Data Mining Techniques for Large-Scale Gene Expression Analysis

by

Nathan Patrick Palmer

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

September 2011

Author .....
Department of Electrical Engineering and Computer Science
August 31, 2011

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

Accepted by .....
Professor Leslie A. Kolodziejski.
Chairman, Department Committee on Graduate Students

Data Mining Techniques for Large-Scale Gene Expression Analysis

by

Nathan Patrick Palmer

Submitted to the Department of Electrical Engineering and Computer Science
on August 31, 2011, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Modern computational biology is awash in large-scale data mining problems. Several high-throughput technologies have been developed that enable us, with relative ease and little expense, to evaluate the coordinated expression levels of tens of thousands of genes, evaluate hundreds of thousands of single-nucleotide polymorphisms, and sequence individual genomes. The data produced by these assays has provided the research and commercial communities with the opportunity to derive improved clinical prognostic indicators, as well as develop an understanding, at the molecular level, of the systemic underpinnings of a variety of diseases.

Aside from the statistical methods used to evaluate these assays, another, more subtle challenge is emerging. Despite the explosive growth in the amount of data being generated and submitted to the various publicly available data repositories, very little attention has been paid to managing the phenotypic characterization of their samples (i.e., managing class labels in a controlled fashion). If sense is to be made of the underlying assay data, the samples' descriptive metadata must first be standardized in a machine-readable format.

In this thesis, we explore these issues, specifically within the context of curating and analyzing a large DNA microarray database. We address three main challenges. First, we acquire a large subset of a publicly available microarray repository and develop a principled method for extracting phenotype information from free-text sample labels, then use that information to generate an index of the sample's medically-relevant annotation. The indexing method we develop, Concordia, incorporates pre-existing expert knowledge relating to the hierarchical relationships

between medical terms, allowing queries of arbitrary specificity to be efficiently answered. Second, we describe a highly flexible approach to answering the question: "Given a previously unseen gene expression sample, how can we compute its similarity to all of the labeled samples in our database, and how can we utilize those similarity scores to predict the phenotype of the new sample?" Third, we describe a method for identifying phenotype-specific transcriptional profiles within the context of this database, and explore a method for measuring the relative strength of those signatures across the rest of the database, allowing us to identify molecular signatures that are shared across various tissues ad diseases. These shared fingerprints may form a quantitative basis for optimal therapy selection and drug repositioning for a variety of diseases.

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

Acknowledgments

The author would like to thank Dr. Bonnie Berger and Dr. Isaac Kohane for many years of thoughtful guidance and support. He would also like to thank his close collaborator, Patrick Schmid, for an enduring friendship forged over many cups of coffee and shared research interests. Finally, he would like to thank his loving parents and brother, and especially his wife, Kaitlyn, whose patience and support made this thesis work possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Modern biology is awash in data mining problems, many of a high-dimensional nature. Several high-throughput technologies have been developed that enable us, with relative ease and little expense, to evaluate the coordinated expression levels of tens of thousands of genes [24, 33], evaluate hundreds of thousands of single-nucleotide polymorphisms [6], and sequence individual genomes [30]. The data produced by these assays has provided the research and commercial communities with the opportunity to derive improved clinical prognostic indicators, as well as develop an understanding, at the molecular level, of the systemic underpinnings of a variety of diseases.

Although several of these technologies have been available for a number of years, there were relatively few samples available, with respect to the number of features assayed. This is beginning to change, with several government-funded biological data repositories boasting sample collections numbering in the hundreds of thousands [82, 14]. The accessibility of such large and phenotypically diverse data resources should eventually enable large-scale data-driven hypothesis generation and testing. This

will necessitate, however, a new generation of data mining techniques that are tuned to the specific signals generated by these assays, and aware of their inherent technical and biological noise properties.

Aside from the statistical methods used to evaluate the assays themselves, another, more subtle challenge is emerging. Despite the explosive growth in the amount of data being generated and submitted to the various publicly-available biological data repositories (or perhaps *because of* that growth), very little work has gone into managing the phenotypic characterization of their samples. If sense is to be made of the underlying assay data, something must first be done to standardize the descriptive nomenclature used to annotate each sample and data set. If we are to bring data mining and machine learning methods to bear on these resources, the sample labeling, which is at present largely based on free-text, must be made machine readable.

In this thesis, we explore these issues, specifically within the context of curating and analyzing a large DNA microarray database [83].

1.1 Microarrays

There are three main classes of macromolecules that form the information and physical superstructures required for life: deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and proteins [4]. The "central dogma of molecular biology" [27] asserts a model of information flow between these molecules that closely reflects the biomechanical operation of the cell. DNA contains all of the fundamental instructions required to form a living organism. Those instructions, encoded primarily as genes, are passed along, via gene *transcription*, through messenger RNA (mRNA) molecules that inform the synthesis of proteins, via *translation*. Those protein end-products

are the organic polymers responsible for directing nearly all of the activity within living cells.

This arrangement is loosely comparable to that of a modern computer executing a program. DNA can be seen as the rough (although inexact) analog to a program's source code: it provides a predefined set of instructions for running a set of algorithms in response to input. Consequently, the state of the machine's memory is fully determined by the combination of that source code along with the input supplied to the program – as the mRNA messages observable at any point in time in a cell are largely the result of the instructions provided by DNA in response to the environment.

Thus, quantification of mRNA concentrations has become a popular mechanism for characterizing the molecular state of a cell [5]. Modern DNA microarrays are capable of measuring the relative expression intensities of tens of thousands of genes simultaneously. A thorough explanation of microarray technology is beyond the scope of this thesis, but can readily be found elsewhere [47]. Briefly, microarrays consist of a set of DNA "probes" bound to a solid substrate over which flourescence-tagged cDNA copy of the sample RNA is hybridized. After sample hybridization, the array is scanned, and the relative flourescence of each probe is recorded. Gene-level transcription levels are then inferred, based on the sequences of the array's probes.

Acceptance of this "high-throughput" technique was inhibited early-on by several high-profile studies citing reproducibility problems [88, 61]. Subsequently, however, many of these inconsistencies were associated with the differences in the cited array technologies and designs, post-processing normalization and statistical analyses [48, 99, 15, 12], and a number of studies have successfully demonstrated biological consistency between molecular phenotype signatures derived from high-throughput array technologies [91].

1.2 Thesis Overview

This thesis addresses three main challenges. First, we acquire a large subset of a publicly-available microarray repository and develop a principled method for extracting phenotype information from free-text sample labels, then use that information to generate an index of the sample's medically-relevant annotation. The indexing method we develop, Concordia, incorporates pre-existing expert knowledge relating to the hierarchical relationships between medical terms, allowing queries of arbitrary specificity to be efficiently answered. Second, we describe a highly-flexible approach to answering the question: "Given a previously unseen gene expression sample, how can we compute its similarity to all of the labeled samples in our database, and how can we utilize those similarity scores to predict the phenotype of the new sample?" Third, we describe a method for identifying phenotype-specific transcriptional profiles within the context of this database, and explore a method for measuring the relative strength of those signatures across the rest of the database.

Chapter 2 discusses generically the problems presented by free-text bio-medical annotations, and details our approach to building a standardized queryable database from them. Chapter 3 describes the application of that framework specifically to the data resources explored in the remainder of this thesis, namely the transcriptional profiles available from NCBI's Gene Expression Omnibus. Chapter 4 presents a method for using the curated database in a nearest-neighbors-like approach to attribute phenotype to new samples. Chapter 5 describes the relationship between normal tissue, malignant tissues and stem cells, from the perspective of a highly-conserved stem cell expression profile derived from our database. Due to the diverse nature of the content in this thesis, related and previous works are presented inline with each topic, rather than together in one section.

Chapter 2

Concordia: A UMLS-based Index of Free-text Documents

The widespread adoption of electronic storage media throughout the medical and biomedical research communities presents significant new challenges and opportunities. Current estimates place United States healthcare IT spending in the range of $7 billion per year [39]. Recent publications have emphasized the utility of these data resources for genomic research, as well as patient care [46, 100, 65]. By recent estimates however, only 17% of doctors and 10% of hospitals are currently utilizing such systems [18]. A variety of programs recently enacted by the US government are 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 [18]. 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 [14]) has already outstripped our ability to perform large-scale, automated discovery of relevant patterns among records with shared phenotype. At the time of writing, GEO contained over 600,000 samples, each associated with a variety of free-text medical and biological descriptions. 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-pollenization would enable sophisticated quantitative clinical diagnostic systems, as well as accelerate the pace of therapeutic innovation.

To our knowledge, there are no open, scalable, standardized systems for cataloging and searching large volumes of medical data that leverage 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.

One of the main contributions of this thesis is a data indexing system that addresses these challenges, called Concordia. Concordia is a scalable standards-based infrastructure for searching multiple disparate textual databases by mapping their contents onto a structured ontology of medical concepts. This 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). While our main focus will remain on indexing the metadata associated

with a large gene expression database (NCBI's GEO), we will mention several other use cases for such a system.

This system may be used to form the cornerstone backend search tool required to build portable applications that leverage the wide variety of data-rich resources that are becoming available. Outside the realm of searching biological sample repositories this may also help address one of the core challenges in personalized healthcare practice: identifying clinically distinct subgroups to which a particular patient belongs [45].

The remainder of this chapter will consist of a brief description of the conceptual methods underlying Concordia, and then a brief discussion of implementation details.

2.1 Designing an ontology-based indexing system

Two major challenges arise when indexing free-text medical literature as it appears in electronic medical records, medical reference volumes or other medical documents: resolving synonyms and identifying conceptual relationships between medical terms. We suggest that both of these challenges can be addressed by building a system around the National Library of Medicines Unified Medical Language System (UMLS) [?, 66]. UMLS is an ontological organization of medical concepts, built from various thesauri, such as SNOMED [73], MeSH [51], and RxNorm [53]. Concordia works by mapping both source documents and user queries into UMLS hierarchically structured space of medical concepts.

2.1.1 Resolving multiple synonymous terms to unified concepts

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 carcinoma both describe the same medical concept, but there is no agreement 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 carcinoma 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.

We address this problem by mapping the text content of each entity in the database to a controlled vocabulary, the UMLS. The UMLS consists of a series of "biomedical vocabularies developed by the US National Library of Medicine" [66, 19].The purpose of these expert-curated vocabularies is to provide a set of thesauri that map multiple synonymous phrases to a single unified concept. The collection of these mappings is called the UMLS Metathesaurus. MetaMap [7], a program that generates these thesaurus correspondences from free text, is available from the National Library of Medicine (NLM), and is the standard tool for such tasks. MetaMap matches simple syntactic noun phrases from an input text to UMLS concepts, effectively standardizing the text to a set of unique concepts.

In our setting, applying MetaMap to the database entities allows us to alleviate the problem of resolving synonymous but textually disparate phrases. One of the major contributions of our approach is the concept 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, allowing us to search for database entities

matching the query in the structured space of standardized UMLS concepts rather than free-text. In addition, when the 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.

2.1.2 Mapping documents and queries onto UMLS ontology of medical concepts

UMLS provides a hierarchical organization of the concepts that it contains. For example, the concepts "white blood cell abnormality," "thrombosis," "anemia," and "hemorrhage" are all among the descendants of the concept "hematological diseases" in the UMLS hierarchy. Concordia processes free-text user queries with the same NLP-driven UMLS mapping tool that we use to process the source documents, thus translating the query task into the simpler job of identifying documents associated with the concept(s) to which it maps (see Figure 2-2). Continuing the above example, any query that maps to the hematological diseases concept should return documents related to any one (or several) of these four subordinate concepts, even though the NLP mapping of the documents source text may not have directly hit hematological diseases. We exploit the expert knowledge encoded in this structure by storing references to the source documents (e.g., medical records, diagnostic tests, medical literature) on top of the UMLS hierarchy.

The ontology is manifested as a directed acyclic graph. Each vertex in this graph stores a pointer to the documents that reference it. This hierarchical structure allows us to efficiently traverse the ontology and retrieve records related to a particular concept and its subordinates. For each concept that a query string maps to, we can

thus efficiently (running time proportional to the number of vertices in the subgraph) return both all of the source documents whose text directly mapped to that concept, as well as those whose text mapped to some subordinate concept.

In scenarios where traversing each subgraph presents a prohibitive impact on system performance, we have also developed a "pre-processing" procedure that enables faster query response. This procedure first creates a hash table mapping each concept to a list of all of its descendants, as well as another hash table mapping each concept to a list of all of its ancestors. These hash tables are then used to generate two additional hash tables: one that maps each concept to a list of all documents referenced by it or one of its subordinates; and another that maps each document to a list of the concepts and their ancestor concepts that are hit by the document. Using these data structures, the query running time for a single concept is constant, rather than dependent on the size of the UMLS subgraph below it. Of course, this comes at the expense of the time required to completely traverse the subgraph reachable from each vertex, as well as the storage space required to persist the results of those graph traversals.

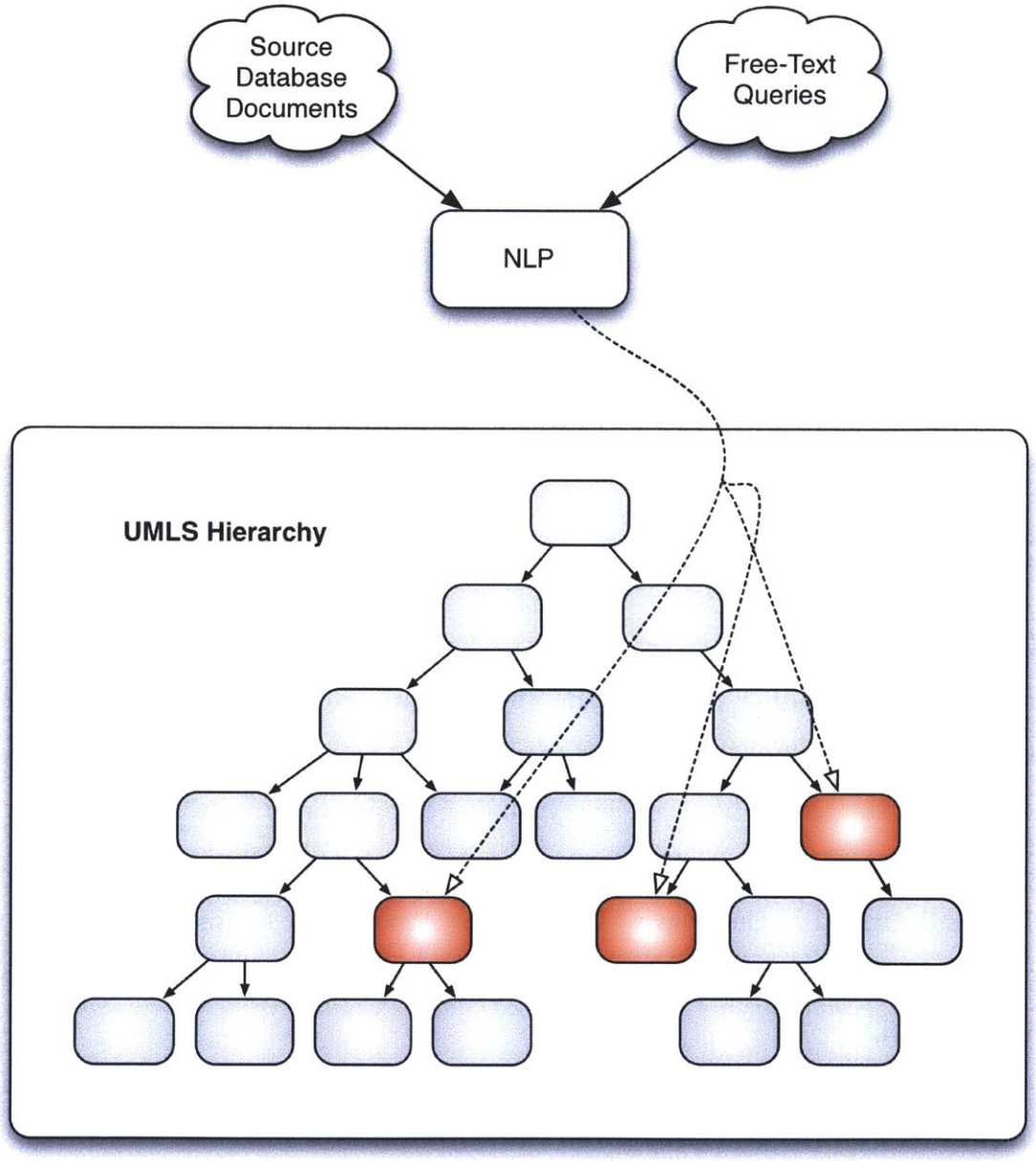

Figure 2-1: Free text source documents are processed with NLP software that maps them to unified medical concepts. Queries against those documents are similarly processed with NLP, reducing the query procedure to matching concepts from the query string against the standardized database.

Utilizing these data structures, we have developed a mechanism in Concordia

that can efficiently aggregate documents that match arbitrarily complex logical combinations of UMLS concepts. We have implemented a standard stack-based algorithm [68] for evaluating infix set logic expressions. Here, the operands are the set operators (INTERSECTION, UNION, DIFFERENCE) and the arguments are 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 enables us to perform free-text queries such as "anemia and cancer" or "lung cancer and metastasis but not smoking" against the library of documents.

2.2 Concordia implementation details

Traditional relational database systems are typically regarded as more flexible than hierarchical databases. Such systems index their records based on lexicographical ordering of key values, irrespective of conceptual relationships that exist between these keys [26]. Here, however, every query (or precomputed traversal) performed against the UMLS index will require a traversal of the UMLS topology similar to the one described in the above. Thus, a hierarchically structured database is more appropriate than a relational model when indexing documents based on an ontology such as UMLS.

We designed software for maintaining the persistent hierarchical database in Java, utilizing Oracles BerkeleyDB JE package. This package allowed us to easily serialize the in-core data structures manipulated by our search algorithms without the communication overhead incurred when interacting with an out-of-core database service.

To make our Concordia databases accessible to a wide variety of applications, we implemented both SOAP [90] and XML-RPC [97] services that wrap the Concordia

functionality. These APIs expose two main pieces of functionality: they present the topology of the UMLS ontology to the client, and present the concept / document associations to the client. API procedures for reporting all ancestor and descendant concepts relative to a particular query concept (as well as their respective minimum hop distance in the UMLS graph), reporting all indexed documents (including those related to descendant concepts) related to a specified query string, and reporting all concepts (including ancestors) related to a particular document are implemented in the current version of the system. We have also implemented an interface between Concordia and the R statistical programming environment [101] via the rJava package [102], avoiding the network-induced overhead of the XML-based APIs.

The XML-based APIs (SOAP, XML-RPC) enable out-of-core queries to non-Java applications. In addition, they allow the database the be hosted on dedicated hardware, freeing client applications to run anywhere the network protocols allow. The R API presents a convenient interface to the popular statistical environment. While rJava is JNI-dependent [40], the overhead incurred therein is less than the XML-based APIs. For this reason, the R API is preferred when performance is at a premium, but it is not feasible to develop the entire analysis workflow solely in Java.

In addition, this design, with standards-based APIs surrounding the Concordia system, allows for a great deal of scalability through data federation. Similar in spirit to Googles MapReduce methodology [28], queries may be initially processed by a head node which in turn requests that multiple worker nodes perform the database search in parallel . Each of these worker nodes would be capable of searching a separate portion of the database. Results would then be returned to the head node, aggregated, and returned to the client.

This infrastructure enables us to scale to meet future needs by simply adding additional worker nodes. Throughput, scalability and fault tolerance may all be

improved by a variety of striping schemes [81]. Although the particular application of Concordia described in this thesis does not directly rely on this functionality, we mention it as possible future work.

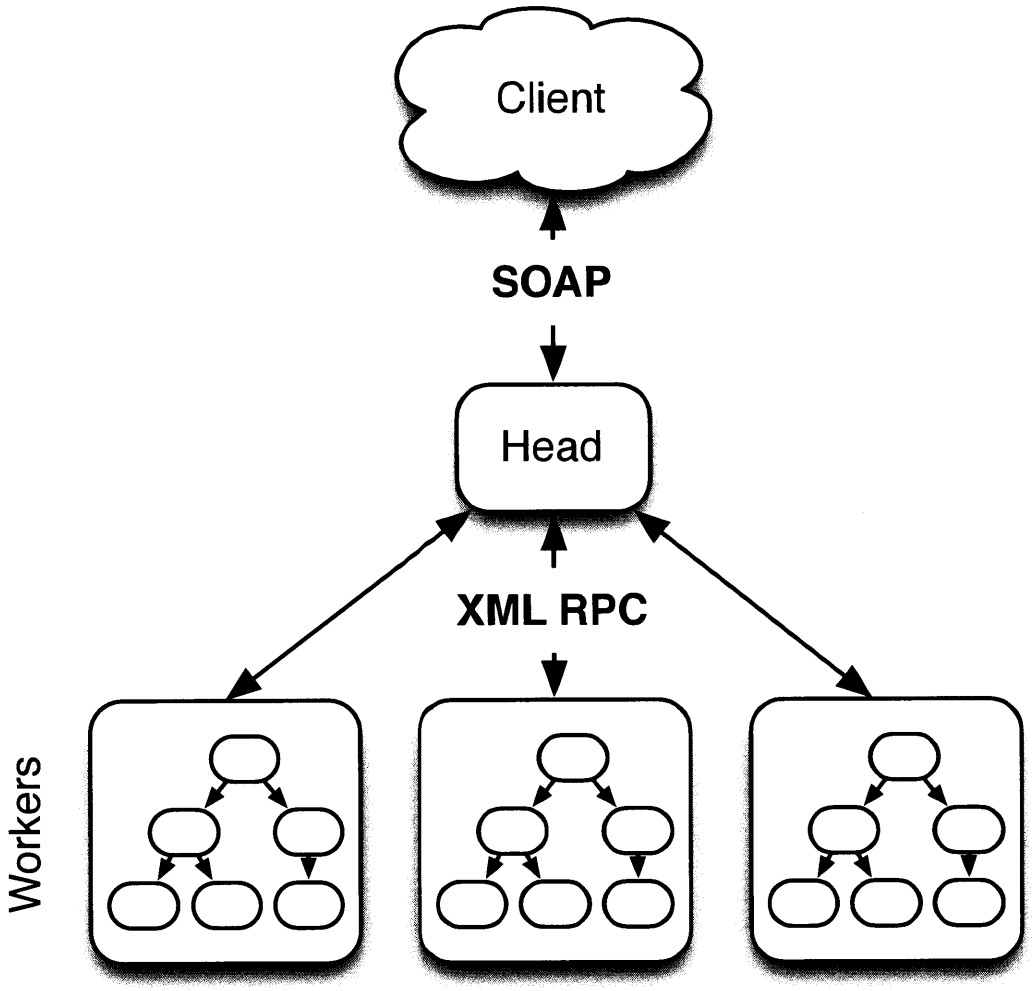

Figure 2-2: The Concordia APIs enable naive data federation to support both fault tolerance and improved total throughput. A query node serves as an intermediary between the client applications and the data-storing nodes that each run a separate Concordia instance. That is, each data-storing "worker" node is responsible for storing a specific subset of the total database. The query node processes client requests and requests that each worker node process the portion of the query representing the data that it is responsible for managing. Each worker node returns its portion of the response, and the multiple worker responses are aggregated at the query node, where they are assembled to be returned to the client application.

2.3 Applications of Concordia

In addition to the database constructed from NCBI's Gene Expression Omnibus that is described in detail later in this thesis, we note several other likely applications.

NCBI maintains a database of gene-centric annotation data [60]. These annotations explicitly connect each annotated gene to both biological processes and medical concepts. We have experimented with using Concordia to index these annotations. Such a resource may be valuable for gene set enrichment-type analyses [95], where one wishes to understand the broad-stroke disease and tissue concepts most strongly associated with a predetermined set of genes. We propose replacing the commonly-used Gene Ontology [8] (or GO) terms in these analyses with UMLS concepts whose associated genes are learned from the application of Concordia described above.

Data mining electronic health records is becoming an increasingly popular method for performing pharmacovigilance [100] and genetic studies [46]. By standardizing the nomenclature of disparate data sources across healthcare IT infrastructure and research databases, Concordia may enable larger-scale studies to be performed with minimal overhead.

Concordia is not limited to generating its index based solely on the UMLS ontology. The system has been designed to be easily reconfigured to generate an index over any ontology or acyclic graph. The possible applications of this technology, therefore span beyond the realm of medical language and knowledge, and it may prove a useful backend datastore for implementing semantic web applications [41].

Chapter 3

Making NCBI's Gene Expression Omnibus Machine Readable with Concordia

This chapter presents a brief overview of NCBI's Gene Expression Omnibus, and describes how we applied the Concordia framework to construct a machine-readable index of it. In addition, we will also discuss the handling of GEO's underlying gene expression intensity data, which will be used throughout the remainder of this thesis.

3.1 Motivation

Our analyses will require a large volume of gene expression (microarray) data acquired from samples spanning a wide range of clinically relevant biological conditions. We have assembled this database from NCBIs Gene Expression Omnibus (GEO) [13]. GEO contains tens of thousands of human microarray samples derived

from extremely diverse experimental conditions. While the sheer volume of data available to the public is promising, bringing these resources to bear on structured analyses presents several challenges [22]. First, the quantitative measurements acquired from the microarrays need to be normalized to be comparable to one another across data sets [47]. Second, because NCBI has not enforced any organization on GEOs free-text description fields, significant work must be done to decipher the relevant phenotype of each sample [22].

3.2 Metadata

GEO serves as a public repository of gene expression microarray data. Each submitted microarray hybridization is, in GEO terminology, called a sample. Each sample has a set of gene expression intensities associated with it, along with a variety of metadata describing the sample phenotype and any relevant treatment information. All of this metadata is free–text. Samples are grouped into data series, which represent a collection of related samples, typically derived from the same study / publication. Further, each sample may be assigned to zero or more data sets, providing an additional level of phenotypic classification. For example, a researcher publishing a paper on colon tumors might submit a series to GEO containing colon cancer samples along with non-malignant colon biopsies to serve as a control (the series metadata would describe the overall experiment design, and the sample–level annotations would be expected to provide malignant vs. non–malignant phenotype). Continuing the example, a data set might be composed over a subset of these samples, comprising only those samples derived from patients with a history of inflammatory bowel disease (this phenotype classification might only be represented at the data set level).

The goal of our automated processing of GEO was to obtain as comprehensive

picture of phenotype as possible for each sample. As such, we associated all of the relevant free text (sample–, series– and data set–level) available to each sample. We then removed erroneous phenotype associations caused by this overly-optimistic strategy manually (see section 3.2.1).

One of the greatest challenges that needs to be addressed when utilizing loosely-curated resources such as GEO is the lack of standardized nomenclature (see Chapter 2). Specifically, our analyses (described later in this thesis) will seek to cast the activity of both individual genes and coupled gene modules within the context of multiple disease and tissue conditions. GEO's samples are described by free-text descriptions relating to the experimental setup and phenotype of the biological material. In order to utilize this data for our analyses, we need to construct an index into it based on a standardized biomedical nomenclature. We employed the previously described Concordia framework to address this challenge. The UMLS thesauri used to construct the Concordia index were MeSH [51] and SNOMED [73].

3.2.1 Generating high-confidence phenotype labelings

The NLP software that was used to map the samples into the UMLS hierarchy tends to be overly sensitive [63], picking up unintended text–to–concept associations. For example, it frequently mistakes certain abbreviations used in a sample's description for concepts that the author hadn't intended (e.g., any text containing the abbreviation "mg." gets mapped by MMTx to the concept that represents "Madagascar", even when the context of the document suggests the author intended it as an abbreviation for "milligram"). These are challenges that can be overcome only by training an NLP algorithm on domain-specific patterns. As a consequence, the concept associations produced by MMTx represent a good first-pass filter on the text, but

are hardly reliable enough to drive analyses where we require data relating to very specific phenotypes. Rather than attempting to refine the NLP procedure, we took a simpler data-driven approach to constructing a high-confidence set of reference samples.

In addition to these homonym errors, the inclusion of high-level descriptive metadata is often a source of false concept associations. For example, the summary text for GEO series GSE9187 reads:

> Transcriptional profiling of human **breast cancer** cell line LM2, a subline of MDA-MB-231 highly metastatic to **lung** when injected to nude mice, to identify the genes that are regulated after the metastasis gene metadherin is knocked down. Keywords: Genetic modification

Because the NLP software identifies tissue concepts related to both the terms "breast cancer" and "lung", we need to filter out the unintended association (in this case lung, since the cell line is actually breast cancer).

In order to assist with the task of manual data curation, we developed a graphical user interface that allows domain experts to examine the NLP results and both verify the concept associations, as well as add new ones that MMTx missed altogether. This tool allows a user to quickly examine and annotate a large number of samples by grouping samples based on their MMTx-derived annotation. Thus, repeat errors are easily caught and corrected. Nevertheless, the process of generating a high-confidence reference set in this manner is tedious.

This tool presents the user with a list of the GEO samples in the database, grouped by GEO series. The user may select one of these sample identifiers, populating the remaining UI elements, including fields derived from the sample's associated series, fields derived from the samples' associated data sets and fields derived from

the sample–level annotations. The UI presents the user with the UMLS concepts associated with various text fragments from each level of metadata, allowing each one to be manually verified (see Figure 3-1). A total of 3030 samples (listed in section 7.1) were manually verified. Those samples comprise the data used throughout the remainder of this thesis.

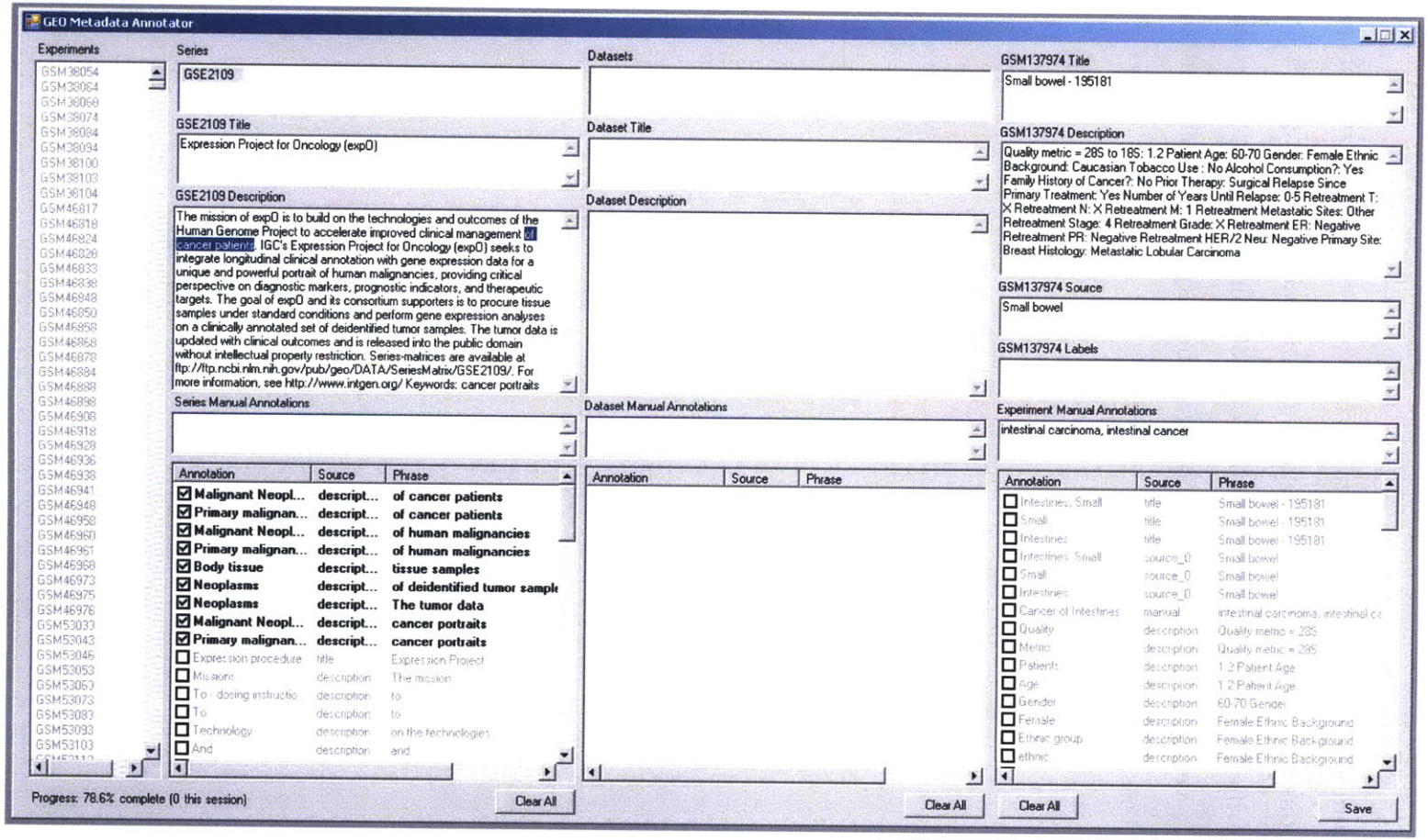

Figure 3-1: A screen shot of the software designed to allow manual validation of the Concordia-derived UMLS annotation. On the extreme left–hand side of the interface, there is a list of the GEO samples in the database, grouped by GEO series. The user may select one of these, populating the remaining fields on the form. the next column to the right contains the GEO fields derived from the sample's associated series. To the right of that are the fields derived from the samples' associated data sets. To the right of the data set entries are the fields derived from the sample–level annotations. At the bottom of each column, a list of check–boxes allows the user to manually validate the UMLS concepts associated with various text fragments.

3.3 Expression Intensity Data

The database is presently comprised exclusively of gene expression samples performed on the Affymetrix HGU-133 Plus 2.0 platform. The original CEL files were downloaded from GEO and Affymetrixs MAS 5.0 normalization procedure [2] was performed on each sample. The probe-level measurements were subsequently summarized (mean) at the gene level. This summarization step both helps to reduce probe-specific noise [47] and to make the database interoperable with a wide variety of gene-level and protein-level analysis tools. The gene-specific measurements for each sample were then processed to generate both rank normalized values, and z-scored log2 transformed values. The ranked data is often useful when mining the database for phenotypic trends across large sets of samples [47] (e.g., clustering samples related to particular tumor pathologies, and comparing our database to external resources, such as data acquired from alternative platforms), whereas the z-scored data can be useful when looking for more subtle gene-level trends [23] (e.g., detecting tightly-coupled gene modules [10]).

The MAS 5.0 procedure operates on individual samples without explicitly modeling any inter-chip relationships, as opposed to, e.g., the loess or RMA procedures [20]. Our approach of rank or z-score normalizing the MAS 5.0 data has proven adequate for a wide variety of analyses, while also enabling fast and reliable integration of new samples into the database without the requirement of reprocessing the entire library of expression values. Because we have generated gene-level data (as opposed to, say, platform-specific probe-level data), the data are readily interpretable within the framework of genetic interactions, protein-protein interaction networks, regulatory models, gene set / pathway analysis, and all of the publicly available data resources devoted to them.

Chapter 4

Mapping the Transcriptional State Space of Cell Identity

The material presented in this chapter is the result of collaboration with Patrick R. Schmid.

This chapter describes a statistical method that uses our curated GEO Concordia database for phenotype prediction. The central problem we address here is: Given a previously unseen gene expression sample, how can we compute its similarity to all of the labeled samples in our Concordia database, and how can we utilize those similarity scores to predict the phenotype of the new sample?

While microarray experiments have become commonplace, the transcriptional landscape of tissue and disease still remains poorly understood. A macroscopic analysis of 3030 human gene expression samples presented here reveals that biologically related tissues are highly co-localized within the transcriptional state space. From a transcriptome-wide perspective, the boundaries between these related tissues overlap

to form a smooth continuum of related phenotypes. By systematically focusing on restricted regions of this landscape, these boundaries become increasingly distinct, and additional biologically-meaningful organization is revealed. We have developed an online resource that uses this structure to provide detailed tissue and disease phenotype enrichment information for user-submitted microarray samples. We also show that tissue specific marker genes are activated in localized subspaces, and that tumor samples, while proximal to their unaffected counterparts, exhibit greater respective expression variance. Furthermore, we see that tumor metastases are often enriched for phenotypes relating to their tissue of origin. In addition, we provide a specificity measure for the conventional clinical classification of tissue and disease, and lay the foundation for large-scale automated clinical prognostication and novel drug-discovery.

4.1 Background and biological motivation

Although the human genetic code has been successfully sequenced, the comprehensive transcriptome-wide landscape representing similarities between various tissues and diseases has yet to be deciphered. Even with the hundreds of thousands of expression arrays available through public repositories such as NCBIs Gene Expression Omnibus (GEO) [13] and EMBIs ArrayExpress [71], transcriptional analyses have generally been limited to isolated pockets of this landscape (e.g., comparison of a diseased tissue class vs. its normal counterpart). While it is known that the activation or repression of specific genetic modules plays a role in creating the phenotypic variances that differentiate various tissue and disease states [87], these programs are frequently shared based on high-level biological similarities [74, 107]. The topology of the transcriptional space, as defined by the common vs. unique activation of

these modules in various tissues and diseases, remains unclear. Insight into the organization of this landscape may provide a quantitative basis for drug repositioning and design, as well as the practice of personalized therapeutics.

Among the principal challenges for large-scale analyses have been the biological and measurement noise and biases that characterize each data set [85], as well as the lack of standardized annotations of tissue and disease characteristics for each sample [22]. Unlike previous efforts such as Oncomine [78], EBIs Human Gene Expression Map [58], TiGER [54], BODYMAP [69], BioGPS, and TiSGeD [105] that address these difficulties by providing gene-centric analyses performed using relative measures of expression, we took a macroscopic view by combining the results from 3030 microarray samples and explored the global transcriptomic landscape.

Indeed, we see that a macroscopic analysis of the human transcriptome shows that tissues reside as close neighbors in a smooth continuum in the expression landscape. Although there are relatively gradual transitions between phenotypes, there is a high degree of localization of individual phenotypes in expression space. A holistic analysis of the global expression landscape reveals that the expression locality of phenotypes provides a key insight to prediction accuracies for clinically meaningful categories using gene expression.

However, the tantalizing goal of leveraging a large, diverse set of data to automatically classify and prognosticate patients based on biological samples, perhaps along lines even more clinically relevant than current diagnostic classes [56], has remained distant. The classification of diseases across studies has either used expression samples from small numbers of data sets [92] or within relatively homogenous disease classes [78]. To this end, we take a robust statistical approach that allows users to obtain detailed tissue and disease labels, as well as an enrichment statistic, for new, unlabeled microarray sample by mapping it to this expression continuum.

By providing a view into this continuum, we see that malignant tumor samples are co-located with their unaffected counterparts and that tumor metastases are also mapped to the location of their primary site.

4.2 Multi-resolution analysis of human transcriptional state space

A multi-resolution investigation of the global transcriptome landscape, as represented by the first two principal components of the gene expression values of 3030 microarray samples obtained from GEO, reveals that the expression continuum is first divided into three distinct landmarks: blood, brain, and soft tissue (Figure 4-1A).

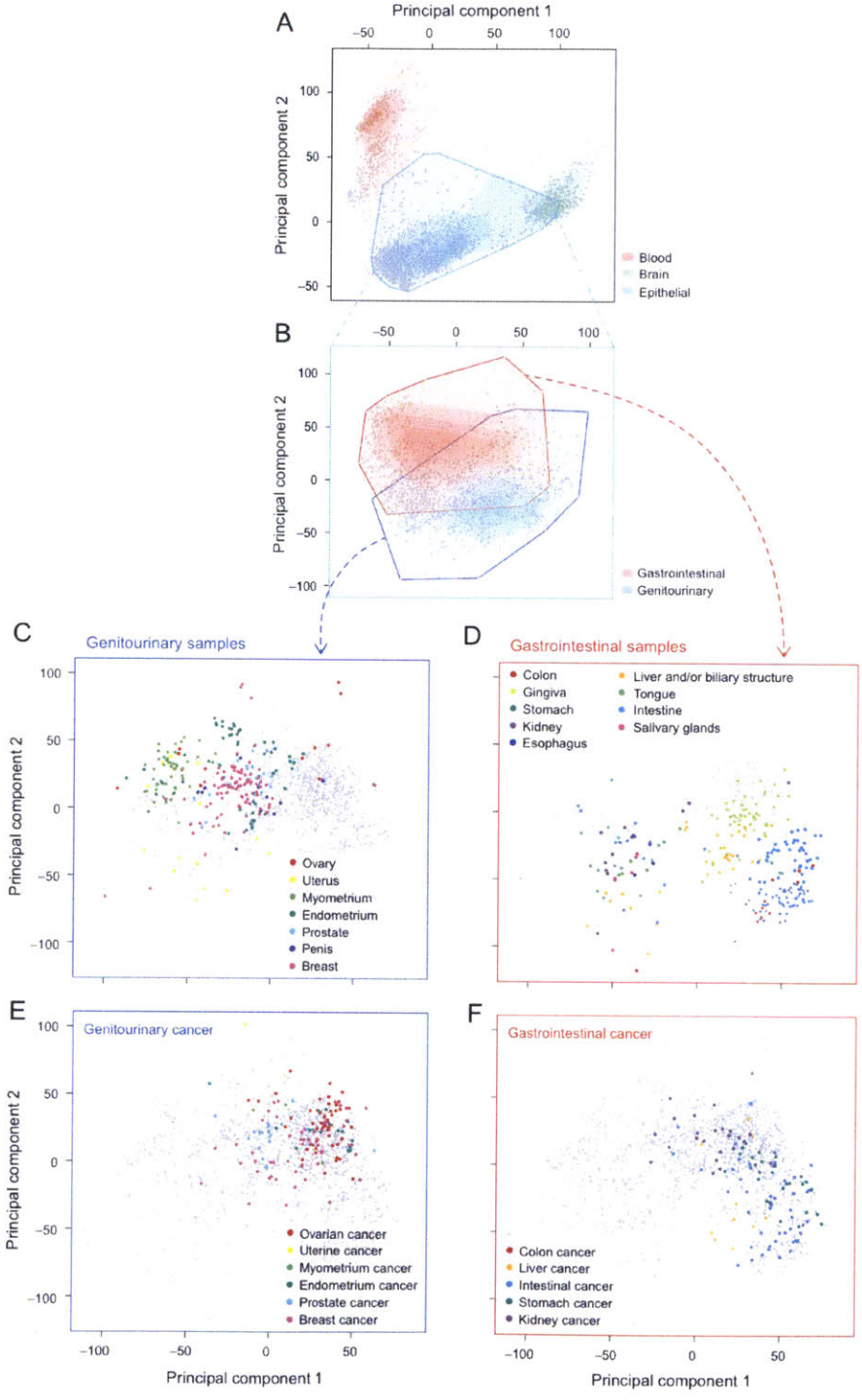
A
Principal component 1
Principal component 2
Blood
Brain
Epithelial
B
Gastrointestinal
Genitourinary
C
Genitourinary samples
Ovary
Uterus
Myometrium
Endometrium
Prostate
Penis
Breast
D
Gastrointestinal samples
Colon
Gingiva
Stomach
Kidney
Esophagus
Liver and/or biliary structure
Tongue
Intestine
Salivary glands
E
Genitourinary cancer
Ovarian cancer
Uterine cancer
Myometrium cancer
Endometrium cancer
Prostate cancer
Breast cancer
F
Gastrointestinal cancer
Colon cancer
Liver cancer
Intestinal cancer
Stomach cancer
Kidney cancer

Figure 4-1: Multi-resolution analysis of the gene expression landscape. (A) The gene expression landscape, as represented by the first two principal components of the expression values of 3030 microarray samples separates into three distinct clusters: blood, brain, and soft tissue. 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. (B) There is a clear separation of genitourinary tissue samples and gastrointestinal samples in the soft tissue cluster. (C) A closer examination of the genitourinary and gastrointestinal sub-clusters shows clear localization of phenotypes. (D) Cancerous tissue samples show greater variance in their expression signals while still remaining proximal to their non-cancerous counterparts.

Tissue specific genes were selected by performing permutation based 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 [17]. We report only the GO terms that had a resampling-based p-value less than 0.05.

As to be expected, when analyzing the tissue specific characteristics of the clusters for marker genes we see the over-expression of genes such as COL3A1, COL6A3, KRT19, KRT14, and CADH1 in the soft tissue cluster and GFAP, APLP1, GRIA2, PLP1, and SLC1A2 in the brain cluster (Figures 4-2 – 4-4). GO enrichment analysis of these tissue specific genes further points to over-enrichment for terms related

to each of the three tissue types (see Supplemental Tables (Chapter 7) 7.2 – 7.4). Interestingly, many spinal cord tissue samples lie at the intersection of the brain and soft tissue clusters, while bone and bone marrow samples lie at the intersection of the blood and soft tissue clusters. Although there have been several reports that data from different datasets are not comparable as the dataset (aka batch) signal is dominant [70, 77], we find that the tissue signal is dominant in this macroscopic view of the transcriptome (see Section 4.4).

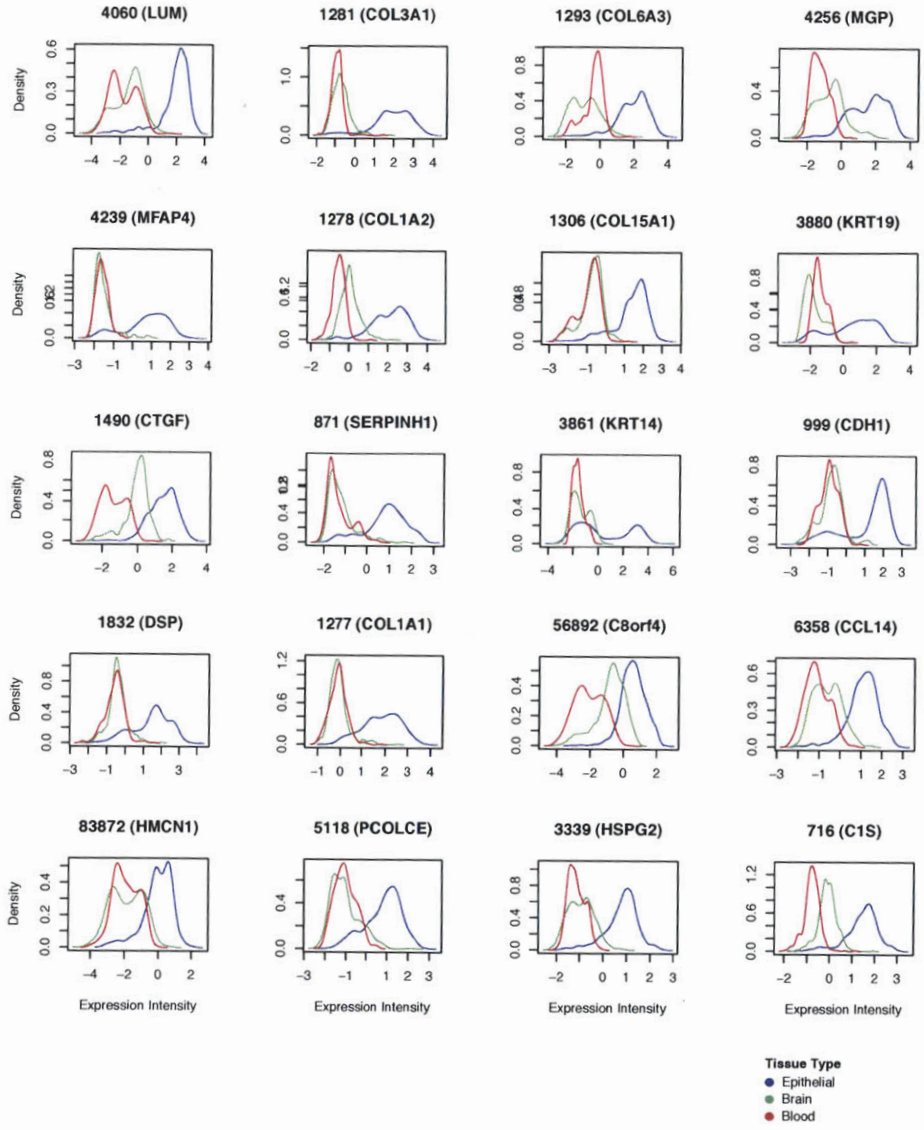

Figure 4-2: Expression intensity distribution of the top 20 overexpressed soft tissue genes. Each plot corresponds to the kernel density estimate of expression values for the gene named above each plot for the three broad tissue types, blood, brain, and soft tissue. We see that the expression values 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.

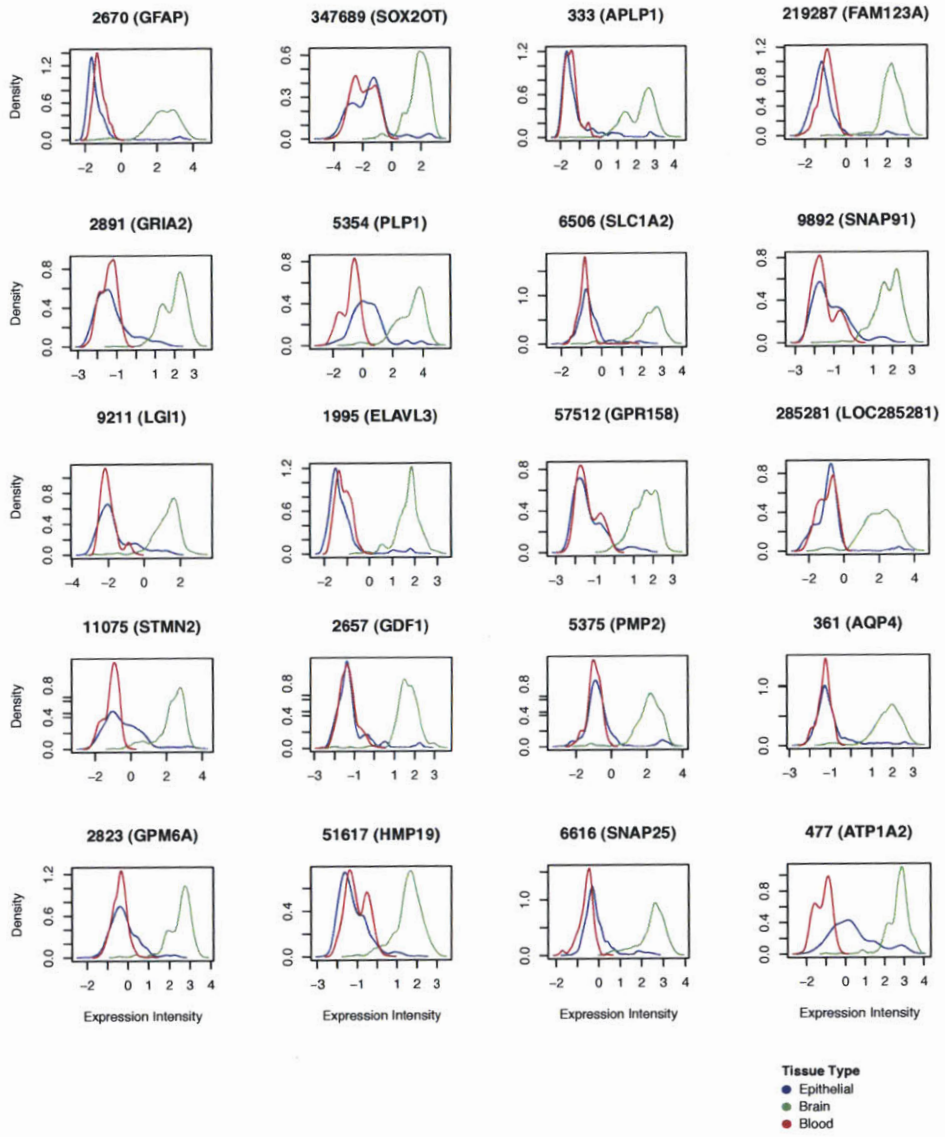

Figure 4-3: Expression intensity distribution of the top 20 overexpressed brain tissue genes. Each plot corresponds to the kernel density estimate of expression values for the gene named above each plot for the three broad tissue types, blood, brain, and soft tissue. We see that the expression values 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.

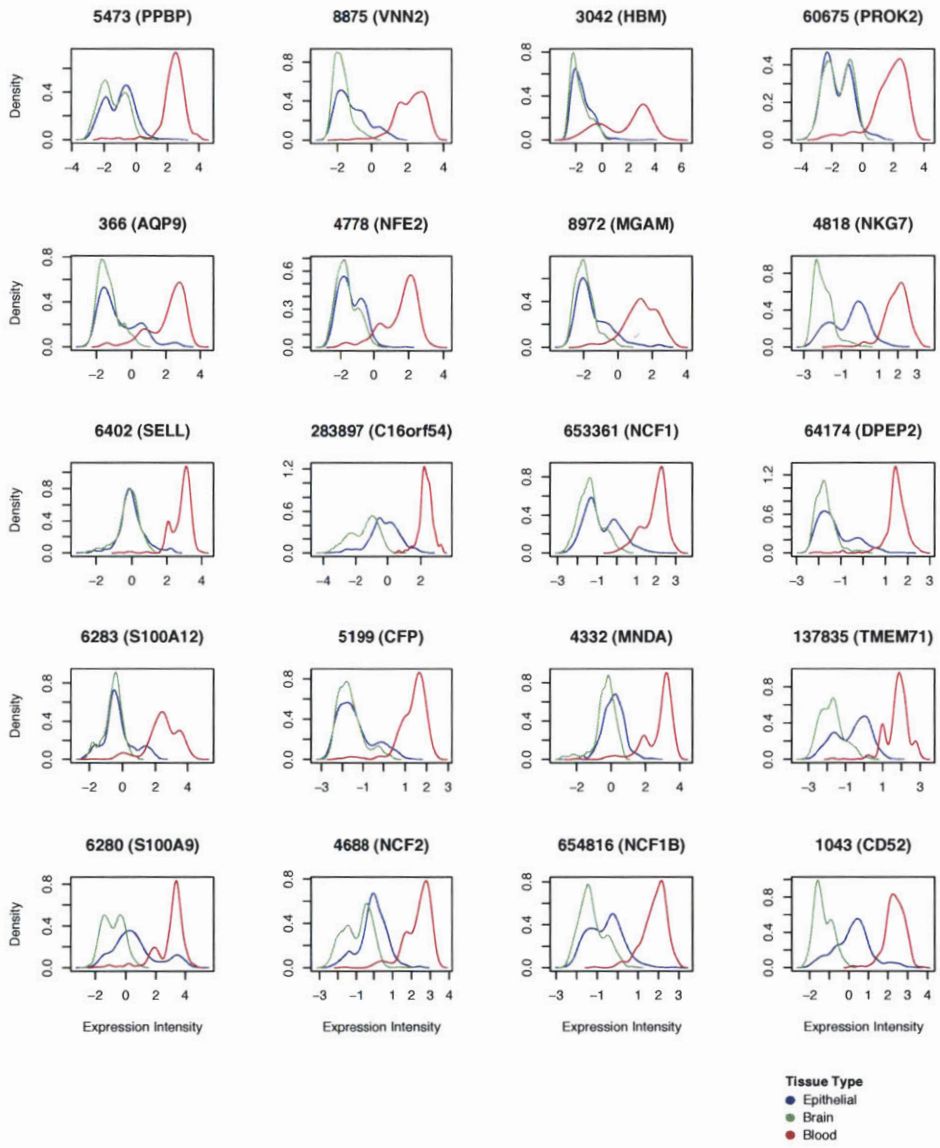

Figure 4-4: Expression intensity distribution of the top 20 overexpressed blood genes. Each plot corresponds to the kernel density estimate of expression values for the gene named above each plot for the three broad tissue types, blood, brain, and soft tissue. We see that that the expression value of 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.

Mirroring the properties of the full transcriptome landscape, a multi-resolution examination of the soft tissue cluster reveals not only that the tissues inhabit distinct

locations in the transcriptome landscape, but also the co-localization of functionally related tissues. This co-localization of related tissues forms a smooth expression continuum in which there are phenotypic hotspots but also a gradient where the tissues mesenchymal in origin are on one side, while the more epithelial tissues are on the other. As depicted in Figure 4-1B, the tissues from the respiratory system, the lower gastrointestinal tract, and oral soft tissue together form a distinct sub-region in the smooth landscape, but also reside in their own confined neighborhoods within that sub-region. Juxtaposed to the soft lining epithelial tissue, which has relatively high rates of cell replication, the less mitotically active tissues of the genitourinary system can be found. Further limiting the investigation to only the soft lining tissue reveals clear co-localization of related tissue such as colon, intestinal mucosa, and stomach in one part of the transcriptomic topography, and tongue, esophagus, and dentition in another.

Unlike phenotypically normal tissue, the genetically altered state of malignant tissue causes a change in the placement of cancerous tissue in the expression landscape within the confines of the three major clusters. Similar to the high-resolution findings of Lukk et al. [58], we see a shift of the tumorous samples away from the fully differentiated state of normal tissue. We further find that, for example, cancerous genitourinary samples no longer are distinctly localized to hotspots at the periphery of the soft tissue cluster, but rather become intermingled with one another, causing greater overlap between the phenotypes in the continuum (Figure 4-1C). Indeed, this lack of specificity is also evident in the decrease in correlation between normal tissue samples and their corresponding cancerous tissue samples (Table 4.1).

Table 4.1: The change in correlation between normal and cancerous tissue samples. Each value in the table corresponds to the mean correlation between all samples with the given phenotype. We see that the correlation between normal tissue samples is generally higher than the corresponding correlation of cancerous tissue pointing toward a loss of differentiation. This loss of correlation may be attributed to the loss differentiation in tumor tissue causing a more variance in gene expression.

	Normal Tissue	**Cancerous Tissue**
Breast	0.86	0.83
Ovary	0.85	0.85
Lung	0.83	0.85
Colon	0.88	0.87
Skin	0.91	0.77
Myometrium	0.89	0.78
Endometrium	0.87	0.85
Intestine	0.87	0.86
Stomach	0.87	0.87
Kidney	0.9	0.83
Prostate	0.91	0.86

4.3 Predicting phenotype with the Concordia GEO database

We developed a principled statistical approach that utilizes the previously discussed high degree of correlation between samples of similar phenotype to associate previously unseen gene expression samples with Unified Medical Language System (UMLS) [19] concepts. In particular, we employ a database of samples that we have previously labeled with UMLS concepts through both an automated, natural language processing method and hand curation (see Chapter 3). 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. Our method then calculates the UMLS concepts most likely to be related to the given input by assessing the over-enrichment of those concepts in nearby database samples (Figure 4-5).

4.3.1 UMLS concept enrichment score calculation

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. Previous work by Adler et al. [1] shows that using correlation yields favorable results for finding similar gene expression samples in the context of computing co-expression. 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. Our enrichment score is similar to the normalized Kolmogorov-Smirnov statistic that lies at the heart of the Connectivity Map [49] ranking of drug similarities and the Gene Set Enrichment Analysis procedure [96].

Algorithmically, the statistic is calculated as follows: First, the database consisting of n curated expression samples $\{s_1, s_2, s_3, ..., s_n\}$ 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 coef-

ficients $\rho_{s_1\prime}, \rho_{s_2\prime}, \rho_{s_3\prime}, , \rho_{s_n\prime}$. 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:

$$x_i = \frac{\frac{1+\rho_{s_i\prime}}{2}}{\sum_{s_j\prime \in S_c} \frac{1+\rho_{s_j\prime}}{2}} \text{ when sample } s_i\prime \text{ is associated with concept } c, \text{ and}$$

$$x_i = \frac{-1}{n - |S_c|} \text{ for all other samples that are not associated with concept } c$$

Intuitively, when ss_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 xi 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:

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

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.

In this manner, we are able to localize new gene expression data in the context of the existing gene expression landscape. It is noteworthy that no hard boundaries are drawn when a new input sample is labeled, but rather the concepts pertinent to the

transcriptomic context for the input sample are reported. To illustrate its function, we provide an online tool that allows users to input their own microarray samples performed on the Affymetrix HGU-133 Plus 2.0 array and obtain their over-enriched tissue and disease concepts.

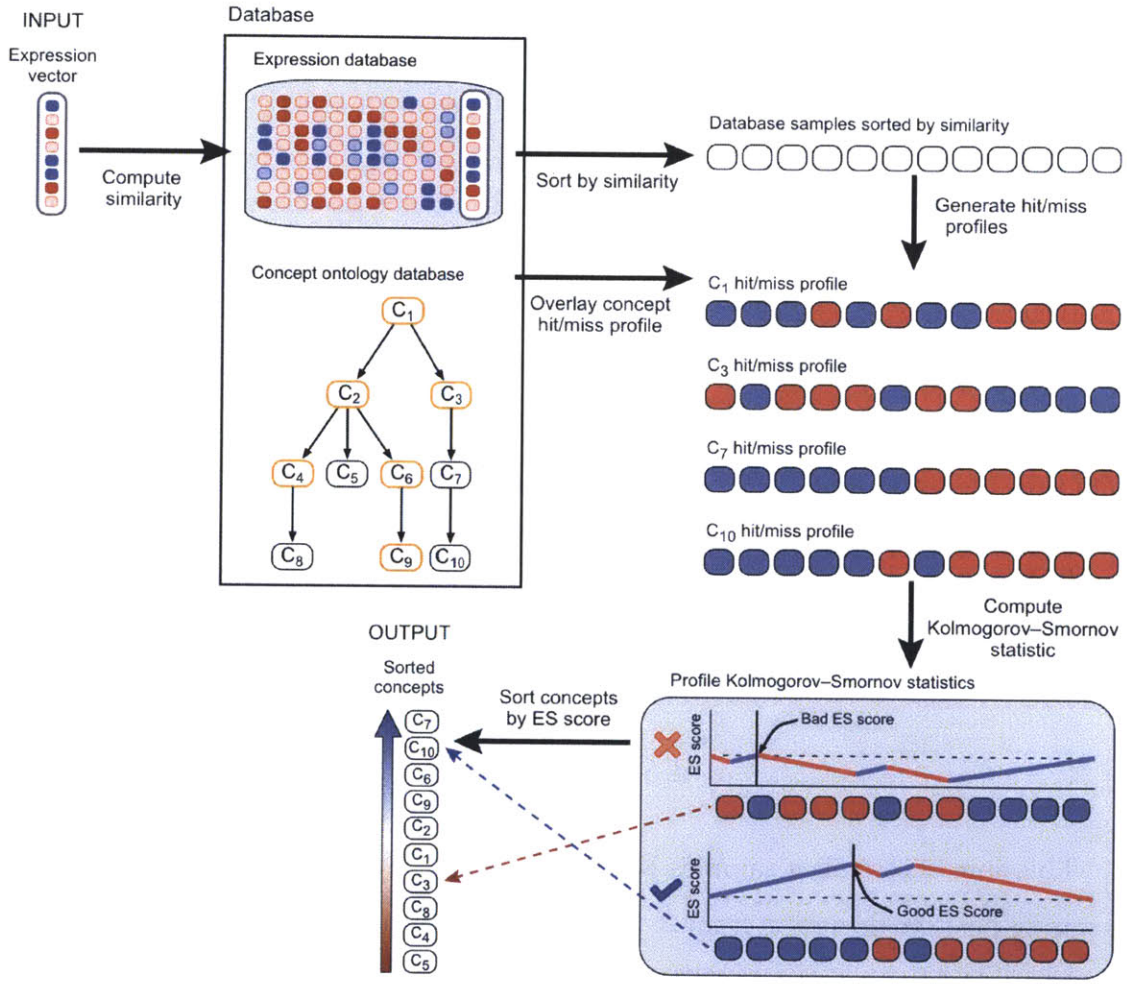

Figure 4-5: A user submits a gene expression profile to the database that then computes the similarity to all other samples in the database. Based on the similarity, an enrichment score is computed for each UMLS concept for which data exists in the database and the concepts are returned to the user in order of statistical significance.

4.3.2 Quantifying performance of the enrichment strategy

To demonstrate the quality of the concept associations produced, we performed leave-one-sample-out cross-validation and computed the receiver operating characteristic (ROC) curve for each of the tissue and disease UMLS concepts. We use the the area under the ROC curve (AUC) as a summary statistic for each concept represented in the database. The ROC curves represent the full spectrum of performance characteristics available under a variety of ES thresholds. We compute the ROC curves for each concept as follows.

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 ith sample in the list is actually labeled with concept c. If the sample is not labeled with concept c, the false-positive (FP) count is incremented. At each position i, the TP and FP are divided by the number of known positives and negatives (respectively) to obtain the true-positive rate (TPR) and false-positive rate (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 system's 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 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 4.2 contains the performance of a selection of UMLS concepts, along with the number of samples and series that were associated with that concept, and the depth of that concept in the UMLS hierarchy. The complete list of performance values can be found in Supplementary Tables (Chapter 7, Table 7.1) . It is important to note that many of these concepts are highly correlated and share some or all of the database samples. As a direct consequence, many of the concepts have similarly high (or low, respectively) AUC values.

Table 4.2: Area under the curve for selected UMLS concepts.

Concept	**UMLS Depth**	**AUC**	**Number of Series**	**Number of Samples**
Malignant Neoplasms	3	0.82	74	855
Malignant neoplasm of breast	3	0.97	9	69
Malignant neoplasm of ovary	4	0.99	4	51
Malignant neoplasm of lung	4	0.97	4	98
Leukemia	3	0.99	13	151
Stem cells	2	0.93	19	179
Pluripotent Stem Cells	3	0.996	5	53
Soft Tissue	2	0.69	98	1513

Breast	2	0.93	13	195
Ovary	3	0.95	8	103
Lung	2	0.95	9	131
Inflammatory disorder	1	0.79	13	91
Rheumatoid Arthritis	3	0.93	7	31
Inflammatory Bowel Diseases	4	0.99	2	24

Our approach employs a non-parametric enrichment statistic that only requires the labeling of the samples in the gene expression database and therefore can be updated in real-time without having to retrain the database. Moreover, as "normal control" is hard to define in a clinical setting [57, 64, 75], we also ensured that this approach does not require case-control type input and, but rather just a single microarray sample.

We see a significant increase in accuracy of the concept associations as more data is added to the underlying database. In an effort to quantify this effect, 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 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 4-6). It is also noteworthy that

the concepts that are the most susceptible to change are "specific" concepts (e.g., "Pluripotent stem cells" [48% to 99% accuracy] and "Myeloid Leukemia" [49% to 99% accuracy]), while the classification accuracy of the broad topics (e.g., "Structure of soft tissues of abdomen" [27% to 56% accuracy]) never improve drastically, as the underlying gene expression values are so diverse.

This implies that the power of this type of macroscopic analysis increases with the amount of underlying data, thereby enabling better biological quantification of the phenotypic signal, and that a system such as this could be deployed in a clinical setting with minimal alteration of normal protocols.

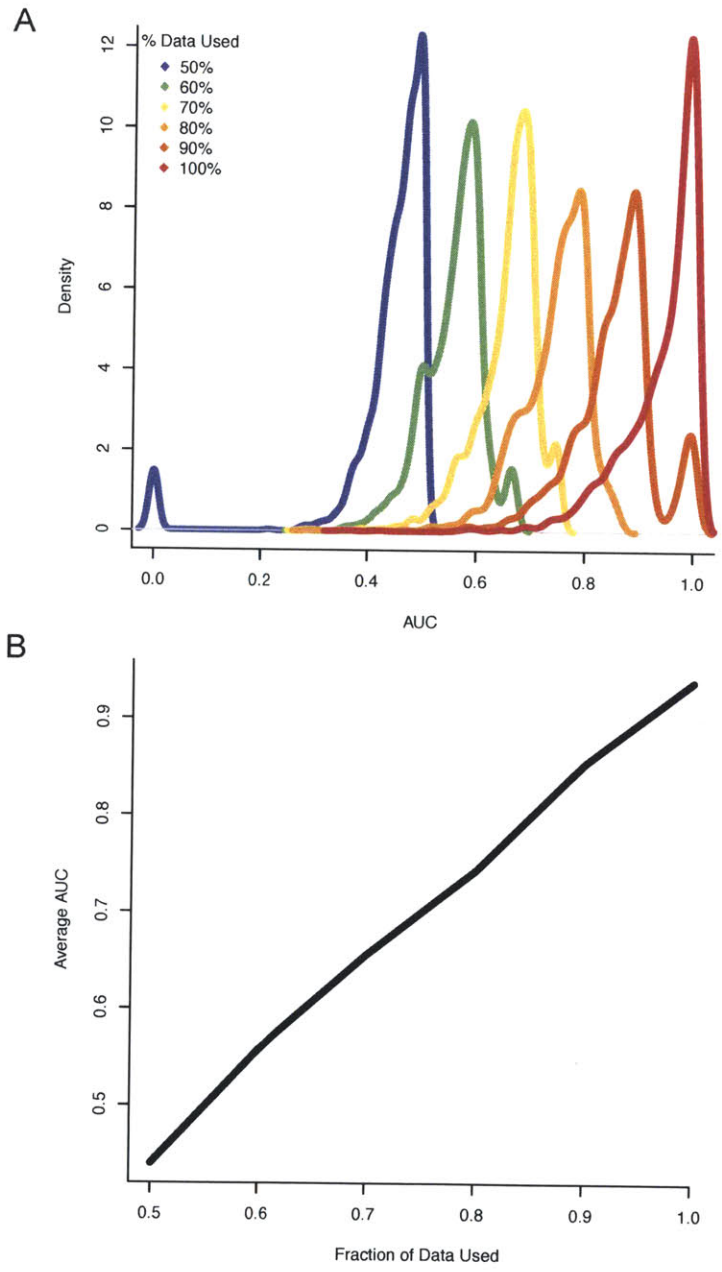

Figure 4-6: Improvement of accuracy of the enrichment statistic with the increase of data in the database. (A) Density estimate of the performance of the method over various amounts of data. (B) The average AUC values over all concepts when varying the amount of data used to compute the enrichment scores. For example, when using only 50% of the data for a given concept, the average AUC drops down to 42%.

4.4 A discussion on "batch effects"

There have been several reports that data from different datasets are not comparable as the dataset (aka batch) signal is dominant [70, 77]. While the localization of phenotypes as seen in the expression landscape (Figure 4-1), regardless of series of origin, depicts the lack of a dataset effect in principal component space, the 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.

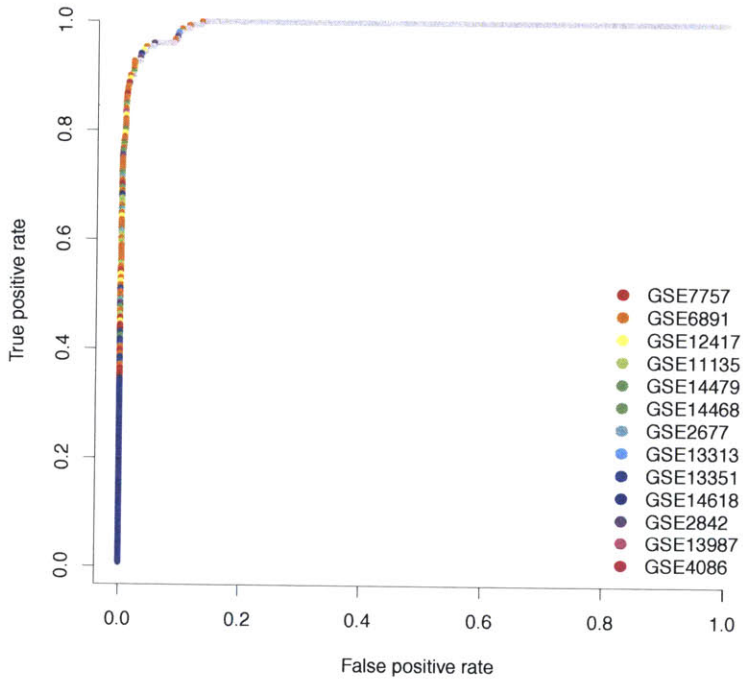

Figure 4-7: The ROC curve for leukemia depicts a lack of batch effect. The colors plotted along the curve correspond to the series of origin for each of the samples used to generate the curve. The intermingling of series points to the robustness of the phenotypic signal: samples with the same phenotype cluster together before all other phenotypes, and samples from different data series are intermingled within a phenotype.

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 4-7) 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-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 intra-series correlation between differing phenotypes is greater than the inter-series correlation between samples corresponding to identical phenotypes. The p-values 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 4.3 provides these numbers for all tissues that

are represented in the tissue relationship network such that a negative batch effect implies that the phenotypic signal dominated the dataset signal

Table 4.3: Area under the curve for selected UMLS concepts.

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

4.5 Using the Concordia-based phenotype enrichment statistics to identify the primary site of tumor metastases

By mapping a tumor metastasis tissue sample onto the gene expression landscape, we provide an unbiased measure of its phenotypic predisposition based on gene expression. It is commonly known by pathologists that tumor metastasis samples resemble the tissue of the primary site rather than that of a tissue in the metastasized location. Nevertheless, the proper identification of the primary site of a metastasis can

be critical in determining the appropriate clinical treatment plan [21, 42, 43, 103]. In conjunction with clinical and pathological investigation, the origin of metastatic tumors can currently be identified by immunohistochemistry (IHC), computed tomography (CT) scans, and position emission tomography (PET) scans. Various primary site prediction methods have been proposed including a panel of antibodies [72], CT and PET scans [76], and other gene expression-based methods. Bridgewater, et al. [21], for example, proposed an algorithmic method for identifying the origin of metastases by employing a nearest-neighbor method to look up the five nearest neighbors of an input gene expression sample in the commercial CupPrint database (www.agendia.com) for carcinomas with an unknown primary site.

Indeed, we find that even without the use of specially tuned primary site detection method as proposed by others [21, 84], metastatic tissue samples lie in the vicinity of their tissue of origin in the global continuum of the gene expression landscape. For instance, using the metastasized breast cancer samples from Zhang, et al. [108] (GSE14107), 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 (Figure 4-8A). Some of the over-enriched UMLS concepts include "White Adipose Tissue," "Subcutaneous Fat," "Subcutaneous Tissue," "Lactiferous duct," "Mammary lobe," and "Glandular structure of breast". 15 of the 17 colorectal cancer samples from GSE10961 (Figure 4-8B) 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."

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 smooth continuum of the transcriptomic landscape. This is particularly evident

within the soft-tissue cluster (bottom left of Figure 4-8A-D), 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. [84], 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 [31]. 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 (Figure 4-8C). 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) (Figure 4-8D).

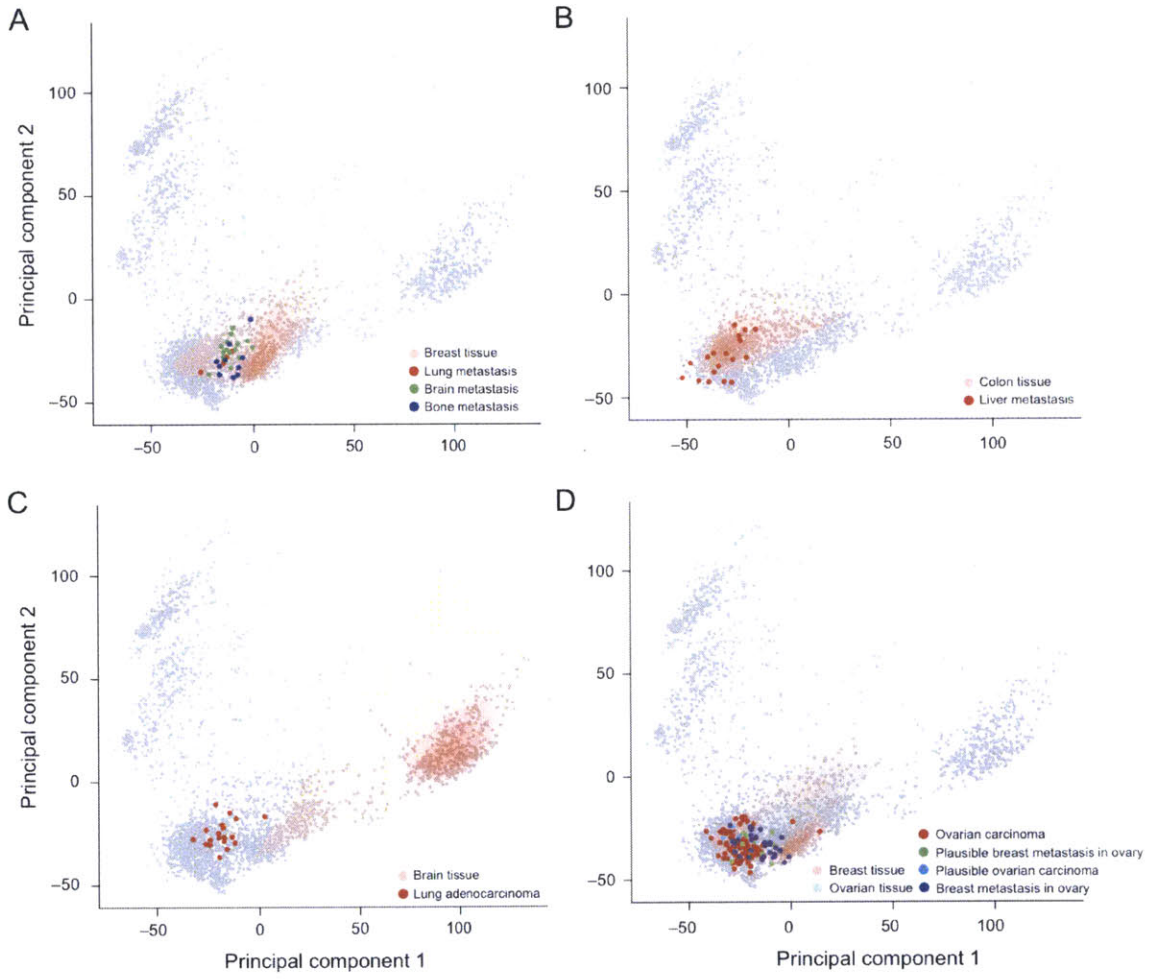

Figure 4-8: Principal component analysis shows that metastatic samples more closely resemble their primary sites. Along with the concept enrichment, the first two principal components of the gene expression data show that the gene expression signature of tumor metastases more closely resembles that of 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 1). (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.

4.6 Implications

We have illustrated a detailed map of the relative locations of diseases and tissues in the smooth continuum of the expression landscape using a publicly-available, broad and heterogeneous expression data set. Given merely gene expression data, we are able to recapitulate the biological and medically relevant concepts relating to the underlying genetic processes. It is also encouraging to see that the power of macroscopic analysis increases with the size of the data, confirming the value of maintaining large, public gene expression databases. The topographical structure of the expression landscape can in and of itself provide key insights into the biomedical similarity of tissues and diseases and can offer valuable knowledge of topics ranging from gene expression study design to treatment of diseases. Clinical applications of this approach could effectively be used to more accurately and consistently annotate clinical samples and provide an impartial view of the landscape of clinico-pathological classification. In addition to 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. The ability to use gene expression to automatically classify and prognosticate patients based on biological samples may be facilitated by a better understanding of the genetic landscape in relation to the conventional clinical classification of disease. Although not meant as a replacement for a skilled clinician or pathologist, this unbiased transcriptome-wide perspective may aid clinicians in identifying the primary sites of metastases and can provide another possible clinical application of this system. Alternatively, it could 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. Since it makes use of an enrichment

statistic that only requires the usual standard of care in the labeling of the sample, this system could be deployed in a clinical setting with minimal alteration of normal procedures.

As suggested by some [56], 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 [31, 35]. It may be the case that the gene expression signatures of disease provide clues to a disease network [11] other than what classical medical knowledge dictates, thus providing new insights into relationships between diseases that previously were unknown. Used as part of a larger diagnostic pipeline that incorporates environmental, genomic, and phenotypic factors as input, mapping expression samples onto the transcriptomic landscape can provide genomic contextual information to the clinician when determining a diagnosis and for developing suitable therapeutics.

Due to the data-driven nature of the method, changing the scope or domain of the labels on the gene expression samples in the database allows the method to be applied in different contexts. Although we have presented the accuracy of the system across various UMLS concepts as a measure of the performance of the system for classification, the primary goal of this method is to provide contextual information of a new gene expression sample. By shifting the current gene expression analysis paradigm of what makes this sample unique? to what other samples am I similar to so that similar treatment can be administered, we hope to address one of the key requirements of the performance and development of personalized medicine.

Chapter 5

Core Stem Cell Transcriptional Activity and Cancer Signatures

This chapter describes an in-depth transcriptional analysis using the Concordia framework. We first identify a highly-conserved stem cell transcriptional profile, then go on to explore the extent to which these expression characteristics are present in a diverse set of normal tissue and malignancy samples. While these results represent an example of the type of large-scale analysis that can be performed with little overhead using the Concordia framework, they also contribute novel biological insight to the ongoing debate about the role played by stem cells in the progression of a variety of cancers.

Understanding the fundamental mechanisms of tumorigenesis remains one of the most pressing problems in modern biology and translational medicine. While the cancer stem cell hypothesis presents a compelling model of self-renewal and partial differentiation, the relationship between tumor cells and normal stem cells has remained unclear. Here, we present a comprehensive analysis of the mRNA expression

profiles derived from a heterogeneous set of normal human tissues, cancers and stem cells. We identify, in an unbiased fashion, transcriptional activity that is associated with pluripotent stem cells and then use this profile to derive a quantitative measure of stem cell-like gene expression activity. We show how this signature stratifies a diverse gene expression dataset according to the samples relative plasticity and differentiation, from ES and iPS cells, to partially committed multipotent stem cells and immortalized cell lines, to solid and hematopoietic cancers, and ultimately fully-differentiated non-diseased tissue. This stem-like signature also serves as a proxy for tumor grade in a variety of solid tumors, suggesting potential therapeutic implications. Our results expand upon the growing link between genes associated with stem cells and their role in a diverse set of cancers.

5.1 Biological overview and motivation

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 [86, 94, 80]. 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 [67]. 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 [16]. 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. Similarly, diverse malignancies have been shown to share broad gene expression programming while also maintaining tissue specific characteristics. Multiple controversies continue to circulate around the

role of particular genes in stem cells vs. differentiated tissues (e.g. N-cadherin [50]), 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 [104]. 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 [44, 29]. 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 [36], leukemic stem cells were observed to express several markers in common with normal stem cells. Subsequently, analogous models have been developed for a number of solid tumors (e.g., brain [89], breast [3], skin [34], ovarian epithelial [9], prostate [25], bone [38], and colon [79] cancers), primarily through the identification of a small population (typically $< 5\%$) 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 at-

tributable 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 [55], while brain cancer and colon cancer stem cells have been isolated among CD133+ cells [89, 79]. Breast cancer stem cells have been defined by a CD44+CD24- phenotype [3], while prostate cancer stem cells have been isolated from minority CD44+/$\alpha_2\beta_1^{hi}$/CD133+ populations [25]. Bone sarcoma cells with proliferative potential have been shown to express activated Stat3 [38]. These cells also expressed a subset of the embryonic stem cell-associated genes (Oct3/4, Nanog) [106], but again, the degree to which these trends may be apparent across other populations of cancer stem cells is unknown.

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 [104]. 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 [52]. Similar methods with broad applicability will pave the way for individually tailored treatment strategies. Enrichment for stem cell-specific hand curated gene sets has been demonstrated in aggressive breast tumors, as well as bladder carcinomas and glioblastomas [16]. Diverse malignant tissue samples have been shown to exhibit a broadly similar trend within a large gene expression database [58], but no specific connection has been made in this context to stem cell-like activity.

Here, we present a structured perspective on a diverse compilation of gene ex-

pression samples that reveals a robust multidimensional continuum from ES / iPS cells to fully differentiated tissues. Our results indicate that, within this functional genomic landscape, cancers display a combination of stem cell-like programming and tissue-specific signatures. We derive a shared molecular measure of pluripotentiality that may help bridge the gap between the disparate cell-type-specific markers associated with individual cancer stem cell populations and their shared proliferative potential. In addition, we demonstrate that our differentiation and pluripotentiality-centric view of gene expression correlates with classical grading systems for a variety of solid tumors, suggesting that our results may form a quantitative axis with practical relevance to personalized medicine.

5.2 Identifying a stem cell gene set

Our first goal was to identify a set of genes whose expression profiles represent a tightly conserved core of transcriptional programming among stem cells. We call this set of genes our stem cell gene set (SCGS). We derived SCGS from a high-quality database called Concordia, representing a significant subset of the NCBIs Gene Expression Omnibus (GEO). Concordia was constructed using a combination of automated textual parsing, human curation and normalization methods (see Chapter 3).

In order to identify a set of genes with highly specific stem cell expression intensities, we used this curated database to identify all of the stem cell samples in our dataset. We then applied a standard signal processing tool, a finite impulse response filter (FIR) [62], to identify those genes with the most highly-conserved expression intensities among the stem cell samples. That is, those genes with a range of expression intensities among the stem cell samples that was most distinct from the

non-stem cell samples scored the highest.

We sought to associate with each gene a measure of how well conserved its expression intensity was over the stem cell samples. Rather than seeking a strict measure of constitutive over- or under-expression of the gene among the stem cell population, our goal was instead to identify individual genes that tightly cluster the stem cell population anywhere along the spectrum of expression intensities.

The input to the FIR scoring procedure is a list of all of the expression samples, sorted according to their intensity for a particular gene. The filter then applies a sliding window to the list and outputs, at each window position, the proportion of stem cell samples within the frame. The maximal value of this sliding window at any position in the list is then taken as that genes score. We use a window equal in size to the total number of stem cell samples in the database, so the interpretation of the filters maximal output is intuitive: If were looking to find all of the stem cell samples within one window frame, what is the greatest fraction that we can localize, given the expression values for this gene across the entire database. Genes with the highest scores are those with most specific stem cell expression intensities.

Binomial p-values (k=number of stem cell samples in a given window frame; n=window frame size; p=proportion of stem cell samples in the entire database) are reported along with these scores.

In contrast to a standard t-test, 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 specific in the stem cell samples, allowing for the diverse population of non-stem cell 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 stem cell marker genes) 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 as only the 24.6% strongest marker gene.

We verified the ability of the SCGS to capture a nuanced measure of stem cell-like gene expression activity by demonstrating the accurate clustering of a series of developing ES cell populations in mouse. This analysis also shows the concordance between the SCGS transcriptional profile and cellular state of differentiation.

To verify that our SCGS captures a quantitative transcriptional measure of differentiation, we used it to examine the expression dynamics of a set of developing mouse ES cells over time (data available in GEO Series GSE12550). This data set consisted of a time course of differentiating mouse ES cells, with gene expression measured at four time points (ES cells, 4 days of differentiation, 8 days of differentiation and 14 days of differentiation). Human genes were mapped to mouse via NCBI's Homologene [14]. Human genes that lacked a unique match in mouse were ignored. Expression intensities were processed in an identical manner to the human data (see Chapter 3), and again summarized by gene. We computed the dominant variance among the differentiating mouse cells via PCA. We used each mouse ES samples coordinates in the first principal basis as a summary value of SCGS gene expression activity.

The dominant expression signal reflected in these genes accurately sorts the samples according to their time point, as shown in Figure 5-1. This supports the hypothesis that our SCGS reflects measurable changes in state of differentiation and pluripotentiality, and reflects the fact that the functional genomic mechanisms associated with stem cell activity are at least partially conserved across species.

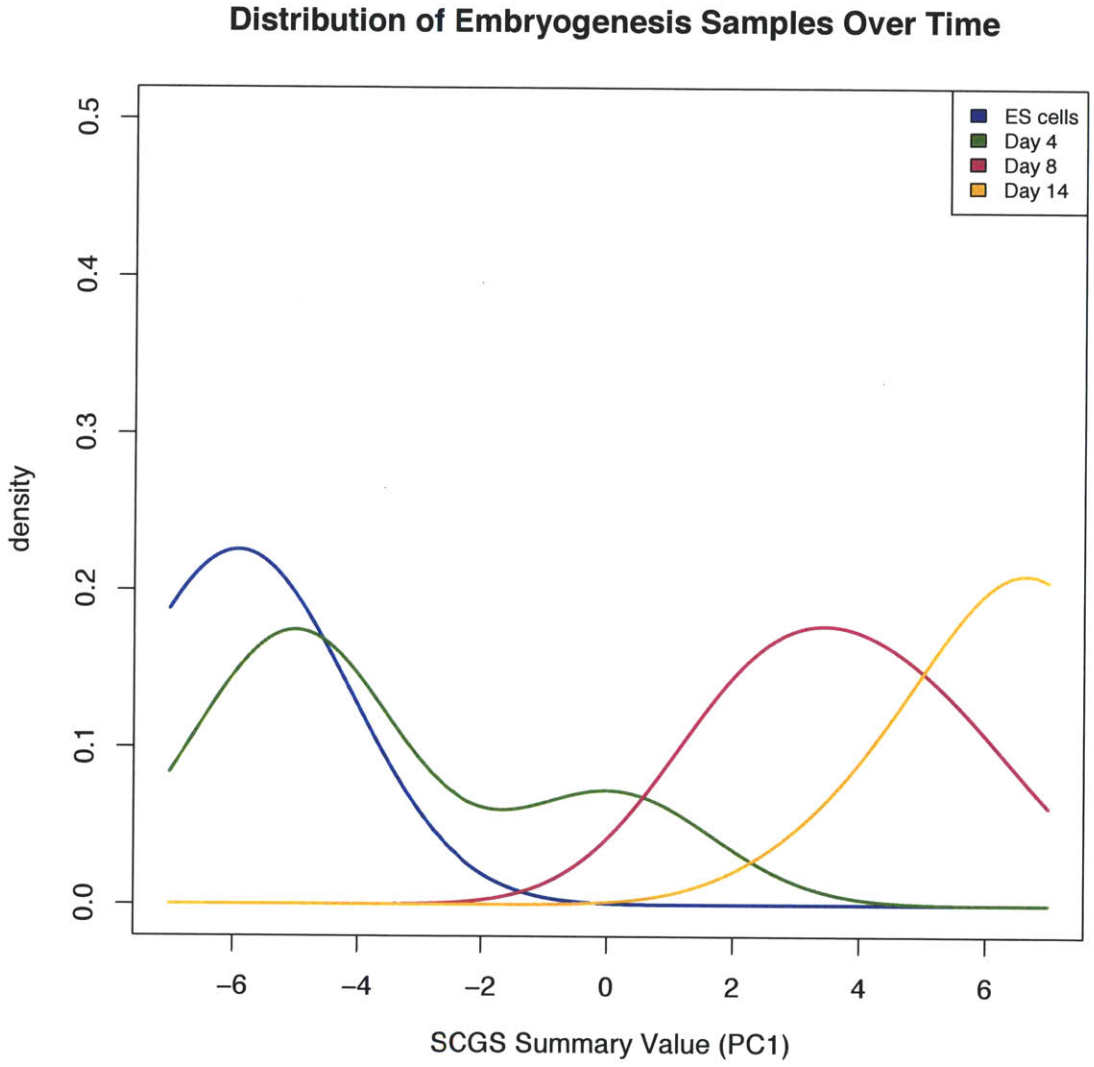

Figure 5-1: Distribution of differentiating mouse ES cells over stem cell signature. Each curve represents the distribution of SCGS summary values for a particular time point. The stem cell signature collocates the four time points samples and clearly separates the early and late stages of differentiation.

Previous studies [16] have examined the expression patterns of literature-curated gene sets relating to ES-like activity among a variety of malignancies. In contrast, we have constructed a gene set that reflects only those transcriptional signals with

the greatest ability to localize the stem cell samples within the spectrum of human tissues and diseases.

The genes comprising the SCGS can be found in Table 5.1.

Table 5.1: Genes comprising the SCGS

Gene Name	Gene ID	Score	Binomial p-value	Percentile
FGF2	2247	0.681564246	7.10E-105	0
ROR1	4919	0.670391061	9.24E-102	4.94E-05
MLLT6	4302	0.664804469	3.21E-100	9.88E-05
GPR176	11245	0.664804469	3.21E-100	9.88E-05
ADH5	128	0.648044693	1.16E-95	0.000197511
ABCG1	9619	0.642458101	3.65E-94	0.000246889
KIF7	374654	0.61452514	7.83E-87	0.000296267
SOHLH2	54937	0.603351955	5.68E-84	0.000345645
RIOK2	55781	0.597765363	1.48E-82	0.000395023
PECAM1	5175	0.597765363	1.48E-82	0.000395023
HLA-DMB	3109	0.597765363	1.48E-82	0.000395023
PDHB	5162	0.586592179	9.33E-80	0.000543156
KDELC1	79070	0.586592179	9.33E-80	0.000543156
AFTPH	54812	0.581005587	2.26E-78	0.000641912
RRAS2	22800	0.581005587	2.26E-78	0.000641912
GIMAP5	55340	0.581005587	2.26E-78	0.000641912
C14orf119	55017	0.575418994	5.37E-77	0.000790045
ARMCX2	9823	0.569832402	1.25E-75	0.000839423
ARMC9	80210	0.569832402	1.25E-75	0.000839423
FKSG49	400949	0.569832402	1.25E-75	0.000839423

FAM118B	79607	0.56424581	2.82E-74	0.000987557
MED20	9477	0.56424581	2.82E-74	0.000987557
PDE4C	5143	0.56424581	2.82E-74	0.000987557
ARRB2	409	0.56424581	2.82E-74	0.000987557
HDX	139324	0.56424581	2.82E-74	0.000987557
CCDC99	54908	0.558659218	6.26E-73	0.001234446
HLA-DPA1	3113	0.558659218	6.26E-73	0.001234446
HLA-DRB5	3127	0.558659218	6.26E-73	0.001234446
HTR7	3363	0.558659218	6.26E-73	0.001234446
RAB3B	5865	0.553072626	1.36E-71	0.001431957
SAT1	6303	0.553072626	1.36E-71	0.001431957
FYB	2533	0.547486034	2.87E-70	0.001530713
TRIM24	8805	0.541899441	5.95E-69	0.001580091
IGF2BP1	10642	0.541899441	5.95E-69	0.001580091
C14orf166	51637	0.541899441	5.95E-69	0.001580091
MTHFD1L	25902	0.541899441	5.95E-69	0.001580091
MMADHC	27249	0.536312849	1.20E-67	0.001777602
FST	10468	0.536312849	1.20E-67	0.001777602
DACT1	51339	0.536312849	1.20E-67	0.001777602
ICMT	23463	0.530726257	2.39E-66	0.001925736
ARPM1	84517	0.530726257	2.39E-66	0.001925736
MCM10	55388	0.525139665	4.62E-65	0.002024491
C3orf21	152002	0.525139665	4.62E-65	0.002024491
PFAS	5198	0.525139665	4.62E-65	0.002024491
CHEK2	11200	0.525139665	4.62E-65	0.002024491

CDC25A	993	0.525139665	4.62E-65	0.002024491
MS4A7	58475	0.525139665	4.62E-65	0.002024491
OSTC	58505	0.525139665	4.62E-65	0.002024491
IPO5	3843	0.525139665	4.62E-65	0.002024491
BOD1	91272	0.519553073	8.75E-64	0.002419514
TRNT1	51095	0.519553073	8.75E-64	0.002419514
SEC11A	23478	0.519553073	8.75E-64	0.002419514
PLRG1	5356	0.519553073	8.75E-64	0.002419514
BCAT1	586	0.519553073	8.75E-64	0.002419514
MCM6	4175	0.51396648	1.62E-62	0.002666403
BBS9	27241	0.51396648	1.62E-62	0.002666403
XPOT	11260	0.51396648	1.62E-62	0.002666403
HAS2	3037	0.51396648	1.62E-62	0.002666403
ALX1	8092	0.51396648	1.62E-62	0.002666403
C1orf135	79000	0.51396648	1.62E-62	0.002666403
DNMT3B	1789	0.508379888	2.94E-61	0.00296267
CCNG1	900	0.508379888	2.94E-61	0.00296267
UBE2K	3093	0.508379888	2.94E-61	0.00296267
CD53	963	0.508379888	2.94E-61	0.00296267
MLL3	58508	0.508379888	2.94E-61	0.00296267
PTPRC	5788	0.502793296	5.21E-60	0.00320956
CCR1	1230	0.502793296	5.21E-60	0.00320956
C13orf15	28984	0.502793296	5.21E-60	0.00320956
SNRPD3	6634	0.502793296	5.21E-60	0.00320956
SNX8	29886	0.502793296	5.21E-60	0.00320956

C10orf128	170371	0.502793296	5.21E-60	0.00320956
DIAPH3	81624	0.502793296	5.21E-60	0.00320956
HLA-DQA1	3117	0.502793296	5.21E-60	0.00320956
LIN28B	389421	0.502793296	5.21E-60	0.00320956
IL10RA	3587	0.502793296	5.21E-60	0.00320956
TNFSF10	8743	0.497206704	9.02E-59	0.003703338
SLC24A1	9187	0.497206704	9.02E-59	0.003703338
MALAT1	378938	0.497206704	9.02E-59	0.003703338
INHBE	83729	0.497206704	9.02E-59	0.003703338
SMARCC2	6601	0.497206704	9.02E-59	0.003703338
GARS	2617	0.497206704	9.02E-59	0.003703338
SAMM50	25813	0.497206704	9.02E-59	0.003703338
CCNA2	890	0.497206704	9.02E-59	0.003703338
IARS	3376	0.497206704	9.02E-59	0.003703338
CCL26	10344	0.491620112	1.53E-57	0.004147738
NDC80	10403	0.491620112	1.53E-57	0.004147738
LOC285431	285431	0.491620112	1.53E-57	0.004147738
DLGAP5	9787	0.491620112	1.53E-57	0.004147738
ZNF167	55888	0.491620112	1.53E-57	0.004147738
RSAD2	91543	0.491620112	1.53E-57	0.004147738
WASH3P	374666	0.491620112	1.53E-57	0.004147738
RARS	5917	0.491620112	1.53E-57	0.004147738
TBCE	6905	0.491620112	1.53E-57	0.004147738
GNL2	29889	0.491620112	1.53E-57	0.004147738
GPX8	493869	0.491620112	1.53E-57	0.004147738

UGGT2	55757	0.48603352	2.54E-56	0.004690895
SLC25A3	5250	0.48603352	2.54E-56	0.004690895
POLE2	5427	0.48603352	2.54E-56	0.004690895
ATP11C	286410	0.48603352	2.54E-56	0.004690895
CTSH	1512	0.48603352	2.54E-56	0.004690895
CTSS	1520	0.48603352	2.54E-56	0.004690895
DBC1	1620	0.48603352	2.54E-56	0.004690895
FAM185B	641808	0.48603352	2.54E-56	0.004690895
KIF11	3832	0.48603352	2.54E-56	0.004690895
HIST1H4C	8364	0.48603352	2.54E-56	0.004690895
HESX1	8820	0.480446927	4.11E-55	0.005184673
GIMAP7	168537	0.480446927	4.11E-55	0.005184673
MSH2	4436	0.480446927	4.11E-55	0.005184673
MXI1	4601	0.480446927	4.11E-55	0.005184673
ORC1L	4998	0.480446927	4.11E-55	0.005184673
CENPI	2491	0.480446927	4.11E-55	0.005184673
GBE1	2632	0.480446927	4.11E-55	0.005184673
XRCC5	7520	0.480446927	4.11E-55	0.005184673
BUB1B	701	0.480446927	4.11E-55	0.005184673
GIMAP6	474344	0.474860335	6.52E-54	0.005629074
DPH3	285381	0.474860335	6.52E-54	0.005629074
KIAA0020	9933	0.474860335	6.52E-54	0.005629074
PRIM1	5557	0.474860335	6.52E-54	0.005629074
CPSF3	51692	0.474860335	6.52E-54	0.005629074
CENPE	1062	0.474860335	6.52E-54	0.005629074

CMBL	134147	0.474860335	6.52E-54	0.005629074
APLP1	333	0.474860335	6.52E-54	0.005629074
EIF2AK4	440275	0.474860335	6.52E-54	0.005629074
GIMAP1	170575	0.474860335	6.52E-54	0.005629074
GIMAP8	155038	0.474860335	6.52E-54	0.005629074
LOC400931	400931	0.474860335	6.52E-54	0.005629074
HAUS1	115106	0.474860335	6.52E-54	0.005629074
CDC123	8872	0.469273743	1.01E-52	0.006270986
EIF2B3	8891	0.469273743	1.01E-52	0.006270986
PAICS	10606	0.469273743	1.01E-52	0.006270986
PLAA	9373	0.469273743	1.01E-52	0.006270986
MTHFD2	10797	0.469273743	1.01E-52	0.006270986
UTP6	55813	0.469273743	1.01E-52	0.006270986
NUSAP1	51203	0.469273743	1.01E-52	0.006270986
SKIV2L2	23517	0.469273743	1.01E-52	0.006270986
MRPL3	11222	0.469273743	1.01E-52	0.006270986
DDX1	1653	0.469273743	1.01E-52	0.006270986
DTD1	92675	0.469273743	1.01E-52	0.006270986
SKP2	6502	0.469273743	1.01E-52	0.006270986
ZNF788	388507	0.469273743	1.01E-52	0.006270986
FBXO22	26263	0.469273743	1.01E-52	0.006270986
HMGA2	8091	0.469273743	1.01E-52	0.006270986
KIF20A	10112	0.463687151	1.53E-51	0.007011653
TSHZ2	128553	0.463687151	1.53E-51	0.007011653
UBXN2A	165324	0.463687151	1.53E-51	0.007011653

DCUN1D5	84259	0.463687151	1.53E-51	0.007011653
CENPH	64946	0.463687151	1.53E-51	0.007011653
PILRB	29990	0.463687151	1.53E-51	0.007011653
MSH6	2956	0.463687151	1.53E-51	0.007011653
C21orf45	54069	0.463687151	1.53E-51	0.007011653
NA	643187	0.458100559	2.28E-50	0.007406676
HAUS6	54801	0.458100559	2.28E-50	0.007406676
TEX10	54881	0.458100559	2.28E-50	0.007406676
MYCN	4613	0.458100559	2.28E-50	0.007406676
NPR3	4883	0.458100559	2.28E-50	0.007406676
C5orf13	9315	0.458100559	2.28E-50	0.007406676
GDF3	9573	0.458100559	2.28E-50	0.007406676
DEPDC1B	55789	0.458100559	2.28E-50	0.007406676
TARDBP	23435	0.458100559	2.28E-50	0.007406676
RAD1	5810	0.458100559	2.28E-50	0.007406676
CDH6	1004	0.458100559	2.28E-50	0.007406676
RPF2	84154	0.458100559	2.28E-50	0.007406676
TET1	80312	0.458100559	2.28E-50	0.007406676
RARS2	57038	0.458100559	2.28E-50	0.007406676
CCDC90B	60492	0.458100559	2.28E-50	0.007406676
ERCC2	2068	0.458100559	2.28E-50	0.007406676
FANCB	2187	0.458100559	2.28E-50	0.007406676
TRPC4	7223	0.458100559	2.28E-50	0.007406676
BCKDHB	594	0.458100559	2.28E-50	0.007406676
CCNB1	891	0.458100559	2.28E-50	0.007406676

HDAC2	3066	0.458100559	2.28E-50	0.007406676
C4orf52	389203	0.458100559	2.28E-50	0.007406676
ZBTB20	26137	0.458100559	2.28E-50	0.007406676
C5orf56	441108	0.458100559	2.28E-50	0.007406676
HAT1	8520	0.458100559	2.28E-50	0.007406676
MCM4	4173	0.452513966	3.30E-49	0.008641122
DIMT1L	27292	0.452513966	3.30E-49	0.008641122
KIF23	9493	0.452513966	3.30E-49	0.008641122
ASCC3	10973	0.452513966	3.30E-49	0.008641122
NMNAT2	23057	0.452513966	3.30E-49	0.008641122
TM2D2	83877	0.452513966	3.30E-49	0.008641122
TBC1D16	125058	0.452513966	3.30E-49	0.008641122
RCSD1	92241	0.452513966	3.30E-49	0.008641122
DHX15	1665	0.452513966	3.30E-49	0.008641122
NA	80047	0.452513966	3.30E-49	0.008641122
FZD2	2535	0.452513966	3.30E-49	0.008641122
ARHGDIB	397	0.452513966	3.30E-49	0.008641122
BMPR1A	657	0.452513966	3.30E-49	0.008641122
JUNB	3726	0.452513966	3.30E-49	0.008641122

5.3 ES-like signature stratifies a diverse expression database by pluripotentiality and malignancy

Via principal component analysis (PCA), we examined the transcriptional profile of the SCGS across the entire collection of normal tissues, cancers and stem cells assembled from GEO. Performing PCA across only the SCGS 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.

This analysis revealed a striking trend apparent in the first two principal components (PCs) of the gene set: 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 5-2. 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). 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 5-2 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:

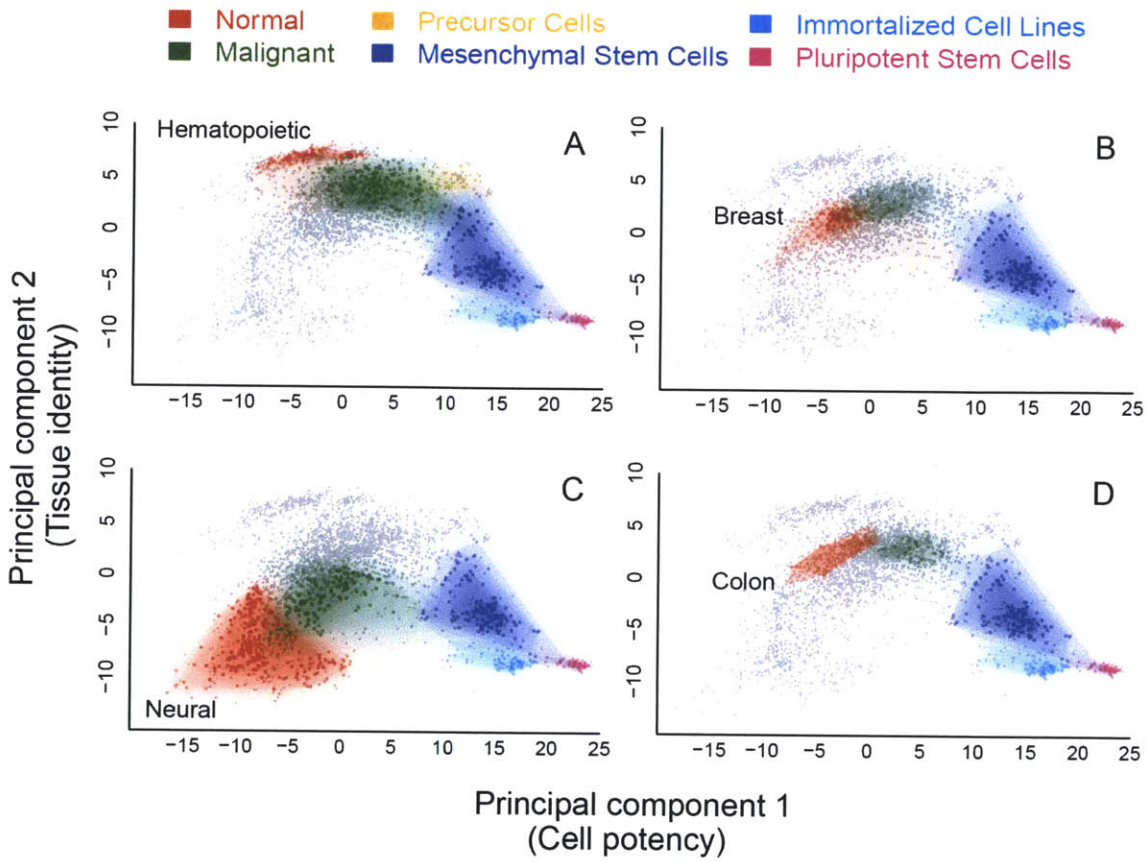

Figure 5-2: The stem cell signature genes stratify a phenotypically diverse database according to pluripotentiality. Each panel shows the entire expression database plotted on the principal coordinates defined by the stem cell signature genes. PC1 is represented on the x-axis of each plot, while PC2 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 (blue) and immortalized cell lines (cyan). Each panel demonstrates 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 the 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. 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, brain and colon all demonstrate the same enhanced stem-like expression activity among their respective malignancies. That is, tumor samples cluster more closely to the pluripotent stem cells than their associated normals

the multipotent stem cells are clustered near the pluripotent stem cells, with the hematopoietic progenitors (the only progenitors in our dataset) slightly farther away (Figure 5-2A).

Further, the hematopoietic, neural and epithelial cancers (shaded green in Figure

5-2A-D) contained in our data all clustered directly between the stem cell populations and their associated normal non-malignant samples. These results suggest that the SCGS captures a kernel of stem cell-like transcriptional activity that concurrently apparent in a variety of malignancies.

5.4 ES-like signature stratifies tumor grade

We used the SCGS 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 identified four independent data series containing expression profiles for graded tumors of various tissue types in GEO (GSE4290, GSE23593, GSE17537, GSE18842) on Affymetrix HGU 133+ 2.0. Each series was pre-processed (MAS5.0 normalized, summarized) as previously described. Within each series, the SCGS 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. For each data set, we computed PCA over the SCGS to identify the dominant differences between the samples within the context of the stem cell genes.

Our analysis revealed a unified molecular signature that correlates with tumor grade for a variety of primary tissues. The first principal component of each data set revealed a strong separation between the high- and low-grade tumor samples in each tissue surveyed. Figure 5-3 shows the distribution of SCGS summary values for the four tissue types graded tumor samples. In each case, the transcriptional activity of the SCGS defines a clear separation between the high- and low-graded tumors, while providing a molecular foundation based on stem-like expression for the clinical difficulty in classifying mid-grade tumors.

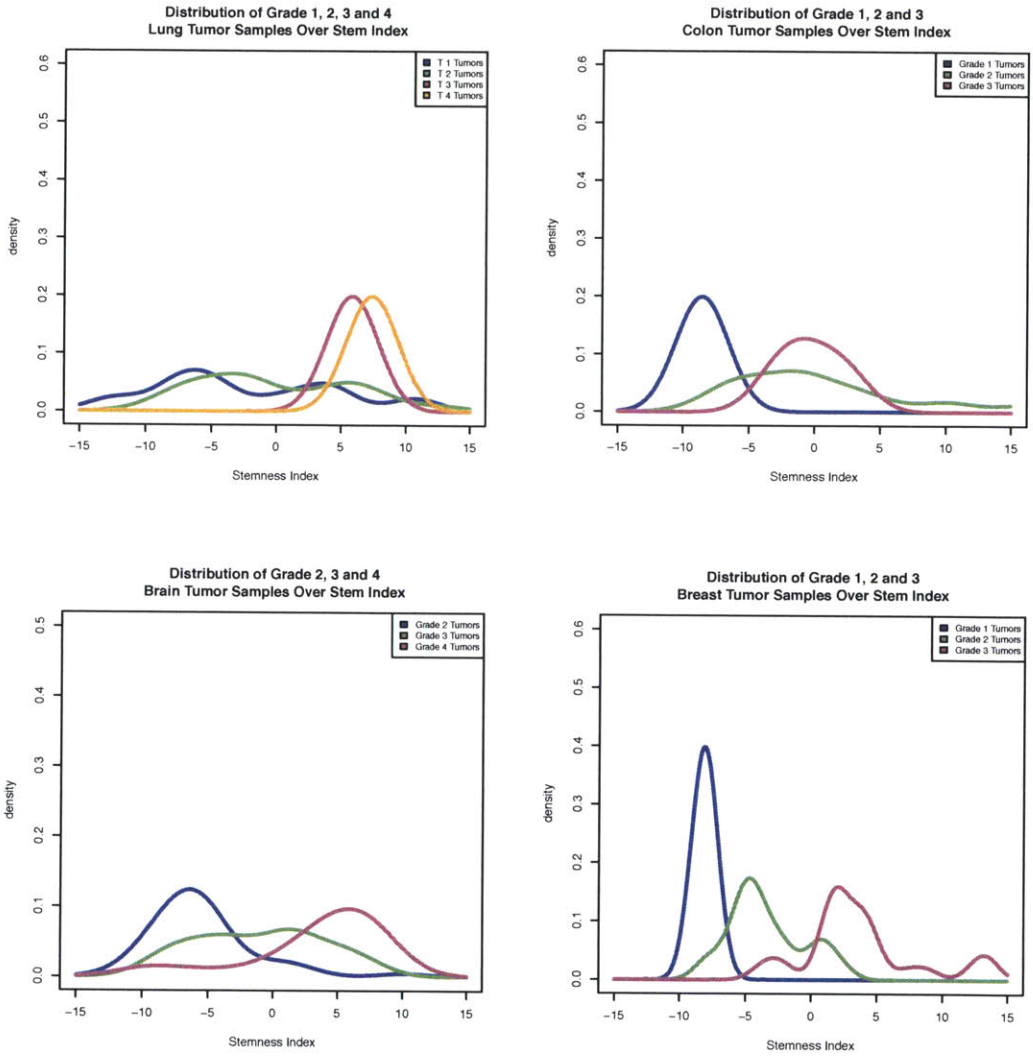

Figure 5-3: Stem cell-like activity correlates with tumor grade in various solid malignancies. Each panel displays the distribution, within the space of the stem cell genes, of graded tumor samples for one particular tissue type. Our molecular index of pluripotentiality and proliferative potential consistently separates high-grade tumors from low grade ones. Based on this transcriptional index, the mid-grade tumors are less well defined.

5.5 Characterizing the functional diversity of the stem cell gene set

Our final goal was to characterize the functional diversity of the genes comprising the SCGS. Hierarchical clustering of these genes transcriptional activity in a popula-

tion of pluripotent stem cells revealed four distinct coexpression modules. For each module, we then identified a set of over-enriched GO biological processes.

To illustrate the gene expression trends apparent within each gene cluster, Figure 5-4 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 Table 7.5 and Table 7.6 for detailed annotations) are most highly expressed in the pluripotent stem cells, and less so, respectively, through increasing levels of cellular differentiation / decreasing pluripotentiality. Genes related to metabolism and hormone signaling (see Table 7.7 for detailed annotations) 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 (see Table 7.8 for detailed annotations) 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.

5.5.1 SCGS genes represented in the Figure 5-4

This section lists, in order of appearance, the genes displayed in 5-4.

Genes represented in the DNA replication / cell cycle cluster

DNMT3B, MCM6, CDC25A, PFAS, MCM4, XRCC5, HAUS6, TET1, IGF2BP1, PLAA, DEPDC1B, TEX10, CCDC99, MSH2, BUB1B, MSH6, DLGAP5, SKIV2L2, CENPE, CHEK2, SOHLH2, CCNB1, RRAS2, PRIM1, PAICS, CCNA2, CPSF3, NUSAP1, LIN28B, IPO5, KIF11, BMPR1A, NDC80, BCAT1, CCNG1, ZNF788,

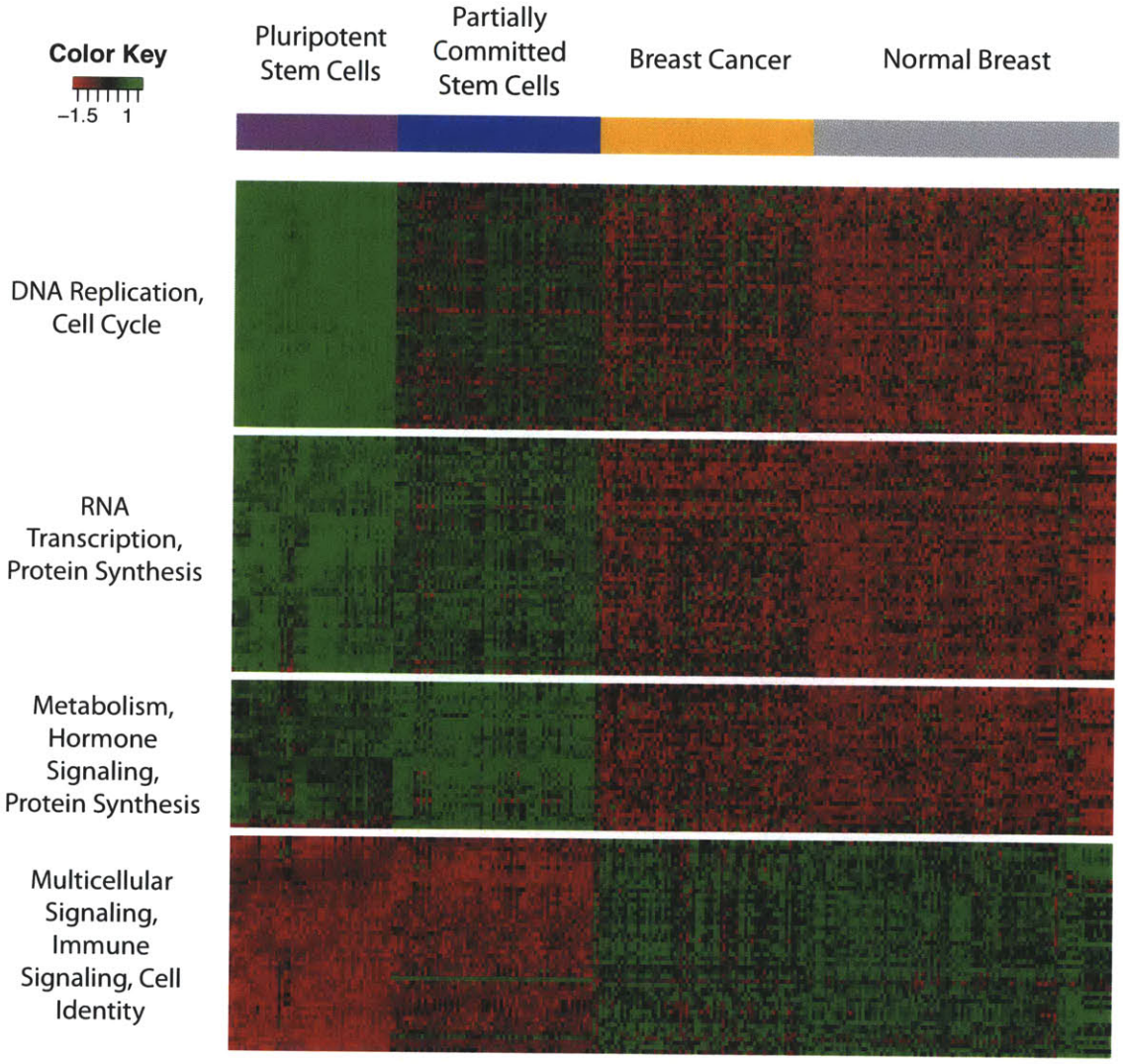

Figure 5-4: Four distinct expression modules are apparent within the stem cell genes. To demonstrate the transcriptome-wide implications of these profiles, this figure shows a series of cell types, ranging from fully differentiated (normal breast), through the associated malignancy, partially committed 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.

ASCC3, FANCB, MCM10, HMGA2, SKP2, TRIM24, ORC1L, HDAC2, HESX1, C1orf135, INHBE, C21orf45, DCUN1D5, POLE2, MRPL3, CENPH, MYCN, HAUS1,

GDF3

Genes represented in the RNA transcription / protein synthesis cluster

TBCE, RIOK2, BCKDHB, RAD1, C5orf13, ADH5, PLRG1, ROR1, RAB3B, LOC285431, DBC1, KIF23, DIAPH3, GNL2, FGF2, TARDBP, NMNAT2, ZNF167, KIF20A, CENPI, DDX1, C3orf21, GPR176, FBXO22, BBS9, C14orf166, BOD1, CDC123, SNRPD3, FAM118B, DPH3, EIF2B3, KDELC1, RPF2, APLP1, DACT1, PDHB, C14orf119, DTD1, SAMM50, CCL26, C4orf52, CCDC90B, MED20, UTP6, RARS2, KIAA0020, ARMCX2, RARS, MTHFD2, DHX15, HTR7, HIST1H4C

Genes represented in the metabolism / hormone signaling cluster

MTHFD1L, ARMC9, XPOT, IARS, HDX, ARPM1, ERCC2, TBC1D16, GARS, KIF7, UBE2K, SLC25A3, ICMT, UGGT2, ATP11C, SLC24A1, EIF2AK4, GPX8, ALX1, OSTC, TRPC4, HAS2, FZD2, TRNT1, MMADHC, SNX8, CDH6, HAT1, SEC11A, DIMT1L, TM2D2, FST, GBE1

Genes represented in the multicellular signaling / immune signaling / cellular identity cluster

MLL3, MXI1, FKSG49, FAM185B, ARRB2, SMARCC2, WASH3P, PILRB, CTSH, SAT1, JUNB, CD53, PECAM1, IL10RA, RCSD1, ARHGDIB, GIMAP5, GIMAP6, HLA-DMB, PTPRC, C10orf128, CMBL, HLA-DRB5, HLA-DPA1, ABCG1, GIMAP7, HLA-DQA1, TSHZ2, C13orf15, CCR1, NPR3, RSAD2, GIMAP1, TNFSF10, AFTPH, MALAT1, UBXN2A, PDE4C, GIMAP8, FYB, MS4A7, C5orf56, LOC400931, MLLT6, CTSS, ZBTB20

5.6 Implications of the stem cell gene set

We have demonstrated conserved stem cell-like transcriptional activity across a wide variety of hematopoietic and solid cancers through a comprehensive molecular survey of malignancy, pluripotent stem cells and normal tissues. Our findings echo several recent developments in the cancer stem cell debate. In particular, our results highlight transcriptional evidence that, despite individual tissue-specific characteristics, a wide range of cancers share a common set of functional genomic mechanisms with each other as well as pluripotent and multipotent stem cells.

The overall landscape of the human transcriptome appears to be organized by a combination of tissue, cell-type and disease-specific features. Previous studies have suggested that the primary factors driving the organization of this landscape are largely attributable to hematopoietic and malignant programming. Our results indicate that while there exists a strong tissue-specific signal, the malignancy signature is more specifically a reflection of the self-renewal and pluripotentiality common to both stem cell populations and heterogeneous tumors.

Although our results neither support nor refute the cancer stem cell hypothesis, they do strongly suggest that there are a large number of stem cell-specific mechanisms 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 identified in this paper.

Controversy remains as to the exact origin of cancer stem cells, with the data often supporting contradictory conclusions for different tissues [55]. Some studies have indicated that cancer stem cells arise from normal pluripotent or multipotent stem cells that have lost the ability to regulate their proliferative activity, while others have indicated that cancer stem cells derive from a committed progenitor cell that has acquired the ability to self renew [104].

Despite these controversies, and what appear to be cell type-specific molecular characteristics, we have demonstrated a specific, unbiased transcriptional signal that is shared among a wide variety of solid and hematopoietic cancers. Moreover, when considered from a transcriptome-wide perspective, this signal is indicative of stem cell-like activity. We have shown how these gene expression patterns are most strongly associated with embryonic and induced pluripotent stem cells, and are successively less apparent in multipotent stem cells, malignancies, and fully differentiated tissues, respectively. In addition, the genes that comprise this signal also reveal a stratification of solid tumors that correlates strongly with classical grading systems.

Recent trends in chemotherapy design have focused not only on regulating cytotoxicity, but also on affecting the differentiation pathways that are apparently impaired in malignant cells. For example, Stegmaier et al. have demonstrated the ability of gefitinib to induce myeloid differentiation in both AML cell lines as well as patient-derived AML blast cells [93]. Indeed, the phenotypic transformation induced by gefitinib was shown to be observable in both cellular morphology and gene expression. The ubiquitous stem cell-like expression patterns described in this paper, as well as those specifically tuned to individual tumor subclasses, may prove useful in screening compounds through the early stages of drug discovery. To these

ends, we have begun investigating the transcriptional effect of known therapeutic compounds on the SCGS via CMAP-like [49] analyses. Understanding the transcriptional changes wrought by these compounds within the context of pluripotentiality and differentiation may be of fundamental value in personalized oncology and therapy selection.

Chapter 6

Concluding Remarks

In this thesis, we have demonstrated how a large set of gene expression samples can be indexed by phenotype using a combination of natural language processing tools, expert curation, and an ontology-based hierarchical database. We described techniques for using the resulting database to characterize the phenotype of new gene expression samples, and demonstrated the ability of that phenotype enrichment scoring scheme to recover clinically relevant information about the tissue of origin for metastasized tumors. Finally, we developed a method that uses our database to identify expression profiles that are highly conserved within particular phenotypes. We used that method to characterize the relative extent to which stem cell-like expression programming is present across the full spectrum of diseases and tissues represented in our database, and highlighted the clinical relevance of those findings.

6.1 Future Work

There are a number of future directions that we intend to explore. These fall into two main categories: 1) analyses that utilize the gene expression resources assembled for this thesis, and 2) future projects that involve repurposing the flexible Concordia ontology-based indexing system for categorizing data in other domaines (e.g., outside the scope of gene expression analyses.

6.1.1 Expanding the expression database

Because the manual curation procedure outlined in this thesis is time consuming and tedious, we have begun developing a set of algorithms that use a reference set such as the Concordia GEO database described earlier as a starting point to bootstrap a large database. Inspired by the reinforcement learning paradigm [98], we take the samples that have been expert labeled, and search for samples that are not already in the database, but have similar gene expression profiles and whose NLP-derived concept associations are similar to those in the database. The goal is to prioritize the introduction of new samples into the database in such a way as to rapidly expand the coverage for one or more particular phenotypes.

For example, if we require a large number of blood samples from type II diabetic patients, but such samples are scarce in our existing dataset (because they have not yet been manually verified), we begin by hand labeling a small number (early cross validation results indicate that as few as 5-10 samples can serve as a reliable starting point). We then ask our algorithm to search for samples whose gene expression and text characteristics are similar to those few initial samples. Using this framework, the system continuously updates its understanding of each phenotype, and dynamically ranks the unlabeled samples phenotypic predisposition in accordance with this belief.

In so doing, the user is constantly provided with high-confidence samples related to their specific interest.

The database, as it is currently constructed, represents a diverse cross-section of both normal and malignant human tissue samples, but is still only a small fraction (roughly 10%) of the total number of HGU133+ 2.0 samples for which GEO makes CEL files available. We intend to use the tools outlined above to expand the breadth of this data set in order to tailor it specifically to our future analyses.

6.1.2 Therapeutic compound expression profiles

The emergence of such large-scale molecular profiling data sets as the one described in this thesis has given rise to a series of techniques for computational drug repositioning [32, 59]. Here, the goal is to search large phenotypically diverse databases for previously un-exploited molecular signatures that appear to be therapeutic targets for existing drug compounds. Because our Concordia database represents a comprehensive hierarchical characterization of the constituent samples, and affords the user effortless querying of this structure, we believe that Concordia is a highly attractive framework for enabling these types of analyses.

In addition, we are also investigating the possibility of assembling a Concordia database containing transcriptional profiles for a collection of tumor biopsies, and associating that database with clinical histories, courses of treatment, and outcomes. We believe that such a resource, if adequate in its breadth and depth, would provide a clinically-relevant method for pairing new patients with a course of treatment that has the optimal expected outcome, based on the molecular profile of their particular tumor.

6.1.3 Adverse drug / event data mining

We have begun early-stage analyses of anonymized patient data to detect adverse events that coincide with a population of patients being prescribed a particular drug. Here, we use Concordia to index the patient "event" database, then assess, for every drug of interest, whether the odds of each event (including those event categorizations implied by the UMLS hierarchy, but not explicitly defined in the raw data) are statistically surprising in the patients who were administered the drug vs. those who were not administered the drug.

6.1.4 Monitoring disease outbreak

The advent of realtime world-wide health monitoring systems like HealthMap [37] has major implications for how we analyze epidemiological data. We anticipate that Concordia's ability to summarize the textual information that underlies such services at arbitrarily specific levels of medical knowledge will prove useful in generating noise-aware statistical models that may eventually be used to track and anticipate disease outbreak.

Chapter 7

Supplemental Tables

This chapter contains supporting material to the main thesis.

7.1 GEO Samples in the Concordia database

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

7.2 Tables

Table 7.1: GO terms associated with the top 250 differentially expressed soft tissue genes

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
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
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
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
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
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
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
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
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
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
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
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
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
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

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
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
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
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
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
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
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
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
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
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

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
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
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

Table 7.2: GO terms associated with the top 250 differentially expressed soft tissue genes

GO ID	**p-value**	**Term**
GO:0005584	0.017	collagen type I
GO:0005583	0	fibrillar collagen
GO:0032964	0	collagen biosynthetic process
GO:0001527	0	microfibril
GO:0043205	0.005	fibril
GO:0030057	0	desmosome
GO:0048407	0	platelet-derived growth factor binding
GO:0030199	0	collagen fibril organization
GO:0005520	0	insulin-like growth factor binding
GO:0005581	0	collagen
GO:0032963	0	collagen metabolic process

GO:0044259	0	multicellular organismal macromolecule metabolic process
GO:0044236	0.001	multicellular organismal metabolic process
GO:0044420	0	extracellular matrix part
GO:0005201	0	extracellular matrix structural constituent
GO:0030198	0	extracellular matrix organization
GO:0005604	0	basement membrane
GO:0043588	0.001	skin development
GO:0005200	0.001	structural constituent of cytoskeleton
GO:0010035	0.033	response to inorganic substance
GO:0001649	0.039	osteoblast differentiation
GO:0009612	0	response to mechanical stimulus
GO:0043062	0	extracellular structure organization
GO:0006956	0.001	complement activation
GO:0070161	0.018	anchoring junction
GO:0002541	0.002	activation of plasma proteins involved in acute inflammatory response
GO:0009987	0.013	cellular process
GO:0005911	0.036	cell-cell junction
GO:0016043	0.048	cellular component organization
GO:0031960	0	response to corticosteroid stimulus
GO:0031012	0	extracellular matrix
GO:0005578	0	proteinaceous extracellular matrix
GO:0016337	0.008	cell-cell adhesion
GO:0019838	0	growth factor binding

GO:0030154	0	cell differentiation
GO:0008201	0	heparin binding
GO:0051384	0	response to glucocorticoid stimulus
GO:0001525	0.017	angiogenesis
GO:0008544	0	epidermis development
GO:0005539	0	glycosaminoglycan binding
GO:0005198	0	structural molecule activity
GO:0006959	0.041	humoral immune response
GO:0001871	0	pattern binding
GO:0030247	0	polysaccharide binding
GO:0030855	0.004	epithelial cell differentiation
GO:0048869	0.017	cellular developmental process
GO:0044421	0	extracellular region part
GO:0009628	0.049	response to abiotic stimulus
GO:0005576	0	extracellular region
GO:0005615	0	extracellular space
GO:0048545	0	response to steroid hormone stimulus
GO:0050896	0.05	response to stimulus
GO:0007584	0.028	response to nutrient
GO:0009888	0	tissue development
GO:0007155	0	cell adhesion
GO:0022610	0	biological adhesion
GO:0009725	0	response to hormone stimulus
GO:0009719	0.008	response to endogenous stimulus
GO:0010033	0	response to organic substance

GO:0009605	0.02	response to external stimulus
GO:0048856	0	anatomical structure development
GO:0042221	0	response to chemical stimulus
GO:0032502	0	developmental process
GO:0006950	0.023	response to stress

Table 7.3: GO terms associated with the top 250 differentially expressed brain genes.

GO ID	**p-value**	**Term**
GO:0045110	0.044	intermediate filament bundle assembly
GO:0005883	0.001	neurofilament
GO:0060052	0.013	neurofilament cytoskeleton organization
GO:0007269	0.02	neurotransmitter secretion
GO:0001505	0	regulation of neurotransmitter levels
GO:0006836	0	neurotransmitter transport
GO:0008021	0.013	synaptic vesicle
GO:0043197	0.032	dendritic spine
GO:0044309	0.032	neuron spine
GO:0033267	0	axon part
GO:0030424	0	axon
GO:0007409	0	axonogenesis
GO:0043005	0	neuron projection
GO:0008509	0.035	anion transmembrane transporter activity
GO:0048812	0	neuron projection morphogenesis
GO:0007417	0	central nervous system development
GO:0048858	0	cell projection morphogenesis

GO:0044456	0	synapse part
GO:0045202	0	synapse
GO:0044463	0	cell projection part
GO:0032990	0.003	cell part morphogenesis
GO:0007268	0	synaptic transmission
GO:0022891	0.018	substrate-specific transmembrane transporter activity
GO:0022857	0.04	transmembrane transporter activity
GO:0005215	0.007	transporter activity
GO:0045211	0.019	postsynaptic membrane
GO:0042995	0	cell projection
GO:0030054	0	cell junction
GO:0007399	0	nervous system development
GO:0048731	0	system development
GO:0022838	0.036	substrate-specific channel activity
GO:0051234	0.02	establishment of localization
GO:0007267	0.021	cell-cell signaling
GO:0006810	0.04	transport
GO:0015075	0.013	ion transmembrane transporter activity
GO:0007154	0.02	cell communication
GO:0006811	0.017	ion transport
GO:0044459	0.003	plasma membrane part
GO:0048856	0.033	anatomical structure development

Table 7.4: GO terms associated with the top 250 differentially expressed blood genes.

GO ID	p-value	Term
GO:0042105	0	alpha-beta T cell receptor complex
GO:0045730	0.008	respiratory burst
GO:0050857	0.041	positive regulation of antigen receptor-mediated signaling pathway
GO:0005833	0	hemoglobin complex
GO:0005344	0.001	oxygen transporter activity
GO:0042101	0.002	T cell receptor complex
GO:0050854	0.005	regulation of antigen receptor-mediated signaling pathway
GO:0031640	0.004	killing of cells of another organism
GO:0045058	0.035	T cell selection
GO:0003823	0	antigen binding
GO:0001906	0.036	cell killing
GO:0050830	0	defense response to Gram-positive bacterium
GO:0009620	0.009	response to fungus
GO:0006968	0	cellular defense response
GO:0001608	0.045	nucleotide receptor activity, G-protein coupled
GO:0045028	0.045	purinergic nucleotide receptor activity, G-protein coupled
GO:0004715	0.036	non-membrane spanning protein tyrosine kinase activity
GO:0042742	0	defense response to bacterium
GO:0031225	0.014	anchored to membrane
GO:0006935	0	chemotaxis

GO:0042330	0	taxis
GO:0050870	0.015	positive regulation of T cell activation
GO:0009617	0	response to bacterium
GO:0042110	0	T cell activation
GO:0006955	0	immune response
GO:0002376	0	immune system process
GO:0050863	0.004	regulation of T cell activation
GO:0040011	0	locomotion
GO:0046649	0	lymphocyte activation
GO:0007626	0	locomotory behavior
GO:0006952	0	defense response
GO:0050867	0.014	positive regulation of cell activation
GO:0045321	0	leukocyte activation
GO:0051707	0	response to other organism
GO:0009897	0.044	external side of plasma membrane
GO:0002684	0	positive regulation of immune system process
GO:0001775	0	cell activation
GO:0051249	0.01	regulation of lymphocyte activation
GO:0050865	0.002	regulation of cell activation
GO:0002694	0.008	regulation of leukocyte activation
GO:0006954	0	inflammatory response
GO:0002682	0	regulation of immune system process
GO:0007610	0.002	behavior
GO:0009607	0	response to biotic stimulus
GO:0030246	0.038	carbohydrate binding

GO:0009611	0	response to wounding
GO:0009605	0.001	response to external stimulus
GO:0005887	0	integral to plasma membrane
GO:0031226	0	intrinsic to plasma membrane
GO:0051704	0.003	multi-organism process
GO:0004872	0	receptor activity
GO:0004871	0	signal transducer activity
GO:0060089	0	molecular transducer activity
GO:0006950	0	response to stress
GO:0050896	0	response to stimulus
GO:0005886	0	plasma membrane
GO:0044459	0	plasma membrane part
GO:0007166	0	cell surface receptor linked signaling pathway
GO:0004888	0.012	transmembrane receptor activity
GO:0023033	0	signaling pathway
GO:0023052	0.003	signaling
GO:0016020	0	membrane
GO:0044425	0	membrane part
GO:0031224	0.002	intrinsic to membrane
GO:0016021	0.012	integral to membrane

Table 7.5: GO terms associated with the DNA replication / cell cycle SCGS expression module

GO ID	**p-value**	**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
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
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
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
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

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
GO ID	**p-value**	**Term**

Table 7.6: GO terms associated with the RNA transcription / protein synthesis SCGS expression module

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
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
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
GO ID	**p-value**	**Term**

Table 7.7: GO terms associated with the metabolism / hormone signaling SCGS expression module

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
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
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
GO ID	**p-value**	**Term**

Table 7.8: GO terms associated with the signaling / cellular identity SCGS expression module

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
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
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
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
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
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
GO ID	**p-value**	**Term**

Bibliography

[1] 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. 2009.

[2] Affymetrix. *Affymetrix Microarray Suite User Guide.* Santa Clara, CA.

[3] M. Al-Hajj, M. S. Wicha, A. Benito-Hernandez, S. J. Morrison, and M. F. Clarke. Prospective identification of tumorigenic breast cancer cells. 100(7):3983–8, 2003.

[4] Bruce Alberts. *Molecular biology of the cell.* Routledge, 1989.

[5] David B Allison, Xiangqin Cui, Grier P Page, and Mahyar Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nature reviews. Genetics*, 7(1):55–65, January 2006.

[6] David Altshuler, Mark J Daly, and Eric S Lander. Genetic mapping in human disease. *Science*, 322(5903):881–888, November 2008.

[7] A R Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 17–21, 2001.

[8] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, May 2000.

[9] S. A. Bapat, A. M. Mali, C. B. Koppikar, and N. K. Kurrey. Stem and progenitor-like cells contribute to the aggressive behavior of human epithelial ovarian cancer. 65(8):3025–9, 2005.

[10] Z Bar-Joseph, GK Gerber, and TI Lee. Computational discovery of gene modules and regulatory networks. *Nature*, 2003.

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

[12] Andrea Barczak, Madeleine Willkom Rodriguez, Kristina Hanspers, Laura L Koth, Yu Chuan Tai, Benjamin M Bolstad, Terence P Speed, and David J Erle. Spotted long oligonucleotide arrays for human gene expression analysis. *Genome research*, 13(7):1775–1785, July 2003.

[13] T. Barrett and R. Edgar. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. 411:352–69, 2006.

[14] 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*, 39(Database issue):D1005–10, January 2011.

[15] C. M. Baum, I. L. Weissman, A. S. Tsukamoto, A. M. Buckle, and B. Peault. Isolation of a candidate human hematopoietic stem-cell population. 89(7):2804–8, 1992.

[16] I. Ben-Porath, M. W. Thomson, V. J. Carey, R. Ge, G. W. Bell, A. Regev, and R. A. Weinberg. An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. 40(5):499–507, 2008.

[17] 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.

[18] David Blumenthal. Stimulating the adoption of health information technology. *The West Virginia medical journal*, 105(3):28–29, April 2009.

[19] O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. 32(Database issue):D267–70–70, 2004.

[20] BM Bolstad, RA Irizarry, and M Åstrand. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 2003.

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

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

[23] C Cheadle, YS Cho-Chung, and KG Becker. Application of z-score transformation to Affymetrix data. - Abstract - UK PubMed Central. *Applied ...*, 2003.

[24] Yanqing Chen, Jun Zhu, Pek Yee Lum, Xia Yang, Shirly Pinto, Douglas J MacNeil, Chunsheng Zhang, John Lamb, Stephen Edwards, Solveig K Sieberts, Amy Leonardson, Lawrence W Castellini, Susanna Wang, Marie-France Champy, Bin Zhang, Valur Emilsson, Sudheer Doss, Anatole Ghazalpour, Steve Horvath, Thomas A Drake, Aldons J Lusis, and Eric E Schadt. Variations in DNA elucidate molecular networks that cause disease. *Nature*, 452(7186):429–435, March 2008.

[25] A. T. Collins, P. A. Berry, C. Hyde, M. J. Stower, and N. J. Maitland. Prospective identification of tumorigenic prostate cancer stem cells. 65(23):10946–51, 2005.

[26] Carlos Coronel, Steven Morris, and Peter Rob. *Database systems.* design, implementation, and management. Course Technology Ptr, November 2009.

[27] Francis Crick. The Central Dogma of Molecular Biology. *Nature*, 227(5258):561–563, August 1970.

[28] Jeffrey Dean and Sanjay Ghemawat. MapReduce. *Communications of the ACM*, 53(1):72, January 2010.

[29] G. Dontu, M. Al-Hajj, W. M. Abdallah, M. F. Clarke, and M. S. Wicha. Stem cells in normal breast development and breast cancer. 36 Suppl 1:59–72, 2003.

[30] Radoje Drmanac, Andrew B Sparks, Matthew J Callow, Aaron L Halpern, Norman L Burns, Bahram G Kermani, Paolo Carnevali, Igor Nazarenko, Geof-

frey B Nilsen, George Yeung, Fredrik Dahl, Andres Fernandez, Bryan Staker, Krishna P Pant, Jonathan Baccash, Adam P Borcherding, Anushka Brownley, Ryan Cedeno, Linsu Chen, Dan Chernikoff, Alex Cheung, Razvan Chirita, Benjamin Curson, Jessica C Ebert, Coleen R Hacker, Robert Hartlage, Brian Hauser, Steve Huang, Yuan Jiang, Vitali Karpinchyk, Mark Koenig, Calvin Kong, Tom Landers, Catherine Le, Jia Liu, Celeste E McBride, Matt Morenzoni, Robert E Morey, Karl Mutch, Helena Perazich, Kimberly Perry, Brock A Peters, Joe Peterson, Charit L Pethiyagoda, Kaliprasad Pothuraju, Claudia Richter, Abraham M Rosenbaum, Shaunak Roy, Jay Shafto, Uladzislau Sharanhovich, Karen W Shannon, Conrad G Sheppy, Michel Sun, Joseph V Thakuria, Anne Tran, Dylan Vu, Alexander Wait Zaranek, Xiaodi Wu, Snezana Drmanac, Arnold R Oliphant, William C Banyai, Bruce Martin, Dennis G Ballinger, George M Church, and Clifford A Reid. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, 327(5961):78–81, January 2010.

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

[32] Joel T Dudley, Marina Sirota, Mohan Shenoy, Reetesh K Pai, Silke Roedder, Annie P Chiang, Alex A Morgan, Minnie M Sarwal, Pankaj Jay Pasricha, and Atul J Butte. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Science translational medicine*, 3(96):96ra76, August 2011.

[33] Valur Emilsson, Gudmar Thorleifsson, Bin Zhang, Amy S Leonardson, Florian Zink, Jun Zhu, Sonia Carlson, Agnar Helgason, G Bragi Walters, Steinunn

Gunnarsdottir, Magali Mouy, Valgerdur Steinthorsdottir, Gudrun H Eiriksdottir, Gyda Bjornsdottir, Inga Reynisdottir, Daniel Gudbjartsson, Anna Helgadottir, Aslaug Jonasdottir, Adalbjorg Jonasdottir, Unnur Styrkarsdottir, Solveig Gretarsdottir, Kristinn P Magnusson, Hreinn Stefansson, Ragnheidur Fossdal, Kristleifur Kristjansson, Hjortur G Gislason, Tryggvi Stefansson, Bjorn G Leifsson, Unnur Thorsteinsdottir, John R Lamb, Jeffrey R Gulcher, Marc L Reitman, Augustine Kong, Eric E Schadt, and Kari Stefansson. Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423–428, March 2008.

[34] D. Fang, T. K. Nguyen, K. Leishear, R. Finko, A. N. Kulp, S. Hotz, P. A. Van Belle, X. Xu, D. E. Elder, and M. Herlyn. A tumorigenic subpopulation with stem cell properties in melanomas. 65(20):9328–37, 2005.

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

[36] P. J. Fialkow. Stem cell origin of human myeloid blood cell neoplasms. 74:43–7, 1990.

[37] Clark C Freifeld, Kenneth D Mandl, Ben Y Reis, and John S Brownstein. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *Journal of the American Medical Informatics Association : JAMIA*, 15(2):150–157, February 2008.

[38] C. P. Gibbs, V. G. Kukekov, J. D. Reith, O. Tchigrinova, O. N. Suslov, E. W. Scott, S. C. Ghivizzani, T. N. Ignatova, and D. A. Steindler. Stem-like cells in bone sarcomas: implications for tumorigenesis. 7(11):967–76, 2005.

[39] Federico Girosi, Robin Meili, and Richard Scoville. Extrapolating evidence of health information technology savings and costs. Technical report, RAND, September 2005.

[40] R Gordon. Essential JNI: Java Native Interface, 1998.

[41] Mark Greaves. Semantic Web 2.0. *IEEE Intelligent Systems*, 22(2):94–96, March 2007.

[42] F Anthony Greco, Howard A Burris III, Joan B Erland, James R Gray, Leonard A Kalman, Marshall T Schreeder, and John D Hainsworth. Carcinoma of Unknown Primary Site: Long Term Follow-Up after Treatment with Paclitaxel, Carboplatin, and Etoposide. 89(12):2655–2660, 2000.

[43] F Anthony Greco, David R Spigel, D. A Yardley, M. G Erlander, X. J Ma, and John D Hainsworth. Molecular Profiling in Unknown Primary Cancer: Accuracy of Tissue of Origin Prediction. 15(5):500–506, 2010.

[44] G. H. Heppner and B. E. Miller. Tumor heterogeneity: biological implications and therapeutic consequences. 2(1):5–23, 1983.

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

[46] Isaac S Kohane. Using electronic health records to drive discovery in disease genomics. *Nature reviews. Genetics*, 12(6):417–428, June 2011.

[47] Isaac S Kohane, Atul J Butte, and Alvin Kho. *Microarrays for an Integrative Genomics.* MIT Press, Cambridge, MA, USA, 2002.

[48] Ravi Kothapalli, Sean J Yoder, Shrikant Mane, and Thomas P Loughran. Microarray results: how accurate are they? *BMC Bioinformatics*, 3:22, August 2002.

[49] J. Lamb. The Connectivity Map: a new tool for biomedical research. 7(1):54–60, 2007.

[50] P. Li and L. I. Zon. Resolving the controversy about N-cadherin and hematopoietic stem cells. 6(3):199–202, 2010.

[51] CE Lipscomb. Medical subject headings (MeSH). *Bulletin of the Medical Library Association*, 2000.

[52] R. Liu, X. Wang, G. Y. Chen, P. Dalerba, A. Gurney, T. Hoey, G. Sherlock, J. Lewicki, K. Shedden, and M. F. Clarke. The prognostic role of a gene signature from tumorigenic breast-cancer cells. 356(3):217–26, 2007.

[53] S Liu, W Ma, R Moore, and V Ganesan. RxNorm: prescription for electronic drug information exchange. *IT professional*, 2005.

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

[55] N. A. Lobo, Y. Shimono, D. Qian, and M. F. Clarke. The biology of cancer stem cells. 23:675–99, 2007.

[56] J. Loscalzo, I. Kohane, and A. L. Barabasi. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. 3:124, 2007.

[57] Jeff Luck, John W Peabody, Timothy R Dresselhaus, Martin Lee, and Peter Glassman. How well does chart abstraction measure quality? A prospective comparison of standardized patients with the medical record. *The American Journal of Medicine*, 108(8):642–649, January 2000.

[58] M. Lukk, M. Kapushesky, J. Nikkilä, H. Parkinson, A. Goncalves, W. Huber, E. Ukkonen, and A. Brazma. A global map of human gene expression. 28(4):322–4, 2010.

[59] Yves A Lussier and James L Chen. The emergence of genome-based drug repositioning. *Science translational medicine*, 3(96):96ps35, August 2011.

[60] Donna Maglott, Jim Ostell, Kim D Pruitt, and Tatiana Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 39(Database issue):D52–7, January 2011.

[61] MAQC Consortium, Leming Shi, Laura H Reid, Wendell D Jones, Richard Shippy, Janet A Warrington, Shawn C Baker, Patrick J Collins, Francoise de Longueville, Ernest S Kawasaki, Kathleen Y Lee, Yuling Luo, Yongming Andrew Sun, James C Willey, Robert A Setterquist, Gavin M Fischer, Weida Tong, Yvonne P Dragan, David J Dix, Felix W Frueh, Frederico M Goodsaid, Damir Herman, Roderick V Jensen, Charles D Johnson, Edward K Lobenhofer, Raj K Puri, Uwe Schrf, Jean Thierry-Mieg, Charles Wang, Mike Wilson, Paul K Wolber, Lu Zhang, Shashi Amur, Wenjun Bao, Catalin C Barbacioru, Anne Bergstrom Lucas, Vincent Bertholet, Cecilie Boysen, Bud Bromley, Donna Brown, Alan Brunner, Roger Canales, Xiaoxi Megan Cao, Thomas A Cebula, James J Chen, Jing Cheng, Tzu-Ming Chu, Eugene Chudin, John Corson, J Christopher Corton, Lisa J Croner, Christopher Davies, Timothy S

Davison, Glenda Delenstarr, Xutao Deng, David Dorris, Aron C Eklund, Xiaohui Fan, Hong Fang, Stephanie Fulmer-Smentek, James C Fuscoe, Kathryn Gallagher, Weigong Ge, Lei Guo, Xu Guo, Janet Hager, Paul K Haje, Jing Han, Tao Han, Heather C Harbottle, Stephen C Harris, Eli Hatchwell, Craig A Hauser, Susan Hester, Huixiao Hong, Patrick Hurban, Scott A Jackson, Hanlee Ji, Charles R Knight, Winston P Kuo, J Eugene LeClerc, Shawn Levy, Quan-Zhen Li, Chunmei Liu, Ying Liu, Michael J Lombardi, Yunqing Ma, Scott R Magnuson, Botoul Maqsodi, Tim McDaniel, Nan Mei, Ola Myklebost, Baitang Ning, Natalia Novoradovskaya, Michael S Orr, Terry W Osborn, Adam Papallo, Tucker A Patterson, Roger G Perkins, Elizabeth H Peters, Ron Peterson, Kenneth L Philips, P Scott Pine, Lajos Pusztai, Feng Qian, Hongzu Ren, Mitch Rosen, Barry A Rosenzweig, Raymond R Samaha, Mark Schena, Gary P Schroth, Svetlana Shchegrova, Dave D Smith, Frank Staedtler, Zhenqiang Su, Hongmei Sun, Zoltan Szallasi, Zivana Tezak, Danielle Thierry-Mieg, Karol L Thompson, Irina Tikhonova, Yaron Turpaz, Beena Vallanat, Christophe Van, Stephen J Walker, Sue Jane Wang, Yonghong Wang, Russ Wolfinger, Alex Wong, Jie Wu, Chunlin Xiao, Qian Xie, Jun Xu, Wen Yang, Liang Zhang, Sheng Zhong, Yaping Zong, and William Slikker. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature biotechnology*, 24(9):1151–1161, September 2006.

[62] James H. McClellan, Ronald W. Schafer, and M. A. Yoder. DSP first : a multimedia approach. *Digital signal processing first*, pages xx, 523 p., 1998. 97036447 James H. McClellan, Ronald W. Schafer, Mark A. Yoder. ill. ; 25 cm. + 1 computer laser optical disc (4 3/4 in.) System requirements for accompa-

nying computer disc: Internet Explorer 3.0 or higher; 8MB RAM of memory for Windows 95; 16 MB RAM of memory for Windows NT4.0. Includes index. MATLAB curriculum series.

[63] Stéphane M Meystre and Peter J Haug. Comparing natural language processing tools to extract medical problems from narrative text. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 525–529, 2005.

[64] 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. 1:148–156, 2007.

[65] Shawn Murphy, Susanne Churchill, Lynn Bry, Henry Chueh, Scott Weiss, Ross Lazarus, Qing Zeng, Anil Dubey, Vivian Gainer, Michael Mendis, John Glaser, and Isaac Kohane. Instrumenting the health care enterprise for discovery research in the genomic era. *Genome research*, 19(9):1675–1681, September 2009.

[66] 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.

[67] K. Naxerova, C. J. Bult, A. Peaston, K. Fancher, B. B. Knowles, S. Kasif, and I. S. Kohane. Analysis of gene expression in a developmental context emphasizes distinct biological leitmotifs in human cancers. 9(7):R108, 2008.

[68] L Nyhoff. C++: an introduction to data structures, 1998.

[69] 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. 34(Database issue):D629–D631, 2006.

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

[71] 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. 39(Database):D1002–D1004, 2010.

[72] N Pavlidis, E Briasoulis, John D Hainsworth, and F Anthony Greco. Diagnostic and therapeutic management of cancer of an unknown primary. 39(14):1990–2005, 2003.

[73] C Price. *SNOMED clinical terms.* British Journal of Healthcare Computing ..., 2000.

[74] Carlos Prieto, Alberto Risueño, Celia Fontanillo, and Javier De Las Rivas. Human Gene Coexpression Landscape: Confident Network Derived from Tissue Transcriptomic Profiles. 3(12):e3911, 2008.

[75] Colin C Pritchard, Li Hsu, Jeffrey Delrow, and Peter S Nelson. Project normal: defining normal variance in mouse gene expression. 98(23):13266–13271, 2001.

[76] D Rades, G Kühnel, I Wildfang, A R Börner, H J Schmoll, and W Knapp. Localised disease in cancer of unknown primary (CUP): the value of positron emission tomography (PET) for individual therapeutic management. *Annals of Oncology*, 12:1605–1609, January 2001.

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

[78] 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. 9(2):166–180, 2007.

[79] L. Ricci-Vitiani, D. G. Lombardi, E. Pilozzi, M. Biffoni, M. Todaro, C. Peschle, and R. De Maria. Identification and expansion of human colon-cancer-initiating cells. 445(7123):111–5, 2007.

[80] 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.

[81] Kenneth Salem and Hector Garcia-Molina. *Disk Striping.* IEEE Computer Society, February 1986.

[82] Eric W Sayers, Tanya Barrett, Dennis A Benson, Evan Bolton, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Scott Federhen, Michael Feolo, Ian M Fingerman, Lewis Y Geer, Wolfgang Helmberg, Yuri Kapustin, David Landsman, David J Lipman, Zhiyong Lu, Thomas L Madden, Tom Madej, Donna R Maglott, Aron Marchler-Bauer, Vadim Miller, Ilene Mizrachi, James Ostell, Anna Panchenko, Lon Phan, Kim D Pruitt, Gregory D Schuler, Edwin Sequeira, Stephen T Sherry, Martin Shumway, Karl Sirotkin, Douglas Slotta, Alexandre Souvorov, Grigory Starchenko, Tatiana A Tatusova, Lukas Wagner, Yanli Wang, W John Wilbur, Eugene Yaschenko, and Jian Ye. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 39(Database issue):D38–51, January 2011.

[83] Asher D Schachter and Isaac S Kohane. Drug target-gene signatures that predict teratogenicity are enriched for developmentally related genes. 2010.

[84] 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. 14:4376–4386, 2003.

[85] Andreas Scherer. *Batch Effects and Noise in Microarray Experiments: Sources and Solutions.* John Wiley & Sons, West Sussex, UK, January 2009.

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

[87] E Segal, N Friedman, D Koller, and A. Regev. A Module Map Showing Conditional Activity of Expression Modules in Cancer. *Nature Genetics*, 36(3):1090–1098, September 2004.

[88] Leming Shi, Gregory Campbell, Wendell D Jones, Fabien Campagne, Zhining Wen, Stephen J Walker, Zhenqiang Su, Tzu-Ming Chu, Federico M Goodsaid, Lajos Pusztai, John D Shaughnessy, André Oberthuer, Russell S Thomas, Richard S Paules, Mark Fielden, Bart Barlogie, Weijie Chen, Pan Du, Matthias Fischer, Cesare Furlanello, Brandon D Gallas, Xijin Ge, Dalila B Megherbi, W Fraser Symmans, May D Wang, John Zhang, Hans Bitter, Benedikt Brors, Pierre R Bushel, Max Bylesjo, Minjun Chen, Jie Cheng, Jing Cheng, Jeff Chou, Timothy S Davison, Mauro Delorenzi, Youping Deng, Viswanath Devanarayan, David J Dix, Joaquin Dopazo, Kevin C Dorff, Fathi Elloumi, Jianqing Fan, Shicai Fan, Xiaohui Fan, Hong Fang, Nina Gonzaludo, Kenneth R Hess, Huixiao Hong, Jun Huan, Rafael A Irizarry, Richard Judson, Dilafruz Juraeva, Samir Lababidi, Christophe G Lambert, Li Li, Yanen Li, Zhen Li, Simon M Lin, Guozhen Liu, Edward K Lobenhofer, Jun Luo, Wen Luo, Matthew N McCall, Yuri Nikolsky, Gene A Pennello, Roger G Perkins, Reena Philip, Vlad Popovici, Nathan D Price, Feng Qian, Andreas Scherer, Tieliu Shi, Weiwei Shi, Jaeyun Sung, Danielle Thierry-Mieg, Jean Thierry-Mieg, Venkata Thodima, Johan

Trygg, Lakshmi Vishnuvajjala, Sue Jane Wang, Jianping Wu, Yichao Wu, Qian Xie, Waleed A Yousef, Liang Zhang, Xuegong Zhang, Sheng Zhong, Yiming Zhou, Sheng Zhu, Dhivya Arasappan, Wenjun Bao, Anne Bergstrom Lucas, Frank Berthold, Richard J Brennan, Andreas Buness, Jennifer G Catalano, Chang Chang, Rong Chen, Yiyu Cheng, Jian Cui, Wendy Czika, Francesca Demichelis, Xutao Deng, Damir Dosymbekov, Roland Eils, Yang Feng, Jennifer Fostel, Stephanie Fulmer-Smentek, James C Fuscoe, Laurent Gatto, Weigong Ge, Darlene R Goldstein, Li Guo, Donald N Halbert, Jing Han, Stephen C Harris, Christos Hatzis, Damir Herman, Jianping Huang, Roderick V Jensen, Rui Jiang, Charles D Johnson, Giuseppe Jurman, Yvonne Kahlert, Sadik A Khuder, Matthias Kohl, Jianying Li, Li Li, Menglong Li, Quan-Zhen Li, Shao Li, Zhiguang Li, Jie Liu, Ying Liu, Zhichao Liu, Lu Meng, Manuel Madera, Francisco Martinez-Murillo, Ignacio Medina, Joseph Meehan, Kelci Miclaus, Richard A Moffitt, David Montaner, Piali Mukherjee, George J Mulligan, Padraic Neville, Tatiana Nikolskaya, Baitang Ning, Grier P Page, Joel Parker, R Mitchell Parry, Xuejun Peng, Ron L Peterson, John H Phan, Brian Quanz, Yi Ren, Samantha Riccadonna, Alan H Roter, Frank W Samuelson, Martin M Schumacher, Joseph D Shambaugh, Qiang Shi, Richard Shippy, Shengzhu Si, Aaron Smalter, Christos Sotiriou, Mat Soukup, Frank Staedtler, Guido Steiner, Todd H Stokes, Qinglan Sun, Pei-Yi Tan, Rong Tang, Zivana Tezak, Brett Thorn, Marina Tsyganova, Yaron Turpaz, Silvia C Vega, Roberto Visintainer, Juergen von Frese, Charles Wang, Eric Wang, Junwei Wang, Wei Wang, Frank Westermann, James C Willey, Matthew Woods, Shujian Wu, Nianqing Xiao, Joshua Xu, Lei Xu, Lun Yang, Xiao Zeng, Jialu Zhang, Li Zhang, Min Zhang, Chen Zhao, Raj K Puri, Uwe Scherf, Weida Tong, Russell D Wolfinger, and MAQC Consortium. The MicroArray Quality Control (MAQC)-II study of

common practices for the development and validation of microarray-based predictive models. *Nature biotechnology*, 28(8):827–838, August 2010.

[89] S. K. Singh, I. D. Clarke, M. Terasaki, V. E. Bonn, C. Hawkins, J. Squire, and P. B. Dirks. Identification of a cancer stem cell in human brain tumors. 63(18):5821–8, 2003.

[90] A Skonnard. *Skonnard: Soap: The simple object access protocol - Google Scholar*. Microsoft Internet Developer, 2000.

[91] Christos Sotiriou, Christos Sotiriou, Martine J Piccart, and Martine J Piccart. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? *Nature Reviews Cancer*, 7(7):545–553, July 2007.

[92] 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. 74:491–503, 2005.

[93] K. Stegmaier, S. M. Corsello, K. N. Ross, J. S. Wong, D. J. Deangelo, and T. R. Golub. Gefitinib induces myeloid differentiation of acute myeloid leukemia. 106(8):2841–8, 2005.

[94] Thorsten Stiewe. The p53 family in differentiation and tumorigenesis. *Nature Reviews Cancer*, 7(3):165–168, March 2007.

[95] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. 102(43):15545–50, 2005.

[96] 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. 102(43):15278–9, 2005.

[97] Lambert M Surhone, Miriam T Timpledon, and Susan F Marseken. *XML - RPC.* Betascript Publishing, July 2010.

[98] Richard S Sutton and Andrew G Barto. *Reinforcement learning.* an introduction. The MIT Press, 1998.

[99] Paul K Tan, Thomas J Downey, Edward L Spitznagel, Pin Xu, Dadin Fu, Dimiter S Dimitrov, Richard A Lempicki, Bruce M Raaka, and Margaret C Cam. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Research*, 31(19):5676–5684, October 2003.

[100] N P Tatonetti, J C Denny, S N Murphy, G H Fernald, G Krishnan, V Castro, P Yue, P S Tsau, I. Kohane, D M Roden, and R B Altman. Detecting Drug Interactions From Adverse-Event Reports: Interaction Between Paroxetine and Pravastatin Increases Blood Glucose Levels. *Clinical pharmacology and therapeutics*, May 2011.

[101] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing.

[102] Simon Urbanek. *rJava: Low-level R to Java interface.*

[103] Gauri R Varadhachary, Dmitri Talantov, Martin N Raber, Christina Meng, Kenneth R Hess, Tim Jatkoe, Renato Lenzi, David R Spigel, Yixin Wang,

F Anthony Greco, James L Abbruzzese, and John D Hainsworth. Molecular profiling of carcinoma of unknown primary and correlation with clinical evaluation. 26(27):4442–4448, 2008.

[104] J. E. Visvader and G. J. Lindeman. Cancer stem cells in solid tumours: accumulating evidence and unresolved questions. 8(10):755–68–768, 2008.

[105] Sheng-Jian Xiao, Chi Zhang, Quan Zou, and Zhi-Liang Ji. TiSGeD: a database for tissue-specific genes. *Bioinformatics*, 26(9):1273–1275, January 2010.

[106] J Yu, M A Vodyanik, K Smuga-Otto, J Antosiewicz-Bourget, J L Frane, S Tian, J Nie, G A Jonsdottir, V Ruotti, R Stewart, I I Slukvin, and J A Thomson. Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells. *Science*, 318(5858):1917–1920, December 2007.

[107] Wen Zhang, Quaid D Morris, Richard Chang, Ofer Shai, Malina A Bakowski, Nicholas Mitsakakis, Naveed Mohammad, Mark D Robinson, Ralph Zirngibl, Eszter Somogyi, Nancy Laurin, Eftekhar Eftekharpour, Eric Sat, Jorg Grigull, Qun Pan, Wen-Tao Peng, Nevan Krogan, Jack Greenblatt, Michael Fehlings, Derek van der Kooy, Jane Aubin, Benoit G Bruneau, Janet Rossant, Benjamin J Blencowe, Brendan J Frey, and Timothy R Hughes. The functional landscape of mouse gene expression. 3, 2004.

[108] Xiang HF Zhang, Qiongqing Wang, William Gerald, Clifford A Hudis, Larry Norton, Marcel Smid, John A Foekens, and Joan Massague. Latent bone metastasis in breast cancer tied to Src-dependent survival signals. 16(1):67–78, 2009.