoma; Transitional epithelial cell; ileum; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Adenocarcinoma, Mucinous; Primary malignant neoplasm of gastrointestinal tract; Exanthema; Disorder of keratinization; Cell-mediated cytotoxic disorder; Cutaneous hypersensitivity; Acquired disorder of keratinization; Histologic type of inflammatory skin disorder; Psoriasis; Other psoriasis; Skin Diseases, Papulosquamous; Inflammatory hyperkeratotic dermatosis; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Urethra; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Carcinoma, Transitional Cell
GSM277477	GSE10961	Liver	Colon	Liver	Other and unspecified gastrointestinal disorders; Constipation; Papillary serous cystadenocarcinoma; Hemoptysis; Adenosquamous carcinoma; Rectum and sigmoid colon; Respiratory tract hemorrhage; Complex epithelial neoplasm; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Exanthema; Disorder of keratinization; Cell-mediated cytotoxic disorder; Cutaneous hypersensitivity; Acquired disorder of keratinization; Histologic type of inflammatory skin disorder; Psoriasis; Other psoriasis; Skin Diseases, Papulosquamous; Inflammatory hyperkeratotic dermatosis; Papillary adenocarcinoma; Squamous epithelial cell; Metastatic Carcinoma; Transitional epithelial cell; [M]Adenocarcinoma, metastatic, NOS; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Stromal Cells; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Urethra; White Adipose Tissue
GSM277478	GSE10961	Liver	Colon	Liver	Other and unspecified gastrointestinal disorders; Constipation; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Adenosquamous carcinoma; Rectum and sigmoid colon; Complex epithelial neoplasm; Hemoptysis; Respiratory tract hemorrhage; Colonic Diseases, Functional; Irritable Bowel Syndrome; Functional disorder of intestine; Urethra; [M]Adenocarcinoma, metastatic, NOS; Adenocarcinoma, Mucinous; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Papillary adenocarcinoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Gastrointestinal Hemorrhage; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Primary malignant neoplasm of gastrointestinal tract; ileum; Anorectal structure; Lower bowel structures; Rectum; Pelvic alimentary structure; Carcinoma, Transitional Cell; Transitional Cell Neoplasm; [M]Transitional cell papilloma or carcinoma NOS; Upper urinary tract structure; Upper genitourinary tract structure; Transitional epithelial cell; Renal collecting system structure; Renal pelvis

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM277479	GSE10961	Liver	Colon	Liver	Other and unspecified gastrointestinal disorders; Colonic Diseases, Functional; Constipation; Papillary serous cystadenocarcinoma; Irritable Bowel Syndrome; Adenosquamous carcinoma; Rectum and sigmoid colon; Functional disorder of intestine; Gastrointestinal Hemorrhage; Hemoptysis; Respiratory tract hemorrhage; Metastatic Carcinoma; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Complex epithelial neoplasm; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; ileum; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Primary malignant neoplasm of gastrointestinal tract; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Upper gastrointestinal disorders; Urethra; Transitional epithelial cell; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Exanthema; Disorder of keratinization; Cell-mediated cytotoxic disorder; Cutaneous hypersensitivity; Acquired disorder of keratinization; Histologic type of inflammatory skin disorder; Psoriasis; Other psoriasis; Skin Diseases, Papulosquamous
GSM277481	GSE10961	Liver	Colon	Liver	Urethra; Adenosquamous carcinoma; Rectum and sigmoid colon; Papillary serous cystadenocarcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Metastatic Carcinoma; Complex epithelial neoplasm; Other and unspecified gastrointestinal disorders; Constipation; Adenocarcinoma, Mucinous; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Skin tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Anorectal structure; Lower bowel structures; Rectum; Pelvic alimentary structure; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; [M]Adenocarcinoma, metastatic, NOS; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Hemoptysis; Respiratory tract hemorrhage; Stromal Cells; Mesenchymal Stem Cells; gastric fundus; Proximal stomach
GSM277494	GSE10961	Liver	Colon	Liver	Other and unspecified gastrointestinal disorders; Urethra; Constipation; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Adenosquamous carcinoma; Rectum and sigmoid colon; [M]Adenocarcinoma, metastatic, NOS; Colonic Diseases, Functional; Irritable Bowel Syndrome; Functional disorder of intestine; Anorectal structure; Lower bowel structures; Rectum; Pelvic alimentary structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Adenocarcinoma, Mucinous; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Hemoptysis; Respiratory tract hemorrhage; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Neoplasm Metastasis; Neoplastic Processes; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Skin tissue; Primary malignant neoplasm of gastrointestinal tract; ileum; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw

Continued on Next Page…

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM277646	GSE10961	Liver	Colon	Liver	Cholelithiasis; Urethra; White Adipose Tissue; Colonic Diseases, Functional; Cholecystolithiasis; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Calculi; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of adrenal cortex; Irritable Bowel Syndrome; Adrenal Cortical Adenoma; Adenosquamous carcinoma; Benign neoplasm of retroperitoneum; Biliary calculi; DISEASES OF THE GALLBLADDER AND BILE DUCTS; Biliary Tract Diseases; Gall Bladder Diseases; Other and unspecified gastrointestinal disorders; Constipation; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Rectum and sigmoid colon; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Functional disorder of intestine; ileum; Proximal stomach; gastric fundus; Gastrointestinal Hemorrhage; Adenocarcinoma, Mucinous; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; HYPERALIMENTATION AND OBESITY; Overnutrition; Obesity; Other endocrine/nutritional/metabolic disorder; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Endocrine tumor morphology; Joint and/or tendon synovial structure
GSM277647	GSE10961	Liver	Colon	Liver	Urethra; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Adenosquamous carcinoma; Metastatic Carcinoma; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Adenocarcinoma, Mucinous; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Rectum and sigmoid colon; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Hemoptysis; Respiratory tract hemorrhage; Skin tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; gastric fundus; Proximal stomach; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face
GSM277648	GSE10961	Liver	Colon	Liver	Other and unspecified gastrointestinal disorders; Urethra; Colonic Diseases, Functional; Constipation; Irritable Bowel Syndrome; Adenosquamous carcinoma; Rectum and sigmoid colon; Functional disorder of intestine; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Adenocarcinoma, Mucinous; Anorectal structure; Lower bowel structures; Rectum; Pelvic alimentary structure; Complex epithelial neoplasm; [M]Adenocarcinoma, metastatic, NOS; ileum; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Skin tissue; Primary malignant neoplasm of gastrointestinal tract; Gastrointestinal Hemorrhage; Hemoptysis; Respiratory tract hemorrhage; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM352095	GSE14017	Lung	Breast	Lung	Urethra; Papillary serous cystadenocarcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Papillary adenocarcinoma; Complex epithelial neoplasm; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Metastatic Carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; [M]Adenocarcinoma, metastatic, NOS; Mammary gland; Ductal Carcinoma; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Skin tissue; Proximal stomach; gastric fundus; Rectum and sigmoid colon; Structure of pyloric portion of stomach
GSM352097	GSE14017	Brain	Breast	Brain	Other and unspecified gastrointestinal disorders; Urethra; Constipation; Papillary serous cystadenocarcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Papillary adenocarcinoma; Gastrointestinal Hemorrhage; Rectum and sigmoid colon; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Proximal stomach; gastric fundus; Complex epithelial neoplasm; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Upper gastrointestinal disorders; Neoplasm Metastasis; Neoplastic Processes; Stomach part; Region of stomach; Superior mediastinum; Ductal Carcinoma; Transitional epithelial cell; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue
GSM352098	GSE14017	Brain	Breast	Brain	Urethra; Papillary serous cystadenocarcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Papillary adenocarcinoma; Metastatic Carcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Complex epithelial neoplasm; Transitional epithelial cell; [M]Adenocarcinoma, metastatic, NOS; Superior mediastinum; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Other and unspecified gastrointestinal disorders; Constipation; Ductal Carcinoma; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY; Disorder of oral soft tissues; gastric fundus; Proximal stomach; Stomach part; Region of stomach; Rectum and sigmoid colon; Lactiferous duct

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM352100	GSE14017	Bone	Breast	Bone	Urethra; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Complex epithelial neoplasm; Lactiferous duct; Mammary lobe; Glandular structure of breast; Adenosquamous carcinoma; Mammary gland; Duct (organ) structure; Papillary serous cystadenocarcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; gastric fundus; Proximal stomach; Metastatic Carcinoma; Vagina; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Rectum and sigmoid colon; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Skin tissue; Adenocarcinoma, Mucinous; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Breast part; Stomach part; Region of stomach; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY; Disorder of oral soft tissues; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland
GSM352101	GSE14017	Brain	Breast	Brain	Transitional epithelial cell; Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Papillary adenocarcinoma; Hemoptysis; Respiratory tract hemorrhage; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Metastatic Carcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Superior mediastinum; Other and unspecified gastrointestinal disorders; Constipation; Ductal Carcinoma; Complex epithelial neoplasm; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Proximal stomach; gastric fundus; [M]Adenocarcinoma, metastatic, NOS; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Stomach part; Region of stomach; Disorder of soft tissue of body cavity; Disorder of soft tissue of head
GSM352103	GSE14017	Bone	Breast	Bone	Urethra; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Leiomyomatous neoplasm - category; mucosa-associated lymphoid tissue lymphoma; Uterine Fibroids; Benign myomatous tumor; Benign neoplasm of female genital organ, site unspecified; Benign neoplasm of body of uterus; Benign neoplasm of uterus NOS; Benign leiomyomatous neoplasm - category; Benign genital neoplasm; Benign neoplasm corpus uteri NEC; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Stromal Cells; gastric fundus; Proximal stomach; Mammary gland; Mesenchymal Stem Cells; Hemoptysis; Respiratory tract hemorrhage; Complex epithelial neoplasm; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Lymph; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; [M]Adenocarcinoma, metastatic, NOS; Superior mediastinum; Stomach part

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM352105	GSE14017	Bone	Breast	Bone	Urethra; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; mucosa-associated lymphoid tissue lymphoma; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Stromal Cells; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Mesenchymal Stem Cells; Metastatic Carcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Superior mediastinum; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Lymph; Ductal Carcinoma; Rectum and sigmoid colon; Mammary gland; gastric fundus
GSM352107	GSE14017	Brain	Breast	Brain	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Papillary adenocarcinoma; Complex epithelial neoplasm; Superior mediastinum; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Rectum and sigmoid colon; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Ductal Carcinoma; Neoplasm Metastasis; Neoplastic Processes; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Disorder of soft tissue of body cavity; Disorder of soft tissue of head
GSM352109	GSE14017	Bone	Breast	Bone	Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Metastatic Carcinoma; Complex epithelial neoplasm; Mammary gland; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; gastric fundus; Proximal stomach; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Rectum and sigmoid colon; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Hemoptysis; Respiratory tract hemorrhage; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Stomach part; Region of stomach; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM352110	GSE14017	Brain	Breast	Brain	Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Rectum and sigmoid colon; Hemoptysis; Respiratory tract hemorrhage; Papillary adenocarcinoma; Complex epithelial neoplasm; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY; Disorder of oral soft tissues; Exanthema; Disorder of keratinization; Cell-mediated cytotoxic disorder; Cutaneous hypersensitivity; Acquired disorder of keratinization; Histologic type of inflammatory skin disorder; Psoriasis; Other psoriasis; Skin Diseases, Papulosquamous; Inflammatory hyperkeratotic dermatosis; Skin tissue; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Esophagus; Other and unspecified gastrointestinal disorders
GSM352111	GSE14017	Brain	Breast	Brain	Transitional epithelial cell; Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Metastatic Carcinoma; Hemoptysis; Respiratory tract hemorrhage; Ductal Carcinoma; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Complex epithelial neoplasm; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Other and unspecified gastrointestinal disorders; Constipation
GSM352113	GSE14017	Brain	Breast	Brain	Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; [M]Adenocarcinoma, metastatic, NOS; Hemoptysis; Respiratory tract hemorrhage; Papillary adenocarcinoma; Rectum and sigmoid colon; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Metastatic Carcinoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Complex epithelial neoplasm; Ductal Carcinoma; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Mammary gland; Breast part; Superior mediastinum; Transitional epithelial cell; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Vagina; gastric fundus; Proximal stomach; White Adipose Tissue; Subcutaneous Fat

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM352114	GSE14017	Lung	Breast	Lung	Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Metastatic Carcinoma; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Papillary adenocarcinoma; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Complex epithelial neoplasm; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; [M]Adenocarcinoma, metastatic, NOS; Hemoptysis; Respiratory tract hemorrhage; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Proximal stomach; gastric fundus; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Rectum and sigmoid colon; Stomach part; Region of stomach; Mammary gland; Gastrointestinal Hemorrhage; Adenocarcinoma, Mucinous; Transitional epithelial cell; Breast part; Ductal Carcinoma
GSM352115	GSE14017	Brain	Breast	Brain	Urethra; Papillary serous cystadenocarcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Metastatic Carcinoma; Complex epithelial neoplasm; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Ductal Carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Rectum and sigmoid colon; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Stromal Cells; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Coughing; Other and unspecified gastrointestinal disorders; Constipation; Neoplasms, Ductal, Lobular, and Medullary
GSM352117	GSE14017	Bone	Breast	Bone	Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Complex epithelial neoplasm; Metastatic Carcinoma; Hemoptysis; Respiratory tract hemorrhage; Rectum and sigmoid colon; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Ductal Carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Mammary gland; Papillary adenocarcinoma; Proximal stomach; gastric fundus; Stomach part; Region of stomach; [M]Adenocarcinoma, metastatic, NOS; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY; Disorder of oral soft tissues; Vagina; Entire viscus; Hollow viscus

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM352119	GSE14017	Bone	Breast	Bone	Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Metastatic Carcinoma; Papillary adenocarcinoma; Hemoptysis; Respiratory tract hemorrhage; Transitional epithelial cell; [M]Adenocarcinoma, metastatic, NOS; Rectum and sigmoid colon; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Complex epithelial neoplasm; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Other and unspecified gastrointestinal disorders; Constipation; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Ductal Carcinoma; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Proximal stomach; gastric fundus; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Stomach part; Region of stomach
GSM352120	GSE14017	Brain	Breast	Brain	Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Stromal Cells; [M]Adenocarcinoma, metastatic, NOS; Hemoptysis; Respiratory tract hemorrhage; Superior mediastinum; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Papillary adenocarcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Mesenchymal Stem Cells; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Complex epithelial neoplasm; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Vagina; Ductal Carcinoma; Transitional epithelial cell; Urinary outflow structure; Lower urinary tract; Bladder and outflow structure; Pelvic cavity urinary structure; Metastatic Carcinoma; Prostatic and/or seminal vesicle structures; Minor pelvis; Male urinary outflow structure; Prostate and vas deferens structures; Prostate
GSM352121	GSE14017	Brain	Breast	Brain	Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Metastatic Carcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Hemoptysis; Respiratory tract hemorrhage; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; [M]Adenocarcinoma, metastatic, NOS; Papillary adenocarcinoma; Rectum and sigmoid colon; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Complex epithelial neoplasm; Superior mediastinum; Ductal Carcinoma; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Proximal stomach; gastric fundus; Stomach part; Region of stomach; Mammary gland; Neoplasm Metastasis; Neoplastic Processes; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM352122	GSE14017	Brain	Breast	Brain	Papillary serous cystadenocarcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Urethra; Metastatic Carcinoma; Papillary adenocarcinoma; Complex epithelial neoplasm; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; [M]Adenocarcinoma, metastatic, NOS; Rectum and sigmoid colon; Superior mediastinum; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY; Disorder of oral soft tissues; Other and unspecified gastrointestinal disorders; Constipation; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Exanthema; Disorder of keratinization
GSM352123	GSE14017	Bone	Breast	Bone	Urethra; Proximal stomach; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Mammary gland; gastric fundus; Metastatic Carcinoma; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Stomach part; Region of stomach; Complex epithelial neoplasm; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; [M]Adenocarcinoma, metastatic, NOS; Hemoptysis; Respiratory tract hemorrhage; Papillary adenocarcinoma; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Rectum and sigmoid colon; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure
GSM352124	GSE14017	Bone	Breast	Bone	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Lactiferous duct; Mammary lobe; Glandular structure of breast; Adenosquamous carcinoma; Duct (organ) structure; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Metastatic Carcinoma; Mammary gland; gastric fundus; Proximal stomach; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Hemoptysis; Respiratory tract hemorrhage; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Rectum and sigmoid colon; Adenocarcinoma, Mucinous; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Stomach part; Region of stomach; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Stromal Cells; Vagina; Ductal Carcinoma

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM352125	GSE14017	Brain	Breast	Brain	Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Hemoptysis; Respiratory tract hemorrhage; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Superior mediastinum; Complex epithelial neoplasm; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Stromal Cells; Transitional epithelial cell; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Rectum and sigmoid colon; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Ductal Carcinoma; Pluripotent Stem Cells
GSM352126	GSE14017	Bone	Breast	Bone	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Stromal Cells; Metastatic Carcinoma; Rectum and sigmoid colon; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Mesenchymal Stem Cells; Mammary gland; Proximal stomach; gastric fundus; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Hemoptysis; Respiratory tract hemorrhage; Vagina; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Skin tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Adenocarcinoma, Mucinous
GSM352127	GSE14017	Lung	Breast	Lung	Urethra; Diffuse low grade B-cell lymphoma; Stomach part; Proximal stomach; Marginal Zone B-Cell Lymphoma; Structure of pyloric portion of stomach; Part of pyloric region; Region of stomach; Pylorus; Adenosquamous carcinoma; gastric fundus; mucosa-associated lymphoid tissue lymphoma; Papillary serous cystadenocarcinoma; Rectum and sigmoid colon; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Complex epithelial neoplasm; Adenocarcinoma, Mucinous; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Campylobacterales; Helicobacter; Helicobacteraceae; Epsilonproteobacteria; Subclass Aerobic-Microaerophilic, Motile Curved Gram-Negative Bacteria; Mammary gland; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Gastrointestinal Hemorrhage; Upper gastrointestinal disorders; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Other and unspecified gastrointestinal disorders; Constipation; Metastatic Carcinoma; Disorder of small intestine; Inflammatory Bowel Diseases; Gastritis

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM352128	GSE14017	Brain	Breast	Brain	HCT116 Cells; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Colonic epithelium; Colonic mucous membrane; Structure of intestinal epithelium; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Hemoptysis; Respiratory tract hemorrhage; Transitional epithelial cell; Pluripotent Stem Cells; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Urethra; Superior mediastinum; Gastrointestinal Hemorrhage; Complex epithelial neoplasm; Stromal Cells; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Metastatic Carcinoma; Other and unspecified gastrointestinal disorders; Constipation; Ductal Carcinoma; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Upper gastrointestinal disorders; Rectum and sigmoid colon; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Structure of pyloric portion of stomach; Part of pyloric region
GSM352129	GSE14017	Brain	Breast	Brain	Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Hemoptysis; Respiratory tract hemorrhage; Metastatic Carcinoma; Papillary adenocarcinoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Gastrointestinal Hemorrhage; Transitional epithelial cell; Other and unspecified gastrointestinal disorders; Constipation; Rectum and sigmoid colon; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; [M]Adenocarcinoma, metastatic, NOS; Complex epithelial neoplasm; Ductal Carcinoma; Upper gastrointestinal disorders; Superior mediastinum; gastric fundus; Proximal stomach; Stomach part; Region of stomach; Esophagus; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Campylobacterales; Helicobacter; Helicobacteraceae
GSM352130	GSE14017	Brain	Breast	Brain	Urethra; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; mucosa-associated lymphoid tissue lymphoma; Metastatic Carcinoma; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Hemoptysis; Respiratory tract hemorrhage; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; gastric fundus; Proximal stomach; Superior mediastinum; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Stomach part; Region of stomach; Transitional epithelial cell; Other and unspecified gastrointestinal disorders; Constipation; Complex epithelial neoplasm; Ductal Carcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Rectum and sigmoid colon; Abdominal bloating

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM352131	GSE14017	Bone	Breast	Bone	Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Metastatic Carcinoma; Rectum and sigmoid colon; [M]Adenocarcinoma, metastatic, NOS; Papillary adenocarcinoma; Hemoptysis; Respiratory tract hemorrhage; Complex epithelial neoplasm; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Skin tissue; Other and unspecified gastrointestinal disorders; Constipation; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Gastrointestinal Hemorrhage; Transitional epithelial cell; Mammary gland; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; gastric fundus; Proximal stomach
GSM352132	GSE14017	Lung	Breast	Lung	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Metastatic Carcinoma; Rectum and sigmoid colon; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Papillary adenocarcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY; Disorder of oral soft tissues; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; [M]Adenocarcinoma, metastatic, NOS; Skin tissue; Squamous epithelial cell; Exanthema; Disorder of keratinization; Cell-mediated cytotoxic disorder; Cutaneous hypersensitivity; Acquired disorder of keratinization; Histologic type of inflammatory skin disorder; Psoriasis; Other psoriasis
GSM354034	GSE14108	Brain	Lung	Brain	Other and unspecified gastrointestinal disorders; Urethra; Constipation; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; Papillary serous cystadenocarcinoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Adenosquamous carcinoma; mucosa-associated lymphoid tissue lymphoma; Metastatic Carcinoma; Rectum and sigmoid colon; Hemoptysis; Respiratory tract hemorrhage; Complex epithelial neoplasm; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Gastrointestinal Hemorrhage; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Proximal stomach; gastric fundus; Stomach part; Region of stomach; Superior mediastinum; Transitional epithelial cell; Carcinoma, Transitional Cell; Transitional Cell Neoplasm; [M]Transitional cell papilloma or carcinoma NOS; Upper urinary tract structure; Upper genitourinary tract structure; Renal collecting system structure; Renal pelvis; Campylobacterales; Helicobacter; Helicobacteraceae; Epsilonproteobacteria; Subclass Aerobic-Microaerophilic, Motile Curved Gram-Negative Bacteria; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM354035	GSE14108	Brain	Lung	Brain	Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Adenosquamous carcinoma; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Urethra; Rectum and sigmoid colon; Other and unspecified gastrointestinal disorders; Constipation; Hemoptysis; Respiratory tract hemorrhage; Papillary adenocarcinoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; [M]Adenocarcinoma, metastatic, NOS; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Squamous epithelial cell; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY; Disorder of oral soft tissues; Primary malignant neoplasm of intrathoracic organs; Primary malignant neoplasm of lung; Primary malignant neoplasm of respiratory tract; Transitional epithelial cell; Esophagus; Adenocarcinoma, Mucinous; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ
GSM354036	GSE14108	Brain	Lung	Brain	Other and unspecified gastrointestinal disorders; Urethra; Constipation; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Papillary adenocarcinoma; Metastatic Carcinoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; Rectum and sigmoid colon; mucosa-associated lymphoid tissue lymphoma; [M]Adenocarcinoma, metastatic, NOS; Hemoptysis; Respiratory tract hemorrhage; Esophagus; Transitional epithelial cell; Gastrointestinal Hemorrhage; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Complex epithelial neoplasm; Superior mediastinum; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Proximal stomach; gastric fundus; Stomach part; Region of stomach; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY; Disorder of oral soft tissues
GSM354037	GSE14108	Brain	Lung	Brain	Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Papillary adenocarcinoma; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Other and unspecified gastrointestinal disorders; Constipation; Hemoptysis; Respiratory tract hemorrhage; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Gastrointestinal Hemorrhage; [M]Adenocarcinoma, metastatic, NOS; Metastatic Carcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Transitional epithelial cell; Sense Organs; Nose; Rhinovirus infection; RNA Virus Infections; Picornaviridae Infections; Complex epithelial neoplasm; Rectum and sigmoid colon; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Campylobacterales; Helicobacter; Helicobacteraceae; Epsilonproteobacteria; Subclass Aerobic-Microaerophilic, Motile Curved Gram-Negative Bacteria; Upper gastrointestinal disorders

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM354038	GSE14108	Brain	Lung	Brain	Other and unspecified gastrointestinal disorders; Constipation; Papillary serous cystadenocarcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Papillary adenocarcinoma; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Rectum and sigmoid colon; Gastrointestinal Hemorrhage; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Complex epithelial neoplasm; Transitional epithelial cell; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Urethra; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Upper gastrointestinal disorders; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Campylobacterales; Helicobacter; Helicobacteraceae; Epsilonproteobacteria; Subclass Aerobic-Microaerophilic, Motile Curved Gram-Negative Bacteria; Primary malignant neoplasm of gastrointestinal tract
GSM354039	GSE14108	Brain	Lung	Brain	HCT116 Cells; Papillary serous cystadenocarcinoma; Hemoptysis; Adenosquamous carcinoma; Colonic epithelium; Colonic mucous membrane; Respiratory tract hemorrhage; Structure of intestinal epithelium; Transitional epithelial cell; Papillary adenocarcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Stromal Cells; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Complex epithelial neoplasm; Metastatic Carcinoma; Squamous epithelial cell; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; [M]Adenocarcinoma, metastatic, NOS; Urethra; Superior mediastinum; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY; Disorder of oral soft tissues; Other and unspecified gastrointestinal disorders; Constipation; Amniotic Fluid
GSM354040	GSE14108	Brain	Lung	Brain	Urethra; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Adenosquamous carcinoma; mucosa-associated lymphoid tissue lymphoma; Complex epithelial neoplasm; Papillary adenocarcinoma; Other and unspecified gastrointestinal disorders; Constipation; Rectum and sigmoid colon; Hemoptysis; Respiratory tract hemorrhage; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; [M]Adenocarcinoma, metastatic, NOS; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY; Disorder of oral soft tissues; Adenocarcinoma, Mucinous; Neoplasm Metastasis; Neoplastic Processes; Proximal stomach

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM354041	GSE14108	Brain	Lung	Brain	Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Metastatic Carcinoma; Papillary adenocarcinoma; Other and unspecified gastrointestinal disorders; Constipation; Hemoptysis; Respiratory tract hemorrhage; Complex epithelial neoplasm; Rectum and sigmoid colon; [M]Adenocarcinoma, metastatic, NOS; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Gastrointestinal Hemorrhage; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; gastric fundus; Proximal stomach; Lactiferous duct; Mammary lobe
GSM354042	GSE14108	Brain	Lung	Brain	Papillary serous cystadenocarcinoma; Hemoptysis; Adenosquamous carcinoma; Stromal Cells; Respiratory tract hemorrhage; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Urethra; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Mesenchymal Stem Cells; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Papillary adenocarcinoma; Complex epithelial neoplasm; HCT116 Cells; Colonic epithelium; Colonic mucous membrane; Structure of intestinal epithelium; Superior mediastinum; [M]Adenocarcinoma, metastatic, NOS; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Neoplasms, Muscle Tissue; Malignant myomatous tumor; Structure of bone (organ); Type of bone; Bone structure of spine and/or pelvis; hip bone; Bone structure of ilium; Bone part; Ilium part; Iliac crest structure; Structure of flat bone; Bone structure of pelvic region and/or thigh; Bony pelvis; Metastatic Carcinoma; Pluripotent Stem Cells
GSM354043	GSE14108	Brain	Lung	Brain	Urethra; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Adenosquamous carcinoma; mucosa-associated lymphoid tissue lymphoma; Papillary adenocarcinoma; Other and unspecified gastrointestinal disorders; Constipation; [M]Adenocarcinoma, metastatic, NOS; Rectum and sigmoid colon; Complex epithelial neoplasm; Hemoptysis; Respiratory tract hemorrhage; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Superior mediastinum; Squamous epithelial cell; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY; Disorder of oral soft tissues; Neoplasm Metastasis; Neoplastic Processes; Proximal stomach; gastric fundus; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Stomach part; Region of stomach

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM354044	GSE14108	Brain	Lung	Brain	Other and unspecified gastrointestinal disorders; Urethra; Constipation; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; Complex epithelial neoplasm; Adenosquamous carcinoma; mucosa-associated lymphoid tissue lymphoma; Papillary serous cystadenocarcinoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Rectum and sigmoid colon; Superior mediastinum; Adenocarcinoma, Mucinous; Metastatic Carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Hemoptysis; Respiratory tract hemorrhage; Papillary adenocarcinoma; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Proximal stomach; gastric fundus; Stomach part; Region of stomach; Gastrointestinal Hemorrhage; Spleen; [M]Adenocarcinoma, metastatic, NOS; Ductal Carcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ
GSM354045	GSE14108	Brain	Lung	Brain	Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Other and unspecified gastrointestinal disorders; Constipation; [M]Adenocarcinoma, metastatic, NOS; Gastrointestinal Hemorrhage; Papillary adenocarcinoma; Rectum and sigmoid colon; Metastatic Carcinoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Transitional epithelial cell; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Complex epithelial neoplasm; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Upper gastrointestinal disorders; Hemoptysis; Respiratory tract hemorrhage; Proximal stomach; gastric fundus
GSM354046	GSE14108	Brain	Lung	Brain	Papillary serous cystadenocarcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Papillary adenocarcinoma; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Complex epithelial neoplasm; Urethra; Rectum and sigmoid colon; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY; Disorder of oral soft tissues; Squamous epithelial cell; Stromal Cells; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Other and unspecified gastrointestinal disorders; Constipation; Transitional epithelial cell; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM354047	GSE14108	Brain	Lung	Brain	Urethra; Diffuse low grade B-cell lymphoma; Stomach part; Proximal stomach; Marginal Zone B-Cell Lymphoma; Papillary serous cystadenocarcinoma; Structure of pyloric portion of stomach; Part of pyloric region; Region of stomach; Pylorus; Adenosquamous carcinoma; gastric fundus; mucosa-associated lymphoid tissue lymphoma; Metastatic Carcinoma; Rectum and sigmoid colon; [M]Adenocarcinoma, metastatic, NOS; Papillary adenocarcinoma; Other and unspecified gastrointestinal disorders; Constipation; Hemoptysis; Respiratory tract hemorrhage; Complex epithelial neoplasm; Superior mediastinum; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Neoplasm Metastasis; Neoplastic Processes; Lymph; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Adenocarcinoma, Mucinous; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Gingival and periodontal disease NOS; Jaw Diseases
GSM354048	GSE14108	Brain	Lung	Brain	Papillary serous cystadenocarcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Metastatic Carcinoma; Papillary adenocarcinoma; Urethra; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; [M]Adenocarcinoma, metastatic, NOS; Complex epithelial neoplasm; Rectum and sigmoid colon; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Transitional epithelial cell; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Esophagus; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY; Disorder of oral soft tissues; Other and unspecified gastrointestinal disorders; Constipation; Superior mediastinum; Squamous epithelial cell; HCT116 Cells; Colonic epithelium; Colonic mucous membrane; Structure of intestinal epithelium
GSM354049	GSE14108	Brain	Lung	Brain	Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Rectum and sigmoid colon; [M]Adenocarcinoma, metastatic, NOS; Papillary adenocarcinoma; Complex epithelial neoplasm; Squamous epithelial cell; Urethra; Transitional epithelial cell; Other and unspecified gastrointestinal disorders; Constipation; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY; Disorder of oral soft tissues; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Stromal Cells; Neoplasm Metastasis; Neoplastic Processes; Esophagus; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Primary malignant neoplasm of intrathoracic organs; Primary malignant neoplasm of lung

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM354050	GSE14108	Brain	Lung	Brain	Other and unspecified gastrointestinal disorders; Constipation; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Gastrointestinal Hemorrhage; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Upper gastrointestinal disorders; Metastatic Carcinoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Hemoptysis; Respiratory tract hemorrhage; Rectum and sigmoid colon; Papillary adenocarcinoma; Urethra; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Transitional epithelial cell; [M]Adenocarcinoma, metastatic, NOS; gastric fundus; Proximal stomach; Stromal Cells; Structure of medulla of kidney; Complex epithelial neoplasm; Stomach part; Region of stomach; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Disorder of small intestine; Inflammatory Bowel Diseases; Gastritis; Gastroenteritis; Primary malignant neoplasm of gastrointestinal tract
GSM354051	GSE14108	Brain	Lung	Brain	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Metastatic Carcinoma; Hemoptysis; Respiratory tract hemorrhage; Rectum and sigmoid colon; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Papillary adenocarcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY; Disorder of oral soft tissues; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; [M]Adenocarcinoma, metastatic, NOS; Other and unspecified gastrointestinal disorders; Constipation; Stromal Cells; Squamous epithelial cell; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Neoplasm Metastasis; Neoplastic Processes; Adenocarcinoma, Mucinous
GSM354052	GSE14108	Brain	Lung	Brain	Other and unspecified gastrointestinal disorders; Urethra; Constipation; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Rectum and sigmoid colon; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Complex epithelial neoplasm; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Papillary adenocarcinoma; Hemoptysis; Gastrointestinal Hemorrhage; Respiratory tract hemorrhage; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; gastric fundus; Proximal stomach; Neoplasm Metastasis; Neoplastic Processes; Primary malignant neoplasm of gastrointestinal tract

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM459858	GSE18462	Colon	NA	NA	Other and unspecified gastrointestinal disorders; Upper gastrointestinal disorders; Urethra; Colonic Diseases, Functional; Constipation; Irritable Bowel Syndrome; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Adenosquamous carcinoma; Rectum and sigmoid colon; Gastrointestinal Hemorrhage; Functional disorder of intestine; Papillary serous cystadenocarcinoma; ileum; Campylobacterales; Helicobacter; Helicobacteraceae; Epsilonproteobacteria; Subclass Aerobic-Microaerophilic, Motile Curved Gram-Negative Bacteria; [M]Adenocarcinoma, metastatic, NOS; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Complex epithelial neoplasm; Primary malignant neoplasm of gastrointestinal tract; Stomach part; Region of stomach; Disorder of small intestine; Inflammatory Bowel Diseases; Gastritis; Gastroenteritis; Metastatic Carcinoma; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Proteobacteria; Gram-Negative Bacteria; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Papillary adenocarcinoma; Adenocarcinoma, Mucinous; Gingival and periodontal disease NOS
GSM459859	GSE18462	Colon	Colon	NA	Cholelithiasis; Other and unspecified gastrointestinal disorders; Urethra; White Adipose Tissue; Colonic Diseases, Functional; Constipation; Cholecystolithiasis; Calculi; Subcutaneous Fat; Subcutaneous Tissue; Irritable Bowel Syndrome; Adenosquamous carcinoma; Biliary calculi; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; DISEASES OF THE GALLBLADDER AND BILE DUCTS; Biliary Tract Diseases; Gall Bladder Diseases; Rectum and sigmoid colon; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Gastrointestinal Hemorrhage; Functional disorder of intestine; Adenocarcinoma, Mucinous; ileum; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; HYPERALIMENTATION AND OBESITY; Overnutrition; Obesity; Other endocrine/nutritional/metabolic disorder; Proximal stomach; gastric fundus; Upper gastrointestinal disorders; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Metastatic Carcinoma
GSM459860	GSE18462	Liver	NA	NA	Cholelithiasis; Other and unspecified gastrointestinal disorders; Urethra; White Adipose Tissue; Colonic Diseases, Functional; Constipation; Cholecystolithiasis; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Calculi; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of adrenal cortex; Irritable Bowel Syndrome; Adrenal Cortical Adenoma; Adenosquamous carcinoma; Benign neoplasm of retroperitoneum; Biliary calculi; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; DISEASES OF THE GALLBLADDER AND BILE DUCTS; Biliary Tract Diseases; Gall Bladder Diseases; Rectum and sigmoid colon; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Complex epithelial neoplasm; Functional disorder of intestine; Papillary serous cystadenocarcinoma; Adenocarcinoma, Mucinous; Gastrointestinal Hemorrhage; ileum; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; HYPERALIMENTATION AND OBESITY; Overnutrition; Obesity; Other endocrine/nutritional/metabolic disorder; gastric fundus; Proximal stomach; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Endocrine tumor morphology; Metastatic Carcinoma

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM459861	GSE18462	Liver	Colon	Liver	Other and unspecified gastrointestinal disorders; Constipation; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Adenosquamous carcinoma; Rectum and sigmoid colon; Gastrointestinal Hemorrhage; Complex epithelial neoplasm; [M]Adenocarcinoma, metastatic, NOS; Hemoptysis; Respiratory tract hemorrhage; Colonic Diseases, Functional; Irritable Bowel Syndrome; Functional disorder of intestine; Papillary adenocarcinoma; Squamous epithelial cell; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Transitional epithelial cell; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Urethra; ileum; Primary malignant neoplasm of gastrointestinal tract; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY; Disorder of oral soft tissues; Upper gastrointestinal disorders; Primary malignant neoplasm of intrathoracic organs; Primary malignant neoplasm of lung; Primary malignant neoplasm of respiratory tract; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Campylobacterales; Helicobacter; Helicobacteraceae; Epsilonproteobacteria
GSM459862	GSE18462	Colon	NA	NA	Other and unspecified gastrointestinal disorders; Urethra; Colonic Diseases, Functional; Constipation; Irritable Bowel Syndrome; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Adenosquamous carcinoma; Rectum and sigmoid colon; Gastrointestinal Hemorrhage; Functional disorder of intestine; Papillary serous cystadenocarcinoma; Upper gastrointestinal disorders; Metastatic Carcinoma; ileum; Complex epithelial neoplasm; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Campylobacterales; Helicobacter; Helicobacteraceae; Epsilonproteobacteria; Subclass Aerobic-Microaerophilic, Motile Curved Gram-Negative Bacteria; Adenocarcinoma, Mucinous; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Anorectal structure; Lower bowel structures; Rectum; Pelvic alimentary structure; Primary malignant neoplasm of gastrointestinal tract; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Skin tissue; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Proximal stomach
GSM459863	GSE18462	Colon	Colon	NA	Other and unspecified gastrointestinal disorders; Urethra; Constipation; Papillary serous cystadenocarcinoma; Hemoptysis; Adenosquamous carcinoma; Rectum and sigmoid colon; Gastrointestinal Hemorrhage; Respiratory tract hemorrhage; Complex epithelial neoplasm; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Colonic Diseases, Functional; Irritable Bowel Syndrome; Functional disorder of intestine; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Papillary adenocarcinoma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Adenocarcinoma, Mucinous; ileum; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Primary malignant neoplasm of gastrointestinal tract; Squamous epithelial cell; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM459864	GSE18462	Liver	NA	NA	Cholelithiasis; Other and unspecified gastrointestinal disorders; Urethra; White Adipose Tissue; Colonic Diseases, Functional; Constipation; Cholecystolithiasis; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Calculi; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of adrenal cortex; Irritable Bowel Syndrome; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Biliary calculi; Adenosquamous carcinoma; DISEASES OF THE GALLBLADDER AND BILE DUCTS; Biliary Tract Diseases; Gall Bladder Diseases; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Rectum and sigmoid colon; Gastrointestinal Hemorrhage; Functional disorder of intestine; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Papillary serous cystadenocarcinoma; Structure of cortex of kidney; Adenocarcinoma, Mucinous; Complex epithelial neoplasm; ileum; Endocrine tumor morphology; Proximal stomach; gastric fundus; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; Anorectal structure; Lower bowel structures; Rectum; Pelvic alimentary structure
GSM459865	GSE18462	Liver	Colon	Liver	Other and unspecified gastrointestinal disorders; Constipation; Papillary serous cystadenocarcinoma; Hemoptysis; Adenosquamous carcinoma; Rectum and sigmoid colon; Respiratory tract hemorrhage; Metastatic Carcinoma; Colonic Diseases, Functional; Irritable Bowel Syndrome; [M]Adenocarcinoma, metastatic, NOS; Functional disorder of intestine; Papillary adenocarcinoma; Complex epithelial neoplasm; Urethra; Gastrointestinal Hemorrhage; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Transitional epithelial cell; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; ileum; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Primary malignant neoplasm of gastrointestinal tract; Carcinoma, Transitional Cell; Transitional Cell Neoplasm; [M]Transitional cell papilloma or carcinoma NOS; Upper urinary tract structure; Upper genitourinary tract structure; Renal collecting system structure; Renal pelvis; Adenocarcinoma, Mucinous; Neoplasm Metastasis; Neoplastic Processes; Upper gastrointestinal disorders; Anorectal structure
GSM516678	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Rectum and sigmoid colon; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Exanthema; Disorder of keratinization; Cell-mediated cytotoxic disorder; Cutaneous hypersensitivity; Acquired disorder of keratinization; Histologic type of inflammatory skin disorder; Psoriasis; Other psoriasis; Skin Diseases, Papulosquamous; Inflammatory hyperkeratotic dermatosis; Ductal Carcinoma; Skin tissue; Neoplasm Metastasis; Neoplastic Processes; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516679	GSE20565	Ovary	Ovary	NA	Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Transitional epithelial cell; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Complex epithelial neoplasm; Metastatic Carcinoma; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Rectum and sigmoid colon; Urethra; Other and unspecified gastrointestinal disorders; Constipation; [M]Adenocarcinoma, metastatic, NOS; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Endometrial Neoplasms; Endometrial disorder
GSM516680	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Ductal Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Rectum and sigmoid colon; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Neoplasm Metastasis; Neoplastic Processes; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Coughing
GSM516681	GSE20565	Ovary	Ovary	NA	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Papillary adenocarcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; [M]Adenocarcinoma, metastatic, NOS; Rectum and sigmoid colon; Ductal Carcinoma; Neoplasm Metastasis; Neoplastic Processes; Other and unspecified gastrointestinal disorders; Constipation; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Adenocarcinoma, Mucinous; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516682	GSE20565	Ovary	Breast	Ovary	Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Complex epithelial neoplasm; Metastatic Carcinoma; Papillary adenocarcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; [M]Adenocarcinoma, metastatic, NOS; Hemoptysis; Respiratory tract hemorrhage; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Rectum and sigmoid colon; Neoplasm Metastasis; Neoplastic Processes; Other and unspecified gastrointestinal disorders; Constipation; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Ductal Carcinoma; Skin tissue; Maintenance chemotherapy; radiotherapy
GSM516683	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Metastatic Carcinoma; Hemoptysis; Respiratory tract hemorrhage; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; [M]Adenocarcinoma, metastatic, NOS; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Rectum and sigmoid colon; Neoplasm Metastasis; Neoplastic Processes; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Skin tissue; Ductal Carcinoma; Other and unspecified gastrointestinal disorders; Constipation; Coughing
GSM516684	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Metastatic Carcinoma; Papillary adenocarcinoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Rectum and sigmoid colon; [M]Adenocarcinoma, metastatic, NOS; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Neoplasm Metastasis; Neoplastic Processes; Endometrial Neoplasms; Endometrial disorder; Ductal Carcinoma; Ovary and/or broad ligament structures; Ovary

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516685	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Complex epithelial neoplasm; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Rectum and sigmoid colon; Neoplasm Metastasis; Neoplastic Processes; Other and unspecified gastrointestinal disorders; Constipation; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Skin tissue; Ductal Carcinoma; Primary malignant neoplasm of male genital organ
GSM516686	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Complex epithelial neoplasm; [M]Adenocarcinoma, metastatic, NOS; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Neoplasm Metastasis; Neoplastic Processes; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Rectum and sigmoid colon; Ductal Carcinoma; Other and unspecified gastrointestinal disorders; Constipation; Sense Organs; Nose
GSM516687	GSE20565	Ovary	Ovary	NA	Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Urethra; [M]Adenocarcinoma, metastatic, NOS; Rectum and sigmoid colon; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Other and unspecified gastrointestinal disorders; Constipation; Ductal Carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Neoplasm Metastasis; Neoplastic Processes; Transitional epithelial cell; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Superior mediastinum; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum

Continued on Next Page…

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516688	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Complex epithelial neoplasm; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Rectum and sigmoid colon; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Neoplasm Metastasis; Neoplastic Processes; Skin tissue; Hemoptysis; Respiratory tract hemorrhage; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Other and unspecified gastrointestinal disorders; Constipation; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid
GSM516689	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Abdominal bloating; Respiratory tract hemorrhage; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Papillary adenocarcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Rectum and sigmoid colon; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Neoplasm Metastasis; Neoplastic Processes; Endometrium; Skin tissue; Ovary and/or broad ligament structures; Ovary; Mammary gland; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid
GSM516690	GSE20565	Ovary	Breast	Ovary	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Rectum and sigmoid colon; Neoplasm Metastasis; Neoplastic Processes; Ductal Carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Other and unspecified gastrointestinal disorders; Constipation; Superior mediastinum; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Endometrioid tumor

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516691	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Rectum and sigmoid colon; Ductal Carcinoma; [M]Adenocarcinoma, metastatic, NOS; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Neoplasm Metastasis; Neoplastic Processes; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Skin tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw
GSM516692	GSE20565	Ovary	Breast	Ovary	Urethra; Benign neoplasm of intra-abdominal organs; Primary malignant neoplasm of male genital organ; Benign neoplasm of adrenal gland; Prostate carcinoma; Papillary serous cystadenocarcinoma; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Adenosquamous carcinoma; Benign neoplasm of retroperitoneum; Primary malignant neoplasm of prostate; [M]Adenocarcinoma, metastatic, NOS; Papillary adenocarcinoma; Metastatic Carcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Mammary gland; Urinary outflow structure; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Complex epithelial neoplasm; Rectum and sigmoid colon; Proximal stomach; gastric fundus; Neoplasm Metastasis; Neoplastic Processes; Lower urinary tract; Bladder and outflow structure; Pelvic cavity urinary structure; Prostatic and/or seminal vesicle structures
GSM516693	GSE20565	Ovary	Breast	Ovary	Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Papillary adenocarcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Metastatic Carcinoma; Transitional epithelial cell; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; [M]Adenocarcinoma, metastatic, NOS; Complex epithelial neoplasm; Rectum and sigmoid colon; Hemoptysis; Respiratory tract hemorrhage; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Urinary outflow structure; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Lower urinary tract; Bladder and outflow structure; Pelvic cavity urinary structure

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516694	GSE20565	Ovary	Ovary	NA	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Papillary adenocarcinoma; Rectum and sigmoid colon; [M]Adenocarcinoma, metastatic, NOS; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Other and unspecified gastrointestinal disorders; Constipation; Neoplasm Metastasis; Neoplastic Processes; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Adenocarcinoma, Mucinous; Ductal Carcinoma; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis
GSM516695	GSE20565	Ovary	Breast	Ovary	Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Metastatic Carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Mammary gland; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Complex epithelial neoplasm; Rectum and sigmoid colon; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; gastric fundus; Proximal stomach; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Skin tissue; Neoplasm Metastasis; Neoplastic Processes; Urinary outflow structure; Hemoptysis; Respiratory tract hemorrhage; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Lower urinary tract
GSM516696	GSE20565	Ovary	Breast	Ovary	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Hemoptysis; Respiratory tract hemorrhage; Complex epithelial neoplasm; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Ductal Carcinoma; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Neoplasm Metastasis; Neoplastic Processes; Other and unspecified gastrointestinal disorders; Constipation; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Transitional epithelial cell; Endometrioid tumor; Malignant endometrioid tumor

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516697	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Rectum and sigmoid colon; [M]Adenocarcinoma, metastatic, NOS; Neoplasm Metastasis; Neoplastic Processes; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Ductal Carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Coughing; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid
GSM516698	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Rectum and sigmoid colon; [M]Adenocarcinoma, metastatic, NOS; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Neoplasm Metastasis; Neoplastic Processes; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Ductal Carcinoma; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Skin tissue; Coughing; Gingival and periodontal disease NOS; Jaw Diseases
GSM516699	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Abdominal bloating; Respiratory tract hemorrhage; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Complex epithelial neoplasm; [M]Adenocarcinoma, metastatic, NOS; Papillary adenocarcinoma; Rectum and sigmoid colon; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Neoplasm Metastasis; Neoplastic Processes; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Ductal Carcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Ovary and/or broad ligament structures; Ovary; Transitional epithelial cell; Other and unspecified gastrointestinal disorders

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516700	GSE20565	Ovary	Ovary	NA	Urethra; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Adenosquamous carcinoma; Complex epithelial neoplasm; Rectum and sigmoid colon; [M]Adenocarcinoma, metastatic, NOS; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Papillary adenocarcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Neoplasm Metastasis; Neoplastic Processes; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Other and unspecified gastrointestinal disorders; Constipation; Hemoptysis; Respiratory tract hemorrhage; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Carcinoma, Transitional Cell; Transitional Cell Neoplasm
GSM516701	GSE20565	Ovary	Breast	Ovary	Transitional epithelial cell; Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Metastatic Carcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; [M]Adenocarcinoma, metastatic, NOS; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Prostatic and/or seminal vesicle structures; Minor pelvis; Male urinary outflow structure; Prostate and vas deferens structures; Prostate; Male internal genital organ; Pelvic cavity male genital structure; Urinary outflow structure; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Benign neoplasm of other endocrine glands and related structures
GSM516702	GSE20565	Ovary	Ovary	NA	Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; medulloblastoma; [M]Adenocarcinoma, metastatic, NOS; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Urethra; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Adrenal mass; Tumors of Adrenal Cortex; Adrenal Cortex Diseases; Adrenal Gland Diseases; Adrenal Gland Neoplasms; Papillary adenocarcinoma; Adrenocortical carcinoma; Endocrine tumor morphology; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Neoplasm Metastasis; Neoplastic Processes; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Metastatic Carcinoma; Pluripotent Stem Cells; Retroperitoneal Neoplasms; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516703	GSE20565	Ovary	Ovary	NA	Other and unspecified gastrointestinal disorders; Urethra; Constipation; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Rectum and sigmoid colon; Gastrointestinal Hemorrhage; Colonic Diseases, Functional; Irritable Bowel Syndrome; Metastatic Carcinoma; Functional disorder of intestine; [M]Adenocarcinoma, metastatic, NOS; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Complex epithelial neoplasm; Papillary adenocarcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Transitional epithelial cell; Upper gastrointestinal disorders; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Hemoptysis; Respiratory tract hemorrhage; Primary malignant neoplasm of gastrointestinal tract; Proximal stomach; gastric fundus; Campylobacterales
GSM516704	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Papillary adenocarcinoma; Pain of digestive structure; Metastatic Carcinoma; Hemoptysis; Respiratory tract hemorrhage; [M]Adenocarcinoma, metastatic, NOS; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Rectum and sigmoid colon; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Neoplasm Metastasis; Neoplastic Processes; Transitional epithelial cell; Complex epithelial neoplasm; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Ovary and/or broad ligament structures; Ovary; Endometrium; Mammalian Oviducts
GSM516705	GSE20565	Ovary	Breast	Ovary	Urethra; Benign neoplasm of intra-abdominal organs; Primary malignant neoplasm of male genital organ; Entire viscus; Benign neoplasm of adrenal gland; Prostate carcinoma; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Benign neoplasm of adrenal cortex; Entire fallopian tube; Adrenal Cortical Adenoma; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Benign neoplasm of retroperitoneum; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Primary malignant neoplasm of prostate; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Metastatic Carcinoma; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Transitional epithelial cell; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Urinary outflow structure; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Lower urinary tract; Bladder and outflow structure; Pelvic cavity urinary structure; Prostatic and/or seminal vesicle structures; Minor pelvis; Male urinary outflow structure; Prostate and vas deferens structures; Prostate; Male internal genital organ; Pelvic cavity male genital structure; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Hemoptysis; Respiratory tract hemorrhage

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516706	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Metastatic Carcinoma; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Complex epithelial neoplasm; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Rectum and sigmoid colon; Neoplasm Metastasis; Neoplastic Processes; Transitional epithelial cell; Skin tissue; Ovary and/or broad ligament structures; Ovary; Endometrial Neoplasms
GSM516707	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Rectum and sigmoid colon; [M]Adenocarcinoma, metastatic, NOS; Complex epithelial neoplasm; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Transitional epithelial cell; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Neoplasm Metastasis; Neoplastic Processes; Other and unspecified gastrointestinal disorders; Constipation; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Skin tissue; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures
GSM516708	GSE20565	Ovary	Ovary	NA	Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Metastatic Carcinoma; Transitional epithelial cell; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Complex epithelial neoplasm; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Rectum and sigmoid colon; [M]Adenocarcinoma, metastatic, NOS; Urethra; Ductal Carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Superior mediastinum; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Endometrial Neoplasms; Endometrial disorder; Other and unspecified gastrointestinal disorders; Constipation

Continued on Next Page…

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516709	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Rectum and sigmoid colon; Complex epithelial neoplasm; [M]Adenocarcinoma, metastatic, NOS; Hemoptysis; Respiratory tract hemorrhage; Papillary adenocarcinoma; Other and unspecified gastrointestinal disorders; Constipation; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Neoplasm Metastasis; Neoplastic Processes; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Adenocarcinoma, Mucinous; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Transitional epithelial cell; Neoplasms, Cystic, Mucinous, and Serous
GSM516710	GSE20565	Ovary	Breast	Ovary	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Metastatic Carcinoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; [M]Adenocarcinoma, metastatic, NOS; Papillary adenocarcinoma; Rectum and sigmoid colon; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Neoplasm Metastasis; Neoplastic Processes; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Transitional epithelial cell; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Ductal Carcinoma; Endometrial Neoplasms; Endometrial disorder; Coughing
GSM516711	GSE20565	Ovary	Ovary	NA	Other and unspecified gastrointestinal disorders; Constipation; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Rectum and sigmoid colon; Respiratory tract hemorrhage; Papillary adenocarcinoma; Complex epithelial neoplasm; Urethra; Transitional epithelial cell; Neoplasm Metastasis; Neoplastic Processes; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Ductal Carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Coughing

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516712	GSE20565	Ovary	Ovary	NA	Rhinovirus infection; Other and unspecified gastrointestinal disorders; Urethra; Constipation; Papillary serous cystadenocarcinoma; RNA Virus Infections; Adenosquamous carcinoma; Picornaviridae Infections; Papillary adenocarcinoma; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Sense Organs; Nose; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Complex epithelial neoplasm; Rectum and sigmoid colon; Transitional epithelial cell; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Hemoptysis; Respiratory tract hemorrhage; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Carcinoma, Transitional Cell; Transitional Cell Neoplasm; [M]Transitional cell papilloma or carcinoma NOS; Upper urinary tract structure; Upper genitourinary tract structure
GSM516713	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Metastatic Carcinoma; Papillary adenocarcinoma; Hemoptysis; Respiratory tract hemorrhage; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; [M]Adenocarcinoma, metastatic, NOS; Complex epithelial neoplasm; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Rectum and sigmoid colon; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Neoplasm Metastasis; Neoplastic Processes; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Transitional epithelial cell; Ductal Carcinoma; Ovary and/or broad ligament structures; Ovary; Skin tissue
GSM516714	GSE20565	Ovary	Breast	Ovary	Urethra; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Papillary serous cystadenocarcinoma; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Adenosquamous carcinoma; Benign neoplasm of retroperitoneum; Papillary adenocarcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Complex epithelial neoplasm; Transitional epithelial cell; Neoplasm Metastasis; Neoplastic Processes; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Hemoptysis; Respiratory tract hemorrhage; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Mammary gland; Stromal Cells; Lactiferous duct

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516715	GSE20565	Ovary	Ovary	NA	Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; [M]Adenocarcinoma, metastatic, NOS; Ductal Carcinoma; Rectum and sigmoid colon; Superior mediastinum; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Neoplasm Metastasis; Neoplastic Processes; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Urethra; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland
GSM516716	GSE20565	Ovary	Breast	Ovary	Urethra; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Primary malignant neoplasm of prostate; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; [M]Adenocarcinoma, metastatic, NOS; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Metastatic Carcinoma; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Complex epithelial neoplasm; Urinary outflow structure; Hemoptysis; Respiratory tract hemorrhage; Rectum and sigmoid colon; Skin tissue; Endometrium; Mammary gland; Lower urinary tract; Bladder and outflow structure; Pelvic cavity urinary structure; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Neoplasm Metastasis
GSM516717	GSE20565	Ovary	Breast	Ovary	Urethra; Benign neoplasm of intra-abdominal organs; Primary malignant neoplasm of male genital organ; Benign neoplasm of adrenal gland; Prostate carcinoma; Papillary serous cystadenocarcinoma; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Adenosquamous carcinoma; Benign neoplasm of retroperitoneum; Primary malignant neoplasm of prostate; Metastatic Carcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; [M]Adenocarcinoma, metastatic, NOS; Papillary adenocarcinoma; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; gastric fundus; Proximal stomach; Mammary gland; Complex epithelial neoplasm; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Urinary outflow structure; Hemoptysis; Respiratory tract hemorrhage; Lower urinary tract; Bladder and outflow structure; Pelvic cavity urinary structure; Rectum and sigmoid colon; Neoplasm Metastasis; Neoplastic Processes; Endometrium; Lactiferous duct; Mammary lobe

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516718	GSE20565	Ovary	Breast	Ovary	Urethra; Benign neoplasm of intra-abdominal organs; Entire viscus; Benign neoplasm of adrenal gland; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Benign neoplasm of adrenal cortex; Entire fallopian tube; Adrenal Cortical Adenoma; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Benign neoplasm of retroperitoneum; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Transitional epithelial cell; Metastatic Carcinoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Adrenal mass; Tumors of Adrenal Cortex; Adrenal Cortex Diseases; Adrenal Gland Diseases; Adrenal Gland Neoplasms; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Urinary outflow structure; Prostatic and/or seminal vesicle structures; Minor pelvis; Male urinary outflow structure; Prostate and vas deferens structures; Prostate
GSM516719	GSE20565	Ovary	Breast	Ovary	Urethra; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Primary malignant neoplasm of prostate; Metastatic Carcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; [M]Adenocarcinoma, metastatic, NOS; gastric fundus; Proximal stomach; Mammary gland; Rectum and sigmoid colon; Complex epithelial neoplasm; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Papillary adenocarcinoma; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Urinary outflow structure; Skin tissue; Lower urinary tract; Bladder and outflow structure; Pelvic cavity urinary structure; Neoplasm Metastasis; Neoplastic Processes; Hemoptysis; Respiratory tract hemorrhage; Endometrium; Lactiferous duct
GSM516720	GSE20565	Ovary	Ovary	NA	Transitional epithelial cell; Other and unspecified gastrointestinal disorders; Urethra; Colonic Diseases, Functional; Constipation; Papillary serous cystadenocarcinoma; Irritable Bowel Syndrome; Adenosquamous carcinoma; Rectum and sigmoid colon; Gastrointestinal Hemorrhage; Functional disorder of intestine; Papillary adenocarcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; [M]Adenocarcinoma, metastatic, NOS; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Hemoptysis; Respiratory tract hemorrhage; Metastatic Carcinoma; Upper gastrointestinal disorders; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Complex epithelial neoplasm; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; ileum; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516721	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; [M]Adenocarcinoma, metastatic, NOS; Rectum and sigmoid colon; Hemoptysis; Respiratory tract hemorrhage; Papillary adenocarcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Other and unspecified gastrointestinal disorders; Constipation; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Neoplasm Metastasis; Neoplastic Processes; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face
GSM516722	GSE20565	Ovary	Breast	Ovary	Urethra; Benign neoplasm of intra-abdominal organs; Primary malignant neoplasm of male genital organ; Benign neoplasm of adrenal gland; Prostate carcinoma; Papillary serous cystadenocarcinoma; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Adenosquamous carcinoma; Benign neoplasm of retroperitoneum; Primary malignant neoplasm of prostate; Mammary gland; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; gastric fundus; Proximal stomach; Metastatic Carcinoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Complex epithelial neoplasm; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Endometrium; Papillary adenocarcinoma; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; [M]Adenocarcinoma, metastatic, NOS; Urinary outflow structure; Skin tissue; Rectum and sigmoid colon; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Lower urinary tract
GSM516723	GSE20565	Ovary	Ovary	NA	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Papillary adenocarcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Rectum and sigmoid colon; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; [M]Adenocarcinoma, metastatic, NOS; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Skin tissue; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Ductal Carcinoma; Neoplasm Metastasis; Neoplastic Processes; Mammary gland; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Adenocarcinoma, Mucinous; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516724	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Ductal Carcinoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Rectum and sigmoid colon; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Neoplasm Metastasis; Neoplastic Processes; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Transitional epithelial cell; Superior mediastinum; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma
GSM516725	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Complex epithelial neoplasm; [M]Adenocarcinoma, metastatic, NOS; Rectum and sigmoid colon; Other and unspecified gastrointestinal disorders; Constipation; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Neoplasm Metastasis; Neoplastic Processes; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Transitional epithelial cell; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Skin tissue; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Gingival and periodontal disease NOS
GSM516726	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Abdominal bloating; Respiratory tract hemorrhage; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Papillary adenocarcinoma; Pain of digestive structure; Metastatic Carcinoma; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; [M]Adenocarcinoma, metastatic, NOS; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Rectum and sigmoid colon; Complex epithelial neoplasm; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Transitional epithelial cell; Neoplasm Metastasis; Neoplastic Processes; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Endometrial Neoplasms; Endometrial disorder; Ductal Carcinoma; Primary malignant neoplasm of male genital organ

Continued on Next Page…

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516727	GSE20565	Ovary	Ovary	NA	Transitional epithelial cell; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Urethra; Complex epithelial neoplasm; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Rectum and sigmoid colon; Other and unspecified gastrointestinal disorders; Constipation; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Neoplasm Metastasis; Neoplastic Processes; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Ductal Carcinoma; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures
GSM516728	GSE20565	Ovary	Ovary	NA	Rhinovirus infection; Other and unspecified gastrointestinal disorders; Urethra; Constipation; Sense Organs; Papillary serous cystadenocarcinoma; Nose; RNA Virus Infections; Adenosquamous carcinoma; Picornaviridae Infections; Papillary adenocarcinoma; Metastatic Carcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; [M]Adenocarcinoma, metastatic, NOS; Complex epithelial neoplasm; Rectum and sigmoid colon; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Hemoptysis; Respiratory tract hemorrhage; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Transitional epithelial cell; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Gastrointestinal Hemorrhage
GSM516729	GSE20565	Ovary	Breast	Ovary	Urethra; Benign neoplasm of intra-abdominal organs; Primary malignant neoplasm of male genital organ; Benign neoplasm of adrenal gland; Prostate carcinoma; Papillary serous cystadenocarcinoma; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Adenosquamous carcinoma; Benign neoplasm of retroperitoneum; Primary malignant neoplasm of prostate; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Papillary adenocarcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; [M]Adenocarcinoma, metastatic, NOS; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Metastatic Carcinoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Complex epithelial neoplasm; Urinary outflow structure; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Lower urinary tract; Bladder and outflow structure; Pelvic cavity urinary structure; Mammary gland; Endocrine tumor morphology; Neoplasm Metastasis; Neoplastic Processes; Prostatic and/or seminal vesicle structures; Minor pelvis; Male urinary outflow structure; Prostate and vas deferens structures; Prostate; Male internal genital organ; Pelvic cavity male genital structure

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516731	GSE20565	Ovary	Ovary	NA	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Adenosquamous carcinoma; Rectum and sigmoid colon; Other and unspecified gastrointestinal disorders; Constipation; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Adenocarcinoma, Mucinous; Hemoptysis; Respiratory tract hemorrhage; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Neoplasms, Cystic, Mucinous, and Serous; Cystic, mucinous AND/OR serous neoplasm; Squamous epithelial cell; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY; Disorder of oral soft tissues; Abdominal bloating
GSM516732	GSE20565	Ovary	Breast	Ovary	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Metastatic Carcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Hemoptysis; Respiratory tract hemorrhage; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Transitional epithelial cell; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; [M]Adenocarcinoma, metastatic, NOS; Complex epithelial neoplasm; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Ductal Carcinoma; Adrenal mass; Tumors of Adrenal Cortex; Adrenal Cortex Diseases; Adrenal Gland Diseases; Adrenal Gland Neoplasms; Malignant neoplasm of female genital organ
GSM516733	GSE20565	Ovary	Breast	Ovary	Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Complex epithelial neoplasm; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Neoplasm Metastasis; Neoplastic Processes; Urethra; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Rectum and sigmoid colon; Ductal Carcinoma; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Gingival and periodontal disease NOS

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516734	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Neoplasm Metastasis; Neoplastic Processes; Rectum and sigmoid colon; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Skin tissue; Other and unspecified gastrointestinal disorders; Constipation
GSM516735	GSE20565	Ovary	Ovary	NA	Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Rectum and sigmoid colon; Other and unspecified gastrointestinal disorders; Constipation; Exanthema; Disorder of keratinization; Cell-mediated cytotoxic disorder; Cutaneous hypersensitivity; Acquired disorder of keratinization; Histologic type of inflammatory skin disorder; Psoriasis; Other psoriasis; Skin Diseases, Papulosquamous; Inflammatory hyperkeratotic dermatosis; Squamous epithelial cell; Neoplasm Metastasis; Neoplastic Processes; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY; Disorder of oral soft tissues; Ductal Carcinoma; ovarian neoplasm
GSM516736	GSE20565	Ovary	Breast	Ovary	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Transitional epithelial cell; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; [M]Adenocarcinoma, metastatic, NOS; Metastatic Carcinoma; Complex epithelial neoplasm; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Urinary outflow structure; Hemoptysis; Respiratory tract hemorrhage; Lower urinary tract; Bladder and outflow structure; Pelvic cavity urinary structure; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516737	GSE20565	Ovary	Breast	Ovary	Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Papillary adenocarcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Metastatic Carcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; [M]Adenocarcinoma, metastatic, NOS; Mammary gland; Rectum and sigmoid colon; Complex epithelial neoplasm; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Hemoptysis; Respiratory tract hemorrhage; Skin tissue; Proximal stomach; gastric fundus; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Neoplasm Metastasis; Neoplastic Processes; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Adenocarcinoma, Mucinous; ovarian neoplasm
GSM516738	GSE20565	Ovary	Breast	Ovary	Transitional epithelial cell; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Hemoptysis; Respiratory tract hemorrhage; Metastatic Carcinoma; Other and unspecified gastrointestinal disorders; Constipation; Ductal Carcinoma; Complex epithelial neoplasm; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; [M]Adenocarcinoma, metastatic, NOS; Urethra; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Rectum and sigmoid colon
GSM516739	GSE20565	Ovary	Ovary	NA	Transitional epithelial cell; Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Metastatic Carcinoma; Rectum and sigmoid colon; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Complex epithelial neoplasm; [M]Adenocarcinoma, metastatic, NOS; Other and unspecified gastrointestinal disorders; Constipation; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Neoplasm Metastasis; Neoplastic Processes; Superior mediastinum; Ductal Carcinoma

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516740	GSE20565	Ovary	Breast	Ovary	Transitional epithelial cell; Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Hemoptysis; Respiratory tract hemorrhage; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Complex epithelial neoplasm; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Other and unspecified gastrointestinal disorders; Constipation
GSM516741	GSE20565	Ovary	Breast	Ovary	Urethra; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Primary malignant neoplasm of prostate; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Metastatic Carcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Mammary gland; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Urinary outflow structure; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Proximal stomach; gastric fundus; Complex epithelial neoplasm; Lower urinary tract; Bladder and outflow structure; Pelvic cavity urinary structure; Prostatic and/or seminal vesicle structures; Minor pelvis; Male urinary outflow structure; Prostate and vas deferens structures; Prostate; Male internal genital organ; Pelvic cavity male genital structure; Rectum and sigmoid colon
GSM516742	GSE20565	Ovary	Ovary	NA	Transitional epithelial cell; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Hemoptysis; Respiratory tract hemorrhage; Metastatic Carcinoma; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Superior mediastinum; [M]Adenocarcinoma, metastatic, NOS; Complex epithelial neoplasm; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Rectum and sigmoid colon; Carcinoma, Transitional Cell; Transitional Cell Neoplasm; [M]Transitional cell papilloma or carcinoma NOS; Upper urinary tract structure; Upper genitourinary tract structure; Ductal Carcinoma; Urethra; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Neoplasm Metastasis; Neoplastic Processes; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516743	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Rectum and sigmoid colon; [M]Adenocarcinoma, metastatic, NOS; Papillary adenocarcinoma; Other and unspecified gastrointestinal disorders; Constipation; Complex epithelial neoplasm; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Hemoptysis; Respiratory tract hemorrhage; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Neoplasm Metastasis; Neoplastic Processes; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Adenocarcinoma, Mucinous; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures
GSM516744	GSE20565	Ovary	Breast	Ovary	Urethra; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Primary malignant neoplasm of prostate; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Papillary adenocarcinoma; gastric fundus; Proximal stomach; Metastatic Carcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Mammary gland; Complex epithelial neoplasm; Rectum and sigmoid colon; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; [M]Adenocarcinoma, metastatic, NOS; Adenocarcinoma, Mucinous; Urinary outflow structure; Hemoptysis; Respiratory tract hemorrhage; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Lower urinary tract; Bladder and outflow structure; Pelvic cavity urinary structure; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma
GSM516745	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Papillary adenocarcinoma; Rectum and sigmoid colon; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Skin tissue; Neoplasm Metastasis; Neoplastic Processes; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516746	GSE20565	Ovary	Breast	Ovary	Transitional epithelial cell; Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Metastatic Carcinoma; Hemoptysis; Respiratory tract hemorrhage; Other and unspecified gastrointestinal disorders; Constipation; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; [M]Adenocarcinoma, metastatic, NOS; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Complex epithelial neoplasm; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Ductal Carcinoma; Gastrointestinal Hemorrhage; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Rectum and sigmoid colon; Malignant neoplasm of female genital organ
GSM516747	GSE20565	Ovary	Ovary	NA	Other and unspecified gastrointestinal disorders; Constipation; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Rectum and sigmoid colon; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Hemoptysis; Respiratory tract hemorrhage; Metastatic Carcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Urethra; Colonic Diseases, Functional; Irritable Bowel Syndrome; Functional disorder of intestine; Transitional epithelial cell; Complex epithelial neoplasm; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Neoplasm Metastasis; Neoplastic Processes; Gastrointestinal Hemorrhage; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Malignant neoplasm of female genital organ
GSM516748	GSE20565	Ovary	Breast	Ovary	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Adenosquamous carcinoma; Rectum and sigmoid colon; [M]Adenocarcinoma, metastatic, NOS; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Hemoptysis; Respiratory tract hemorrhage; Other and unspecified gastrointestinal disorders; Constipation; Skin tissue; Papillary adenocarcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Adenocarcinoma, Mucinous; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Carcinoma, Transitional Cell; Transitional Cell Neoplasm; [M]Transitional cell papilloma or carcinoma NOS; Upper urinary tract structure; Upper genitourinary tract structure; Transitional epithelial cell; Renal collecting system structure; Renal pelvis

Continued on Next Page…

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516749	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Rectum and sigmoid colon; [M]Adenocarcinoma, metastatic, NOS; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Skin tissue; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Neoplasm Metastasis; Neoplastic Processes; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Ductal Carcinoma; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face
GSM516750	GSE20565	Ovary	Ovary	NA	Other and unspecified gastrointestinal disorders; Urethra; Constipation; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Rectum and sigmoid colon; Metastatic Carcinoma; Colonic Diseases, Functional; Irritable Bowel Syndrome; Complex epithelial neoplasm; [M]Adenocarcinoma, metastatic, NOS; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Functional disorder of intestine; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Gastrointestinal Hemorrhage; Anorectal structure; Lower bowel structures; Rectum; Pelvic alimentary structure; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Proximal stomach; gastric fundus; Stomach part; Region of stomach; Adenocarcinoma, Mucinous; Papillary adenocarcinoma; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon
GSM516751	GSE20565	Ovary	Ovary	NA	Other and unspecified gastrointestinal disorders; Urethra; Constipation; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Rectum and sigmoid colon; Metastatic Carcinoma; Complex epithelial neoplasm; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Anorectal structure; Lower bowel structures; Rectum; Pelvic alimentary structure; [M]Adenocarcinoma, metastatic, NOS; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Hemoptysis; Respiratory tract hemorrhage; Primary malignant neoplasm of large intestine; Colon Carcinoma; Primary malignant neoplasm of colon; Primary malignant neoplasm of intestinal tract; Adenocarcinoma, Mucinous; Primary malignant neoplasm of gastrointestinal tract; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Gastrointestinal Hemorrhage; Functional disorder of intestine; Colonic Diseases, Functional; Irritable Bowel Syndrome

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516752	GSE20565	Ovary	Breast	Ovary	Urethra; Lactiferous duct; Mammary lobe; Glandular structure of breast; Adenosquamous carcinoma; Duct (organ) structure; Papillary serous cystadenocarcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Mammary gland; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Complex epithelial neoplasm; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; gastric fundus; Proximal stomach; Metastatic Carcinoma; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Rectum and sigmoid colon; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Skin tissue; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Papillary adenocarcinoma; Endometrium; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Hemoptysis; Respiratory tract hemorrhage
GSM516753	GSE20565	Ovary	Breast	Ovary	Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Papillary adenocarcinoma; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Rectum and sigmoid colon; Complex epithelial neoplasm; Neoplasm Metastasis; Neoplastic Processes; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Hemoptysis; Respiratory tract hemorrhage; Endometrium; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Urinary outflow structure; Malignant neoplasm of female genital organ
GSM516754	GSE20565	Ovary	Breast	Ovary	Urethra; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Primary malignant neoplasm of prostate; Papillary adenocarcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; [M]Adenocarcinoma, metastatic, NOS; Metastatic Carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Urinary outflow structure; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Transitional epithelial cell; Prostatic and/or seminal vesicle structures; Minor pelvis; Male urinary outflow structure; Prostate and vas deferens structures; Prostate; Male internal genital organ; Pelvic cavity male genital structure; Lower urinary tract; Bladder and outflow structure; Pelvic cavity urinary structure; Endometrium; Neoplasm Metastasis; Neoplastic Processes; Complex epithelial neoplasm

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516755	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Metastatic Carcinoma; Hemoptysis; Respiratory tract hemorrhage; [M]Adenocarcinoma, metastatic, NOS; Complex epithelial neoplasm; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Skin tissue; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Rectum and sigmoid colon; Neoplasm Metastasis; Neoplastic Processes; Other and unspecified gastrointestinal disorders
GSM516756	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Complex epithelial neoplasm; Metastatic Carcinoma; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; [M]Adenocarcinoma, metastatic, NOS; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Rectum and sigmoid colon; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Transitional epithelial cell; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Neoplasm Metastasis; Neoplastic Processes; Endometrial Neoplasms; Endometrial disorder; Ductal Carcinoma; Skin tissue
GSM516757	GSE20565	Ovary	Breast	Ovary	Urethra; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Metastatic Carcinoma; Papillary adenocarcinoma; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Complex epithelial neoplasm; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; [M]Adenocarcinoma, metastatic, NOS; Hemoptysis; Respiratory tract hemorrhage; Rectum and sigmoid colon; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Adenocarcinoma, Mucinous; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Skin tissue; Neoplasm Metastasis; Neoplastic Processes; Mammary gland; Proximal stomach; gastric fundus; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Urinary outflow structure

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516758	GSE20565	Ovary	Ovary	NA	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Rectum and sigmoid colon; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Ductal Carcinoma; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; [M]Adenocarcinoma, metastatic, NOS; Adenocarcinoma, Mucinous; Neoplasm Metastasis; Neoplastic Processes; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Skin tissue; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY
GSM516759	GSE20565	Ovary	Breast	Ovary	Urethra; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Papillary serous cystadenocarcinoma; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Adenosquamous carcinoma; Benign neoplasm of retroperitoneum; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Mammary gland; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Complex epithelial neoplasm; Proximal stomach; gastric fundus; Papillary adenocarcinoma; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Neoplasm Metastasis; Neoplastic Processes; Rectum and sigmoid colon; Endometrium; Hemoptysis; Respiratory tract hemorrhage; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Urinary outflow structure
GSM516760	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Abdominal bloating; Respiratory tract hemorrhage; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Papillary adenocarcinoma; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Rectum and sigmoid colon; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Neoplasm Metastasis; Neoplastic Processes; Endometrium; Ovary and/or broad ligament structures; Ovary; Primary malignant neoplasm of male genital organ; Prostate carcinoma

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516761	GSE20565	Ovary	Ovary	NA	SPECIFIC ENDOMETRIOSES; Urethra; Endometriosis, site unspecified; White Adipose Tissue; Uterine Fibroids; Endometriosis of uterus; Benign myomatous tumor; Benign neoplasm of trunk; Benign neoplasm of intra-abdominal organs; Benign neoplasm of other endocrine glands and related structures; Benign neoplasm of adrenal gland; Benign neoplasm of female genital organ, site unspecified; Benign neoplasm of body of uterus; Benign neoplasm of uterus NOS; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Benign leiomyomatous neoplasm - category; Leiomyomatous neoplasm - category; Benign genital neoplasm; Benign neoplasm of abdomen; Endometriosis of pelvis; Disorder characterized by pain; Benign neoplasm corpus uteri NEC; Benign tumor of endocrine gland; Adenosquamous carcinoma; Papillary serous cystadenocarcinoma; Myomatous neoplasm; Endocrine tumor morphology; Layer of adrenal gland; Endocrine gland part; Adrenal part; Adrenal Cortex; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Mammary gland; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; gastric fundus; Stromal Cells; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ
GSM516762	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Rectum and sigmoid colon; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Neoplasm Metastasis; Neoplastic Processes; Other and unspecified gastrointestinal disorders; Constipation; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Ductal Carcinoma; Skin tissue; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Diffuse low grade B-cell lymphoma
GSM516763	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; [M]Adenocarcinoma, metastatic, NOS; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Rectum and sigmoid colon; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Neoplasm Metastasis; Neoplastic Processes; Other and unspecified gastrointestinal disorders; Constipation; Ductal Carcinoma; Transitional epithelial cell; Skin tissue

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516764	GSE20565	Ovary	Ovary	NA	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Rectum and sigmoid colon; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Other and unspecified gastrointestinal disorders; Constipation; Adenocarcinoma, Mucinous; Neoplasms, Cystic, Mucinous, and Serous; Cystic, mucinous AND/OR serous neoplasm; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Papillary adenocarcinoma; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; [M]Adenocarcinoma, metastatic, NOS; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY; Disorder of oral soft tissues; Ductal Carcinoma; Coughing; Squamous epithelial cell
GSM516765	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Abdominal bloating; Respiratory tract hemorrhage; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Metastatic Carcinoma; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Rectum and sigmoid colon; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Complex epithelial neoplasm; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Neoplasm Metastasis; Neoplastic Processes; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Ductal Carcinoma; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Transitional epithelial cell
GSM516766	GSE20565	Ovary	Breast	Ovary	Urethra; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Primary malignant neoplasm of prostate; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Metastatic Carcinoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; [M]Adenocarcinoma, metastatic, NOS; Mammary gland; Papillary adenocarcinoma; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Proximal stomach; gastric fundus; Complex epithelial neoplasm; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Urinary outflow structure; Lower urinary tract; Bladder and outflow structure; Pelvic cavity urinary structure; Endometrium; Neoplasm Metastasis; Neoplastic Processes; Rectum and sigmoid colon; Skin tissue; Lactiferous duct; Mammary lobe; Glandular structure of breast

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516767	GSE20565	Ovary	Ovary	NA	Other and unspecified gastrointestinal disorders; Urethra; Constipation; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Rectum and sigmoid colon; Metastatic Carcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; [M]Adenocarcinoma, metastatic, NOS; Papillary adenocarcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Complex epithelial neoplasm; Hemoptysis; Respiratory tract hemorrhage; Neoplasm Metastasis; Neoplastic Processes; Gastrointestinal Hemorrhage; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Functional disorder of intestine; Colonic Diseases, Functional; Irritable Bowel Syndrome; Benign neoplasm of other endocrine glands and related structures
GSM516768	GSE20565	Ovary	Ovary	NA	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Hemoptysis; Adenosquamous carcinoma; Stromal Cells; Respiratory tract hemorrhage; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Metastatic Carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Mesenchymal Stem Cells; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Rectum and sigmoid colon; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Squamous epithelial cell; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Skin tissue; Bone structure of spine and/or pelvis; hip bone; Bone structure of ilium; Bone part; Ilium part; Iliac crest structure; Structure of flat bone; Bone structure of pelvic region and/or thigh; Bony pelvis; Structure of bone (organ); Type of bone; Neoplasm Metastasis; Neoplastic Processes
GSM516769	GSE20565	Ovary	Ovary	NA	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Metastatic Carcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Rectum and sigmoid colon; [M]Adenocarcinoma, metastatic, NOS; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Papillary adenocarcinoma; Skin tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Hemoptysis; Respiratory tract hemorrhage; Other and unspecified gastrointestinal disorders; Constipation; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Neoplasm Metastasis; Neoplastic Processes; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Mammary gland; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516770	GSE20565	Ovary	Ovary	NA	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Metastatic Carcinoma; Rectum and sigmoid colon; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Skin tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Neoplasm Metastasis; Neoplastic Processes; Ductal Carcinoma; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Mammary gland; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Adenocarcinoma, Mucinous; Other and unspecified gastrointestinal disorders; Constipation; Lactiferous duct
GSM516771	GSE20565	Ovary	Ovary	NA	Other and unspecified gastrointestinal disorders; Constipation; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Metastatic Carcinoma; Transitional epithelial cell; Hemoptysis; Respiratory tract hemorrhage; Complex epithelial neoplasm; [M]Adenocarcinoma, metastatic, NOS; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Urethra; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Rectum and sigmoid colon; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Gastrointestinal Hemorrhage; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Malignant neoplasm of female genital organ
GSM516772	GSE20565	Ovary	Ovary	NA	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Papillary adenocarcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Rectum and sigmoid colon; Ductal Carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Adenocarcinoma, Mucinous; Other and unspecified gastrointestinal disorders; Constipation; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; [M]Adenocarcinoma, metastatic, NOS; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Skin tissue; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY; Disorder of oral soft tissues; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal)

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516773	GSE20565	Ovary	Breast	Ovary	Urethra; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Primary malignant neoplasm of prostate; gastric fundus; Proximal stomach; Complex epithelial neoplasm; Metastatic Carcinoma; Rectum and sigmoid colon; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; [M]Adenocarcinoma, metastatic, NOS; Mammary gland; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Adenocarcinoma, Mucinous; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Papillary adenocarcinoma; Hemoptysis; Respiratory tract hemorrhage; Urinary outflow structure; Neoplasm Metastasis; Neoplastic Processes; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Lactiferous duct; Mammary lobe
GSM516774	GSE20565	Ovary	Breast	Ovary	Urethra; Benign neoplasm of intra-abdominal organs; Primary malignant neoplasm of male genital organ; Benign neoplasm of adrenal gland; Prostate carcinoma; Papillary serous cystadenocarcinoma; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Adenosquamous carcinoma; Benign neoplasm of retroperitoneum; Primary malignant neoplasm of prostate; Papillary adenocarcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; [M]Adenocarcinoma, metastatic, NOS; Metastatic Carcinoma; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Hemoptysis; Respiratory tract hemorrhage; Complex epithelial neoplasm; Urinary outflow structure; Neoplasm Metastasis; Neoplastic Processes; Endometrium; Endocrine tumor morphology; Lower urinary tract; Bladder and outflow structure; Pelvic cavity urinary structure; Stromal Cells; Mammary gland; Lactiferous duct; Mammary lobe; Glandular structure of breast
GSM516775	GSE20565	Ovary	Breast	Ovary	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Metastatic Carcinoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Complex epithelial neoplasm; [M]Adenocarcinoma, metastatic, NOS; Hemoptysis; Respiratory tract hemorrhage; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Rectum and sigmoid colon; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Neoplasm Metastasis; Neoplastic Processes; Ductal Carcinoma; Transitional epithelial cell; Endometrium; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Endometrioid tumor; Malignant endometrioid tumor

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516776	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Metastatic Carcinoma; Papillary adenocarcinoma; Rectum and sigmoid colon; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; [M]Adenocarcinoma, metastatic, NOS; Mammary gland; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Hemoptysis; Respiratory tract hemorrhage; Proximal stomach; gastric fundus; Neoplasm Metastasis; Neoplastic Processes; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Adenocarcinoma, Mucinous; Skin tissue; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Primary malignant neoplasm of male genital organ; Prostate carcinoma
GSM516777	GSE20565	Ovary	Ovary	NA	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Rectum and sigmoid colon; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Ductal Carcinoma; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Papillary adenocarcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; [M]Adenocarcinoma, metastatic, NOS; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Disorder of soft tissue of body cavity; Disorder of soft tissue of head; Mouth Diseases; DISEASES OF THE SALIVARY GLANDS AND ORAL CAVITY; Disorder of oral soft tissues; Neoplasm Metastasis; Neoplastic Processes; Coughing
GSM516778	GSE20565	Ovary	Breast	Ovary	Urethra; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Primary malignant neoplasm of prostate; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Metastatic Carcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Mammary gland; gastric fundus; Proximal stomach; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Papillary adenocarcinoma; Urinary outflow structure; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Complex epithelial neoplasm; [M]Adenocarcinoma, metastatic, NOS; Lower urinary tract; Bladder and outflow structure; Pelvic cavity urinary structure; Prostatic and/or seminal vesicle structures; Minor pelvis; Male urinary outflow structure; Prostate and vas deferens structures; Prostate; Male internal genital organ; Pelvic cavity male genital structure; Lactiferous duct

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516779	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Metastatic Carcinoma; Rectum and sigmoid colon; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Ductal Carcinoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Neoplasm Metastasis; Neoplastic Processes; Skin tissue; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Coughing; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis
GSM516780	GSE20565	Ovary	Ovary	NA	Transitional epithelial cell; Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Complex epithelial neoplasm; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Rectum and sigmoid colon; Other and unspecified gastrointestinal disorders; Constipation; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Neoplasm Metastasis; Neoplastic Processes; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Superior mediastinum; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Diffuse low grade B-cell lymphoma
GSM516781	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Hemoptysis; Respiratory tract hemorrhage; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; [M]Adenocarcinoma, metastatic, NOS; Complex epithelial neoplasm; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Rectum and sigmoid colon; Metastatic Carcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Transitional epithelial cell; Skin tissue; Endometrial Neoplasms; Endometrial disorder; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516782	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Other and unspecified gastrointestinal disorders; Constipation; Rectum and sigmoid colon; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Superior mediastinum; Ductal Carcinoma; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs
GSM516783	GSE20565	Ovary	Ovary	NA	Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Urethra; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Metastatic Carcinoma; Complex epithelial neoplasm; Transitional epithelial cell; [M]Adenocarcinoma, metastatic, NOS; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Rectum and sigmoid colon; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Endometrial Neoplasms; Endometrial disorder; Ductal Carcinoma; Pluripotent Stem Cells; Other and unspecified gastrointestinal disorders; Constipation
GSM516784	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Rectum and sigmoid colon; Ductal Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Neoplasm Metastasis; Neoplastic Processes; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Superior mediastinum; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Other and unspecified gastrointestinal disorders; Constipation; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Squamous epithelial cell

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516785	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Abdominal bloating; Respiratory tract hemorrhage; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Papillary adenocarcinoma; Pain of digestive structure; [M]Adenocarcinoma, metastatic, NOS; Complex epithelial neoplasm; Rectum and sigmoid colon; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Other and unspecified gastrointestinal disorders; Constipation; Neoplasm Metastasis; Neoplastic Processes; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Transitional epithelial cell; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland
GSM516786	GSE20565	Ovary	Ovary	NA	Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Metastatic Carcinoma; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Urethra; Transitional epithelial cell; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; [M]Adenocarcinoma, metastatic, NOS; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Complex epithelial neoplasm; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Neoplasm Metastasis; Neoplastic Processes; Endometrial Neoplasms; Endometrial disorder; Rectum and sigmoid colon; Ductal Carcinoma; Other and unspecified gastrointestinal disorders
GSM516787	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; [M]Adenocarcinoma, metastatic, NOS; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Complex epithelial neoplasm; Rectum and sigmoid colon; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Metastatic Carcinoma; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Pluripotent Stem Cells; Transitional epithelial cell; Neoplasms, Muscle Tissue; Malignant myomatous tumor; Ovary and/or broad ligament structures; Ovary; Neoplasm Metastasis

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516788	GSE20565	Ovary	Ovary	NA	Transitional epithelial cell; Other and unspecified gastrointestinal disorders; Urethra; Constipation; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Rectum and sigmoid colon; Gastrointestinal Hemorrhage; Colonic Diseases, Functional; Irritable Bowel Syndrome; [M]Adenocarcinoma, metastatic, NOS; Functional disorder of intestine; Papillary adenocarcinoma; Metastatic Carcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; gastric fundus; Proximal stomach; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Stomach part; Region of stomach; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Upper gastrointestinal disorders; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Complex epithelial neoplasm; Campylobacterales; Helicobacter; Helicobacteraceae; Epsilonproteobacteria; Subclass Aerobic-Microaerophilic, Motile Curved Gram-Negative Bacteria; Primary malignant neoplasm of large intestine
GSM516789	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Rectum and sigmoid colon; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Other and unspecified gastrointestinal disorders; Constipation; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Neoplasm Metastasis; Neoplastic Processes; Ductal Carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Endometrioid tumor; Malignant endometrioid tumor
GSM516790	GSE20565	Ovary	Breast	Ovary	Urethra; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Primary malignant neoplasm of prostate; Metastatic Carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; gastric fundus; Proximal stomach; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Mammary gland; [M]Adenocarcinoma, metastatic, NOS; Complex epithelial neoplasm; Papillary adenocarcinoma; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Hemoptysis; Respiratory tract hemorrhage; Urinary outflow structure; Skin tissue; Adenocarcinoma, Mucinous; Rectum and sigmoid colon; Neoplasm Metastasis; Neoplastic Processes; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Lactiferous duct

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516791	GSE20565	Ovary	Breast	Ovary	Urethra; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Papillary serous cystadenocarcinoma; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Adenosquamous carcinoma; Benign neoplasm of retroperitoneum; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Mammary gland; Metastatic Carcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; [M]Adenocarcinoma, metastatic, NOS; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Papillary adenocarcinoma; Endocrine tumor morphology; Complex epithelial neoplasm; Proximal stomach; gastric fundus; Rectum and sigmoid colon; Layer of adrenal gland; Endocrine gland part; Adrenal part; Adrenal Cortex; Urinary outflow structure; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Endometrium; Neoplasm Metastasis; Neoplastic Processes; Lactiferous duct; Mammary lobe
GSM516792	GSE20565	Ovary	Breast	Ovary	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Metastatic Carcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; [M]Adenocarcinoma, metastatic, NOS; Hemoptysis; Respiratory tract hemorrhage; Neoplasm Metastasis; Neoplastic Processes; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Adenocarcinoma, Mucinous; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Rectum and sigmoid colon; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Endometrium; Malignant neoplasm of female genital organ
GSM516793	GSE20565	Ovary	Breast	Ovary	Urethra; Benign neoplasm of intra-abdominal organs; Entire viscus; Benign neoplasm of adrenal gland; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Benign neoplasm of adrenal cortex; Entire fallopian tube; Adrenal Cortical Adenoma; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Benign neoplasm of retroperitoneum; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Hemoptysis; Respiratory tract hemorrhage; Complex epithelial neoplasm; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Transitional epithelial cell; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Neoplasm Metastasis; Neoplastic Processes; Other and unspecified gastrointestinal disorders; Constipation; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Endocrine tumor morphology; Adrenal mass; Tumors of Adrenal Cortex

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516794	GSE20565	Ovary	Breast	Ovary	Urethra; Papillary serous cystadenocarcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Papillary adenocarcinoma; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Other and unspecified gastrointestinal disorders; Constipation; Complex epithelial neoplasm; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Ductal Carcinoma; Neoplasm Metastasis; Neoplastic Processes; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Rectum and sigmoid colon; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Transitional epithelial cell; Malignant neoplasm of female genital organ
GSM516795	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Rectum and sigmoid colon; Ductal Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Neoplasm Metastasis; Neoplastic Processes; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Superior mediastinum; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures
GSM516796	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; [M]Adenocarcinoma, metastatic, NOS; Complex epithelial neoplasm; Hemoptysis; Respiratory tract hemorrhage; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Rectum and sigmoid colon; Metastatic Carcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Pluripotent Stem Cells; Endometrium; Transitional epithelial cell; Stromal Cells; Superior mediastinum; Mesenchymal Stem Cells; Ovary and/or broad ligament structures

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516797	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Complex epithelial neoplasm; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Other and unspecified gastrointestinal disorders; Constipation; Transitional epithelial cell; Neoplasm Metastasis; Neoplastic Processes; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Sense Organs; Nose; Rhinovirus infection; RNA Virus Infections; Picornaviridae Infections; Ductal Carcinoma; Superior mediastinum; Benign neoplasm of intra-abdominal organs
GSM516798	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Rectum and sigmoid colon; [M]Adenocarcinoma, metastatic, NOS; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Neoplasm Metastasis; Neoplastic Processes; Ductal Carcinoma; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Other and unspecified gastrointestinal disorders
GSM516799	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Complex epithelial neoplasm; [M]Adenocarcinoma, metastatic, NOS; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Superior mediastinum; Rectum and sigmoid colon; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Neoplasm Metastasis; Neoplastic Processes; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516800	GSE20565	Ovary	Breast	Ovary	Urethra; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Primary malignant neoplasm of prostate; gastric fundus; Proximal stomach; Metastatic Carcinoma; Mammary gland; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Complex epithelial neoplasm; Urinary outflow structure; [M]Adenocarcinoma, metastatic, NOS; Rectum and sigmoid colon; Lower urinary tract; Bladder and outflow structure; Pelvic cavity urinary structure; Prostatic and/or seminal vesicle structures; Minor pelvis; Male urinary outflow structure; Prostate and vas deferens structures; Prostate; Male internal genital organ; Pelvic cavity male genital structure; Adenocarcinoma, Mucinous; Benign neoplasm of other endocrine glands and related structures
GSM516801	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Papillary adenocarcinoma; Rectum and sigmoid colon; [M]Adenocarcinoma, metastatic, NOS; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Ductal Carcinoma; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Neoplasm Metastasis; Neoplastic Processes; Superior mediastinum; Other and unspecified gastrointestinal disorders; Constipation; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Gingival and periodontal disease NOS
GSM516802	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Metastatic Carcinoma; Complex epithelial neoplasm; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Hemoptysis; Respiratory tract hemorrhage; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Rectum and sigmoid colon; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Neoplasm Metastasis; Neoplastic Processes; Skin tissue; Endometrium; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Ovary and/or broad ligament structures

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516803	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Complex epithelial neoplasm; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Transitional epithelial cell; Neoplasm Metastasis; Neoplastic Processes; Other and unspecified gastrointestinal disorders; Constipation; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Ductal Carcinoma; Rectum and sigmoid colon; Endometrial Neoplasms
GSM516804	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Rectum and sigmoid colon; Papillary adenocarcinoma; Ductal Carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; [M]Adenocarcinoma, metastatic, NOS; Neoplasm Metastasis; Neoplastic Processes; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Adenocarcinoma, Mucinous; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Mammary gland; Disorder of soft tissue of body cavity; Disorder of soft tissue of head
GSM516805	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Metastatic Carcinoma; Complex epithelial neoplasm; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; [M]Adenocarcinoma, metastatic, NOS; Sense Organs; Nose; Rhinovirus infection; RNA Virus Infections; Picornaviridae Infections; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Rectum and sigmoid colon; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Hemoptysis; Respiratory tract hemorrhage; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Neoplasm Metastasis; Neoplastic Processes

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516806	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Hemoptysis; Respiratory tract hemorrhage; [M]Adenocarcinoma, metastatic, NOS; Rectum and sigmoid colon; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Skin tissue; Exanthema; Disorder of keratinization; Cell-mediated cytotoxic disorder; Cutaneous hypersensitivity; Acquired disorder of keratinization; Histologic type of inflammatory skin disorder; Psoriasis; Other psoriasis; Skin Diseases, Papulosquamous; Inflammatory hyperkeratotic dermatosis; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Other and unspecified gastrointestinal disorders; Constipation; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma
GSM516807	GSE20565	Ovary	Breast	Ovary	Urethra; Proximal stomach; Lactiferous duct; Mammary lobe; Glandular structure of breast; Adenosquamous carcinoma; Mammary gland; gastric fundus; Duct (organ) structure; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Metastatic Carcinoma; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Rectum and sigmoid colon; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Adenocarcinoma, Mucinous; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Skin tissue; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Stomach part; Region of stomach
GSM516808	GSE20565	Ovary	Ovary	NA	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Metastatic Carcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Rectum and sigmoid colon; Hemoptysis; Respiratory tract hemorrhage; [M]Adenocarcinoma, metastatic, NOS; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Skin tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Neoplasm Metastasis; Neoplastic Processes; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Mammary gland; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Ductal Carcinoma; ovarian neoplasm; Ovarian Diseases

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516809	GSE20565	Ovary	Breast	Ovary	Urethra; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Papillary adenocarcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; [M]Adenocarcinoma, metastatic, NOS; Complex epithelial neoplasm; Rectum and sigmoid colon; Other and unspecified gastrointestinal disorders; Constipation; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Neoplasm Metastasis; Neoplastic Processes; Ductal Carcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Skin tissue; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Transitional epithelial cell; Endometrioid tumor; Malignant endometrioid tumor
GSM516810	GSE20565	Ovary	Breast	Ovary	Urethra; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Primary malignant neoplasm of prostate; Metastatic Carcinoma; [M]Adenocarcinoma, metastatic, NOS; Complex epithelial neoplasm; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Papillary adenocarcinoma; Rectum and sigmoid colon; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Hemoptysis; Respiratory tract hemorrhage; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Neoplasm Metastasis; Neoplastic Processes; Mammary gland; Urinary outflow structure; Proximal stomach; gastric fundus; Endometrium; Skin tissue; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure
GSM516811	GSE20565	Ovary	Ovary	NA	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Rectum and sigmoid colon; Other and unspecified gastrointestinal disorders; Constipation; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; [M]Adenocarcinoma, metastatic, NOS; Ductal Carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Neoplasm Metastasis; Neoplastic Processes; Adenocarcinoma, Mucinous; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Superior mediastinum; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Coughing; Structure of pyloric portion of stomach

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516812	GSE20565	Ovary	Breast	Ovary	Urethra; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Primary malignant neoplasm of prostate; Metastatic Carcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; [M]Adenocarcinoma, metastatic, NOS; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Mammary gland; Papillary adenocarcinoma; Complex epithelial neoplasm; Carcinoma of genital organs NOS; Carcinoma of genitourinary organ; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Proximal stomach; gastric fundus; Urinary outflow structure; Rectum and sigmoid colon; Hemoptysis; Respiratory tract hemorrhage; Lower urinary tract; Bladder and outflow structure; Pelvic cavity urinary structure; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Neoplasm Metastasis
GSM516813	GSE20565	Ovary	Breast	Ovary	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; Rectum and sigmoid colon; [M]Adenocarcinoma, metastatic, NOS; Neoplasm Metastasis; Neoplastic Processes; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Ductal Carcinoma; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Gingival and periodontal disease NOS
GSM516814	GSE20565	Ovary	Breast	Ovary	Urethra; Entire viscus; Hollow viscus; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Abdominal organ; Metastatic Carcinoma; Entire fallopian tube; Hemoptysis; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Respiratory tract hemorrhage; Papillary adenocarcinoma; [M]Adenocarcinoma, metastatic, NOS; Rectum and sigmoid colon; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Endometrioid tumor; Malignant endometrioid tumor; Carcinoma, Endometrioid; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Neoplasm Metastasis; Neoplastic Processes; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Other and unspecified gastrointestinal disorders; Constipation; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Ductal Carcinoma; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM516815	GSE20565	Ovary	Breast	Ovary	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Hemoptysis; Adenosquamous carcinoma; Respiratory tract hemorrhage; Metastatic Carcinoma; Papillary adenocarcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; [M]Adenocarcinoma, metastatic, NOS; Rectum and sigmoid colon; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Ductal Carcinoma; Skin tissue; Neoplasm Metastasis; Neoplastic Processes; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Mammary gland; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases
GSM516816	GSE20565	Ovary	Breast	Ovary	Urethra; Benign neoplasm of intra-abdominal organs; Entire viscus; Benign neoplasm of adrenal gland; Hollow viscus; Papillary serous cystadenocarcinoma; Abdominal organ; Benign neoplasm of adrenal cortex; Entire fallopian tube; Adrenal Cortical Adenoma; Adenosquamous carcinoma; Entire pelvic organ; Entire female internal genital organ; Benign neoplasm of retroperitoneum; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Papillary adenocarcinoma; Metastatic Carcinoma; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; [M]Adenocarcinoma, metastatic, NOS; Primary malignant neoplasm of male genital organ; Prostate carcinoma; Primary malignant neoplasm of prostate; Complex epithelial neoplasm; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Hemoptysis; Respiratory tract hemorrhage; Endocrine tumor morphology; Adrenal mass; Tumors of Adrenal Cortex; Adrenal Cortex Diseases; Adrenal Gland Diseases; Adrenal Gland Neoplasms; Neoplasm Metastasis; Neoplastic Processes; Transitional epithelial cell; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Cancer of ovary and other female genital organs
GSM516817	GSE20565	Ovary	Ovary	NA	Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Stromal Cells; Complex epithelial neoplasm; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Hemoptysis; Respiratory tract hemorrhage; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; [M]Adenocarcinoma, metastatic, NOS; Mesenchymal Stem Cells; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Urethra; Hereditary disorder by system; Metastatic Carcinoma; Adrenal mass; Tumors of Adrenal Cortex; Adrenal Cortex Diseases; Adrenal Gland Diseases; Adrenal Gland Neoplasms; Bone structure of spine and/or pelvis; hip bone; Bone structure of ilium; Bone part; Ilium part; Iliac crest structure; Structure of flat bone; Bone structure of pelvic region and/or thigh; Bony pelvis; Structure of bone (organ); Type of bone

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM359477	GSE14378	Lung	Kidney	Lung	Urethra; White Adipose Tissue; Complex epithelial neoplasm; Subcutaneous Fat; Subcutaneous Tissue; Adenosquamous carcinoma; Papillary serous cystadenocarcinoma; Rectum and sigmoid colon; Metastatic Carcinoma; Stromal Cells; Adenocarcinoma, Mucinous; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Skin tissue; Hemoptysis; Respiratory tract hemorrhage; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Other and unspecified gastrointestinal disorders; Constipation; [M]Adenocarcinoma, metastatic, NOS; Mesenchymal Stem Cells; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw
GSM359478	GSE14378	Lung	Kidney	Lung	Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Stromal Cells; Complex epithelial neoplasm; Metastatic Carcinoma; Hemoptysis; Respiratory tract hemorrhage; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; [M]Adenocarcinoma, metastatic, NOS; Papillary adenocarcinoma; Mesenchymal Stem Cells; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Rectum and sigmoid colon; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Neoplasm Metastasis; Neoplastic Processes; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Transitional epithelial cell; Urethra; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Malignant neoplasm of female genital organ
GSM359479	GSE14378	Lung	Kidney	Lung	Urethra; White Adipose Tissue; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Subcutaneous Fat; Subcutaneous Tissue; Adenosquamous carcinoma; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; Metastatic Carcinoma; Rectum and sigmoid colon; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Adenocarcinoma, Mucinous; Stromal Cells; [M]Adenocarcinoma, metastatic, NOS; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Hemoptysis; Respiratory tract hemorrhage; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Mesenchymal Stem Cells; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Joint and/or tendon synovial structure; Synovial joint structure

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM359480	GSE14378	Lung	Kidney	Lung	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Metastatic Carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Hemoptysis; Respiratory tract hemorrhage; Rectum and sigmoid colon; Adenocarcinoma, Mucinous; Other and unspecified gastrointestinal disorders; Constipation; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Papillary adenocarcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum
GSM359481	GSE14378	Lung	Kidney	Lung	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Metastatic Carcinoma; Rectum and sigmoid colon; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; Proximal stomach; gastric fundus; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Adenocarcinoma, Mucinous; Other and unspecified gastrointestinal disorders; Constipation; Hemoptysis; Respiratory tract hemorrhage; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Stomach part; Region of stomach; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases
GSM359482	GSE14378	Lung	Kidney	Lung	Urethra; White Adipose Tissue; Papillary serous cystadenocarcinoma; Subcutaneous Fat; Subcutaneous Tissue; Adenosquamous carcinoma; Complex epithelial neoplasm; Metastatic Carcinoma; Rectum and sigmoid colon; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; Adenocarcinoma, Mucinous; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Proximal stomach; gastric fundus; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Hemoptysis; Respiratory tract hemorrhage; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Other and unspecified gastrointestinal disorders; Constipation; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Stomach part; Region of stomach; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Anorectal structure; Lower bowel structures

Continued on Next Page...

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM359483	GSE14378	Lung	Kidney	Lung	Other and unspecified gastrointestinal disorders; Urethra; Constipation; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Metastatic Carcinoma; Rectum and sigmoid colon; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Adenocarcinoma, Mucinous; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Hemoptysis; Respiratory tract hemorrhage; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Functional disorder of intestine; Skin tissue; Colonic Diseases, Functional; Irritable Bowel Syndrome; Papillary adenocarcinoma; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures
GSM359484	GSE14378	Lung	Kidney	Lung	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Adenosquamous carcinoma; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Hemoptysis; Respiratory tract hemorrhage; Rectum and sigmoid colon; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Adenocarcinoma, Mucinous; [M]Adenocarcinoma, metastatic, NOS; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Papillary adenocarcinoma; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Stromal Cells; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint
GSM359485	GSE14378	Lung	Kidney	Lung	Urethra; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Stromal Cells; Metastatic Carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Rectum and sigmoid colon; Hemoptysis; Respiratory tract hemorrhage; Mesenchymal Stem Cells; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Papillary adenocarcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Adenocarcinoma, Mucinous; Skin tissue; [M]Adenocarcinoma, metastatic, NOS; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Coughing; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM359486	GSE14378	Lung	Kidney	Lung	Urethra; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Adenosquamous carcinoma; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; gastric fundus; Proximal stomach; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Skin tissue; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Metastatic Carcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; Mammary gland; Adenocarcinoma, Mucinous; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Stomach part; Region of stomach; Lymph
GSM359487	GSE14378	Lung	Kidney	Lung	Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Papillary adenocarcinoma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Complex epithelial neoplasm; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Stromal Cells; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Metastatic Carcinoma; Hemoptysis; Respiratory tract hemorrhage; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; [M]Adenocarcinoma, metastatic, NOS; Rectum and sigmoid colon; Cancer of ovary and other female genital organs; Malignant neoplasm of ovary; Urethra; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; ovarian neoplasm; Ovarian Diseases; Gonadal Disorders; Neoplasm of uterine adnexa; Adnexal Diseases; Other and unspecified gastrointestinal disorders; Constipation; Malignant neoplasm of female genital organ; Malignant neoplasm of other and unspecified female genital organs; Mesenchymal Stem Cells
GSM359488	GSE14378	Lung	Kidney	Lung	Urethra; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Adenosquamous carcinoma; Maintenance chemotherapy; radiotherapy; mucosa-associated lymphoid tissue lymphoma; Chemotherapy Regimen; Metastatic Carcinoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Hemoptysis; Respiratory tract hemorrhage; Adenocarcinoma, Mucinous; Rectum and sigmoid colon; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Other and unspecified gastrointestinal disorders; Constipation; Papillary adenocarcinoma; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; Coughing; Proximal stomach; gastric fundus

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM359489	GSE14378	Lung	Kidney	Lung	Urethra; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Adenosquamous carcinoma; Metastatic Carcinoma; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Rectum and sigmoid colon; Skin tissue; Adenocarcinoma, Mucinous; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; gastric fundus; Proximal stomach; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Hemoptysis; Respiratory tract hemorrhage
GSM359490	GSE14378	Lung	Kidney	Lung	Urethra; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Structure of medulla of kidney; Adenosquamous carcinoma; Complex epithelial neoplasm; Papillary serous cystadenocarcinoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Rectum and sigmoid colon; Colonic Diseases, Functional; Irritable Bowel Syndrome; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; Metastatic Carcinoma; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Proximal stomach; gastric fundus; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Mammary gland; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Structure of layer of kidney; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Other and unspecified gastrointestinal disorders; Constipation; ileum; Structure of cortex of kidney; Adenocarcinoma, Mucinous; Malignant neoplasm of kidney; Skin tissue; Lymph
GSM359491	GSE14378	Lung	Kidney	Lung	Urethra; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Structure of medulla of kidney; Adenosquamous carcinoma; Papillary serous cystadenocarcinoma; gastric fundus; Proximal stomach; Complex epithelial neoplasm; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Adenocarcinoma, Mucinous; Rectum and sigmoid colon; Metastatic Carcinoma; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; Mammary gland; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Abdominal bloating

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM359492	GSE14378	Lung	Kidney	Lung	Adenocarcinoma of pelvis; Urethra; White Adipose Tissue; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Subcutaneous Fat; Subcutaneous Tissue; Structure of medulla of kidney; Adenosquamous carcinoma; Malignant tumor of kidney parenchyma; Complex epithelial neoplasm; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Rectum and sigmoid colon; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Adenocarcinoma, Mucinous; Proximal stomach; gastric fundus; Skin tissue; Malignant neoplasm of kidney; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Benign neoplasm of other endocrine glands and related structures; Benign tumor of endocrine gland; Mammary gland; Breast part; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ
GSM359493	GSE14378	Lung	Kidney	Lung	Urethra; White Adipose Tissue; Complex epithelial neoplasm; Subcutaneous Fat; Subcutaneous Tissue; Adenosquamous carcinoma; Papillary serous cystadenocarcinoma; Metastatic Carcinoma; Rectum and sigmoid colon; Stromal Cells; Adenocarcinoma, Mucinous; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Skin tissue; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Hemoptysis; Respiratory tract hemorrhage; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; Entire viscus; Hollow viscus; Abdominal organ; Entire fallopian tube; Entire pelvic organ; Entire female internal genital organ; Entire pelvic viscus; Entire female genital organ; Intra-abdominal genital structure; Abdominal bloating; Flatulence, eructation, and gas pain; [D]Gas pain (abdominal); Pain of digestive structure
GSM359494	GSE14378	Lung	Kidney	Lung	Urethra; White Adipose Tissue; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; Subcutaneous Fat; Subcutaneous Tissue; Adenosquamous carcinoma; mucosa-associated lymphoid tissue lymphoma; Papillary serous cystadenocarcinoma; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Metastatic Carcinoma; gastric fundus; Proximal stomach; Complex epithelial neoplasm; Stomach part; Region of stomach; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Adenocarcinoma, Mucinous; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Lymph; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Skin tissue; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma

Continued on Next Page. . .

Table C.2 – Continued

Sample ID	Series ID	Tissue	Pri	Met	Top 50 Concepts
GSM359495	GSE14378	Lung	Kidney	Lung	Urethra; Adenosquamous carcinoma; Papillary serous cystadenocarcinoma; Complex epithelial neoplasm; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Metastatic Carcinoma; Rectum and sigmoid colon; Skin tissue; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; mucosa-associated lymphoid tissue lymphoma; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Adenocarcinoma, Mucinous; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Other and unspecified gastrointestinal disorders; Constipation; Maintenance chemotherapy; radiotherapy; Chemotherapy Regimen; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; gastric fundus
GSM359496	GSE14378	Lung	Kidney	Lung	Urethra; Diffuse low grade B-cell lymphoma; Marginal Zone B-Cell Lymphoma; Adenosquamous carcinoma; mucosa-associated lymphoid tissue lymphoma; White Adipose Tissue; Subcutaneous Fat; Subcutaneous Tissue; Metastatic Carcinoma; Papillary serous cystadenocarcinoma; Complex epithelial neoplasm; Joint and/or tendon synovial structure; Synovial joint structure; Structure of synovial tissue of joint; Adenocarcinoma, Mucinous; Adenocarcinoma of pelvis; Primary malignant neoplasm of kidney; Renal glomerular disease; RENAL GLOMERULAR AND TUBULOINTERSTITIAL DISEASES; Renal Cell Carcinoma; Malignant tumor of kidney parenchyma; Structure of pyloric portion of stomach; Part of pyloric region; Pylorus; Proximal stomach; gastric fundus; Skin tissue; Gingival and periodontal disease NOS; Jaw Diseases; Inflammatory disorder of jaw; Inflammatory disorder of head; Disorder of teeth AND/OR supporting structures; Chronic disease of teeth AND/OR supporting structures; Chronic digestive system disorder; Disorder of face; Periodontal Diseases; Periodontitis; Benign neoplasm of intra-abdominal organs; Benign neoplasm of adrenal gland; Benign neoplasm of adrenal cortex; Adrenal Cortical Adenoma; Benign neoplasm of retroperitoneum; Stomach part; Region of stomach; Rectum and sigmoid colon; Lactiferous duct; Mammary lobe; Glandular structure of breast; Duct (organ) structure; Mammary gland

Appendix D

Marker genes their and over-enriched GO concepts

D.1 Over-enriched GO concepts for breast tissue marker genes

Table D.1:

GO ID	GO Term	P Value
GO:0048513	organ development	0
GO:0032502	developmental process	0
GO:0007275	multicellular organismal development	0
GO:0009888	tissue development	0
GO:0048856	anatomical structure development	0
GO:0032501	multicellular organismal process	0
GO:0048731	system development	0
GO:0022612	gland morphogenesis	0
GO:0060429	epithelium development	0
GO:0048729	tissue morphogenesis	0
GO:0060512	prostate gland morphogenesis	0
GO:0002009	morphogenesis of an epithelium	0
GO:0009887	organ morphogenesis	0
GO:0001763	morphogenesis of a branching structure	0
GO:0009725	response to hormone stimulus	0
GO:0061138	morphogenesis of a branching epithelium	0
GO:0035239	tube morphogenesis	0
GO:0048608	reproductive structure development	0
GO:0045444	fat cell differentiation	0
GO:0001655	urogenital system development	0
GO:0008544	epidermis development	0
GO:0060525	prostate glandular acinus development	0
GO:0009719	response to endogenous stimulus	0
GO:0030850	prostate gland development	0
GO:0009653	anatomical structure morphogenesis	0

Continued on Next Page...

Table D.1 – Continued

GO ID	GO Term	P Value
GO:0048732	gland development	0
GO:0060740	prostate gland epithelium morphogenesis	0.00001
GO:0043627	response to estrogen stimulus	0.00001
GO:0030855	epithelial cell differentiation	0.00001
GO:0032355	response to estradiol stimulus	0.00001
GO:0060562	epithelial tube morphogenesis	0.00001
GO:0048511	rhythmic process	0.00002
GO:0010033	response to organic substance	0.00002
GO:0003401	axis elongation	0.00002
GO:0008593	regulation of Notch signaling pathway	0.00002
GO:0010466	negative regulation of peptidase activity	0.00003
GO:0045137	development of primary sexual characteristics	0.00003
GO:0035282	segmentation	0.00003
GO:0051239	regulation of multicellular organismal process	0.00004
GO:0048545	response to steroid hormone stimulus	0.00005
GO:0060993	kidney morphogenesis	0.00005
GO:0003006	developmental process involved in reproduction	0.00006
GO:0035295	tube development	0.00007
GO:0045747	positive regulation of Notch signaling pathway	0.00007
GO:0070995	NADPH oxidation	0.00009
GO:0072086	specification of loop of Henle identity	0.00009
GO:0072272	proximal/distal pattern formation involved in metanephric nephron development	0.00009
GO:0060560	developmental growth involved in morphogenesis	0.00010
GO:2000026	regulation of multicellular organismal development	0.00010
GO:2000027	regulation of organ morphogenesis	0.00011
GO:0007548	sex differentiation	0.00011
GO:0010771	negative regulation of cell morphogenesis involved in differentiation	0.00014
GO:0009954	proximal/distal pattern formation	0.00014
GO:0018108	peptidyl-tyrosine phosphorylation	0.00014
GO:0022414	reproductive process	0.00015
GO:0046545	development of primary female sexual characteristics	0.00015
GO:0046546	development of primary male sexual characteristics	0.00015
GO:0048646	anatomical structure formation involved in morphogenesis	0.00015
GO:0000003	reproduction	0.00016
GO:0018212	peptidyl-tyrosine modification	0.00016
GO:0042221	response to chemical stimulus	0.00016
GO:0050673	epithelial cell proliferation	0.00018
GO:0016331	morphogenesis of embryonic epithelium	0.00018
GO:0060688	regulation of morphogenesis of a branching structure	0.00019
GO:0046660	female sex differentiation	0.00019
GO:0050730	regulation of peptidyl-tyrosine phosphorylation	0.00019
GO:0051346	negative regulation of hydrolase activity	0.00021
GO:0046661	male sex differentiation	0.00022
GO:0044057	regulation of system process	0.00023
GO:0006415	translational termination	0.00024
GO:0010647	positive regulation of cell communication	0.00025
GO:0007389	pattern specification process	0.00026
GO:0023056	positive regulation of signaling	0.00026
GO:0001649	osteoblast differentiation	0.00027
GO:0048807	female genitalia morphogenesis	0.00027
GO:0060648	mammary gland bud morphogenesis	0.00027
GO:0071481	cellular response to X-ray	0.00027
GO:0072047	proximal/distal pattern formation involved in nephron development	0.00027
GO:0072081	specification of nephron tubule identity	0.00027
GO:0072268	pattern specification involved in metanephros development	0.00027
GO:2000040	regulation of planar cell polarity pathway involved in axis elongation	0.00027
GO:2000041	negative regulation of planar cell polarity pathway involved in axis elongation	0.00027
GO:0048584	positive regulation of response to stimulus	0.00029

Continued on Next Page. . .

Table D.1 – Continued

GO ID	GO Term	P Value
GO:0052548	regulation of endopeptidase activity	0.00029
GO:0040007	growth	0.00030
GO:0030278	regulation of ossification	0.00031
GO:0010951	negative regulation of endopeptidase activity	0.00032
GO:0045927	positive regulation of growth	0.00033
GO:0001736	establishment of planar polarity	0.00037
GO:0044058	regulation of digestive system process	0.00037
GO:0072210	metanephric nephron development	0.00037
GO:0050793	regulation of developmental process	0.00037
GO:0071845	cellular component disassembly at cellular level	0.00037
GO:0052547	regulation of peptidase activity	0.00038
GO:0031667	response to nutrient levels	0.00038
GO:0048754	branching morphogenesis of a tube	0.00038
GO:0022411	cellular component disassembly	0.00040
GO:0031016	pancreas development	0.00041
GO:0048546	digestive tract morphogenesis	0.00042
GO:0007164	establishment of tissue polarity	0.00045
GO:0060572	morphogenesis of an epithelial bud	0.00045
GO:0072088	nephron epithelium morphogenesis	0.00045
GO:0006414	translational elongation	0.00045
GO:0043624	cellular protein complex disassembly	0.00046
GO:0043241	protein complex disassembly	0.00050
GO:0009967	positive regulation of signal transduction	0.00053
GO:0030154	cell differentiation	0.00053
GO:0008584	male gonad development	0.00053
GO:0048610	cellular process involved in reproduction	0.00053
GO:0043616	keratinocyte proliferation	0.00054
GO:0003402	planar cell polarity pathway involved in axis elongation	0.00055
GO:0060028	convergent extension involved in axis elongation	0.00055
GO:0061004	pattern specification involved in kidney development	0.00055
GO:0072048	renal system pattern specification	0.00055
GO:0072070	loop of Henle development	0.00055
GO:2000051	negative regulation of non-canonical Wnt receptor signaling pathway	0.00055
GO:0035148	tube formation	0.00056
GO:0008406	gonad development	0.00057
GO:0002064	epithelial cell development	0.00058
GO:0001503	ossification	0.00062
GO:0048468	cell development	0.00062
GO:0035019	somatic stem cell maintenance	0.00064
GO:0072028	nephron morphogenesis	0.00064
GO:0048565	digestive tract development	0.00064
GO:0009991	response to extracellular stimulus	0.00068
GO:0022602	ovulation cycle process	0.00068
GO:0045995	regulation of embryonic development	0.00069
GO:0034623	cellular macromolecular complex disassembly	0.00075
GO:0010165	response to X-ray	0.00076
GO:0060571	morphogenesis of an epithelial fold	0.00076
GO:0042127	regulation of cell proliferation	0.00079
GO:0032984	macromolecular complex disassembly	0.00081
GO:0006469	negative regulation of protein kinase activity	0.00081
GO:0001656	metanephros development	0.00083
GO:0061180	mammary gland epithelium development	0.00083
GO:0048869	cellular developmental process	0.00083
GO:0008283	cell proliferation	0.00085
GO:0072009	nephron epithelium development	0.00088
GO:0006928	cellular component movement	0.00090
GO:0030540	female genitalia development	0.00090
GO:2000095	regulation of Wnt receptor signaling pathway, planar cell polarity pathway	0.00090

Continued on Next Page...

Table D.1 – Continued

GO ID	GO Term	P Value
GO:0065008	regulation of biological quality	0.00094
GO:0031018	endocrine pancreas development	0.00097
GO:0042698	ovulation cycle	0.00098
GO:0072001	renal system development	0.00102
GO:0001738	morphogenesis of a polarized epithelium	0.00102
GO:0060445	branching involved in salivary gland morphogenesis	0.00102
GO:0033673	negative regulation of kinase activity	0.00106
GO:0022600	digestive system process	0.00107
GO:0055123	digestive system development	0.00110
GO:0009790	embryo development	0.00122
GO:0032101	regulation of response to external stimulus	0.00130
GO:0071478	cellular response to radiation	0.00134
GO:0010950	positive regulation of endopeptidase activity	0.00135
GO:0034695	response to prostaglandin E stimulus	0.00135
GO:0060526	prostate glandular acinus morphogenesis	0.00135
GO:0060527	prostate epithelial cord arborization involved in prostate glandular acinus morphogenesis	0.00135
GO:0090244	Wnt receptor signaling pathway involved in somitogenesis	0.00135
GO:2000050	regulation of non-canonical Wnt receptor signaling pathway	0.00135
GO:0051348	negative regulation of transferase activity	0.00142
GO:0048762	mesenchymal cell differentiation	0.00144
GO:0043434	response to peptide hormone stimulus	0.00146
GO:0035270	endocrine system development	0.00152
GO:0060603	mammary gland duct morphogenesis	0.00153
GO:0072073	kidney epithelium development	0.00153
GO:0043407	negative regulation of MAP kinase activity	0.00156
GO:0007155	cell adhesion	0.00160
GO:0022610	biological adhesion	0.00160
GO:0050873	brown fat cell differentiation	0.00172
GO:0003002	regionalization	0.00173
GO:0030879	mammary gland development	0.00185
GO:0002067	glandular epithelial cell differentiation	0.00187
GO:0009404	toxin metabolic process	0.00187
GO:0060174	limb bud formation	0.00187
GO:0060687	regulation of branching involved in prostate gland morphogenesis	0.00187
GO:0072079	nephron tubule formation	0.00187
GO:0090178	regulation of establishment of planar polarity involved in neural tube closure	0.00187
GO:0090179	planar cell polarity pathway involved in neural tube closure	0.00187
GO:0045667	regulation of osteoblast differentiation	0.00192
GO:0051094	positive regulation of developmental process	0.00204
GO:0048589	developmental growth	0.00209
GO:0022603	regulation of anatomical structure morphogenesis	0.00211
GO:0032103	positive regulation of response to external stimulus	0.00213
GO:0019080	viral genome expression	0.00225
GO:0019083	viral transcription	0.00225
GO:0007584	response to nutrient	0.00228
GO:0044092	negative regulation of molecular function	0.00231
GO:0048598	embryonic morphogenesis	0.00233
GO:0060485	mesenchyme development	0.00233
GO:0007435	salivary gland morphogenesis	0.00240
GO:0010719	negative regulation of epithelial to mesenchymal transition	0.00248
GO:0034694	response to prostaglandin stimulus	0.00248
GO:0060693	regulation of branching involved in salivary gland morphogenesis	0.00248
GO:0072078	nephron tubule morphogenesis	0.00248
GO:0090177	establishment of planar polarity involved in neural tube closure	0.00248
GO:0043405	regulation of MAP kinase activity	0.00250
GO:0016477	cell migration	0.00261
GO:0045595	regulation of cell differentiation	0.00271
GO:0007586	digestion	0.00279

Continued on Next Page...

Table D.1 – Continued

GO ID	GO Term	P Value
GO:0043193	positive regulation of gene-specific transcription	0.00280
GO:0034097	response to cytokine stimulus	0.00289
GO:0045596	negative regulation of cell differentiation	0.00290
GO:0035107	appendage morphogenesis	0.00291
GO:0035108	limb morphogenesis	0.00291
GO:0030307	positive regulation of cell growth	0.00298
GO:0043069	negative regulation of programmed cell death	0.00310
GO:0010470	regulation of gastrulation	0.00317
GO:0030916	otic vesicle formation	0.00317
GO:0050872	white fat cell differentiation	0.00317
GO:0060487	lung epithelial cell differentiation	0.00317
GO:0060513	prostatic bud formation	0.00317
GO:0061333	renal tubule morphogenesis	0.00317
GO:0071599	otic vesicle development	0.00317
GO:0071600	otic vesicle morphogenesis	0.00317
GO:0007431	salivary gland development	0.00323
GO:0019827	stem cell maintenance	0.00323
GO:0090263	positive regulation of canonical Wnt receptor signaling pathway	0.00323
GO:0010552	positive regulation of gene-specific transcription from RNA polymerase II promoter	0.00329
GO:0001838	embryonic epithelial tube formation	0.00354
GO:0019748	secondary metabolic process	0.00354
GO:0048736	appendage development	0.00359
GO:0060173	limb development	0.00359
GO:2000241	regulation of reproductive process	0.00359
GO:0009605	response to external stimulus	0.00368
GO:0072175	epithelial tube formation	0.00374
GO:0060548	negative regulation of cell death	0.00382
GO:0003208	cardiac ventricle morphogenesis	0.00386
GO:0051216	cartilage development	0.00389
GO:0042249	establishment of planar polarity of embryonic epithelium	0.00394
GO:0043508	negative regulation of JUN kinase activity	0.00394
GO:0060479	lung cell differentiation	0.00394
GO:0060601	lateral sprouting from an epithelium	0.00394
GO:0070741	response to interleukin-6	0.00394
GO:0045793	positive regulation of cell size	0.00395
GO:0050731	positive regulation of peptidyl-tyrosine phosphorylation	0.00395
GO:0043086	negative regulation of catalytic activity	0.00415
GO:0048638	regulation of developmental growth	0.00416
GO:0048864	stem cell development	0.00421
GO:0071214	cellular response to abiotic stimulus	0.00421
GO:0072006	nephron development	0.00421
GO:0051270	regulation of cellular component movement	0.00422
GO:0071900	regulation of protein serine/threonine kinase activity	0.00422
GO:0072358	cardiovascular system development	0.00428
GO:0072359	circulatory system development	0.00428
GO:0071901	negative regulation of protein serine/threonine kinase activity	0.00439
GO:0032569	gene-specific transcription from RNA polymerase II promoter	0.00447
GO:0060443	mammary gland morphogenesis	0.00457
GO:0070555	response to interleukin-1	0.00457
GO:0048870	cell motility	0.00459
GO:0051674	localization of cell	0.00459
GO:0007219	Notch signaling pathway	0.00462
GO:0030099	myeloid cell differentiation	0.00465
GO:0006111	regulation of gluconeogenesis	0.00479
GO:0031581	hemidesmosome assembly	0.00479
GO:0035112	genitalia morphogenesis	0.00479
GO:0046689	response to mercury ion	0.00479
GO:0050732	negative regulation of peptidyl-tyrosine phosphorylation	0.00479

Continued on Next Page. . .

Table D.1 – Continued

GO ID	GO Term	P Value
GO:0050930	induction of positive chemotaxis	0.00479
GO:0060742	epithelial cell differentiation involved in prostate gland development	0.00479
GO:0009913	epidermal cell differentiation	0.00486
GO:0008285	negative regulation of cell proliferation	0.00488
GO:0045598	regulation of fat cell differentiation	0.00495
GO:0001568	blood vessel development	0.00502
GO:0001822	kidney development	0.00507
GO:0060541	respiratory system development	0.00545
GO:0002076	osteoblast development	0.00571
GO:0060343	trabecula formation	0.00571
GO:0060602	branch elongation of an epithelium	0.00571
GO:0061383	trabecula morphogenesis	0.00571
GO:0046888	negative regulation of hormone secretion	0.00577
GO:0008585	female gonad development	0.00590
GO:0010212	response to ionizing radiation	0.00619
GO:0060349	bone morphogenesis	0.00620
GO:0010038	response to metal ion	0.00622
GO:0007178	transmembrane receptor protein serine/threonine kinase signaling pathway	0.00639
GO:0043067	regulation of programmed cell death	0.00660
GO:0035272	exocrine system development	0.00666
GO:0048145	regulation of fibroblast proliferation	0.00666
GO:0002065	columnar/cuboidal epithelial cell differentiation	0.00670
GO:0060442	branching involved in prostate gland morphogenesis	0.00670
GO:0048514	blood vessel morphogenesis	0.00674
GO:0051048	negative regulation of secretion	0.00709
GO:0002062	chondrocyte differentiation	0.00713
GO:0003231	cardiac ventricle development	0.00713
GO:0007044	cell-substrate junction assembly	0.00713
GO:0048144	fibroblast proliferation	0.00713
GO:0001944	vasculature development	0.00714
GO:0032868	response to insulin stimulus	0.00719
GO:0016049	cell growth	0.00730
GO:0014031	mesenchymal cell development	0.00740
GO:0006355	regulation of transcription, DNA-dependent	0.00746
GO:0010941	regulation of cell death	0.00752
GO:0016337	cell-cell adhesion	0.00760
GO:0030177	positive regulation of Wnt receptor signaling pathway	0.00763
GO:0048705	skeletal system morphogenesis	0.00767
GO:0003338	metanephros morphogenesis	0.00777
GO:0007379	segment specification	0.00777
GO:0010631	epithelial cell migration	0.00777
GO:0035121	tail morphogenesis	0.00777
GO:0060026	convergent extension	0.00777
GO:0060071	Wnt receptor signaling pathway, planar cell polarity pathway	0.00777
GO:0071479	cellular response to ionizing radiation	0.00777
GO:0072080	nephron tubule development	0.00777
GO:0090132	epithelium migration	0.00777
GO:0090175	regulation of establishment of planar polarity	0.00777
GO:0001756	somitogenesis	0.00815
GO:0030334	regulation of cell migration	0.00841
GO:0043066	negative regulation of apoptosis	0.00841
GO:0003206	cardiac chamber morphogenesis	0.00868
GO:0007267	cell-cell signaling	0.00876
GO:0051271	negative regulation of cellular component movement	0.00877
GO:0003151	outflow tract morphogenesis	0.00891
GO:0042517	positive regulation of tyrosine phosphorylation of Stat3 protein	0.00891
GO:0045600	positive regulation of fat cell differentiation	0.00891
GO:0048745	smooth muscle tissue development	0.00891

Continued on Next Page...

Table D.1 – Continued

GO ID	GO Term	P Value
GO:0061326	renal tubule development	0.00891
GO:0071453	cellular response to oxygen levels	0.00891
GO:0071456	cellular response to hypoxia	0.00891
GO:2000145	regulation of cell motility	0.00897
GO:0051093	negative regulation of developmental process	0.00920
GO:0060606	tube closure	0.00924
GO:0040011	locomotion	0.00961
GO:0001835	blastocyst hatching	0.00964
GO:0009957	epidermal cell fate specification	0.00964
GO:0010804	negative regulation of tumor necrosis factor-mediated signaling pathway	0.00964
GO:0021594	rhombomere formation	0.00964
GO:0021660	rhombomere 3 formation	0.00964
GO:0021664	rhombomere 5 morphogenesis	0.00964
GO:0021666	rhombomere 5 formation	0.00964
GO:0032605	hepatocyte growth factor production	0.00964
GO:0032646	regulation of hepatocyte growth factor production	0.00964
GO:0033210	leptin-mediated signaling pathway	0.00964
GO:0034115	negative regulation of heterotypic cell-cell adhesion	0.00964
GO:0034699	response to luteinizing hormone stimulus	0.00964
GO:0035188	hatching	0.00964
GO:0035690	cellular response to drug	0.00964
GO:0044343	canonical Wnt receptor signaling pathway involved in regulation of type B pancreatic cell proliferation	0.00964
GO:0044345	stromal-epithelial cell signaling involved in prostate gland development	0.00964
GO:0044346	fibroblast apoptosis	0.00964
GO:0045738	negative regulation of DNA repair	0.00964
GO:0048175	hepatocyte growth factor biosynthetic process	0.00964
GO:0048176	regulation of hepatocyte growth factor biosynthetic process	0.00964
GO:0048178	negative regulation of hepatocyte growth factor biosynthetic process	0.00964
GO:0050674	urothelial cell proliferation	0.00964
GO:0050675	regulation of urothelial cell proliferation	0.00964
GO:0050677	positive regulation of urothelial cell proliferation	0.00964
GO:0050902	leukocyte adhesive activation	0.00964
GO:0051040	regulation of calcium-independent cell-cell adhesion	0.00964
GO:0051041	positive regulation of calcium-independent cell-cell adhesion	0.00964
GO:0060432	lung pattern specification process	0.00964
GO:0060436	bronchiole morphogenesis	0.00964
GO:0060495	cell-cell signaling involved in lung development	0.00964
GO:0060496	mesenchymal-epithelial cell signaling involved in lung development	0.00964
GO:0060649	mammary gland bud elongation	0.00964
GO:0060659	nipple sheath formation	0.00964
GO:0060661	submandibular salivary gland formation	0.00964
GO:0060668	regulation of branching involved in salivary gland morphogenesis by extracellular matrix-epithelial cell signaling	0.00964
GO:0060741	prostate gland stromal morphogenesis	0.00964
GO:0060876	semicircular canal formation	0.00964
GO:0060879	semicircular canal fusion	0.00964
GO:0061115	lung proximal/distal axis specification	0.00964
GO:0070103	regulation of interleukin-6-mediated signaling pathway	0.00964
GO:0070104	negative regulation of interleukin-6-mediated signaling pathway	0.00964
GO:0070106	interleukin-27-mediated signaling pathway	0.00964
GO:0070346	positive regulation of fat cell proliferation	0.00964
GO:0070352	positive regulation of white fat cell proliferation	0.00964
GO:0070541	response to platinum ion	0.00964
GO:0071104	response to interleukin-9	0.00964
GO:0071105	response to interleukin-11	0.00964
GO:0071335	hair follicle cell proliferation	0.00964
GO:0071336	regulation of hair follicle cell proliferation	0.00964

Continued on Next Page. . .

Table D.1 – Continued

GO ID	GO Term	P Value
GO:0071338	positive regulation of hair follicle cell proliferation	0.00964
GO:0071684	organism emergence from protective structure	0.00964
GO:0071772	response to BMP stimulus	0.00964
GO:0071773	cellular response to BMP stimulus	0.00964
GO:0090245	axis elongation involved in somitogenesis	0.00964
GO:0090246	convergent extension involved in somitogenesis	0.00964
GO:2000035	regulation of stem cell division	0.00964
GO:2000079	regulation of canonical Wnt receptor signaling pathway involved in controlling type B pancreatic cell proliferation	0.00964
GO:2000080	negative regulation of canonical Wnt receptor signaling pathway involved in controlling type B pancreatic cell proliferation	0.00964
GO:2000269	regulation of fibroblast apoptosis	0.00964
GO:2000270	negative regulation of fibroblast apoptosis	0.00964
GO:2000271	positive regulation of fibroblast apoptosis	0.00964
GO:2000278	regulation of DNA biosynthetic process	0.00964
GO:2000279	negative regulation of DNA biosynthetic process	0.00964
GO:0008361	regulation of cell size	0.00973
GO:0050729	positive regulation of inflammatory response	0.00981
GO:0061053	somite development	0.00981

D.2 Over-enriched GO concepts for breast cancer marker genes

Table D.2:

GO ID	GO Term	P Value
GO:0035239	tube morphogenesis	0
GO:0035295	tube development	0
GO:0060562	epithelial tube morphogenesis	0
GO:0048754	branching morphogenesis of a tube	0
GO:0010677	negative regulation of cellular carbohydrate metabolic process	1.00E-05
GO:0045912	negative regulation of carbohydrate metabolic process	1.00E-05
GO:0006357	regulation of transcription from RNA polymerase II promoter	1.00E-05
GO:0001763	morphogenesis of a branching structure	2.00E-05
GO:0046546	development of primary male sexual characteristics	3.00E-05
GO:2000026	regulation of multicellular organismal development	3.00E-05
GO:0050793	regulation of developmental process	4.00E-05
GO:0046661	male sex differentiation	4.00E-05
GO:0060444	branching involved in mammary gland duct morphogenesis	4.00E-05
GO:0048731	system development	5.00E-05
GO:0002009	morphogenesis of an epithelium	5.00E-05
GO:0030539	male genitalia development	6.00E-05
GO:0048856	anatomical structure development	9.00E-05
GO:0045884	regulation of survival gene product expression	9.00E-05
GO:0048513	organ development	0.0001
GO:0033148	positive regulation of estrogen receptor signaling pathway	0.00011
GO:0061138	morphogenesis of a branching epithelium	0.00012
GO:0030520	estrogen receptor signaling pathway	0.00012
GO:0006366	transcription from RNA polymerase II promoter	0.00013
GO:0060603	mammary gland duct morphogenesis	0.00014
GO:0009725	response to hormone stimulus	0.00015
GO:0007548	sex differentiation	0.00018
GO:0033145	positive regulation of steroid hormone receptor signaling pathway	0.00018
GO:0048808	male genitalia morphogenesis	0.00018
GO:0060740	prostate gland epithelium morphogenesis	0.0002
GO:0048732	gland development	0.00021
GO:0060512	prostate gland morphogenesis	0.00022
GO:0048729	tissue morphogenesis	0.00024
GO:0048806	genitalia development	0.00025
GO:0010871	negative regulation of receptor biosynthetic process	0.00027
GO:0031953	negative regulation of protein autophosphorylation	0.00027
GO:0060745	mammary gland branching involved in pregnancy	0.00027
GO:0045595	regulation of cell differentiation	0.00027
GO:0001501	skeletal system development	0.00027
GO:0009719	response to endogenous stimulus	0.00032
GO:0007275	multicellular organismal development	0.00034
GO:0022612	gland morphogenesis	0.00039
GO:0003006	developmental process involved in reproduction	0.0004
GO:0030154	cell differentiation	0.00043
GO:0060443	mammary gland morphogenesis	0.00044
GO:0030500	regulation of bone mineralization	0.00048
GO:0008634	negative regulation of survival gene product expression	0.00049
GO:0001655	urogenital system development	0.00051
GO:0006629	lipid metabolic process	0.00052
GO:0048869	cellular developmental process	0.00059
GO:0030879	mammary gland development	0.0006
GO:0033146	regulation of estrogen receptor signaling pathway	0.00063

Continued on Next Page...

Table D.2 – Continued

GO ID	GO Term	P Value
GO:0009887	organ morphogenesis	0.00064
GO:0032502	developmental process	0.00065
GO:0070167	regulation of biomineral tissue development	0.00065
GO:0030278	regulation of ossification	0.00071
GO:0045137	development of primary sexual characteristics	0.00071
GO:0030850	prostate gland development	0.00075
GO:0009888	tissue development	0.00078
GO:0060736	prostate gland growth	0.00079
GO:0061180	mammary gland epithelium development	0.00086
GO:2000112	regulation of cellular macromolecule biosynthetic process	0.00091
GO:0010906	regulation of glucose metabolic process	0.00092
GO:0060429	epithelium development	0.00095
GO:0035112	genitalia morphogenesis	0.00096
GO:0060525	prostate glandular acinus development	0.00096
GO:0060742	epithelial cell differentiation involved in prostate gland development	0.00096
GO:0051239	regulation of multicellular organismal process	0.00106
GO:0009653	anatomical structure morphogenesis	0.00109
GO:0030730	sequestering of triglyceride	0.00115
GO:0010556	regulation of macromolecule biosynthetic process	0.00121
GO:0006109	regulation of carbohydrate metabolic process	0.00133
GO:0010675	regulation of cellular carbohydrate metabolic process	0.00133
GO:0051171	regulation of nitrogen compound metabolic process	0.00134
GO:0010745	negative regulation of macrophage derived foam cell differentiation	0.00135
GO:0010869	regulation of receptor biosynthetic process	0.00135
GO:0060749	mammary gland alveolus development	0.00135
GO:0061377	mammary gland lobule development	0.00135
GO:0001503	ossification	0.00146
GO:0022603	regulation of anatomical structure morphogenesis	0.0015
GO:0030282	bone mineralization	0.00157
GO:0060135	maternal process involved in female pregnancy	0.00157
GO:0080090	regulation of primary metabolic process	0.00173
GO:0043401	steroid hormone mediated signaling pathway	0.00181
GO:0006355	regulation of transcription, DNA-dependent	0.0019
GO:0034339	regulation of transcription from RNA polymerase II promoter by nuclear hormone receptor	0.00192
GO:0016042	lipid catabolic process	0.00194
GO:0031952	regulation of protein autophosphorylation	0.00207
GO:0031323	regulation of cellular metabolic process	0.0021
GO:0045449	regulation of transcription	0.00224
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	0.00231
GO:0032800	receptor biosynthetic process	0.00233
GO:0045599	negative regulation of fat cell differentiation	0.00233
GO:0019219	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	0.00238
GO:0031326	regulation of cellular biosynthetic process	0.00244
GO:0051252	regulation of RNA metabolic process	0.00257
GO:0010551	regulation of gene-specific transcription from RNA polymerase II promoter	0.00257
GO:0045893	positive regulation of transcription, DNA-dependent	0.0026
GO:0032569	gene-specific transcription from RNA polymerase II promoter	0.00265
GO:0009889	regulation of biosynthetic process	0.00269
GO:0019216	regulation of lipid metabolic process	0.00275
GO:0051254	positive regulation of RNA metabolic process	0.00283
GO:0032868	response to insulin stimulus	0.00304
GO:0008584	male gonad development	0.00316
GO:0019222	regulation of metabolic process	0.0032
GO:0010628	positive regulation of gene expression	0.00334
GO:0006916	anti-apoptosis	0.0034
GO:2000113	negative regulation of cellular macromolecule biosynthetic process	0.00353
GO:0031214	biomineral tissue development	0.00357
GO:0010552	positive regulation of gene-specific transcription from RNA polymerase II promoter	0.0038

Continued on Next Page...

Table D.2 – Continued

GO ID	GO Term	P Value
GO:0010033	response to organic substance	0.0039
GO:0042551	neuron maturation	0.00391
GO:0007399	nervous system development	0.00391
GO:0048598	embryonic morphogenesis	0.00397
GO:0030182	neuron differentiation	0.00407
GO:0048469	cell maturation	0.00415
GO:0045596	negative regulation of cell differentiation	0.00419
GO:0007497	posterior midgut development	0.00427
GO:0010804	negative regulation of tumor necrosis factor-mediated signaling pathway	0.00427
GO:0019102	male somatic sex determination	0.00427
GO:0021506	anterior neuropore closure	0.00427
GO:0021995	neuropore closure	0.00427
GO:0032788	saturated monocarboxylic acid metabolic process	0.00427
GO:0032789	unsaturated monocarboxylic acid metabolic process	0.00427
GO:0034115	negative regulation of heterotypic cell-cell adhesion	0.00427
GO:0035690	cellular response to drug	0.00427
GO:0060514	prostate induction	0.00427
GO:0060520	activation of prostate induction by androgen receptor signaling pathway	0.00427
GO:0060741	prostate gland stromal morphogenesis	0.00427
GO:0072363	regulation of glycolysis by positive regulation of transcription from an RNA polymerase II promoter	0.00427
GO:0072366	regulation of cellular ketone metabolic process by positive regulation of transcription from an RNA polymerase II promoter	0.00427
GO:0072369	regulation of lipid transport by positive regulation of transcription from an RNA polymerase II promoter	0.00427
GO:2000278	regulation of DNA biosynthetic process	0.00427
GO:2000279	negative regulation of DNA biosynthetic process	0.00427
GO:0008209	androgen metabolic process	0.00427
GO:0010558	negative regulation of macromolecule biosynthetic process	0.00441
GO:2000027	regulation of organ morphogenesis	0.00463
GO:0045923	positive regulation of fatty acid metabolic process	0.00465
GO:0060255	regulation of macromolecule metabolic process	0.00471
GO:0033143	regulation of steroid hormone receptor signaling pathway	0.00504
GO:0050873	brown fat cell differentiation	0.00504
GO:0048545	response to steroid hormone stimulus	0.00518
GO:0031327	negative regulation of cellular biosynthetic process	0.00523
GO:0032501	multicellular organismal process	0.00525
GO:0006350	transcription	0.00539
GO:0010743	regulation of macrophage derived foam cell differentiation	0.00544
GO:0030518	steroid hormone receptor signaling pathway	0.0055
GO:0032583	regulation of gene-specific transcription	0.0058
GO:0009890	negative regulation of biosynthetic process	0.00582
GO:0016331	morphogenesis of embryonic epithelium	0.00587
GO:0006351	transcription, DNA-dependent	0.00598
GO:0048699	generation of neurons	0.00601
GO:0032774	RNA biosynthetic process	0.0061
GO:0045444	fat cell differentiation	0.00625
GO:0045776	negative regulation of blood pressure	0.0063
GO:0010742	macrophage derived foam cell differentiation	0.00674
GO:0090077	foam cell differentiation	0.00674
GO:0060688	regulation of morphogenesis of a branching structure	0.00721
GO:0022414	reproductive process	0.0076
GO:0000003	reproduction	0.00787
GO:0045941	positive regulation of transcription	0.0081
GO:0043255	regulation of carbohydrate biosynthetic process	0.00817
GO:0060284	regulation of cell development	0.00822
GO:0048608	reproductive structure development	0.0084
GO:0006710	androgen catabolic process	0.00851

Continued on Next Page...

Table D.2 – Continued

GO ID	GO Term	P Value
GO:0010803	regulation of tumor necrosis factor-mediated signaling pathway	0.00851
GO:0018993	somatic sex determination	0.00851
GO:0030505	inorganic diphosphate transport	0.00851
GO:0031444	slow-twitch skeletal muscle fiber contraction	0.00851
GO:0032275	luteinizing hormone secretion	0.00851
GO:0033034	positive regulation of myeloid cell apoptosis	0.00851
GO:0033211	adiponectin-mediated signaling pathway	0.00851
GO:0045719	negative regulation of glycogen biosynthetic process	0.00851
GO:0045820	negative regulation of glycolysis	0.00851
GO:0048386	positive regulation of retinoic acid receptor signaling pathway	0.00851
GO:0060599	lateral sprouting involved in mammary gland duct morphogenesis	0.00851
GO:0060738	epithelial-mesenchymal signaling involved in prostate gland development	0.00851
GO:0072361	regulation of glycolysis by regulation of transcription from an RNA polymerase II promoter	0.00851
GO:0072364	regulation of cellular ketone metabolic process by regulation of transcription from an RNA polymerase II promoter	0.00851
GO:0072367	regulation of lipid transport by regulation of transcription from an RNA polymerase II promoter	0.00851
GO:0022008	neurogenesis	0.0086
GO:0030324	lung development	0.00865
GO:0009755	hormone-mediated signaling pathway	0.00868
GO:0046324	regulation of glucose import	0.00868
GO:0010468	regulation of gene expression	0.00869
GO:0045664	regulation of neuron differentiation	0.00871
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	0.00876
GO:0050772	positive regulation of axonogenesis	0.00919
GO:0030323	respiratory tube development	0.00938
GO:0030522	intracellular receptor mediated signaling pathway	0.00938
GO:0051093	negative regulation of developmental process	0.00942
GO:0043193	positive regulation of gene-specific transcription	0.00952
GO:0048468	cell development	0.00958
GO:0043467	regulation of generation of precursor metabolites and energy	0.00972

D.3 Stem cell marker genes

Although there are a total of 189 genes in the stem cell marker gene set, they have been broken down into four main functional groups: “DNA replication / cell cycle,” “RNA transcription / protein synthesis,” “metabolism / hormone signaling / protein synthesis,” and “ multicellular signaling / immune signaling / cell identity.”

D.3.1 Genes in the DNA replication / cell cycle module

Table D.3:

Gene Name	Gene ID	Score	Binomial P Value	Percentile
DNMT3B	1789	0.508379888	2.94E-61	0.00296267
MCM6	4175	0.51396648	1.62E-62	0.002666403
CDC25A	993	0.525139665	4.62E-65	0.002024491
PFAS	5198	0.525139665	4.62E-65	0.002024491
MCM4	4173	0.452513966	3.30E-49	0.008641122
XRCC5	7520	0.480446927	4.11E-55	0.005184673
HAUS6	54801	0.458100559	2.28E-50	0.007406676
TET1	80312	0.458100559	2.28E-50	0.007406676
IGF2BP1	10642	0.541899441	5.95E-69	0.001580091
PLAA	9373	0.469273743	1.01E-52	0.006270986
DEPDC1B	55789	0.458100559	2.28E-50	0.007406676
TEX10	54881	0.458100559	2.28E-50	0.007406676
CCDC99	54908	0.558659218	6.26E-73	0.001234446
MSH2	4436	0.480446927	4.11E-55	0.005184673
BUB1B	701	0.480446927	4.11E-55	0.005184673
MSH6	2956	0.463687151	1.53E-51	0.007011653
DLGAP5	9787	0.491620112	1.53E-57	0.004147738
SKIV2L2	23517	0.469273743	1.01E-52	0.006270986
CENPE	1062	0.474860335	6.52E-54	0.005629074
CHEK2	11200	0.525139665	4.62E-65	0.002024491
SOHLH2	54937	0.603351955	5.68E-84	0.000345645
CCNB1	891	0.458100559	2.28E-50	0.007406676
RRAS2	22800	0.581005587	2.26E-78	0.000641912
PRIM1	5557	0.474860335	6.52E-54	0.005629074
PAICS	10606	0.469273743	1.01E-52	0.006270986
CCNA2	890	0.497206704	9.02E-59	0.003703338
CPSF3	51692	0.474860335	6.52E-54	0.005629074
NUSAP1	51203	0.469273743	1.01E-52	0.006270986
LIN28B	389421	0.502793296	5.21E-60	0.00320956
IPO5	3843	0.525139665	4.62E-65	0.002024491
KIF11	3832	0.48603352	2.54E-56	0.004690895
BMPR1A	657	0.452513966	3.30E-49	0.008641122
NDC80	10403	0.491620112	1.53E-57	0.004147738
BCAT1	586	0.519553073	8.75E-64	0.002419514
CCNG1	900	0.508379888	2.94E-61	0.00296267
ZNF788	388507	0.469273743	1.01E-52	0.006270986
ASCC3	10973	0.452513966	3.30E-49	0.008641122
FANCB	2187	0.458100559	2.28E-50	0.007406676
MCM10	55388	0.525139665	4.62E-65	0.002024491
HMGA2	8091	0.469273743	1.01E-52	0.006270986
SKP2	6502	0.469273743	1.01E-52	0.006270986

Continued on Next Page...

Table D.3 – Continued

Gene Name	Gene ID	Score	Binomial P Value	Percentile
TRIM24	8805	0.541899441	5.95E-69	0.001580091
ORC1L	4998	0.480446927	4.11E-55	0.005184673
HDAC2	3066	0.458100559	2.28E-50	0.007406676
HESX1	8820	0.480446927	4.11E-55	0.005184673
C1orf135	79000	0.51396648	1.62E-62	0.002666403
INHBE	83729	0.497206704	9.02E-59	0.003703338
C21orf45	54069	0.463687151	1.53E-51	0.007011653
DCUN1D5	84259	0.463687151	1.53E-51	0.007011653
POLE2	5427	0.48603352	2.54E-56	0.004690895
MRPL3	11222	0.469273743	1.01E-52	0.006270986
CENPH	64946	0.463687151	1.53E-51	0.007011653
MYCN	4613	0.458100559	2.28E-50	0.007406676
HAUS1	115106	0.474860335	6.52E-54	0.005629074
GDF3	9573	0.458100559	2.28E-50	0.007406676

D.3.2 Stem cell genes in the RNA transcription / protein synthesis module

Table D.4:

Gene Name	Gene ID	Score	Binomial P Value	Percentile
TBCE	6905	0.491620112	1.53E-57	0.004147738
RIOK2	55781	0.597765363	1.48E-82	0.000395023
BCKDHB	594	0.458100559	2.28E-50	0.007406676
RAD1	5810	0.458100559	2.28E-50	0.007406676
C5orf13	9315	0.458100559	2.28E-50	0.007406676
ADH5	128	0.648044693	1.16E-95	0.000197511
PLRG1	5356	0.519553073	8.75E-64	0.002419514
ROR1	4919	0.670391061	9.24E-102	4.94E-05
RAB3B	5865	0.553072626	1.36E-71	0.001431957
LOC285431	285431	0.491620112	1.53E-57	0.004147738
DBC1	1620	0.48603352	2.54E-56	0.004690895
KIF23	9493	0.452513966	3.30E-49	0.008641122
DIAPH3	81624	0.502793296	5.21E-60	0.00320956
GNL2	29889	0.491620112	1.53E-57	0.004147738
FGF2	2247	0.681564246	7.10E-105	0
TARDBP	23435	0.458100559	2.28E-50	0.007406676
NMNAT2	23057	0.452513966	3.30E-49	0.008641122
ZNF167	55888	0.491620112	1.53E-57	0.004147738
KIF20A	10112	0.463687151	1.53E-51	0.007011653
CENPI	2491	0.480446927	4.11E-55	0.005184673
DDX1	1653	0.469273743	1.01E-52	0.006270986
C3orf21	152002	0.525139665	4.62E-65	0.002024491
GPR176	11245	0.664804469	3.21E-100	9.88E-05
FBXO22	26263	0.469273743	1.01E-52	0.006270986
BBS9	27241	0.51396648	1.62E-62	0.002666403
C14orf166	51637	0.541899441	5.95E-69	0.001580091
BOD1	91272	0.519553073	8.75E-64	0.002419514
CDC123	8872	0.469273743	1.01E-52	0.006270986
SNRPD3	6634	0.502793296	5.21E-60	0.00320956
FAM118B	79607	0.56424581	2.82E-74	0.000987557
DPH3	285381	0.474860335	6.52E-54	0.005629074
EIF2B3	8891	0.469273743	1.01E-52	0.006270986
KDELC1	79070	0.586592179	9.33E-80	0.000543156
RPF2	84154	0.458100559	2.28E-50	0.007406676
APLP1	333	0.474860335	6.52E-54	0.005629074
DACT1	51339	0.536312849	1.20E-67	0.001777602
PDHB	5162	0.586592179	9.33E-80	0.000543156
C14orf119	55017	0.575418994	5.37E-77	0.000790045
DTD1	92675	0.469273743	1.01E-52	0.006270986
SAMM50	25813	0.497206704	9.02E-59	0.003703338
CCL26	10344	0.491620112	1.53E-57	0.004147738
C4orf52	389203	0.458100559	2.28E-50	0.007406676
CCDC90B	60492	0.458100559	2.28E-50	0.007406676
MED20	9477	0.56424581	2.82E-74	0.000987557
UTP6	55813	0.469273743	1.01E-52	0.006270986
RARS2	57038	0.458100559	2.28E-50	0.007406676
KIAA0020	9933	0.474860335	6.52E-54	0.005629074
ARMCX2	9823	0.569832402	1.25E-75	0.000839423
RARS	5917	0.491620112	1.53E-57	0.004147738
MTHFD2	10797	0.469273743	1.01E-52	0.006270986
DHX15	1665	0.452513966	3.30E-49	0.008641122
HTR7	3363	0.558659218	6.26E-73	0.001234446

Continued on Next Page. . .

Table D.4 – Continued

Gene Name	Gene ID	Score	Binomial P Value	Percentile
HIST1H4C	8364	0.48603352	2.54E-56	0.004690895

D.3.3 Genes in the metabolism / hormone signaling / protein synthesis module

Table D.5:

Gene Name	Gene ID	Score	Binomial P Value	Percentile
MTHFD1L	25902	0.541899441	5.95E-69	0.001580091
ARMC9	80210	0.569832402	1.25E-75	0.000839423
XPOT	11260	0.51396648	1.62E-62	0.002666403
IARS	3376	0.497206704	9.02E-59	0.003703338
HDX	139324	0.56424581	2.82E-74	0.000987557
ARPM1	84517	0.530726257	2.39E-66	0.001925736
ERCC2	2068	0.458100559	2.28E-50	0.007406676
TBC1D16	125058	0.452513966	3.30E-49	0.008641122
GARS	2617	0.497206704	9.02E-59	0.003703338
KIF7	374654	0.61452514	7.83E-87	0.000296267
UBE2K	3093	0.508379888	2.94E-61	0.00296267
SLC25A3	5250	0.48603352	2.54E-56	0.004690895
ICMT	23463	0.530726257	2.39E-66	0.001925736
UGGT2	55757	0.48603352	2.54E-56	0.004690895
ATP11C	286410	0.48603352	2.54E-56	0.004690895
SLC24A1	9187	0.497206704	9.02E-59	0.003703338
EIF2AK4	440275	0.474860335	6.52E-54	0.005629074
GPX8	493869	0.491620112	1.53E-57	0.004147738
ALX1	8092	0.51396648	1.62E-62	0.002666403
OSTC	58505	0.525139665	4.62E-65	0.002024491
TRPC4	7223	0.458100559	2.28E-50	0.007406676
HAS2	3037	0.51396648	1.62E-62	0.002666403
FZD2	2535	0.452513966	3.30E-49	0.008641122
TRNT1	51095	0.519553073	8.75E-64	0.002419514
MMADHC	27249	0.536312849	1.20E-67	0.001777602
SNX8	29886	0.502793296	5.21E-60	0.00320956
CDH6	1004	0.458100559	2.28E-50	0.007406676
HAT1	8520	0.458100559	2.28E-50	0.007406676
SEC11A	23478	0.519553073	8.75E-64	0.002419514
DIMT1L	27292	0.452513966	3.30E-49	0.008641122
TM2D2	83877	0.452513966	3.30E-49	0.008641122
FST	10468	0.536312849	1.20E-67	0.001777602
GBE1	2632	0.480446927	4.11E-55	0.005184673

D.3.4 Genes in the multicellular signaling / immune signaling / cell identity module

Table D.6:

Gene Name	Gene ID	Score	Binomial P Value	Percentile
NA	80047	0.452513966	3.30E-49	0.008641122
MLL3	58508	0.508379888	2.94E-61	0.00296267
MXI1	4601	0.480446927	4.11E-55	0.005184673
FKSG49	400949	0.569832402	1.25E-75	0.000839423
FAM185B	641808	0.48603352	2.54E-56	0.004690895
ARRB2	409	0.56424581	2.82E-74	0.000987557
SMARCC2	6601	0.497206704	9.02E-59	0.003703338
WASH3P	374666	0.491620112	1.53E-57	0.004147738
PILRB	29990	0.463687151	1.53E-51	0.007011653
CTSH	1512	0.48603352	2.54E-56	0.004690895
SAT1	6303	0.553072626	1.36E-71	0.001431957
JUNB	3726	0.452513966	3.30E-49	0.008641122
CD53	963	0.508379888	2.94E-61	0.00296267
PECAM1	5175	0.597765363	1.48E-82	0.000395023
IL10RA	3587	0.502793296	5.21E-60	0.00320956
RCSD1	92241	0.452513966	3.30E-49	0.008641122
ARHGDIB	397	0.452513966	3.30E-49	0.008641122
GIMAP5	55340	0.581005587	2.26E-78	0.000641912
GIMAP6	474344	0.474860335	6.52E-54	0.005629074
HLA-DMB	3109	0.597765363	1.48E-82	0.000395023
PTPRC	5788	0.502793296	5.21E-60	0.00320956
C10orf128	170371	0.502793296	5.21E-60	0.00320956
CMBL	134147	0.474860335	6.52E-54	0.005629074
HLA-DRB5	3127	0.558659218	6.26E-73	0.001234446
HLA-DPA1	3113	0.558659218	6.26E-73	0.001234446
ABCG1	9619	0.642458101	3.65E-94	0.000246889
GIMAP7	168537	0.480446927	4.11E-55	0.005184673
HLA-DQA1	3117	0.502793296	5.21E-60	0.00320956
TSHZ2	128553	0.463687151	1.53E-51	0.007011653
C13orf15	28984	0.502793296	5.21E-60	0.00320956
CCR1	1230	0.502793296	5.21E-60	0.00320956
NPR3	4883	0.458100559	2.28E-50	0.007406676
RSAD2	91543	0.491620112	1.53E-57	0.004147738
GIMAP1	170575	0.474860335	6.52E-54	0.005629074
TNFSF10	8743	0.497206704	9.02E-59	0.003703338
AFTPH	54812	0.581005587	2.26E-78	0.000641912
NA	643187	0.458100559	2.28E-50	0.007406676
MALAT1	378938	0.497206704	9.02E-59	0.003703338
UBXN2A	165324	0.463687151	1.53E-51	0.007011653
PDE4C	5143	0.56424581	2.82E-74	0.000987557
GIMAP8	155038	0.474860335	6.52E-54	0.005629074
FYB	2533	0.547486034	2.87E-70	0.001530713
MS4A7	58475	0.525139665	4.62E-65	0.002024491
C5orf56	441108	0.458100559	2.28E-50	0.007406676
LOC400931	400931	0.474860335	6.52E-54	0.005629074
MLLT6	4302	0.664804469	3.21E-100	9.88E-05
CTSS	1520	0.48603352	2.54E-56	0.004690895
ZBTB20	26137	0.458100559	2.28E-50	0.007406676

D.3.5 GO terms associated with the DNA replication / cell cycle module

Table D.7:

GO ID	P Value	GO Term
GO:0000280	7.52E-14	nuclear division
GO:0007067	7.52E-14	mitosis
GO:0048285	1.22E-13	organelle fission
GO:0000087	1.28E-13	M phase of mitotic cell cycle
GO:0022403	3.70E-13	cell cycle phase
GO:0000279	1.26E-12	M phase
GO:0000278	1.92E-12	mitotic cell cycle
GO:0022402	2.78E-12	cell cycle process
GO:0051301	3.40E-12	cell division
GO:0007049	3.88E-12	cell cycle
GO:0000070	6.02E-09	mitotic sister chromatid segregation
GO:0000819	7.13E-09	sister chromatid segregation
GO:0000226	2.29E-08	microtubule cytoskeleton organization
GO:0006996	4.19E-08	organelle organization
GO:0007059	6.75E-08	chromosome segregation
GO:0007051	7.94E-08	spindle organization
GO:0051276	8.06E-08	chromosome organization
GO:0000075	1.92E-07	cell cycle checkpoint
GO:0051656	3.08E-07	establishment of organelle localization
GO:0050000	4.99E-07	chromosome localization
GO:0051303	4.99E-07	establishment of chromosome localization
GO:0051726	9.53E-07	regulation of cell cycle
GO:0007017	1.09E-06	microtubule-based process
GO:0007093	1.63E-06	mitotic cell cycle checkpoint
GO:0051640	1.78E-06	organelle localization
GO:0006259	1.81E-06	DNA metabolic process
GO:0008608	3.22E-06	attachment of spindle microtubules to kinetochore
GO:0051313	3.22E-06	attachment of spindle microtubules to chromosome
GO:0007346	4.21E-06	regulation of mitotic cell cycle
GO:0040001	4.82E-06	establishment of mitotic spindle localization
GO:0006261	9.11E-06	DNA-dependent DNA replication
GO:0007080	9.42E-06	mitotic metaphase plate congression
GO:0051293	9.42E-06	establishment of spindle localization
GO:0051653	9.42E-06	spindle localization
GO:0007079	1.53E-05	mitotic chromosome movement towards spindle pole
GO:0051984	1.53E-05	positive regulation of chromosome segregation
GO:0051987	1.53E-05	positive regulation of attachment of spindle microtubules to kinetochore
GO:0051329	1.58E-05	interphase of mitotic cell cycle
GO:0051310	1.62E-05	metaphase plate congression
GO:0051325	2.26E-05	interphase
GO:0034453	2.57E-05	microtubule anchoring
GO:0010564	3.29E-05	regulation of cell cycle process
GO:0010638	3.35E-05	positive regulation of organelle organization
GO:0006260	3.41E-05	DNA replication
GO:0006189	4.59E-05	’de novo’ IMP biosynthetic process
GO:0045842	4.59E-05	positive regulation of mitotic metaphase/anaphase transition
GO:0051305	4.59E-05	chromosome movement towards spindle pole
GO:0051988	4.59E-05	regulation of attachment of spindle microtubules to kinetochore
GO:0042770	5.20E-05	DNA damage response, signal transduction
GO:0070925	6.40E-05	organelle assembly
GO:0007052	7.38E-05	mitotic spindle organization
GO:0000077	8.44E-05	DNA damage checkpoint

Continued on Next Page...

Table D.7 – Continued

GO ID	P Value	GO Term
GO:0045840	8.53E-05	positive regulation of mitosis
GO:0051225	8.53E-05	spindle assembly
GO:0051785	8.53E-05	positive regulation of nuclear division
GO:0006188	9.16E-05	IMP biosynthetic process
GO:0046040	9.16E-05	IMP metabolic process
GO:0031570	0.000102493	DNA integrity checkpoint
GO:0006270	0.000126262	DNA-dependent DNA replication initiation
GO:0045787	0.000138788	positive regulation of cell cycle
GO:0007095	0.000152304	mitotic cell cycle G2/M transition DNA damage checkpoint
GO:0034501	0.000152304	protein localization to kinetochore
GO:0043570	0.000152304	maintenance of DNA repeat elements
GO:0051096	0.000152304	positive regulation of helicase activity
GO:0071780	0.000152304	mitotic cell cycle G2/M transition checkpoint
GO:0007010	0.000158535	cytoskeleton organization
GO:0006974	0.000162218	response to DNA damage stimulus
GO:0002566	0.000227877	somatic diversification of immune receptors via somatic mutation
GO:0016446	0.000227877	somatic hypermutation of immunoglobulin genes
GO:0051383	0.000227877	kinetochore organization
GO:0000086	0.000242661	G2/M transition of mitotic cell cycle
GO:0031123	0.000242661	RNA 3'-end processing
GO:0000132	0.00031822	establishment of mitotic spindle orientation
GO:0051095	0.00031822	regulation of helicase activity
GO:0051294	0.00031822	establishment of spindle orientation
GO:0051297	0.00052015	centrosome organization
GO:0008340	0.000542761	determination of adult lifespan
GO:0010389	0.000542761	regulation of G2/M transition of mitotic cell cycle
GO:0045910	0.000542761	negative regulation of DNA recombination
GO:0031023	0.000559652	microtubule organizing center organization
GO:0090068	0.000644305	positive regulation of cell cycle process
GO:0016043	0.000661968	cellular component organization
GO:0090304	0.000751504	nucleic acid metabolic process
GO:0051716	0.000765834	cellular response to stimulus
GO:0006268	0.000825026	DNA unwinding involved in replication
GO:0051983	0.000987526	regulation of chromosome segregation
GO:0010259	0.001164124	multicellular organismal aging
GO:0031058	0.001164124	positive regulation of histone modification
GO:0071174	0.001164124	mitotic cell cycle spindle checkpoint
GO:0006139	0.001184437	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
GO:0033554	0.001264272	cellular response to stress
GO:0071103	0.001274869	DNA conformation change
GO:0034641	0.001471331	cellular nitrogen compound metabolic process
GO:0007088	0.001545082	regulation of mitosis
GO:0051783	0.001545082	regulation of nuclear division
GO:0032507	0.001787196	maintenance of protein location in cell
GO:0009127	0.00200931	purine nucleoside monophosphate biosynthetic process
GO:0009168	0.00200931	purine ribonucleoside monophosphate biosynthetic process
GO:0031577	0.00200931	spindle checkpoint
GO:0000082	0.002145096	G1/S transition of mitotic cell cycle
GO:0051130	0.002169458	positive regulation of cellular component organization
GO:0045185	0.002241011	maintenance of protein location
GO:0032392	0.002254764	DNA geometric change
GO:0032508	0.002254764	DNA duplex unwinding
GO:0006807	0.002269381	nitrogen compound metabolic process
GO:0051651	0.002440746	maintenance of location in cell
GO:0033043	0.002513612	regulation of organelle organization
GO:0016458	0.002651184	gene silencing
GO:0006298	0.002785911	mismatch repair
GO:0031572	0.002785911	G2/M transition DNA damage checkpoint

Continued on Next Page...

Table D.7 – Continued

GO ID	P Value	GO Term
GO:0009126	0.003071393	purine nucleoside monophosphate metabolic process
GO:0009167	0.003071393	purine ribonucleoside monophosphate metabolic process
GO:0031056	0.003071393	regulation of histone modification
GO:0031124	0.003071393	mRNA 3'-end processing
GO:0000710	0.003955576	meiotic mismatch repair
GO:0003272	0.003955576	endocardial cushion formation
GO:0007100	0.003955576	mitotic centrosome separation
GO:0010610	0.003955576	regulation of mRNA stability involved in response to stress
GO:0021998	0.003955576	neural plate mediolateral regionalization
GO:0033129	0.003955576	positive regulation of histone phosphorylation
GO:0043146	0.003955576	spindle stabilization
GO:0043148	0.003955576	mitotic spindle stabilization
GO:0046680	0.003955576	response to DDT
GO:0048338	0.003955576	mesoderm structural organization
GO:0048352	0.003955576	paraxial mesoderm structural organization
GO:0060623	0.003955576	regulation of chromosome condensation
GO:0071281	0.003955576	cellular response to iron ion
GO:0071283	0.003955576	cellular response to iron(III) ion
GO:0002204	0.004006215	somatic recombination of immunoglobulin genes involved in immune response
GO:0002208	0.004006215	somatic diversification of immunoglobulins involved in immune response
GO:0007091	0.004006215	mitotic metaphase/anaphase transition
GO:0009156	0.004006215	ribonucleoside monophosphate biosynthetic process
GO:0030010	0.004006215	establishment of cell polarity
GO:0030071	0.004006215	regulation of mitotic metaphase/anaphase transition
GO:0031576	0.004006215	G2/M transition checkpoint
GO:0045190	0.004006215	isotype switching
GO:0010605	0.004216709	negative regulation of macromolecule metabolic process
GO:0008283	0.004296653	cell proliferation
GO:0002381	0.004343602	immunoglobulin production involved in immunoglobulin mediated immune response
GO:0006342	0.004693708	chromatin silencing
GO:0030261	0.004693708	chromosome condensation
GO:0051129	0.004995788	negative regulation of cellular component organization
GO:0009161	0.005431668	ribonucleoside monophosphate metabolic process
GO:0016447	0.005431668	somatic recombination of immunoglobulin gene segments
GO:0000018	0.005819321	regulation of DNA recombination
GO:0045814	0.005819321	negative regulation of gene expression, epigenetic
GO:0040029	0.005896798	regulation of gene expression, epigenetic
GO:0006281	0.006387647	DNA repair
GO:0009892	0.006597795	negative regulation of metabolic process
GO:0010639	0.006626223	negative regulation of organelle organization
GO:0016445	0.006631468	somatic diversification of immunoglobulins
GO:0008630	0.007492078	DNA damage response, signal transduction resulting in induction of apoptosis
GO:0000236	0.007895805	mitotic prometaphase
GO:0003203	0.007895805	endocardial cushion morphogenesis
GO:0009082	0.007895805	branched chain family amino acid biosynthetic process
GO:0010041	0.007895805	response to iron(III) ion
GO:0010424	0.007895805	DNA methylation on cytosine within a CG sequence
GO:0032776	0.007895805	DNA methylation on cytosine
GO:0033127	0.007895805	regulation of histone phosphorylation
GO:0048369	0.007895805	lateral mesoderm morphogenesis
GO:0048370	0.007895805	lateral mesoderm formation
GO:0048371	0.007895805	lateral mesodermal cell differentiation
GO:0048372	0.007895805	lateral mesodermal cell fate commitment
GO:0048377	0.007895805	lateral mesodermal cell fate specification
GO:0048378	0.007895805	regulation of lateral mesodermal cell fate specification
GO:0048382	0.007895805	mesendoderm development
GO:0051571	0.007895805	positive regulation of histone H3-K4 methylation
GO:0060897	0.007895805	neural plate regionalization

Continued on Next Page. . .

Table D.7 – Continued

GO ID	P Value	GO Term
GO:0070562	0.007895805	regulation of vitamin D receptor signaling pathway
GO:0090307	0.007895805	spindle assembly involved in mitosis
GO:0032269	0.008382756	negative regulation of cellular protein metabolic process
GO:0002562	0.008872146	somatic diversification of immune receptors via germline recombination within a single locus
GO:0016444	0.008872146	somatic cell DNA recombination
GO:0048477	0.008872146	oogenesis
GO:0051235	0.009127171	maintenance of location
GO:0050767	0.009727988	regulation of neurogenesis
GO:0002200	0.009850495	somatic diversification of immune receptors
GO:0048863	0.010356874	stem cell differentiation
GO:0051248	0.010368518	negative regulation of protein metabolic process
GO:0006344	0.011820745	maintenance of chromatin silencing
GO:0010586	0.011820745	miRNA metabolic process
GO:0010587	0.011820745	miRNA catabolic process
GO:0031442	0.011820745	positive regulation of mRNA 3'-end processing
GO:0046499	0.011820745	S-adenosylmethioninamine metabolic process
GO:0048368	0.011820745	lateral mesoderm development
GO:0050685	0.011820745	positive regulation of mRNA processing
GO:0051299	0.011820745	centrosome separation
GO:0051573	0.011820745	negative regulation of histone H3-K9 methylation
GO:0060896	0.011820745	neural plate pattern specification
GO:0060914	0.011820745	heart formation
GO:0070507	0.011943695	regulation of microtubule cytoskeleton organization
GO:0031324	0.012021243	negative regulation of cellular metabolic process
GO:0006310	0.012383973	DNA recombination
GO:0033044	0.012494885	regulation of chromosome organization
GO:0051960	0.013012966	regulation of nervous system development
GO:0051053	0.013630083	negative regulation of DNA metabolic process
GO:0002377	0.015413557	immunoglobulin production
GO:0000089	0.015730456	mitotic metaphase
GO:0000281	0.015730456	cytokinesis after mitosis
GO:0001880	0.015730456	Mullerian duct regression
GO:0006269	0.015730456	DNA replication, synthesis of RNA primer
GO:0006346	0.015730456	methylation-dependent chromatin silencing
GO:0031062	0.015730456	positive regulation of histone methylation
GO:0031440	0.015730456	regulation of mRNA 3'-end processing
GO:0042661	0.015730456	regulation of mesodermal cell fate specification
GO:0045347	0.015730456	negative regulation of MHC class II biosynthetic process
GO:0051570	0.015730456	regulation of histone H3-K9 methylation
GO:0060218	0.015730456	hemopoietic stem cell differentiation
GO:0060236	0.015730456	regulation of mitotic spindle organization
GO:0070561	0.015730456	vitamin D receptor signaling pathway
GO:0072132	0.015730456	mesenchyme morphogenesis
GO:0032886	0.016029199	regulation of microtubule-based process
GO:0051495	0.017291676	positive regulation of cytoskeleton organization
GO:0040007	0.017363157	growth
GO:0042493	0.017388016	response to drug
GO:0031400	0.01786688	negative regulation of protein modification process
GO:0008629	0.017938333	induction of apoptosis by intracellular signals
GO:0060284	0.019513871	regulation of cell development
GO:0009628	0.01952189	response to abiotic stimulus
GO:0003197	0.019624993	endocardial cushion development
GO:0007501	0.019624993	mesodermal cell fate specification
GO:0010870	0.019624993	positive regulation of receptor biosynthetic process
GO:0030916	0.019624993	otic vesicle formation
GO:0031061	0.019624993	negative regulation of histone methylation
GO:0031573	0.019624993	intra-S DNA damage checkpoint

Continued on Next Page...

Table D.7 – Continued

GO ID	P Value	GO Term
GO:0051382	0.019624993	kinetochore assembly
GO:0051569	0.019624993	regulation of histone H3-K4 methylation
GO:0070934	0.019624993	CRD-mediated mRNA stabilization
GO:0071305	0.019624993	cellular response to vitamin D
GO:0071398	0.019624993	cellular response to fatty acid
GO:0071453	0.019624993	cellular response to oxygen levels
GO:0071456	0.019624993	cellular response to hypoxia
GO:0071599	0.019624993	otic vesicle development
GO:0071600	0.019624993	otic vesicle morphogenesis
GO:0090224	0.019624993	regulation of spindle organization
GO:0007163	0.019938926	establishment or maintenance of cell polarity
GO:0014070	0.021040728	response to organic cyclic substance
GO:0009987	0.022113253	cellular process
GO:0044260	0.022685343	cellular macromolecule metabolic process
GO:0032268	0.022850588	regulation of cellular protein metabolic process
GO:0006398	0.023504417	histone mRNA 3'-end processing
GO:0031054	0.023504417	pre-microRNA processing
GO:0033762	0.023504417	response to glucagon stimulus
GO:0046498	0.023504417	S-adenosylhomocysteine metabolic process
GO:0051567	0.023504417	histone H3-K9 methylation
GO:0060033	0.023504417	anatomical structure regression
GO:0000079	0.024205165	regulation of cyclin-dependent protein kinase activity
GO:0009411	0.024205165	response to UV
GO:0031323	0.024229028	regulation of cellular metabolic process
GO:0016570	0.025724865	histone modification
GO:0002440	0.026466249	production of molecular mediator of immune response
GO:0006302	0.026466249	double-strand break repair
GO:0031145	0.026466249	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process
GO:0016569	0.026555857	covalent chromatin modification
GO:0016310	0.026882049	phosphorylation
GO:0034661	0.027368783	ncRNA catabolic process
GO:0051323	0.027368783	metaphase
GO:0060391	0.027368783	positive regulation of SMAD protein nuclear translocation
GO:0071396	0.027368783	cellular response to lipid
GO:0007292	0.028019516	female gamete generation
GO:0032270	0.028347257	positive regulation of cellular protein metabolic process
GO:0030900	0.029134926	forebrain development
GO:0010212	0.029608727	response to ionizing radiation
GO:0051439	0.029608727	regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
GO:0032880	0.030472794	regulation of protein localization
GO:0044237	0.03110202	cellular metabolic process
GO:0009113	0.031218149	purine base biosynthetic process
GO:0010224	0.031218149	response to UV-B
GO:0017085	0.031218149	response to insecticide
GO:0019047	0.031218149	provirus integration
GO:0030069	0.031218149	lysogeny
GO:0031060	0.031218149	regulation of histone methylation
GO:0034508	0.031218149	centromere complex assembly
GO:0048340	0.031218149	paraxial mesoderm morphogenesis
GO:0048532	0.031218149	anatomical structure arrangement
GO:0048853	0.031218149	forebrain morphogenesis
GO:0055015	0.031218149	ventricular cardiac muscle cell development
GO:0060045	0.031218149	positive regulation of cardiac muscle cell proliferation
GO:0060390	0.031218149	regulation of SMAD protein nuclear translocation
GO:0071407	0.031218149	cellular response to organic cyclic substance
GO:0016064	0.031233241	immunoglobulin mediated immune response
GO:0019724	0.032058539	B cell mediated immunity

Continued on Next Page. . .

Table D.7 – Continued

GO ID	P Value	GO Term
GO:0007420	0.032187216	brain development
GO:0051247	0.033532315	positive regulation of protein metabolic process
GO:0009950	0.035052572	dorsal/ventral axis specification
GO:0010453	0.035052572	regulation of cell fate commitment
GO:0010470	0.035052572	regulation of gastrulation
GO:0016572	0.035052572	histone phosphorylation
GO:0031503	0.035052572	protein complex localization
GO:0033205	0.035052572	cell cycle cytokinesis
GO:0042659	0.035052572	regulation of cell fate specification
GO:0010243	0.036312306	response to organic nitrogen
GO:0051641	0.037096512	cellular localization
GO:0045786	0.037642407	negative regulation of cell cycle
GO:0051246	0.038616306	regulation of protein metabolic process
GO:0001710	0.03887211	mesodermal cell fate commitment
GO:0006301	0.03887211	postreplication repair
GO:0006303	0.03887211	double-strand break repair via nonhomologous end joining
GO:0006349	0.03887211	regulation of gene expression by genetic imprinting
GO:0006378	0.03887211	mRNA polyadenylation
GO:0010869	0.03887211	regulation of receptor biosynthetic process
GO:0031057	0.03887211	negative regulation of histone modification
GO:0043584	0.03887211	nose development
GO:0045346	0.03887211	regulation of MHC class II biosynthetic process
GO:0071241	0.03887211	cellular response to inorganic substance
GO:0071248	0.03887211	cellular response to metal ion
GO:0071514	0.03887211	genetic imprinting
GO:0046661	0.041686743	male sex differentiation
GO:0051438	0.041686743	regulation of ubiquitin-protein ligase activity
GO:0048015	0.042610059	phosphoinositide-mediated signaling
GO:0006379	0.042676819	mRNA cleavage
GO:0045342	0.042676819	MHC class II biosynthetic process
GO:0048333	0.042676819	mesodermal cell differentiation
GO:0055012	0.042676819	ventricular cardiac muscle cell differentiation
GO:0051128	0.043302372	regulation of cellular component organization
GO:0051340	0.044479666	regulation of ligase activity
GO:0048519	0.045547242	negative regulation of biological process
GO:0034645	0.045691844	cellular macromolecule biosynthetic process
GO:0007281	0.046379426	germ cell development
GO:0031099	0.046379426	regeneration
GO:0001556	0.046466754	oocyte maturation
GO:0002021	0.046466754	response to dietary excess
GO:0007076	0.046466754	mitotic chromosome condensation
GO:0007094	0.046466754	mitotic cell cycle spindle assembly checkpoint
GO:0009083	0.046466754	branched chain family amino acid catabolic process
GO:0010714	0.046466754	positive regulation of collagen metabolic process
GO:0032967	0.046466754	positive regulation of collagen biosynthetic process
GO:0046112	0.046466754	nucleobase biosynthetic process
GO:0051568	0.046466754	histone H3-K4 methylation
GO:0051094	0.046704657	positive regulation of developmental process
GO:0006950	0.047411532	response to stress

D.3.6 GO terms associated with the RNA transcription / protein synthesis module

Table D.8:

GO ID	P Value	GO Term
GO:0006420	2.84E-05	arginyl-tRNA aminoacylation
GO:0018198	0.000197338	peptidyl-cysteine modification
GO:0009108	0.001505193	coenzyme biosynthetic process
GO:0008380	0.002033993	RNA splicing
GO:0006397	0.002458656	mRNA processing
GO:0022613	0.002766281	ribonucleoprotein complex biogenesis
GO:0007192	0.003118819	activation of adenylate cyclase activity by serotonin receptor signaling pathway
GO:0017014	0.003118819	protein amino acid nitrosylation
GO:0018119	0.003118819	peptidyl-cysteine S-nitrosylation
GO:0042660	0.003118819	positive regulation of cell fate specification
GO:0046294	0.003118819	formaldehyde catabolic process
GO:0048936	0.003118819	peripheral nervous system neuron axonogenesis
GO:0044281	0.003169195	small molecule metabolic process
GO:0051188	0.004581947	cofactor biosynthetic process
GO:0006520	0.005315717	cellular amino acid metabolic process
GO:0016071	0.005476853	mRNA metabolic process
GO:0000022	0.006228148	mitotic spindle elongation
GO:0000189	0.006228148	nuclear translocation of MAPK
GO:0019478	0.006228148	D-amino acid catabolic process
GO:0042699	0.006228148	follicle-stimulating hormone signaling pathway
GO:0046185	0.006228148	aldehyde catabolic process
GO:0046292	0.006228148	formaldehyde metabolic process
GO:0051231	0.006228148	spindle elongation
GO:0060128	0.006228148	adrenocorticotropin hormone secreting cell differentiation
GO:0060591	0.006228148	chondroblast differentiation
GO:0009987	0.006259244	cellular process
GO:0006396	0.00728534	RNA processing
GO:0006446	0.007904176	regulation of translational initiation
GO:0017157	0.008264316	regulation of exocytosis
GO:0006418	0.008631734	tRNA aminoacylation for protein translation
GO:0043038	0.008631734	amino acid activation
GO:0043039	0.008631734	tRNA aminoacylation
GO:0019752	0.009318116	carboxylic acid metabolic process
GO:0043436	0.009318116	oxoacid metabolic process
GO:0014889	0.009328015	muscle atrophy
GO:0017182	0.009328015	peptidyl-diphthamide metabolic process
GO:0017183	0.009328015	peptidyl-diphthamide biosynthetic process from peptidyl-histidine
GO:0018125	0.009328015	peptidyl-cysteine methylation
GO:0046416	0.009328015	D-amino acid metabolic process
GO:0060129	0.009328015	thyroid-stimulating hormone-secreting cell differentiation
GO:0070935	0.009328015	3'-UTR-mediated mRNA stabilization
GO:0044282	0.009730879	small molecule catabolic process
GO:0006082	0.009845979	organic acid metabolic process
GO:0042180	0.010395066	cellular ketone metabolic process
GO:0006732	0.012350571	coenzyme metabolic process
GO:0048511	0.012350571	rhythmic process
GO:0007008	0.012418447	outer mitochondrial membrane organization
GO:0043922	0.012418447	negative regulation by host of viral transcription
GO:0048935	0.012418447	peripheral nervous system neuron development
GO:0051409	0.012418447	response to nitrosative stress
GO:0070096	0.012418447	mitochondrial outer membrane translocase complex assembly
GO:0006413	0.014514097	translational initiation

Continued on Next Page...

Table D.8 – Continued

GO ID	P Value	GO Term
GO:0044106	0.014817902	cellular amine metabolic process
GO:0021534	0.015499473	cell proliferation in hindbrain
GO:0021924	0.015499473	cell proliferation in the external granule layer
GO:0021930	0.015499473	granule cell precursor proliferation
GO:0032057	0.015499473	negative regulation of translational initiation in response to stress
GO:0048934	0.015499473	peripheral nervous system neuron differentiation
GO:0006067	0.018571121	ethanol metabolic process
GO:0006069	0.018571121	ethanol oxidation
GO:0007210	0.018571121	serotonin receptor signaling pathway
GO:0032055	0.018571121	negative regulation of translation in response to stress
GO:0032897	0.018571121	negative regulation of viral transcription
GO:0034308	0.018571121	monohydric alcohol metabolic process
GO:0060644	0.018571121	mammary gland epithelial cell differentiation
GO:0009063	0.019515168	cellular amino acid catabolic process
GO:0043921	0.021633418	modulation by host of viral transcription
GO:0046668	0.021633418	regulation of retinal cell programmed cell death
GO:0051775	0.021633418	response to redox state
GO:0052312	0.021633418	modulation of transcription in other organism involved in symbiotic interaction
GO:0052472	0.021633418	modulation by host of symbiont transcription
GO:0022618	0.022249871	ribonucleoprotein complex assembly
GO:0010001	0.022814877	glial cell differentiation
GO:0051301	0.023268534	cell division
GO:0006519	0.02370024	cellular amino acid and derivative metabolic process
GO:0009396	0.024686392	folic acid and derivative biosynthetic process
GO:0009435	0.024686392	NAD biosynthetic process
GO:0018202	0.024686392	peptidyl-histidine modification
GO:0043558	0.024686392	regulation of translational initiation in response to stress
GO:0046653	0.024686392	tetrahydrofolate metabolic process
GO:0046666	0.024686392	retinal cell programmed cell death
GO:0060045	0.024686392	positive regulation of cardiac muscle cell proliferation
GO:0009310	0.025133766	amine catabolic process
GO:0042698	0.025728003	ovulation cycle
GO:0051186	0.026128322	cofactor metabolic process
GO:0034622	0.026162461	cellular macromolecular complex assembly
GO:0002042	0.027730071	cell migration involved in sprouting angiogenesis
GO:0010453	0.027730071	regulation of cell fate commitment
GO:0019359	0.027730071	nicotinamide nucleotide biosynthetic process
GO:0021936	0.027730071	regulation of granule cell precursor proliferation
GO:0021940	0.027730071	positive regulation of granule cell precursor proliferation
GO:0030815	0.027730071	negative regulation of cAMP metabolic process
GO:0030818	0.027730071	negative regulation of cAMP biosynthetic process
GO:0042659	0.027730071	regulation of cell fate specification
GO:0043555	0.027730071	regulation of translation in response to stress
GO:0007188	0.028161812	G-protein signaling, coupled to cAMP nucleotide second messenger
GO:0042063	0.03068472	gliogenesis
GO:0030800	0.030764483	negative regulation of cyclic nucleotide metabolic process
GO:0030803	0.030764483	negative regulation of cyclic nucleotide biosynthetic process
GO:0030809	0.030764483	negative regulation of nucleotide biosynthetic process
GO:0043537	0.030764483	negative regulation of blood vessel endothelial cell migration
GO:0006412	0.03284547	translation
GO:0007128	0.033789655	meiotic prophase I
GO:0021984	0.033789655	adenohypophysis development
GO:0032855	0.033789655	positive regulation of Rac GTPase activity
GO:0051324	0.033789655	prophase
GO:0051851	0.033789655	modification by host of symbiont morphology or physiology
GO:0034660	0.03423083	ncRNA metabolic process
GO:0045761	0.034630745	regulation of adenylate cyclase activity
GO:0009308	0.035832323	amine metabolic process

Continued on Next Page...

Table D.8 – Continued

GO ID	P Value	GO Term
GO:0000377	0.035987987	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
GO:0000398	0.035987987	nuclear mRNA splicing, via spliceosome
GO:0031279	0.035987987	regulation of cyclase activity
GO:0051339	0.036674296	regulation of lyase activity
GO:0006086	0.036805614	acetyl-CoA biosynthetic process from pyruvate
GO:0009083	0.036805614	branched chain family amino acid catabolic process
GO:0010510	0.036805614	regulation of acetyl-CoA biosynthetic process from pyruvate
GO:0045980	0.036805614	negative regulation of nucleotide metabolic process
GO:0051046	0.03692867	regulation of secretion
GO:0019933	0.038062107	cAMP-mediated signaling
GO:0010608	0.038117727	posttranscriptional regulation of gene expression
GO:0018193	0.038921335	peptidyl-amino acid modification
GO:0043536	0.039812388	positive regulation of blood vessel endothelial cell migration
GO:0045947	0.039812388	negative regulation of translational initiation
GO:0046782	0.039812388	regulation of viral transcription
GO:0055021	0.039812388	regulation of cardiac muscle tissue growth
GO:0055024	0.039812388	regulation of cardiac muscle tissue development
GO:0060043	0.039812388	regulation of cardiac muscle cell proliferation
GO:0044237	0.040070335	cellular metabolic process
GO:0000375	0.042344467	RNA splicing, via transesterification reactions
GO:0006085	0.042810004	acetyl-CoA biosynthetic process
GO:0006700	0.042810004	C21-steroid hormone biosynthetic process
GO:0006760	0.042810004	folic acid and derivative metabolic process
GO:0051193	0.042810004	regulation of cofactor metabolic process
GO:0051196	0.042810004	regulation of coenzyme metabolic process
GO:0034621	0.043195956	cellular macromolecular complex subunit organization
GO:0030817	0.045295615	regulation of cAMP biosynthetic process
GO:0014003	0.04579849	oligodendrocyte development
GO:0017158	0.04579849	regulation of calcium ion-dependent exocytosis
GO:0019080	0.04579849	viral genome expression
GO:0019083	0.04579849	viral transcription
GO:0019363	0.04579849	pyridine nucleotide biosynthetic process
GO:0060420	0.04579849	regulation of heart growth
GO:0006171	0.046799216	cAMP biosynthetic process
GO:0030814	0.046799216	regulation of cAMP metabolic process
GO:0051726	0.047999309	regulation of cell cycle
GO:0007018	0.048321133	microtubule-based movement
GO:0050709	0.048777871	negative regulation of protein secretion
GO:0051702	0.048777871	interaction with symbiont
GO:0006399	0.049088873	tRNA metabolic process
GO:0007187	0.04986109	G-protein signaling, coupled to cyclic nucleotide second messenger

D.3.7 GO terms associated with the metabolism / hormone signaling module

Table D.9:

GO ID	P Value	GO Term
GO:0034660	0.001322169	ncRNA metabolic process
GO:0006399	0.001776558	tRNA metabolic process
GO:0042278	0.002085852	purine nucleoside metabolic process
GO:0046128	0.002085852	purine ribonucleoside metabolic process
GO:0006409	0.002129925	tRNA export from nucleus
GO:0009642	0.002129925	response to light intensity
GO:0015957	0.002129925	bis(5'-nucleosidyl) oligophosphate biosynthetic process
GO:0015960	0.002129925	diadenosine polyphosphate biosynthetic process
GO:0015965	0.002129925	diadenosine tetraphosphate metabolic process
GO:0015966	0.002129925	diadenosine tetraphosphate biosynthetic process
GO:0032289	0.002129925	myelin formation in the central nervous system
GO:0051031	0.002129925	tRNA transport
GO:0001942	0.003573516	hair follicle development
GO:0022404	0.003573516	molting cycle process
GO:0022405	0.003573516	hair cycle process
GO:0006418	0.00409276	tRNA aminoacylation for protein translation
GO:0042303	0.00409276	molting cycle
GO:0042633	0.00409276	hair cycle
GO:0043038	0.00409276	amino acid activation
GO:0043039	0.00409276	tRNA aminoacylation
GO:0006348	0.004255476	chromatin silencing at telomere
GO:0006426	0.004255476	glycyl-tRNA aminoacylation
GO:0006428	0.004255476	isoleucyl-tRNA aminoacylation
GO:0006481	0.004255476	C-terminal protein amino acid methylation
GO:0015942	0.004255476	formate metabolic process
GO:0018410	0.004255476	peptide or protein carboxyl-terminal blocking
GO:0042780	0.004255476	tRNA 3'-end processing
GO:0009119	0.004836233	ribonucleoside metabolic process
GO:0055086	0.005692612	nucleobase, nucleoside and nucleotide metabolic process
GO:0006475	0.00637666	internal protein amino acid acetylation
GO:0015956	0.00637666	bis(5'-nucleosidyl) oligophosphate metabolic process
GO:0015959	0.00637666	diadenosine polyphosphate metabolic process
GO:0022010	0.00637666	myelination in the central nervous system
GO:0032291	0.00637666	ensheathment of axons in the central nervous system
GO:0035315	0.00637666	hair cell differentiation
GO:0043628	0.00637666	ncRNA 3'-end processing
GO:0046499	0.00637666	S-adenosylmethioninamine metabolic process
GO:0051798	0.00637666	positive regulation of hair follicle development
GO:0009116	0.007645128	nucleoside metabolic process
GO:0007199	0.008493487	G-protein signaling, coupled to cGMP nucleotide second messenger
GO:0032276	0.008493487	regulation of gonadotropin secretion
GO:0032277	0.008493487	negative regulation of gonadotropin secretion
GO:0040016	0.008493487	embryonic cleavage
GO:0046880	0.008493487	regulation of follicle-stimulating hormone secretion
GO:0046882	0.008493487	negative regulation of follicle-stimulating hormone secretion
GO:0051797	0.008493487	regulation of hair follicle development
GO:0060218	0.008493487	hemopoietic stem cell differentiation
GO:0035264	0.009928836	multicellular organism growth
GO:0032288	0.010605965	myelin assembly
GO:0032926	0.010605965	negative regulation of activin receptor signaling pathway
GO:0042634	0.010605965	regulation of hair cycle
GO:0006283	0.012714102	transcription-coupled nucleotide-excision repair

Continued on Next Page...

Table D.9 – Continued

GO ID	P Value	GO Term
GO:0032274	0.012714102	gonadotropin secretion
GO:0046498	0.012714102	S-adenosylhomocysteine metabolic process
GO:0046884	0.012714102	follicle-stimulating hormone secretion
GO:0070509	0.012714102	calcium ion import
GO:0070588	0.012714102	calcium ion transmembrane transport
GO:0000154	0.014817908	rRNA modification
GO:0030825	0.014817908	positive regulation of cGMP metabolic process
GO:0033683	0.014817908	nucleotide-excision repair, DNA incision
GO:0044237	0.016838242	cellular metabolic process
GO:0006465	0.01691739	signal peptide processing
GO:0009396	0.01691739	folic acid and derivative biosynthetic process
GO:0043249	0.01691739	erythrocyte maturation
GO:0043558	0.01691739	regulation of translational initiation in response to stress
GO:0045684	0.01691739	positive regulation of epidermis development
GO:0046653	0.01691739	tetrahydrofolate metabolic process
GO:0044281	0.017394375	small molecule metabolic process
GO:0009163	0.019012558	nucleoside biosynthetic process
GO:0019934	0.019012558	cGMP-mediated signaling
GO:0042451	0.019012558	purine nucleoside biosynthetic process
GO:0042455	0.019012558	ribonucleoside biosynthetic process
GO:0043555	0.019012558	regulation of translation in response to stress
GO:0044060	0.019012558	regulation of endocrine process
GO:0046129	0.019012558	purine ribonucleoside biosynthetic process
GO:0009650	0.021103419	UV protection
GO:0018196	0.021103419	peptidyl-asparagine modification
GO:0018279	0.021103419	protein amino acid N-linked glycosylation via asparagine
GO:0048820	0.021103419	hair follicle maturation
GO:0030823	0.023189983	regulation of cGMP metabolic process
GO:0060986	0.023189983	endocrine hormone secretion
GO:0007164	0.025272258	establishment of tissue polarity
GO:0006486	0.026347976	protein amino acid glycosylation
GO:0043413	0.026347976	macromolecule glycosylation
GO:0070085	0.026347976	glycosylation
GO:0032925	0.027350252	regulation of activin receptor signaling pathway
GO:0048821	0.027350252	erythrocyte development
GO:0044249	0.027781463	cellular biosynthetic process
GO:0044260	0.028257369	cellular macromolecule metabolic process
GO:0006760	0.029423975	folic acid and derivative metabolic process
GO:0034645	0.030926132	cellular macromolecule biosynthetic process
GO:0001502	0.031493433	cartilage condensation
GO:0014003	0.031493433	oligodendrocyte development
GO:0006730	0.032794344	one-carbon metabolic process
GO:0046483	0.032943656	heterocycle metabolic process
GO:0006725	0.033244252	cellular aromatic compound metabolic process
GO:0032924	0.033558636	activin receptor signaling pathway
GO:0009058	0.034305782	biosynthetic process
GO:0009416	0.03460864	response to light stimulus
GO:0002244	0.035619593	hemopoietic progenitor cell differentiation
GO:0043616	0.035619593	keratinocyte proliferation
GO:0071695	0.035619593	anatomical structure maturation
GO:0009059	0.035896956	macromolecule biosynthetic process
GO:0008152	0.036403368	metabolic process
GO:0010558	0.036475033	negative regulation of macromolecule biosynthetic process
GO:0031069	0.037676311	hair follicle morphogenesis
GO:0006519	0.038301916	cellular amino acid and derivative metabolic process
GO:0031327	0.040019133	negative regulation of cellular biosynthetic process
GO:0030968	0.041777065	endoplasmic reticulum unfolded protein response
GO:0034620	0.041777065	cellular response to unfolded protein

Continued on Next Page…

Table D.9 – Continued

GO ID	P Value	GO Term
GO:0043009	0.041931225	chordate embryonic development
GO:0009890	0.042699542	negative regulation of biosynthetic process
GO:0009792	0.043082223	embryo development ending in birth or egg hatching
GO:0000718	0.043821118	nucleotide-excision repair, DNA damage removal
GO:0007223	0.043821118	Wnt receptor signaling pathway, calcium modulating pathway
GO:0045682	0.043821118	regulation of epidermis development
GO:0046068	0.043821118	cGMP metabolic process
GO:0009987	0.045108181	cellular process
GO:0009101	0.045768921	glycoprotein biosynthetic process
GO:0042558	0.045860967	pteridine and derivative metabolic process
GO:0006412	0.049386928	translation
GO:0045055	0.049928082	regulated secretory pathway
GO:0048730	0.049928082	epidermis morphogenesis

D.3.8 GO terms associated with the signaling / cellular identity module

Table D.10:

GO ID	P Value	GO Term
GO:0006955	1.69E-08	immune response
GO:0002376	2.37E-08	immune system process
GO:0002504	4.25E-06	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II
GO:0001910	2.04E-05	regulation of leukocyte mediated cytotoxicity
GO:0001911	3.22E-05	negative regulation of leukocyte mediated cytotoxicity
GO:0031341	3.34E-05	regulation of cell killing
GO:0031342	5.36E-05	negative regulation of cell killing
GO:0042492	5.36E-05	gamma-delta T cell differentiation
GO:0045586	5.36E-05	regulation of gamma-delta T cell differentiation
GO:0045588	5.36E-05	positive regulation of gamma-delta T cell differentiation
GO:0046643	5.36E-05	regulation of gamma-delta T cell activation
GO:0046645	5.36E-05	positive regulation of gamma-delta T cell activation
GO:0001909	6.18E-05	leukocyte mediated cytotoxicity
GO:0002704	0.00011219	negative regulation of leukocyte mediated immunity
GO:0002707	0.00011219	negative regulation of lymphocyte mediated immunity
GO:0002925	0.00011219	positive regulation of humoral immune response mediated by circulating immunoglobulin
GO:0033687	0.00011219	osteoblast proliferation
GO:0046629	0.00011219	gamma-delta T cell activation
GO:0002922	0.000149366	positive regulation of humoral immune response
GO:0002923	0.000149366	regulation of humoral immune response mediated by circulating immunoglobulin
GO:0002706	0.000215899	regulation of lymphocyte mediated immunity
GO:0019882	0.000271484	antigen processing and presentation
GO:0002714	0.000292106	positive regulation of B cell mediated immunity
GO:0002891	0.000292106	positive regulation of immunoglobulin mediated immune response
GO:0001906	0.000302434	cell killing
GO:0002703	0.00035299	regulation of leukocyte mediated immunity
GO:0002920	0.000413044	regulation of humoral immune response
GO:0065007	0.000531015	biological regulation
GO:0050789	0.000672523	regulation of biological process
GO:0002715	0.000715957	regulation of natural killer cell mediated immunity
GO:0042269	0.000715957	regulation of natural killer cell mediated cytotoxicity
GO:0001912	0.00080427	positive regulation of leukocyte mediated cytotoxicity
GO:0002698	0.00080427	negative regulation of immune effector process
GO:0050794	0.000941615	regulation of cellular process
GO:0050896	0.001113031	response to stimulus
GO:0031343	0.001207177	positive regulation of cell killing
GO:0046635	0.001207177	positive regulation of alpha-beta T cell activation
GO:0002683	0.001214137	negative regulation of immune system process
GO:0002712	0.001438112	regulation of B cell mediated immunity
GO:0002889	0.001438112	regulation of immunoglobulin mediated immune response
GO:0002252	0.001521832	immune effector process
GO:0002228	0.001560873	natural killer cell mediated immunity
GO:0042267	0.001560873	natural killer cell mediated cytotoxicity
GO:0002697	0.001840539	regulation of immune effector process
GO:0002824	0.001958061	positive regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains
GO:0050777	0.001958061	negative regulation of immune response
GO:0002449	0.00205033	lymphocyte mediated immunity
GO:0002821	0.002100019	positive regulation of adaptive immune response
GO:0045582	0.002100019	positive regulation of T cell differentiation
GO:0002705	0.002246722	positive regulation of leukocyte mediated immunity
GO:0002708	0.002246722	positive regulation of lymphocyte mediated immunity

Continued on Next Page...

Table D.10 – Continued

GO ID	P Value	GO Term
GO:0002158	0.002358132	osteoclast proliferation
GO:0002361	0.002358132	CD4-positive, CD25-positive, alpha-beta regulatory T cell differentiation
GO:0002370	0.002358132	natural killer cell cytokine production
GO:0002727	0.002358132	regulation of natural killer cell cytokine production
GO:0002729	0.002358132	positive regulation of natural killer cell cytokine production
GO:0009720	0.002358132	detection of hormone stimulus
GO:0009726	0.002358132	detection of endogenous stimulus
GO:0032829	0.002358132	regulation of CD4-positive, CD25-positive, alpha-beta regulatory T cell differentiation
GO:0032831	0.002358132	positive regulation of CD4-positive, CD25-positive, alpha-beta regulatory T cell differentiation
GO:0034436	0.002358132	glycoprotein transport
GO:0045838	0.002358132	positive regulation of membrane potential
GO:0050904	0.002358132	diapedesis
GO:0060448	0.002358132	dichotomous subdivision of terminal units involved in lung branching
GO:0045621	0.002398149	positive regulation of lymphocyte differentiation
GO:0046634	0.002398149	regulation of alpha-beta T cell activation
GO:0002455	0.003404688	humoral immune response mediated by circulating immunoglobulin
GO:0007204	0.003545142	elevation of cytosolic calcium ion concentration
GO:0002443	0.003699526	leukocyte mediated immunity
GO:0065008	0.004027722	regulation of biological quality
GO:0002700	0.004167465	regulation of production of molecular mediator of immune response
GO:0051480	0.004272108	cytosolic calcium ion homeostasis
GO:0001915	0.004710882	negative regulation of T cell mediated cytotoxicity
GO:0002716	0.004710882	negative regulation of natural killer cell mediated immunity
GO:0034314	0.004710882	Arp2/3 complex-mediated actin nucleation
GO:0045591	0.004710882	positive regulation of regulatory T cell differentiation
GO:0045953	0.004710882	negative regulation of natural killer cell mediated cytotoxicity
GO:0050855	0.004710882	regulation of B cell receptor signaling pathway
GO:0051607	0.004786756	defense response to virus
GO:0002699	0.005221786	positive regulation of immune effector process
GO:0060402	0.005221786	calcium ion transport into cytosol
GO:0046631	0.005445889	alpha-beta T cell activation
GO:0060401	0.005674356	cytosolic calcium ion transport
GO:0045580	0.005907169	regulation of T cell differentiation
GO:0002822	0.006385745	regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains
GO:0032879	0.006415683	regulation of localization
GO:0002819	0.006631468	regulation of adaptive immune response
GO:0002032	0.007058262	desensitization of G-protein coupled receptor protein signaling pathway by arrestin
GO:0002378	0.007058262	immunoglobulin biosynthetic process
GO:0045542	0.007058262	positive regulation of cholesterol biosynthetic process
GO:0045589	0.007058262	regulation of regulatory T cell differentiation
GO:0045896	0.007058262	regulation of transcription, mitotic
GO:0045897	0.007058262	positive regulation of transcription, mitotic
GO:0046021	0.007058262	regulation of transcription from RNA polymerase II promoter, mitotic
GO:0046022	0.007058262	positive regulation of transcription from RNA polymerase II promoter, mitotic
GO:0006917	0.00726145	induction of apoptosis
GO:0012502	0.007337971	induction of programmed cell death
GO:0045619	0.007923631	regulation of lymphocyte differentiation
GO:0048878	0.008359535	chemical homeostasis
GO:0045088	0.009319878	regulation of innate immune response
GO:0002710	0.009400284	negative regulation of T cell mediated immunity
GO:0033688	0.009400284	regulation of osteoblast proliferation
GO:0034113	0.009400284	heterotypic cell-cell adhesion
GO:0090205	0.009400284	positive regulation of cholesterol metabolic process
GO:0002440	0.009906968	production of molecular mediator of immune response
GO:0002521	0.010351705	leukocyte differentiation
GO:0006874	0.010942755	cellular calcium ion homeostasis

Continued on Next Page. . .

Table D.10 – Continued

GO ID	P Value	GO Term
GO:2000021	0.011129305	regulation of ion homeostasis
GO:0045010	0.011736959	actin nucleation
GO:0045019	0.011736959	negative regulation of nitric oxide biosynthetic process
GO:0045066	0.011736959	regulatory T cell differentiation
GO:0050857	0.011736959	positive regulation of antigen receptor-mediated signaling pathway
GO:0016064	0.011764243	immunoglobulin mediated immune response
GO:0055074	0.012023642	calcium ion homeostasis
GO:0019724	0.012087588	B cell mediated immunity
GO:0006875	0.012668084	cellular metal ion homeostasis
GO:0050870	0.013762313	positive regulation of T cell activation
GO:0001916	0.0140683	positive regulation of T cell mediated cytotoxicity
GO:0007171	0.0140683	activation of transmembrane receptor protein tyrosine kinase activity
GO:0010887	0.0140683	negative regulation of cholesterol storage
GO:0031953	0.0140683	negative regulation of protein amino acid autophosphorylation
GO:0032366	0.0140683	intracellular sterol transport
GO:0032367	0.0140683	intracellular cholesterol transport
GO:0045059	0.0140683	positive thymic T cell selection
GO:0048304	0.0140683	positive regulation of isotype switching to IgG isotypes
GO:0055091	0.0140683	phospholipid homeostasis
GO:0060136	0.0140683	embryonic process involved in female pregnancy
GO:0055065	0.014365205	metal ion homeostasis
GO:0002573	0.015170568	myeloid leukocyte differentiation
GO:0010740	0.015260172	positive regulation of intracellular protein kinase cascade
GO:0006959	0.015531987	humoral immune response
GO:0001914	0.016394319	regulation of T cell mediated cytotoxicity
GO:0002031	0.016394319	G-protein coupled receptor internalization
GO:0006198	0.016394319	cAMP catabolic process
GO:0032689	0.016394319	negative regulation of interferon-gamma production
GO:0045060	0.016394319	negative thymic T cell selection
GO:0045824	0.016394319	negative regulation of innate immune response
GO:0060600	0.016394319	dichotomous subdivision of an epithelial terminal unit
GO:0035556	0.01664198	intracellular signal transduction
GO:0019221	0.017777681	cytokine-mediated signaling pathway
GO:0023036	0.017777681	initiation of signal transduction
GO:0023038	0.017777681	signal initiation by diffusible mediator
GO:0023049	0.017777681	signal initiation by protein/peptide mediator
GO:0043410	0.017777681	positive regulation of MAPKKK cascade
GO:0010872	0.018715026	regulation of cholesterol esterification
GO:0032365	0.018715026	intracellular lipid transport
GO:0043011	0.018715026	myeloid dendritic cell differentiation
GO:0043368	0.018715026	positive T cell selection
GO:0043383	0.018715026	negative T cell selection
GO:0046641	0.018715026	positive regulation of alpha-beta T cell proliferation
GO:0048302	0.018715026	regulation of isotype switching to IgG isotypes
GO:0030005	0.018740757	cellular di-, tri-valent inorganic cation homeostasis
GO:0006952	0.019140405	defense response
GO:0050776	0.01936046	regulation of immune response
GO:0030217	0.020972695	T cell differentiation
GO:0002820	0.021030435	negative regulation of adaptive immune response
GO:0002823	0.021030435	negative regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains
GO:0009214	0.021030435	cyclic nucleotide catabolic process
GO:0010893	0.021030435	positive regulation of steroid biosynthetic process
GO:0042987	0.021030435	amyloid precursor protein catabolic process
GO:0043372	0.021030435	positive regulation of CD4-positive, alpha beta T cell differentiation
GO:0045540	0.021030435	regulation of cholesterol biosynthetic process
GO:0045830	0.021030435	positive regulation of isotype switching
GO:0046902	0.021030435	regulation of mitochondrial membrane permeability

Continued on Next Page. . .

Table D.10 – Continued

GO ID	P Value	GO Term
GO:0048291	0.021030435	isotype switching to IgG isotypes
GO:0045597	0.021730044	positive regulation of cell differentiation
GO:0055066	0.021730044	di-, tri-valent inorganic cation homeostasis
GO:0043065	0.021732802	positive regulation of apoptosis
GO:0043068	0.022200664	positive regulation of programmed cell death
GO:0007165	0.022734777	signal transduction
GO:0010942	0.022994253	positive regulation of cell death
GO:0001913	0.023340555	T cell mediated cytotoxicity
GO:0030146	0.023340555	diuresis
GO:0033700	0.023340555	phospholipid efflux
GO:0034374	0.023340555	low-density lipoprotein particle remodeling
GO:0045911	0.023340555	positive regulation of DNA recombination
GO:0030003	0.024489935	cellular cation homeostasis
GO:0051251	0.024830961	positive regulation of lymphocyte activation
GO:0001773	0.0256454	myeloid dendritic cell activation
GO:0002029	0.0256454	desensitization of G-protein coupled receptor protein signaling pathway
GO:0002720	0.0256454	positive regulation of cytokine production involved in immune response
GO:0010634	0.0256454	positive regulation of epithelial cell migration
GO:0022401	0.0256454	negative adaptation of signaling pathway
GO:0023058	0.0256454	adaptation of signaling pathway
GO:0031648	0.0256454	protein destabilization
GO:0031952	0.0256454	regulation of protein amino acid autophosphorylation
GO:0034433	0.0256454	steroid esterification
GO:0034434	0.0256454	sterol esterification
GO:0034435	0.0256454	cholesterol esterification
GO:0045061	0.0256454	thymic T cell selection
GO:0045123	0.0256454	cellular extravasation
GO:0050732	0.0256454	negative regulation of peptidyl-tyrosine phosphorylation
GO:0050853	0.0256454	B cell receptor signaling pathway
GO:0046907	0.026085117	intracellular transport
GO:0009967	0.026679788	positive regulation of signal transduction
GO:0051235	0.027090738	maintenance of location
GO:0023056	0.027940783	positive regulation of signaling process
GO:0001960	0.027944981	negative regulation of cytokine-mediated signaling pathway
GO:0002711	0.027944981	positive regulation of T cell mediated immunity
GO:0003091	0.027944981	renal water homeostasis
GO:0009125	0.027944981	nucleoside monophosphate catabolic process
GO:0010885	0.027944981	regulation of cholesterol storage
GO:0046640	0.027944981	regulation of alpha-beta T cell proliferation
GO:0046697	0.027944981	decidualization
GO:0090181	0.027944981	regulation of cholesterol metabolic process
GO:0002460	0.02943091	adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains
GO:0002696	0.02990841	positive regulation of leukocyte activation
GO:0007187	0.02990841	G-protein signaling, coupled to cyclic nucleotide second messenger
GO:0001829	0.030239309	trophectodermal cell differentiation
GO:0006607	0.030239309	NLS-bearing substrate import into nucleus
GO:0010745	0.030239309	negative regulation of macrophage derived foam cell differentiation
GO:0010878	0.030239309	cholesterol storage
GO:0043370	0.030239309	regulation of CD4-positive, alpha beta T cell differentiation
GO:0045191	0.030239309	regulation of isotype switching
GO:0045577	0.030239309	regulation of B cell differentiation
GO:0050891	0.030239309	multicellular organismal water homeostasis
GO:0002250	0.030389025	adaptive immune response
GO:0050863	0.030872742	regulation of T cell activation
GO:0048585	0.03234233	negative regulation of response to stimulus
GO:0050867	0.03234233	positive regulation of cell activation
GO:0002717	0.032528396	positive regulation of natural killer cell mediated immunity

Continued on Next Page...

Table D.10 – Continued

GO ID	P Value	GO Term
GO:0010631	0.032528396	epithelial cell migration
GO:0010632	0.032528396	regulation of epithelial cell migration
GO:0010888	0.032528396	negative regulation of lipid storage
GO:0034375	0.032528396	high-density lipoprotein particle remodeling
GO:0042147	0.032528396	retrograde transport, endosome to Golgi
GO:0042994	0.032528396	cytoplasmic sequestering of transcription factor
GO:0045954	0.032528396	positive regulation of natural killer cell mediated cytotoxicity
GO:0050854	0.032528396	regulation of antigen receptor-mediated signaling pathway
GO:0050995	0.032528396	negative regulation of lipid catabolic process
GO:0060716	0.032528396	labyrinthine layer blood vessel development
GO:0090132	0.032528396	epithelium migration
GO:0055080	0.032742446	cation homeostasis
GO:0046058	0.032838285	cAMP metabolic process
GO:0001893	0.034812254	maternal placenta development
GO:0002702	0.034812254	positive regulation of production of molecular mediator of immune response
GO:0032091	0.034812254	negative regulation of protein binding
GO:0046633	0.034812254	alpha-beta T cell proliferation
GO:0070661	0.034852141	leukocyte proliferation
GO:0019216	0.036393627	regulation of lipid metabolic process
GO:0051649	0.036897528	establishment of localization in cell
GO:0002709	0.037090894	regulation of T cell mediated immunity
GO:0042982	0.037090894	amyloid precursor protein metabolic process
GO:0046676	0.037090894	negative regulation of insulin secretion
GO:0051208	0.037090894	sequestering of calcium ion
GO:0090130	0.037090894	tissue migration
GO:0030097	0.03765206	hemopoiesis
GO:0030098	0.03796129	lymphocyte differentiation
GO:0045595	0.038541331	regulation of cell differentiation
GO:0032844	0.039020736	regulation of homeostatic process
GO:0043691	0.039364327	reverse cholesterol transport
GO:0045058	0.039364327	T cell selection
GO:0045940	0.039364327	positive regulation of steroid metabolic process
GO:0090278	0.039364327	negative regulation of peptide hormone secretion
GO:0006606	0.039554713	protein import into nucleus
GO:0019935	0.0406311	cyclic-nucleotide-mediated signaling
GO:0042592	0.040906208	homeostatic process
GO:0010627	0.041021136	regulation of intracellular protein kinase cascade
GO:0051170	0.041173479	nuclear import
GO:0002792	0.041632566	negative regulation of peptide secretion
GO:0006516	0.041632566	glycoprotein catabolic process
GO:0030104	0.041632566	water homeostasis
GO:0030838	0.041632566	positive regulation of actin filament polymerization
GO:0046638	0.041632566	positive regulation of alpha-beta T cell differentiation
GO:0051220	0.041632566	cytoplasmic sequestering of protein
GO:0051412	0.041632566	response to corticosterone stimulus
GO:0060441	0.041632566	epithelial tube branching involved in lung morphogenesis
GO:0019222	0.042224827	regulation of metabolic process
GO:0031400	0.042817175	negative regulation of protein modification process
GO:0048534	0.043888965	hemopoietic or lymphoid organ development
GO:0001825	0.043895621	blastocyst formation
GO:0002718	0.043895621	regulation of cytokine production involved in immune response
GO:0042992	0.043895621	negative regulation of transcription factor import into nucleus
GO:0043029	0.043895621	T cell homeostasis
GO:0060674	0.043895621	placenta blood vessel development
GO:0009187	0.044485396	cyclic nucleotide metabolic process
GO:0043367	0.046153505	CD4-positive, alpha beta T cell differentiation
GO:0006810	0.04615684	transport
GO:0007243	0.046177765	intracellular protein kinase cascade

Continued on Next Page. . .

Table D.10 – Continued

GO ID	P Value	GO Term
GO:0023014	0.046177765	signal transmission via phosphorylation event
GO:0051094	0.046521539	positive regulation of developmental process
GO:0042308	0.048406228	negative regulation of protein import into nucleus
GO:0045744	0.048406228	negative regulation of G-protein coupled receptor protein signaling pathway
GO:0015031	0.048818151	protein transport
GO:0034504	0.049050825	protein localization in nucleus
GO:0051707	0.049921612	response to other organism

Appendix E

Transcriptomic analysis of drugs

E.1 CMAP sample breakdown

	HL60	MCF7	PC3	SKMEL5	ssMCF7	Total
Control	177	492	277	5	5	956
Treatment	1229	3095	1741	17	18	6100
Total	1406	3587	2018	22	23	7056

Table E.1: Cross-tab of the number of CMAP samples that were controls and treatments and the corresponding cell lines.

	HL60	MCF7	PC3	SKMEL5	ssMCF7	Total
HG-U133A	396	218	148	22	23	807
High Throughput HG-U133A	1010	3149	1870	0	0	6029
High Throughput HG-U133A EA	0	220	0	0	0	220
Total	1406	3587	2018	22	23	7056

Table E.2: Cross-tab of the number of CMAP samples that were performed on the various gene expression platforms and the corresponding cell lines.

	DMSO	ethanol	medium	Total
HG-U133A	732	15	60	807
High Throughput HG-U133A	6029	0	0	6029
High Throughput HG-U133A EA	220	0	0	220
Total	6981	15	60	7056

Table E.3: Cross-tab of the number of CMAP samples that were performed on the various gene expression platforms and the corresponding treatment mediums.

	HL60	MCF7	PC3	SKMEL5	ssMCF7	Total
DMSO	1401	3532	2008	22	18	6981
ethanol	0	15	0	0	0	15
medium	5	40	10	0	5	60
Total	1406	3587	2018	22	23	7056

Table E.4: Cross-tab of the number of CMAP samples that were performed using the various treatment mediums and the corresponding cell lines.

E.2 Stem cell marker genes

We showed in Section 4.4 how we can make use of 189 stem cell marker genes to not only stratify pluripotentiality and malignancy, but also to provide clinical gradings for various types of tumors. Naturally, one would inquire as to how these genes fair in the context of the Connectivity Map [69]. As the stem cell marker genes were derived from data performed on the HG-U133 Plus 2.0 array, there unfortunately is not a complete overlap with the set of genes for which data is available in CMAP (which was performed using the HG-U133A array). As such, the CMAP based stem cell analyses were performed using only the 140 genes out of the 189 genes that were common to both platforms.

Also, as CMAP requires a list of up- and down-regulated genes for an input query signature, we computed mean difference of expression for the each of the 140 genes as compared to the background expression intensity. For example, one of the marker genes in the list is FGF2 fibroblast growth factor. To compute whether FGF2 is up- or down-regulated in stem cells, we took all samples associated with stem cells (the same ones used to derive the stem cell marker gene set) and computed the mean expression for FGF2. Similarly, using all samples not associated with stem cells we computed the mean background expression for FGF2. The set of up-regulated stem cell genes was thus the ones that had a mean expression level greater than the background, and conversely, the set of down-regulated genes were those that had a mean expression that was lower than the background. Table E.5 contains the set of 140 genes along with their respective differences from the background distributions.

Table E.5:

Gene ID	Gene Name	Mean Difference From Background
9787	DLGAP5	6020.48
10112	KIF20A	6001.78
8091	HMGA2	5875.82
9493	KIF23	5715.13
10403	NDC80	5586.72
55388	MCM10	5314.1
1062	CENPE	5306.63
3832	KIF11	5297.51
701	BUB1B	5113.58
586	BCAT1	5000.79

Continued on Next Page...

Table E.5 – Continued

Gene ID	Gene Name	Mean Difference From Background
891	CCNB1	4927.05
3037	HAS2	4799.31
2247	FGF2	4503.43
4998	ORC1L	4473.67
54908	CCDC99	4463.26
79070	KDELC1	4391.97
993	CDC25A	4314.51
2535	FZD2	4265.59
11200	CHEK2	4221.95
890	CCNA2	4127.98
10468	FST	4105.38
51203	NUSAP1	3885.31
1789	DNMT3B	3772.67
5427	POLE2	3749.71
8092	ALX1	3453.42
5865	RAB3B	3430.61
7223	TRPC4	3400
4613	MYCN	3317.84
4173	MCM4	3313.38
51339	DACT1	3247.99
5198	PFAS	3238.2
2068	ERCC2	3181.68
80210	ARMC9	3180.88
4436	MSH2	3162.11
4883	NPR3	3154.66
6502	SKP2	3061.68
4919	ROR1	3026.45
83729	INHBE	2982.7
22800	RRAS2	2916.42
2491	CENPI	2895.62
5917	RARS	2865.68
8820	HESX1	2775.39
125058	TBC1D16	2744.89
11245	GPR176	2727.35
55781	RIOK2	2723.37
5557	PRIM1	2710.14
9315	C5orf13	2612.31
10797	MTHFD2	2562.45
10606	PAICS	2540.7
10973	ASCC3	2536.84
54069	C21orf45	2532.17
79000	C1orf135	2447.35
8805	TRIM24	2435.68
27241	BBS9	2432.53
29889	GNL2	2415.06
9573	GDF3	2408.4
2956	MSH6	2407.1
54801	HAUS6	2395.74
594	BCKDHB	2377.5
55888	ZNF167	2365.16
9373	PLAA	2328.64
6905	TBCE	2295.12
54937	SOHLH2	2157.48
9477	MED20	2151.66
9823	ARMCX2	2090
3843	IPO5	2088.34
4175	MCM6	2064.02
54881	TEX10	2059.02

Continued on Next Page...

Table E.5 – Continued

Gene ID	Gene Name	Mean Difference From Background
8891	EIF2B3	1995.56
1004	CDH6	1966.76
8520	HAT1	1963.74
9933	KIAA0020	1881.71
657	BMPR1A	1867.47
9187	SLC24A1	1801.74
23463	ICMT	1641.3
23057	NMNAT2	1622.73
8364	HIST1H4C	1595.1
27292	DIMT1L	1589.94
55813	UTP6	1571.41
55757	UGCGL2	1461.34
5810	RAD1	1386.42
3363	HTR7	1375.31
3093	UBE2K	1341.05
2632	GBE1	1320.45
25813	SAMM50	1238.9
11260	XPOT	1233.17
60492	CCDC90B	1211
23517	SKIV2L2	1195.19
128	ADH5	1177.49
8872	CDC123	1126.14
333	APLP1	971.1
1620	DBC1	952.37
3066	HDAC2	934.57
900	CCNG1	917.08
5162	PDHB	895
81624	DIAPH3	846.75
6634	SNRPD3	815.01
2617	GARS	809.04
3376	IARS	788.81
11222	MRPL3	716.37
1653	DDX1	661.45
23435	TARDBP	613.56
7520	XRCC5	558.37
26263	FBXO22	526.16
1665	DHX15	469.57
23478	SEC11A	456.88
27249	MMADHC	403.36
51637	C14orf166	402.42
5250	SLC25A3	140.02
6303	SAT1	-1250.14
409	ARRB2	-1378.76
128553	TSHZ2	-1542.25
6601	SMARCC2	-1937.41
4601	MXI1	-1989.15
54812	AFTPH	-1995.69
400949	FKSG49	-2440.06
29990	PILRB	-2529.84
2533	FYB	-2829.75
NA	NA	-2937.06
1230	CCR1	-3038.63
91543	RSAD2	-3138.36
5143	PDE4C	-3337.06
3726	JUNB	-3373.97
55340	GIMAP5	-3646.81
9619	ABCG1	-3850.84
3587	IL10RA	-3974.35

Continued on Next Page...

Table E.5 – Continued

Gene ID	Gene Name	Mean Difference From Background
3109	HLA-DMB	-3996.46
26137	ZBTB20	-4120.89
397	ARHGDIB	-4343.9
1520	CTSS	-4407.48
1512	CTSH	-4462.69
474344	GIMAP6	-4492.8
3127	HLA-DRB5	-4613.64
963	CD53	-4678.29
5175	PECAM1	-4755.39
3113	HLA-DPA1	-4782.27
5788	PTPRC	-5178.57
3117	HLA-DQA1	-5783.23
8743	TNFSF10	-5867.07
28984	C13orf15	-6212.39

Bibliography

[1] ClinicalTrials.gov. US National Institute of Health.

[2] Expression Project for Oncology (expO). International Genomics Consortium.

[3] Priit Adler, Raivo Kolde, Meelis Kull, Aleksandr Tkachenko, Hedi Peterson, Jüri Reimand, and Jaak Vilo. Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. *Genome Biology*, 10(12):R139, 2009.

[4] S Akiyama, N Shiraishi, Y Kuratomi, M Nakagawa, and M Kuwano. Circumvention of multiple-drug resistance in human cancer cells by thioridazine, trifluoperazine, and chlorpromazine. *Journal of the National Cancer Institute*, 76(5):839–844, May 1986.

[5] Muhammad Al-Hajj, Max S. Wicha, Adalberto Benito-Hernandez, Sean J Morrison, and Michael F. Clarke. Prospective identification of tumorigenic breast cancer cells. *Proceedings of the National Academy of Sciences*, 100(7):3983–3988, April 2003.

[6] Ash A. Alizadeh, Michael B. Eisen, R. Eric Davis, Chi Ma, Izidore S. Lossos, Andreas Rosenwald, Jennifer C. Boldrick, Hajeer Sabet, Truc Tran, Xin Yu, John I. Powell, Liming Yang, Gerald E. Marti, Troy Moore, James Hudson, Lisheng Lu, David B. Lewis, Robert Tibshirani, Gavin Sherlock, Wing C. Chan, Timothy C. Greiner, Dennis D. Weisenburger, James O. Armitage, Roger Warnke, Ronald Levy, Wyndham Wilson, Michael R. Grever, John C. Byrd, David Botstein, Patrick O. Brown, and Louis M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, February 2000.

[7] Alan R Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium*, 2001.

[8] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori Aa Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for

the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29, May 2000.

[9] G Aubel-Sadron and D Londos-Gagliardi. Daunorubicin and doxorubicin, anthracycline antibiotics, a physicochemical and biological review. *Biochimie*, 66(5):333–352, May 1984.

[10] Tyler W. H. Backman, Yiqun Cao, and Thomas Girke. ChemMine tools: an online service for analyzing and clustering small molecules. *Nucleic Acids Research*, May 2011.

[11] Sharmila A Bapat, Avinash M Mali, Chaitanyananda B Koppikar, and Nawneet K Kurrey. Stem and progenitor-like cells contribute to the aggressive behavior of human epithelial ovarian cancer. *Cancer Research*, 65(8):3025–3029, April 2005.

[12] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.

[13] Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Rolf N Muertter, Michelle Holko, Oluwabukunmi Ayanbule, Andrey Yefanov, and Alexandra Soboleva. NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Research*, pages 1–6, November 2010.

[14] Ittai Ben-Porath, Matthew W Thomson, Vincent J Carey, Ruping Ge, George W Bell, Aviv Regev, and Robert A Weinberg. An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors. *Nature Genetics*, 40(5):499–507, May 2008.

[15] Gabriel F Berriz, John E Beaver, Can Cenik, Murat Tasan, and Frederick P Roth. Next generation software for functional trend analysis. *Bioinformatics*, 25(22):3043–3044, August 2009.

[16] David Blumenthal. Stimulating the adoption of health information technology. *The New England Journal of Medicine*, 105(3):28–29, April 2009.

[17] Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database):D267–D270, 2004.

[18] J Bridgewater, R van Laar, and L Van'T Veer. Gene expression profiling may improve diagnosis in patients with carcinoma of unknown primary. *British Journal of Cancer*, 98:1425–1430, February 2008.

[19] Atul J Butte and Isaac S Kohane. Creation and implications of a phenome-genome network. *Nature Biotechnology*, 24(1):55–62, 2006.

[20] Stephen J Campbell, Anna Gaulton, Jason Marshall, Dmitri Bichko, Sid Martin, Cory Brouwer, and Lee Harland. Visualizing the drug target landscape. *Drug Discovery Today*, 15(1-2):3–15, 2010.

[21] Raymond E. Carhart, Dennis H. Smith, and R. Venkataraghavan. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *Journal of Chemical Information and Modeling*, 25:64–73, 1985.

[22] Jay Chang, Cecile Lee, Ki-Baik Hahm, Youngsuk Yi, Shin-Geon Choi, and Seong-Jin Kim. Over-expression of ERT(ESX/ESE-1/ELF3), an ets-related transcription factor, induces endogenous TGF-beta type II receptor expression and restores the TGF-beta signaling pathway in Hs578t human breast cancer cells. *Oncogene*, 19(1):151–154, January 2000.

[23] Anne T Collins, Paul A Berry, Catherine Hyde, Michael J Stower, and Norman J Maitland. Prospective identification of tumorigenic prostate cancer stem cells. *Cancer Research*, 65(23):10946–10951, December 2005.

[24] Jeffery Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. *Usenix SDI*, 2004.

[25] Vikas P Deshpande, Robert F Erbacher, and Chris Harris. 2007 IEEE SMC Information Assurance and Security Workshop. In *2007 IEEE SMC Information Assurance and Security Workshop*, pages 333–340. IEEE, 2007.

[26] Preet K Dhillon, Kathryn L Penney, Fredrick Schumacher, Jennifer R Rider, Howard D Sesso, Michael Pollack, Michelangelo Fiorentino, Stephen Finn, Massimo Loda, Nader Rifai, Lorelei A Mucci, Edward L Giovannucci, Meir J Stampfer, and Jing Ma. Common polymorphisms in the adiponectin and its receptor genes, adiponectin levels and the risk of prostate cancer. *Cancer Epidemiology, Biomarkers & Prevention*, September 2011.

[27] Gabriela Dontu, Muhammad Al-Hajj, Wissam M. Abdallah, Michael F. Clarke, and Max S. Wicha. Stem cells in normal breast development and breast cancer. *Cell Proliferation*, 36:59–72, September 2003.

[28] J T Dudley, M Sirota, M Shenoy, R K Pai, S Roedder, A P Chiang, A A Morgan, M M Sarwal, P J Pasricha, and A J Butte. Computational Repositioning of the Anticonvulsant Topiramate for Inflammatory Bowel Disease. *Science Translational Medicine*, 3(96):96ra76–96ra76, August 2011.

[29] Joel T Dudley and Atul J Butte. Biomarker and Drug Discovery for Gastroenterology Through Translational Bioinformatics. *Gastroenterology*, 139(3):735–741, 2010.

[30] Joel T Dudley, Robert Tibshirani, Tarangini Deshpande, and Atul J Butte. Disease signatures are robust across tissues and experiments. *Molecular Systems Biology*, 5:1–8, September 2009.

[31] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

[32] S Falcon and R Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–258, January 2007.

[33] Dong Fang, Thiennga K Nguyen, Kim Leishear, Rena Finko, Angela N Kulp, Susan Hotz, Patricia A Van Belle, Xiaowei Xu, David E Elder, and Meenhard Herlyn. A tumorigenic subpopulation with stem cell properties in melanomas. *Cancer Research*, 65(20):9328–9337, October 2005.

[34] Eric R Fearon. Human Cancer Syndromes: Clues to the Origin and Nature of Cancer. *Science*, 278(5340):1043–1050, November 1997.

[35] Marc Feldmann. Development of anti-TNF therapy for rheumatoid arthritis. *Nature Reviews Immunology*, 2(5):364–371, 2002.

[36] PJ Fialkow. Stem cell origin of human myeloid blood cell neoplasms. *Verhandlungen der Deutschen Gesellschaft fur Pathologie*, 74:43–47, 1990.

[37] S P Fodor, J L Read, M C Pirrung, L Stryer, A T Lu, and D Solas. Light-directed, spatially addressable parallel chemical synthesis. *Science*, 251(4995):767–773, February 1991.

[38] Nicolas O Fortunel, Hasan H Otu, Huck-Hui Ng, Jinhui Chen, Xiuqian Mu, Timothy Chevassut, Xiaoyu Li, Marie Joseph, Charles Bailey, Jacques A Hatzfeld, Antoinette Hatzfeld, Fatih Usta, Vinsensius B Vega, Philip M Long, Towia A Libermann, and Bing Lim. Comment on " 'Stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature". *Science*, 302(5644):393; author reply 393, October 2003.

[39] Wataru Fujibuchi, Larisa Kiseleva, Takeaki Taniguchi, Hajime Harada, and Paul Horton. CellMontage: similar expression profile search server. *Bioinformatics*, 23(22):3103–3104, November 2007.

[40] Gregory N Fuller, Cristian Mircean, Ioan Tabus, Ellen Taylor, Raymond Sawaya, Janet M Bruner, Ilya Shmulevich, and Wei Zhang. Molecular voting for glioma classification reflecting heterogeneity in the continuum of cancer progression. *Oncology reports*, 14(3):651–656, September 2005.

[41] M Teresa Garcia-Unzueta, Andres Herran, Deirdre Sierra-Biddle, J Antonio Amado, J Luis Vázquez-Barquero, and Concepción Alvarez. Alterations of liver function test in patients treated with antipsychotics. *Journal of Clinical Laboratory Analysis*, 17(6):216–218, 2003.

[42] Katherine S Garman, Chaitanya R Acharya, Elena Edelman, Marian Grade, Jochen Gaedcke, Shivani Sud, William Barry, Anna Mae Diehl, Dawn Provenzale, Geoffrey S Ginsburg, B Michael Ghadimi, Thomas Ried, Joseph R Nevins,

Sayan Mukherjeea, David Hsua, and Anil Potti. A genomic approach to colon cancer risk stratification yields biologic insights into therapeutic opportunities. *Proceedings of the National Academy of Sciences*, 105(49):19432–19437, 2008.

[43] C Parker Gibbs, Valery G Kukekov, John D Reith, Olga Tchigrinova, Oleg N Suslov, Edward W Scott, Steven C Ghivizzani, Tatyana N Ignatova, and Dennis A Steindler. Stem-like cells in bone sarcomas: implications for tumorigenesis. *Neoplasia*, 7(11):967–976, November 2005.

[44] PS Gill, J Wernz, DT Scadden, P Cohen, Mukwaya GM, JH von Roenn, M Jacobs, S Kempin, I Silverberg, G Gonzales, MU Rarick, AM Myers, F Shepherd, C Sawka, MC Pike, and ME Ross. Randomized phase III trial of liposomal daunorubicin versus doxorubicin, bleomycin, and vincristine in AIDS-related Kaposi's sarcoma. *Journal of Clinical Oncology*, 14(8):2353–2364, August 1996.

[45] Annuska M Glas, Arno Floore, Leonie J M J Delahaye, Anke T. Witteveen, Rob C F Pover, Niels Bakx, Jaana S T Lahti-Domenici, Tako J Bruinsma, Marc O Warmoes, Rene Bernards, Lodewyk F A Wessels, and Laura J van't Veer. Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC Genomics*, 7:278, 2006.

[46] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, May 2007.

[47] A K Golembesky, M D Gammon, K E North, J T Bensen, J C Schroeder, S L Teitelbaum, A I Neugut, and R M Santella. Peroxisome proliferator-activated receptor-alpha (PPARA) genetic polymorphisms and breast cancer risk: a Long Island ancillary study. *Carcinogenesis*, 29(10):1944–1949, March 2008.

[48] Todd R Golub, Donna K Slonim, Pablo Tamayo, C Huard, M Gaasenbeek, Jill P Mesirov, H Coller, M L Loh, J R Downing, M A Caligiuri, C D Bloomfield, and Eric S Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, October 1999.

[49] Assaf Gottlieb, Gideon Y Stein, Eytan Ruppin, and Roded Sharan. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular Systems Biology*, 7:1–9, June 2011.

[50] Piyush B Gupta, Christine M Fillmore, Guozhi Jiang, Sagi D Shapira, Kai Tao, Charlotte Kuperwasser, and Eric S Lander. Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell*, 146(4):633–644, August 2011.

[51] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *Intelligent Systems*, 2009.

[52] Erez Hartuv and Ron Shamir. A Clustering Algorithm based on Graph Connectivity. *Journal Information Processing Letters*, 76(4-6):1–10, December 2000.

[53] Duane C Hassane, Monica L Guzman, Cheryl Corbett, Xiaojie Li, Ramzi Abboud, Fay Young, Jane L Liesveld, Martin Carroll, and Craig T Jordan. Discovery of agents that eradicate leukemia stem cells using an in silico screen of public gene expression data. *Blood*, 111(12):5654–5662, June 2008.

[54] Gloria H. Heppner and Bonnie E. Miller. Tumor heterogeneity: biological implications and therapeutic consequences. *Cancer Metastasis Reviews*, 2(1):5–23, 1983.

[55] J Hiraga, A Katsumi, T Iwasaki, A Abe, H Kiyoi, T Matsushita, T Kinoshita, and T Naoe. Prognostic analysis of aberrant somatic hypermutation of RhoH gene in diffuse large B cell lymphoma. *Leukemia*, 21(8):1846–1847, August 2007.

[56] Halyan Huang, Chun-Chi Liu, and Xianghong Jasmine Zhou. Bayesian approach to transforming public gene expression repositories into disease Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. *Proceedings of the National Academy of Sciences*, 107(15):6823–6828, 2010.

[57] Kyu-Baek Hwang, Sek Won Kong, Steve A Greenberg, and Peter J Park. Combining gene expression data from different generations of oligonucleotide arrays. *BMC Bioinformatics*, 5:159, October 2004.

[58] Francesco Iorio, Roberto Tagliaferri, and Diego di Bernardo. Identifying Network of Drug Mode of Action by Gene Expression Profiling. *Journal of Computational Biology*, 16(2):241–251, February 2009.

[59] H.V. Jagadish, Laks V.S. Kaskhmanana, and Divesh Srivastava. Hierarchical or Relational? A Case for a Modern Hierarchical Data Model. *IEEE Workshop on Knowledge and Data Exchange*, 1999.

[60] A F Javier, Z Bata-Csorgo, C N Ellis, S Kang, J J Voorhees, and K D Cooper. Rapamycin (sirolimus) inhibits proliferating cell nuclear antigen expression and blocks cell cycle in the G1 phase in human keratinocyte stem cells. *The Journal of Clinical Investigation*, 99(9):2094–2099, May 1997.

[61] Virginia Kaklamani, Nengjun Yi, Kui Zhang, Maureen Sadim, Kenneth Offit, Carole Oddoux, Harry Ostrer, Christos Mantzoros, and Boris Pasche. Polymorphisms of ADIPOQ and ADIPOR1 and prostate cancer risk. *Metabolism*, 60(9):1234–1243, September 2011.

[62] Gerhard Klebe, editor. *Virtual Screening: An Alternative or Complement to High Throughput Screening*. Springer, softcover reprint of hardcover 1st ed. 2001 edition, December 2010.

[63] Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, Yannick Djoumbou, Roman Eisner, An Chi Guo, and David S Wishart. DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Research*, 39(Database):D1035–D1041, 2011.

[64] Isaac S Kohane. The twin questions of personalized medicine: who are you and whom do you most resemble? *Genome Medicine*, 1(1):4, 2009.

[65] Isaac S Kohane, Alvin Kho, and Atul J Butte. *Microarrays for an Integrative Genomics*. The MIT Press, 1 edition, August 2002.

[66] Isaac S Kohane, Daniel R Masys, and Russ B Altman. The incidentalome: a threat to genomic medicine. *The Journal of the American Medical Association*, 296(2):212–215, July 2006.

[67] Ilya Kupershmidt, Qiaojuan Jane Su, Anoop Grewal, Suman Sundaresh, Inbal Halperin, James Flynn, Mamatha Shekar, Helen Wang, Jenny Park, Wenwu Cui, Gregory D Wall, Robert Wisotzkey, Satnam Alag, Saeid Akhtari, and Mostafa Ronaghi. Ontology-based meta-analysis of global collections of high-throughput public data. *PloS one*, 5(9), 2010.

[68] Justin Lamb. The Connectivity Map: a new tool for biomedical research. *Nature Reviews Cancer*, 7:54–60, 2007.

[69] Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N Ross, Michael Reich, Haley Hieronymus, Guo Wei, Scott A Armstrong, Stephen J Haggarty, Paul A Clemons, Ru Wei, Steven A Carr, Eric S Lander, and Todd R Golub. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science*, 313:1929–1935, September 2006.

[70] Joo-Young Lee, Hikari Hashizaki, Tsuyoshi Goto, Tomoya Sakamoto, Nobuyuki Takahashi, and Teruo Kawada. Activation of peroxisome proliferator-activated receptor-Î± enhances fatty acid oxidation in human adipocytes. *Biochemical and Biophysical Research Communications*, 407(4):818–822, April 2011.

[71] Pulin Li and Leonard I Zon. Resolving the controversy about N-cadherin and hematopoietic stem cells. *Cell Stem Cell*, 6(3):199–202, March 2010.

[72] Simon Liu, Wei Ma, Robin Moore, Vikraman Ganesan, and Stuart Nelson. RxNorm: Prescription for Electronic Drug Information Exchange. *IT Professional*, 7:17–23, September 2005.

[73] Xiong Liu, Xueping Yu, Donald J Zack, Heng Zhu, and Jiang Qian. TiGER: A database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, 9(271), 2008.

[74] Neethan A Lobo, Yohei Shimono, Dalong Qian, and Michael F. Clarke. The Biology of Cancer Stem Cells. *Annual Review of Cell and Developmental Biology*, 23(1):675–699, November 2007.

[75] Joseph Loscalzo, Isaac S Kohane, and Albert-László Barabási. Human disease classification in the postgenomic era: A complex systems approach to human pathobiology. *Molecular Systems Biology*, 3(124), 2007.

[76] Rui Lui, Xinhao Wang, Grace Y. Chen, Piero Dalerba, Austin Gurney, Timothy Hoey, Gavin Sherlock, John Lewicki, Kerby Shedden, and Michael F. Clarke. The Prognostic Role of a Gene Signature from Tumorigenic Breast-Cancer Cells. *The New England Journal of Medicine*, 356:217–226, January 2007.

[77] Margus Lukk, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen, and Alvis Brazma. A global map of human gene expression. *Nature Biotechnology*, 28:322–324, April 2010.

[78] Traci R Lyons, Jenean O'Brien, Virginia F Borges, Matthew W Conklin, Patricia J Keely, Kevin W Eliceiri, Andriy Marusyk, Aik-Choon Tan, and Pepper Schedin. Postpartum mammary gland involution drives progression of ductal carcinoma in situ through collagen and COX-2. *Nature Medicine*, 17(9):1109–1115, September 2011.

[79] David Maier. *Theory of Relational Databases.* Computer Science Press, 1983.

[80] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval.* Cambridge Univ Press, 2008.

[81] Matthew N McCall, Karan Uppal, Harris A Jaffee, Michael J Zilliox, and Rafael A Irizarry. The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Research*, 39(Database):D1011–D1015, 2011.

[82] James H McClellen, Ronald W Schafer, and Mark A Yoder. *DSP First: A Multimedia Approach.* Prentice Hall, 1998.

[83] A T McCray, S Srinivasan, and A C Browne. Lexical methods for managing variation in biomedical terminologies. *Proceedings / the ... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care*, pages 235–239, 1994.

[84] US National Library of Medicine. *Genetics Home Reference.* Lister Hill National Center for Biomedical Communications, November 2011.

[85] Karin B. Michels, Caren G. Solomon, Frank B. Hu, Bernard A. Rosner, Susan E. Hankinson, Graham A. Colditz, and JoAnn E. Manson. Type 2 Diabetes and Subsequent Incidence of Breast Cancer in the Nurses' Health Study. *Diabetes Care*, 26(6):1752–1758, June 2003.

[86] Harald Mischak, Rolf Apweiler, Rosamonde E Banks, Mark Conaway, Joshua Coon, Anna Dominiczak, Jochen H H Ehrich, Danilo Fliser, Mark Girolami, Henning Hermjakob, Denis Hochstrasser, Joachim Jankowski, Bruce A Julian, Walter Kolch, Ziad A Massy, Christian Neusuess, Jan Novak, Karlheinz Peter, Kasper Rossing, Joost Schanstra, O John Semmes, Dan Theodorescu, Visith Thongboonkerd, Eva M Weissinger, Jennifer E Van Eyk, and Tadashi Yamamoto. Clinical proteomics: a need to define the field and to begin to set adequate standards. *PROTEOMICS - Clinical Applications*, 1:148–156, 2007.

[87] P Nadkarni, R Chen, and C Brandt. UMLS concept indexing for production databases: a feasibility study. *Journal of the American Medical Informatics Association : JAMIA*, 8(1):80–91, 2001.

[88] Gautam Naik. Scientists' Elusive Goal: Reproducing Study Results, December 2011.

[89] Rajesh Narang. *Database Management Systems*. PHI Learning Pvt. Ltd., March 2006.

[90] Kamila Naxerova, Carol J Bult, Anne Peaston, Karen Fancher, Barbara B Knowles, Simon Kasif, and Isaac S Kohane. Analysis of gene expression in a developmental context emphasizes distinct biological leitmotifs in human cancers. *Genome Biology*, 9(7):R108, 2008.

[91] Larry R Nyhoff. *C++; An Introduction To Data Structures.* Prentice Hall, Upper Saddle River, NJ, United States, 1999.

[92] Nadine Obier, Christoph F Uhlemann, and Albrecht M Müller. Inhibition of histone deacetylases by Trichostatin A leads to a HoxB4-independent increase of hematopoietic progenitor/stem cell frequencies as a result of selective survival. *Cytotherapy*, 12(7):899–908, November 2010.

[93] Osamu Ogasawara, Makiko Otsuji, Kouji Watanabe, Takayasu Iizuka, Takuro Tamura, Teruyoshi Hishiki, Shoko Kawamoto, and Kousaku Okubo. BodyMap-Xs: anatomical breakdown of 17 million animal ESTs for cross-species comparison of gene expression. *Nucleic Acids Research*, 34(Database issue):D629–D631, 2006.

[94] Kouros Owzar, William T Barry, Sin-Ho Jung, Insuk Sohn, and Stephen L George. Statistical challenges in preprocessing in microarray experiments in cancer. *Clinical Cancer Research*, 14(19):5959–5966, October 2008.

[95] Soonmyung Paik, Steven Shak, Gong Tang, Chungyeul Kim, Joffre Baker, Maureen Cronin, Frederick L Baehner, Michael G Walker, Drew Watson, Taesung Park, William Hiller, Edwin R Fisher, D Lawrence Wickerham, John Bryant, and Norman Wolmark. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *The New England Journal of Medicine*, 351(27):2817–2826, 2004.

[96] D Park, R Singh, M Baym, C S Liao, and B Berger. IsoBase: a database of functionally related proteins across PPI networks. *Nucleic Acids Research*, 39(Database):D295–D300, December 2010.

[97] Helen Parkinson, Ugis Sarkans, Nikolay Kolesnikov, Niran Abeygunawardena, Tony Burdett, Miroslaw Dylag, Ibrahim Emam, Anna Farne, Emma Hastings, Ele Holloway, Natalja Kurbatova, Margus Lukk, James Malone, Roby Mani, Ekaterina Pilicheva, Gabriella Rustici, Anjan Sharma, Eleanor Williams, Tomasz Adamusiak, Marco Brandizi, Nataliya Sklyar, and Alvis Brazma. ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Research*, 39(Database):D1002–D1004, November 2010.

[98] A G Polischouk, A Holgersson, D Zong, B Stenerlow, H L Karlsson, L Moller, K Viktorsson, and R Lewensohn. The antipsychotic drug trifluoperazine inhibits DNA repair and sensitizes non small cell lung carcinoma cells to DNA double-strand break induced cell death. *Molecular Cancer Therapeutics*, 6(8):2303–2309, August 2007.

[99] Colin C Pritchard, Li Hsu, Jeffrey Delrow, and Peter S Nelson. Project normal: defining normal variance in mouse gene expression. *Proceedings of the National Academy of Sciences*, 98(23):13266–13271, November 2001.

[100] Paolo P Provenzano, David R Inman, Kevin W Eliceiri, Justin G Knittel, Long Yan, Curtis T Rueden, John G White, and Patricia J Keely. Collagen density promotes mammary tumor initiation and progression. *BMC Medicine*, 6:11, 2008.

[101] Miguel Ramalho-Santos, Soonsang Yoon, Yumi Matsuzaki, Richard C Mulligan, and Douglas A Melton. ”Stemness”: transcriptional profiling of embryonic and adult stem cells. *Science*, 298(5593):597–600, October 2002.

[102] David F Ransohoff. Bias as a threat to the validity of cancer molecular-marker research. *Nature Reviews Cancer*, 5:142–149, 2005.

[103] Daniel R Rhodes, Shanker Kalyana-Sundaram, Vasudeva Mahavisno, Radhika Varambally, Jianjun Yu, Benjamin B Briggs, Terrence R Barrette, Matthew J Anstet, Colleen Kincead-Beal, Prakash Kulkarni, Sooryanaryana Varambally, Debashis Ghosh, and Arul M Chinnaiyan. Oncomine 3.0: Genes, Pathways, and Networks in a Collection of 18,000 Cancer Gene Expression Profiles. *Neoplasia*, 9(2):166–180, February 2007.

[104] Daniel R Rhodes, Jianjun Yu, K. Shanker, Nandan Deshpande, Radhika Varambally, Debashis Ghosh, Terrence Barrette, Akhilesh Pandey, and Arul M Chinnaiyan. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *PNAS*, 101(25):9309–9314, June 2004.

[105] Lucia Ricci-Vitiani, Dario G Lombardi, Emanuela Pilozzi, Mauro Biffoni, Matilde Todaro, Cesare Peschle, and Ruggero De Maria. Identification and expansion of human colon-cancer-initiating cells. *Nature*, 445(7123):111–115, January 2007.

[106] Richard F Riedel, Alessandro Porrello, Emily Pontzer, Emily J Chenette, David S Hsu, Bala Balakumaran, Anil Potti, Joseph Nevins, and Phillip G Febbo. A genomic approach to identify molecular pathways associated with chemotherapy resistance. *Molecular Cancer Therapeutics*, 7(10):3141–3149, October 2008.

[107] Miguel N Rivera and Daniel A Haber. Wilms' tumour: connecting tumorigenesis and organ development in the kidney. *Nature Reviews Cancer*, 5(9):699–712, September 2005.

[108] Lior Rokach, Roni Romano, and Oded Maimon. Negation recognition in medical narrative reports. *Information Retrieval*, 11(6):499–538, June 2009.

[109] Roni Romano, Lior Rokach, and Oded Maimon. Cascaded Data Mining Methods for Text Understanding, with Medical Case Study. In *International Conference on Data Mining*, pages 1–5, 2006.

[110] Jennifer M Rosenbluth, Deborah J Mays, Maria F Pino, Luo Jia Tang, and Jennifer A Pietenpol. A Gene Signature-Based Approach Identifies mTOR as a Regulator of p73. *Molecular and Cellular Biology*, 28(19):5951–5964, October 2008.

[111] Marci E Schaner, Douglas T Ross, Giuseppe Ciaravino, Therese Sørlie, Olga G Troyanskaya, Maximilian Diehn, Yan C Wang, George E Duran, Thomas L Sikic, Sandra Caldeira, Hanne Skomedal, I-Ping Tu, Tina Hernandez-Boussard, Steven W Johnson, Peter J O'Dwyer, Michael J Fero, Gunnar B Kristensen, Anne-Lise Børresen-Dale, Trevor Hastie, Robert Tibshirani, Matt van de Rijn, Nelson N Teng, Teri A Longacre, David Botstein, Patrick O. Brown, and Branimir I Sikic. Gene Expression Patterns in Ovarian Carcinomas. *Molecular Biology of the Cell*, 14:4376–4386, November 2003.

[112] Mark Schena, Dari Shalon, Ronald W. Davis, and Patrick O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, October 1995.

[113] Russell Schwartz and Stanley E Shackney. Applying unmixing to gene expression data for tumor phylogeny inference. *BMC Bioinformatics*, 11(42), 2010.

[114] Paul J Scotting, David A Walker, and Giorgio Perilongo. Childhood solid tumours: a developmental disorder. *Nature Reviews Cancer*, 5(6):481–488, June 2005.

[115] Eran Segal, Nir Friedman, Daphne Koller, and Aviv Regev. A Module Map Showing Conditional Activity of Expression Modules in Cancer. *Nature Genetics*, 26(10):1090–1098, October 2004.

[116] Eran Segal, Michael Shapira, Aviv Regev, Dana Pe'er, David Botstein, Daphne Koller, and Nir Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2):166–176, June 2003.

[117] Burr Settles. Active Learning Literature Survey. *SciencesNew York*, 15(2), 2010.

[118] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13:2498–2504, 2003.

[119] Zhiao Shi, Catherine K Derow, and Bing Zhang. Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression. *BMC Systems Biology*, 4(1):74, 2010.

[120] Rohit Singh, Nathan Palmer, David Gifford, Bonnie Berger, and Ziv Bar-Joseph. Proceedings of the 22nd international conference on Machine learning - ICML '05. In *the 22nd international conference*, pages 832–839, New York, New York, USA, 2005. ACM Press.

[121] Seila K. Singh, Ian D. Clarke, Mizuhiko Terasaki, Victoria E. Bonn, Cynthia Hawkins, Jeremy Squire, and Peter B. Dirks. Identification of a Cancer Stem Cell in Human Brain Tumors. *Cancer Research*, 2003.

[122] M Sirota, J T Dudley, J Kim, A P Chiang, A A Morgan, A Sweet-Cordero, J Sage, and A J Butte. Discovery and Preclinical Validation of Drug Indications Using Compendia of Public Gene Expression Data. *Science Translational Medicine*, 3(96):96ra77–96ra77, August 2011.

[123] Erik L. L. Sonnhammer, Sean R. Eddy, and Richard Durbin. Pfam: A Comprehensive Database of Protein Domain Families Based on Seed Alignments. *PROTEINS: Structure, Function, and Genetics*, 28(3):405–420, July 1997.

[124] Alexander Statnikov, Ioannis Tsamardinos, Yerbolat Dosbayev, and Constantin F Aliferis. GEMS: a system for automated cancer diagnosis and biomarker discovery from microarray gene expression data. *International Journal of Medical Informatics*, 74(7-8):491–503, August 2005.

[125] Julie Steenhuysen. PSA test for prostate cancer not recommended: panel. *Reuters*, pages 1–2, October 2011.

[126] Kimberly Stegmaier, Steven M Corsello, Kenneth N Ross, Jenny S Wong, Daniel J Deangelo, and Todd R Golub. Gefitinib induces myeloid differentiation of acute myeloid leukemia. *Blood*, 106(8):2841–2848, October 2005.

[127] Kimberly Stegmaier, Kenneth N Ross, Sierra A Colavito, Shawn OMalley, Brent R Stockwell, and Todd R Golub. Gene expression–based high-throughput screening (GE-HTS) and application to leukemia differentiation. *Nature Genetics*, 36(3):257–263, March 2004.

[128] Thorsten Stiewe. The p53 family in differentiation and tumorigenesis. *Nature Reviews Cancer*, 7(3):165–167, February 2007.

[129] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting geneome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15278–15279, October 2005.

[130] Hitoshi Takizawa, Roland R Regoes, Chandra S Boddupalli, Sebastian Bonhoeffer, and Markus G Manz. Dynamic variation in cycling of hematopoietic stem cells in steady state and inflammation. *The Journal of Experimental Medicine*, 208(2):273–284, February 2011.

[131] Min Tang, Xiaoqiang Luo, and Salim Roukos. Active learning for statistical natural language parsing. In *Proceedings of the 40th Annual Meeting on ...*. unknown, 2002.

[132] Georg C. Terstappen and Angelo Reggiani. In silico research in drug discovery. *Trends in Pharmacological Sciences*, January 2001.

[133] Ze Tian, Nathan Palmer, Patrick Schmid, Hui Yao, Michal Galdzicki, Bonnie Berger, Erxi Wu, and Isaac S Kohane. A practical platform for blood biomarker study by using global gene expression profiling of peripheral whole blood. *PloS one*, 4(4):e5157, 2009.

[134] J C Tonn and Manfred Westphal. *Neuro-oncology of CNS tumors*. Springer Verlag, 2006.

[135] A Traverse-Glehen, Verney A, L Beseggio, P Felman, E Callet-Bauchu, C Thieblemont, M Ffrench, J-P Magaud, B Coiffier, F Berger, and G Salles. Analysis of BCL-6, CD95, PIM1, RHO/TTF and PAX5 mutations in splenic and nodal marginal zone B-cell lymphomas suggests a particular B-cell origin. *Leukemia*, 21(8):1821–1824, April 2007.

[136] Dolf Trieschnigg, Piotr Pezik, Vivian Lee, Franciska de Jong, Wessel Kraaij, and Dietrich Rebholz-Schuhmann. MeSH Up: effective MeSH text classification for improved document retrieval. *Bioinformatics*, 25(11):1412–1418, June 2009.

[137] Arzu Umar, Hyuk Kang, Annemieke M. Timmermans, Maxime P. Look, Marion E Meijer-van Gelder, Michael A. den Bakker, Navdeep Jaitly, John W. M. Martens, Theo M. Luider, John A. Foekens, and Ljiljana Paša-Tolić. Identification of a putative protein profile associated with tamoxifen therapy resistance in breast cancer. *Molecular & Cellular Proteomics*, 8(6):1278–1294, June 2009.

[138] Jane E Visvader and Geoffrey J Lindeman. Cancer stem cells in solid tumours: accumulating evidence and unresolved questions. *Nature Reviews Cancer*, 8(10):755–768, October 2008.

[139] Lisa D. Wilsbacher and Joseph S. Takahashi. Circadian rhythms: molecular basis of the clock. *Current Opinion in Genetics & Development*, 8(5):595–602, October 1998.

[140] David J Wong, Helen Liu, Todd W Ridky, David Cassarino, Eran Segal, and Howard Y Chang. Module map of stem cell genes guides creation of epithelial cancer stem cells. *Cell Stem Cell*, 2(4):333–344, April 2008.

[141] Vijay K Yechoor, Mary-Elizabeth Patti, Kohjiro Ueki, Palle G Laustsen, Robert Saccone, Ravi Rauniyar, and C Ronald Kahn. Distinct pathways of insulin-regulated versus diabetes-regulated gene expression: an in vivo analysis in MIRKO mice. *Proceedings of the National Academy of Sciences*, 101(47):16525–16530, November 2004.

[142] Junying Yu, Maxim A. Vodyanik, Kim Smuga-Otto, Jessica Antosiewicz-Bourget, Jennifer L. Frane, Shulan Tian, Jeff Nie, Gudrun A. Jonsdottir, Victor Ruotti, Ron Stewart, Igor I. Slukvin, and James A. Thomson. Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells. *Science*, 2007.

[143] Hongjuan Zhao, Börje Ljungberg, Kjell Grankvist, Torgny Rasmuson, Robert Tibshirani, and James D Brooks. Gene expression profiling predicts survival in conventional renal cell carcinoma. *PLoS Medicine*, 3(1):e13, January 2006.