Using Co-expression to Redefine Functional Gene Sets for Gene Set Enrichment Analysis

by

Yuliya Kodysh

S.B., Computer Science, MIT, 2006

Submitted to the Department of Electrical Engineering and Computer Science

in partial fulfillment of the requirements for the degree of

Master of Engineering in Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2007

© Massachusetts Institute of Technology 2007. All rights reserved.
Using Co-expression to Redefine Functional Gene Sets for 
Gene Set Enrichment Analysis

by

Yuliya Kodysh

Submitted to the Department of Electrical Engineering and Computer Science 
on May 29, 2007, in partial fulfillment of the 
requirements for the degree of 
Master of Engineering in Computer Science

Abstract

Manually curated gene sets related to a biological function often contain genes that are 
not tightly co-regulated transcriptionally, which obscures the evidence of coordinated 
differential expression of these gene sets in relevant experiments. To address this 
problem, we explored strategies to refine the manually curated subcollection of the 
Molecular Signatures Database (MSigDB) for use with Gene Set Enrichment Analysis 
(GSEA). We examined the manually curated gene sets in context of an atlas of gene 
expression of many normal human tissues. To refine gene sets, we clustered the 
genes in each set based on co-expression across the tissues to produce more tightly 
co-regulated children gene sets that are also likely more accurate representations 
of the biological process or processes described by the gene set. We evaluated the 
performance of the clustering algorithms by refining gene sets in the context of several 
published GSEA analyses and verifying that the children gene sets score higher with 
GSEA than do the parents. We created and annotated a new, refined version of a 
large portion of the manually curated component of MSigDB, which we hope will be 
a resource for the GSEA community.

Thesis Supervisor: Jill P. Mesirov 
Title: Director, Computational Biology and Bioinformatics, Broad Institute of MIT 
and Harvard
Acknowledgments

I would like to thank Jill Mesirov for guidance with this project. Secondly, I would like to thank Pablo Tamayo for guidance with statistics, an implementation of the NMF consensus clustering algorithm, and helpful comments on this manuscript. I would also like to thank Dan Arlow for suggestions on visualization techniques, assistance with the R language, and for helpful discussions and comments on this manuscript. Finally, I would like to thank Aravind Subramanian for help with the GSEA application.
## Contents

1 Introduction                                                                 19
   1.1 Differential gene expression analysis ........................................ 20
   1.2 Motivation for gene set analysis .............................................. 20
   1.3 GSEA and MSigDB ........................................................................ 21
   1.4 Challenges with Manually Curated Gene Sets .................................... 24
   1.5 Contributions ............................................................................. 26

2 Methods                                                                    27
   2.1 Manually curated sub-collection of MSigDB .................................... 28
   2.2 GNF normal tissue compendium ..................................................... 29
   2.3 Clustering methods ...................................................................... 29
      2.3.1 K-means clustering .................................................................. 30
      2.3.2 NMF consensus clustering using all GNF tissues ....................... 31
      2.3.3 Leading edge clustering ........................................................ 32
   2.4 Refinement method validation ....................................................... 33
   2.5 Overlap significance testing ......................................................... 34

3 Results                                                                    37
   3.1 Clustering identifies a co-regulated subset of P53_SIGNALING that is
downregulated in mutant p53 cells ..................................................... 37
   3.2 Evidence for three distinct co-regulated subsets of P53_SIGNALING .. 38
      3.2.1 Biological significance of P53_SIGNALING clusters .................. 42
   3.3 Choosing a clustering algorithm .................................................... 43
3.3.1 Choosing optimal number of clusters ........................................... 44
3.3.2 Choosing best algorithm overall ............................................. 48
3.3.3 Refining all manually curated gene sets of a certain size ........ 56

4 Discussion ................................................................................. 57
4.1 Gene expression compendium for refinement ............................... 57
4.2 Data transformation ................................................................. 58
  4.2.1 GSEA leading edge data ..................................................... 59
  4.2.2 Row-standardized GNF data .............................................. 59
4.3 Clustering algorithms ............................................................... 60
  4.3.1 K-means vs. NMF consensus ............................................. 61
  4.3.2 NMF consensus on different types of data ......................... 61
4.4 Method validation ................................................................. 62
4.5 Future directions ..................................................................... 63

A NES Enrichment Plots For Different Gene Set Size Ranges .... 65
List of Figures

1-1 A graphical description of the GSEA method, reprinted from Subramanian et al. [1] with permission from the authors. ........................................ 22

1-2 A GSEA plot of the running sum used to compute the enrichment score of the gene set P53_SIGNALING which is downregulated in cells that have the mutant p53 gene. ........................................ 25

2-1 Histogram of sizes of gene sets in the manually curated sub-collection of MSigDB. All sizes to the right of the red line, which occurs at 14, were considered appropriate for refinement and were used to test the refinement algorithms. ........................................ 28

3-1 A heatmap of the P53_SIGNALING gene set representing rows of the standardized clipped GNF expression matrix. The rows/genes are ordered according to the 5-clustering using “zero-out” NMF consensus clustering method, and the colors to the left of the heatmap mark the cluster boundaries. ........................................ 39
3-2 Pattern of NES improvement ratios in the context of the p53+/p53-
GSEA analysis for the NMF consensus clustering algorithm using shifted
row-standardized GNF data with different values of $k$, where $k$ is the
desired number of clusters. The columns represent gene sets, and the
color-coded tick marks represent the NES improvement ratios for all
child gene sets for each $k$. The colored lines track the progress of the
associated $k$-cluster algorithm, following the maximum NES ratio clus-
ter for each gene set. Columns that contain fewer than $k$ tick marks for
a $k$-cluster algorithm signify gene sets for which one or more resulting
clusters contained less than 10 genes and was discarded. The gray line
at 1 is a reference point to help distinguish successful refinements from
failures. For better visibility, the columns are ordered by increasing
maximum NES improvement ratio of any single cluster generated from
that gene set by any of the displayed algorithms. Each gene set shown
is manually curated and contains between 100 and 500 genes.

3-3 Pattern of NES improvement ratios in the context of the ALL/AML
GSEA analysis for the NMF consensus clustering algorithm using shifted
standardized GNF data with different values of $k$.

3-4 Pattern of NES improvement ratios in the context of the GSEA analysis
comparing good vs. poor clinical outcome of lung adenocarcinoma, for
the NMF consensus clustering algorithm using shifted standardized
GNF data with different values of $k$.

3-5 NES ratio plot for gene sets containing between 100 and 500 genes, in
the context of the p53+/p53- GSEA analysis. The results for all algorithms
using 5 clusters are shown.

3-6 NES ratio plot for gene sets containing between 100 and 500 genes,
in the context of the ALL/AML GSEA analysis. The results for all
algorithms with 5 clusters are shown.
3-7 NES ratio plot for gene sets containing between 100 and 500 genes, in the context of the lung adenocarcinoma good vs. poor outcome GSEA analysis. The results for all algorithms with 5 clusters are shown. 51

3-8 NES ratio plot for gene sets containing between 85 and 99 genes, in context of the p53+/- GSEA analysis. The results for each algorithm with its associated best k are shown: $k = 4$ for k-means on row-standardized GNF data and NMF consensus clustering on leading edge data, and $k = 5$ for NMF consensus clustering on shifted and “zero-out” row-standardized GNF data. 52

3-9 NES ratio plot for gene sets containing between 85 and 99 genes, in context of the leukemia ALL/AML GSEA analysis. The results for each algorithm with its associated best k are shown. 53

3-10 NES ratio plot for gene sets containing between 85 and 99 genes, in context of the GSEA analysis involving good vs. poor outcomes of lung adenocarcinoma. The results for each algorithm with its associated best k are shown. 54

A-1 In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for k-means clustering, leading edge NMF consensus clustering, and consensus clustering with 2 different types of processed GNF data. Shown for $k = 2$, where the original gene sets contained between 25 and 29 genes. 66

A-2 In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for k-means clustering, leading edge NMF consensus clustering, and consensus clustering with 2 different types of processed GNF data. Shown for $k = 2$, where the original gene sets contained between 30 and 34 genes. 67
A-3 In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for k-means clustering, shown for \( k = 2, 3, 4, 5 \). The original gene sets contained between 35 and 39 genes.

A-4 In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on leading edge data, shown for \( k = 2, 3, 4, 5 \). The original gene sets contained between 35 and 39 genes.

A-5 In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on shifted row-standardized GNF data, shown for \( k = 2, 3, 4, 5 \). The original gene sets contained between 35 and 39 genes.

A-6 In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on “zero-out” row-standardized GNF data, shown for \( k = 2, 3, 4, 5 \). The original gene sets contained between 35 and 39 genes.

A-7 In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for k-means clustering, leading edge NMF consensus clustering, and consensus clustering with 2 different types of processed GNF data. Shown for \( k = 2 \), which was the optimal cluster number for each algorithm shown. The original gene sets contained between 35 and 39 genes.

A-8 In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for k-means clustering, shown for \( k = 2, 3, 4, 5 \). The original gene sets contained between 40 and 49 genes.
In the context of GSEA analyses of p53+/−, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on leading edge data, shown for \( k = 2, 3, 4, 5 \). The original gene sets contained between 40 and 49 genes.

In the context of GSEA analyses of p53+/−, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on shifted row-standardized GNF data, shown for \( k = 2, 3, 4, 5 \). The original gene sets contained between 40 and 49 genes.

In the context of GSEA analyses of p53+/−, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on "zero-out" row-standardized GNF data, shown for \( k = 2, 3, 4, 5 \). The original gene sets contained between 40 and 49 genes.

In the context of GSEA analyses of p53+/−, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for k-means clustering, leading edge NMF consensus clustering, and consensus clustering with 2 different types of processed GNF data. NES ratios shown for \( k = 2 \), which was the optimal cluster number for each algorithm shown. The original gene sets contained between 40 and 49 genes.

In the context of GSEA analyses of p53+/−, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for k-means clustering, shown for \( k = 2, 3, 4, 5 \). The original gene sets contained between 50 and 84 genes.

In the context of GSEA analyses of p53+/−, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on leading edge data, shown for \( k = 2, 3, 4, 5 \). The original gene sets contained between 50 and 84 genes.
A-15 In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on shifted row-standardized GNF data, shown for $k = 2, 3, 4, 5$. The original gene sets contained between 50 and 84 genes.

A-16 In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on “zero-out” row-standardized GNF data, shown for $k = 2, 3, 4, 5$. The original gene sets contained between 50 and 84 genes.

A-17 In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for k-means clustering, leading edge NMF consensus clustering, and consensus clustering with 2 different types of processed GNF data. NES ratios shown for $k = 3$ in the cases of NMF consensus clustering on “zero-out” GNF data and on leading edge data, and for $k = 4$ in the cases of k-means clustering on raw data and NMF consensus clustering on shifted GNF data. These values of $k$ were the optimal cluster numbers for their associated algorithms. The original gene sets contained between 50 and 84 genes.

A-18 In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for k-means clustering, shown for $k = 2, 3, 4, 5$. The original gene sets contained between 85 and 99 genes.

A-19 In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on leading edge data, shown for $k = 2, 3, 4, 5$. The original gene sets contained between 85 and 99 genes.
In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on shifted row-standardized GNF data, shown for $k = 2, 3, 4, 5$. The original gene sets contained between 85 and 99 genes.

In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on “zero-out” row-standardized GNF data, shown for $k = 2, 3, 4, 5$. The original gene sets contained between 85 and 99 genes.

In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for k-means clustering, leading edge NMF consensus clustering, and consensus clustering with 2 different types of processed GNF data. NES ratios shown for $k = 4$ in the cases of NMF consensus clustering on “zero-out” GNF data and shifted GNF data, and for $k = 5$ in the cases of k-means clustering on raw data and NMF consensus clustering on leading edge data. These values of $k$ were the optimal cluster numbers for their associated algorithms. The original gene sets contained between 85 and 99 genes.
List of Tables

3.1 P53.SIGNALING cluster performance in context of NCI-60 p53 wild-type and mutant cells. The original gene set and most clusters are enriched in the wild-type phenotype, shown as positive NES scores. Negative values for NES indicate enrichment score in the opposite phenotype. The cluster numbers correspond to those from Fig. 3-1. ........................ 40

3.2 P53.SIGNALING cluster performance in context of ALL/AML data. The original gene set and most clusters are enriched in the ALL phenotype, shown as positive NES scores. ........................ 41

3.3 P53.SIGNALING cluster performance in context of lung adenocarcinoma data. The original gene set and most clusters are enriched in the poor outcome phenotype, shown as positive NES scores. ........................ 41

3.4 Guidelines for refining manually curated MSigDB gene sets, broken down by gene set size range. ........................ 55

3.5 Counts of the number of gene sets replaced by $k$ clusters for the new version of the manually curated subcollection, for $k$ between 1 and 5. ........................ 56
Chapter 1

Introduction

Cells regulate gene expression in response to environmental stimuli. We can monitor these gene expression changes using genome-wide transcriptional profiling. Studying these changes in expression can give us information about how cells respond to changes in their environment. However, gene expression measurements are often plagued by noise. Thus, it is often difficult to draw conclusions about the differential expression of individual genes. A modern approach to analyzing differential expression gains statistical power by analyzing coordinated differential expression of functionally related sets of genes, such as biochemical or signaling pathways. However, the gene set approach to differential expression is only advantageous when most of the genes in the set are coordinately differentially expressed, even if this differential expression is not statistically significant at the level of individual genes. In practice, many functionally related gene sets contain only a minority of genes that are coordinately differentially expressed in a context where the biological processes represented in these sets are differentially regulated, which undermines the gene set-based approach. In this thesis, we addressed this problem by creating a method for identifying distinctly co-regulated subsets of functionally related gene sets by looking at patterns of co-expression across a compendium of profiles. We applied our method to a collection of manually curated gene sets from the Molecular Signatures Database (MSigDB) [1] and showed that our refined gene sets consistently show more coherent differential expression in several cancer-related data sets than do the original gene sets. Our collection of re-
fined gene sets should serve as a resource for more accurate analysis of differential gene expression.

In this chapter, we provide background on the problem. We describe one gene set-level method of differential expression analysis, Gene Set Enrichment Analysis (GSEA), as well as the accompanying database of gene sets, MSigDB [1]. We elaborate on the limitations of the manually curated subcollection of gene sets from MSigDB, and address the idea of a refinement of that subcollection through clustering gene sets based on co-expression across a large compendium.

1.1 Differential gene expression analysis

By examining differences in transcript abundance between two populations of cells using DNA microarrays, we can learn about differences in cellular state and activity [2]. For example, differences in gene expression between cells treated with a compound and cells treated with vehicle only can shed light on the effects of the compound. In other cases, we may want to compare two different tissues to learn about biological processes that are specific to each. In general, genes that are differentially expressed are expected to be relevant to the biological processes that account for the differences between the two samples. In addition, one can often learn about the processes that are relevant to the difference between the samples if one identifies differentially expressed genes that are known to be involved in those processes.

1.2 Motivation for gene set analysis

One approach to deriving information from a gene expression analysis is simply to rank genes based on their differential expression between two tissues, and hypothesize that the genes at the top of the list are the most important ones for the process in question. However, there are several drawbacks to this approach. Firstly, gene expression data is very noisy due to variability between samples and the lack of sufficient sample replicates from which one may confidently extract a signal. For these reasons,
it may be difficult to distinguish a gene that is really associated with the process from a false positive. Secondly, it is unclear how to choose the cutoff below which genes are not considered sufficiently differentially expressed to be important. Thus, one can miss the importance of biologically significant genes that are just below the cutoff, or ones that have a low differential expression score due to measurement error. As such, inferring that a post hoc collection of individual genes comprise a biological process relevant to the differential expression may obscure important properties of the real process responsible for the expression differences. Instead, examining differential expression of sets of genes may yield information about biological pathways involved in the process.

Testing a priori defined gene sets, rather than individual genes, for significant differential expression addresses the above problems. While one is often unsure that a single gene is significantly expressed, the chance that a whole set of genes is comprised of false positives is much smaller. In addition, if a set of genes defining some pathway is moderately but coordinately differentially expressed, while none of the individual genes in the set are highly differentially expressed, one can see a pattern using a gene set analysis but might otherwise miss it with individual-gene analysis [3]. Such modest coordinate differential expression can be biologically significant. Finally, using annotated gene sets can yield insight into the biological processes involved in an experiment.

1.3 GSEA and MSigDB

We base our current work on an existing algorithm and application, GSEA [3, 1]. The goal of the GSEA algorithm is, given a ranking of genes by differential expression and a particular a priori biologically significant gene set, to determine whether genes in that set are significantly represented at the top (or bottom) of the gene ranking, or whether they are randomly distributed through the list. In this way, GSEA performs microarray analysis at the gene set level and is designed to detect gene sets containing genes that are coordinately differentially expressed. A gene set receives a high score
if it is judged to be significantly up- or down-regulated, as a set. To score a gene set, GSEA uses a weighted variant of the Kolmogorov-Smirnov running sum statistic. The running sum is calculated by walking down the ranked gene list; if a gene in the set is encountered, the sum value is increased by an amount that is weighted in accordance with the position of the gene in the ranked list; if any other gene is encountered, the value is decreased. The enrichment score (ES) of the gene set, is the value of the running sum at the point of maximum deviation from zero. The significance of the ES is described by a nominal p-value, calculated from an empirical null distribution generated by permuting either gene labels or phenotype labels. The normalized enrichment score (NES) of the gene set is then calculated to take into account the gene set size. The significance of the NES is described by a false-discovery-rate (FDR) q-value, also calculated from an empirical null distribution of NES scores [1].

Because the weighting in the running sum is greater for genes that are more differentially expressed, gene sets containing a relatively large number of genes toward the extremes of the differential expression ranking list will score highly because they will
have a strong deviation from the zero point. In this way, gene sets containing genes that are coordinately differentially expressed (either up or down) will be identified.

There exists a database of annotated gene sets, MSigDB [1], for use with GSEA. A sub-collection of MSigDB is C2, which contains 506 expert-curated gene sets as well as gene sets that are differentially expressed in various chemical and genetic perturbations. Ideally, GSEA combined with MSigDB can determine which biological processes are differentially regulated in an experiment, assuming that MSigDB contains a comprehensive collection of different gene sets describing biological processes. For this reason, the success of GSEA as combined with MSigDB depends on the quality of the MSigDB gene sets.

Because the score of a gene set in GSEA is defined as the running sum value at maximum deviation, genes in the set that fall past the point of maximum deviation in the ranked gene list contribute less to the enrichment score. Therefore, in high-scoring gene sets, it may be useful to think specifically of the core subset of genes that drives the score. That subset is defined as the gene sets “leading edge”, in the context of a particular experiment. The leading edge of a gene set in the context of a GSEA analysis examining differential expression between phenotypes A and B is a good approximation to the subset of genes that is meaningfully differentially expressed. Examining a gene set’s leading edge can therefore be informative because it may illuminate a subprocess or otherwise separate the genes that are actually important to the process in question from the other genes in the set in the context of the given experiment.

The GSEA statistic is designed to identify coherent, or coordinately expressed, sets of genes in a particular biological context. However, gene sets in C2 that are manually curated rather than created by clustering observed expression data are often noisy and not ideally coherent, when examined in the context of real expression data. Such gene sets may receive low GSEA scores in expression contexts where their associated biological processes are relevant, because they may not be coherent enough.
1.4 Challenges with Manually Curated Gene Sets

There are several ways in which a gene set may score poorly in a GSEA analysis, even though the biological process associated with the set could actually be differentially regulated in the experiment. Many biological pathways in humans are affected in non-uniform ways by different perturbations, because the human transcriptional regulatory network is very complicated. The current knowledge is not very finely grained, and many gene sets contain genes that will not generally be fully coordinately expressed.

In some gene sets, a subset of genes may be coordinately up- or down-regulated in the test phenotype while the other genes remain largely unchanged. In this case, the leading edge of the gene set (the driving force behind the enrichment score) will be fairly small relative to the size of the entire gene set. The presence of a large number of non-contributing genes may reduce the normalized score of the gene set and obscure the activity of the biological process represented by the gene set. In other cases, one subset of genes may be up-regulated while another may be down-regulated.

The existence of several co-regulated subsets of genes within a gene set that have distinct expression patterns in one or more experimental contexts may indicate that the gene set represents several biological subprocesses. In a GSEA analysis, such compound gene sets will likely score lower than their coexpressed subsets. An example is the MSigDB geneset P53_SIGNALING [3, 1], which appears to contain at least two co-regulated subgroups of genes, one of which is up-regulated, and the other is down-regulated in the GSEA analysis comparing wildtype and mutant p53 cells from the NCI-60 collection of cancer cell lines [4]. This separation is evident from the enrichment plot in Fig. 1-2, where we observe a large deviation of the running sum statistic from zero both in the + and - direction.

When P53_SIGNALING is evaluated with GSEA along with the entire C2 database in the context of the p53+/- experiment, it is identified as enriched in the p53+ phenotype if one judges by its ES, which is significant at $p < .05$. However, when taking into account issues of multiple hypothesis testing, P53_SIGNALING scores at FDR
Figure 1-2: A GSEA plot of the running sum used to compute the enrichment score of the gene set P53_SIGNALING which is downregulated in cells that have the mutant p53 gene.

$q = .962$, which means that 96% of the gene sets receiving a higher NES for the same phenotype are expected to be false positives. The significance threshold for a $q$ value in Subramanian et al. [1] is 0.25; therefore, P53_SIGNALING is not significant by these standards. It is reasonable to expect that if P53_SIGNALING were separated into several parts corresponding to the overall-induced and overall-downregulated subsets in this experiment, these subsets would result in more significant enrichment scores than the original gene set. We discuss this hypothesis in the Results chapter.

As is evident from this and other examples, it may be useful to identify co-regulated subsets of genes within a gene set. One way of doing that is to group genes within a set based on similarity in their behavior to other genes in the set across a compendium of expression data. These clusters may better approximate aspects of the biological process represented in the gene set than does the original gene set, because they are based on actual expression data.

Since the goal is to study co-regulated subsets of biologically significant gene sets,
we need to examine a large amount of differential expression data to find patterns of co-regulation of genes within these sets. We attempt a first approximation at this generalization using a normal human tissue expression compendium compiled by the Genomics Institute of the Novartis Research Foundation (GNF) [5].

One way to study differential expression across a large amount of expression data involves GSEA. Across a large set of GSEA analyses comparing a pair of samples, the patterns of differential expression of genes in a given gene set may become apparent. For instance, we may see that some genes in a set repeatedly contribute to the enrichment score, while others repeatedly do not; or we may see a gene set split similar to that shown in Fig. 1-2, corresponding to the behavior of P53_SIGNALING in context of the p53+/p53- experiment. In addition to examining gene sets in the context of the GSEA leading edge statistic, it is also informative to look at differential expression data under other transformations, discussed in the Methods chapter.

1.5 Contributions

There are many manually curated gene sets within the C2 subcollection of MSigDB that suffer from the problems described above, either containing some genes that are usually not differentially expressed at all or containing several fairly distinct subsets of genes that are differentially expressed in different situations or in different directions. Based on this knowledge, we refined some of the gene sets in the expert-curated section of C2 by examining their expression across a compendium of normal tissues. We tested several clustering techniques and chose appropriate clustering methods for various gene sets to refine these sets. We created a new version of the manually curated subcollection of C2, where each of the original gene sets within a certain size range is replaced by one or more disjoint subsets of the original.
Chapter 2

Methods

Since most manually curated gene sets are not coherent in context of actual expression data, we explore the clustering of genes by co-expression as a way to extract more coherent subsets of existing gene sets. These refined subsets will likely score higher in GSEA in contexts where the biological processes represented in these subsets are differentially regulated. There are many clustering algorithms one could apply in this problem, where some may be better suited to handle the biological data that we are working with.

In this chapter, we describe the methodology we used to refine the MSigDB manually curated gene sets. We briefly describe the gene sets that we attempt to refine in this work. We discuss the choice of a gene expression compendium to use for refinement of gene sets. We then detail the clustering algorithms that we implemented to partition genes based on co-regulation in the context of our refinement expression compendium, and we describe the procedure we used to choose the clustering algorithm(s) that would be applied to all manually curated MSigDB gene sets within a certain size range. We conclude with a description of the method we used to annotate the resulting refined gene sets.
Histogram of sizes of manually curated gene sets in MSigDB

Figure 2-1: Histogram of sizes of gene sets in the manually curated sub-collection of MSigDB. All sizes to the right of the red line, which occurs at 14, were considered appropriate for refinement and were used to test the refinement algorithms.

2.1 Manually curated sub-collection of MSigDB

There are 506 expert-curated gene sets in MSigDB. Since gene sets containing less than 10 genes rarely score highly in GSEA, we only considered gene sets containing at least 15 genes. The largest manually curated gene set is MITOCHONDRIA [3, 1], containing 487 genes. We have therefore tested our methodology on manually curated gene sets containing between 15 and 487 genes, where the average gene set size is approximately 43 genes. There are 325 such gene sets. The distribution of gene set sizes in the manually curated component of MSigDB is shown in Fig. 2-1.
2.2 GNF normal tissue compendium

We chose the human normal tissue compendium as the overall context for examination of the behavior of MSigDB gene sets. Taking the refinement of the OXPHOS gene set as described in Mootha et al. [3] as an example, we know that many gene sets have characteristic co-regulation patterns across tissues. We note that our methodology is best applied in the presence of very large amounts of expression data. Using more expression data increases the likelihood that our conclusions about gene set behavior and coherence is based on a representative sample of biological processes and that we will be able to identify all truly co-regulated subsets of genes within each set.

The GNF expression compendium contains expression profiles of 79 human tissues on Affymetrix HG-U133A and (custom) GNF1H chip. The data from both chips were summarized using the MAS 5.0 algorithm. We did not use data from the GNF1H chip, because we were unable to find a suitable annotation for the mapping between probe sets and gene names.

Using the HG-U133A chip annotation, we created an expression matrix of the GNF normal tissues, where the rows represented genes, and the columns represented tissue samples. In each case, there were two replicates per tissue, or two columns representing the same tissue from two different individuals. The expression matrix represented 13,154 genes and 158 samples (two per tissue). In cases where multiple probe sets mapped to the same gene, one was chosen essentially at random.

2.3 Clustering methods

We implemented several methods to cluster genes within a set, anticipating that some perform better than others at extracting coherent clusters from the manually curated sets. We later applied these algorithms to all manually curated gene sets in C2 and evaluated them for the quality of refinement of these gene sets. In this section, we describe in detail the application of these algorithms to single gene sets.
2.3.1 K-means clustering

K-means [6] is one of the most popular clustering algorithms. It is also quite easy to implement, and was a natural choice for an algorithm to evaluate as part of this analysis.

Data pre-processing

In order for the GNF data to be usable for our purposes, we had to convert it to a form of differential expression. We standardized the rows of the GNF expression matrix, so that the mean value of each row was 0 and the standard deviation was 1. We thus created a standardized GNF expression matrix, $S$, where $S_{i,j}$ represented the relative expression of gene $i$ in tissue sample $j$. A value $S_{i,j}$ was positive if gene $i$ was relatively upregulated in tissue sample $j$, and negative if it was downregulated. This centering around 0 allowed us to make comparisons between GNF tissues for each gene. To remove extreme values that could dominate differential expression in any gene set or clustering analysis, values in $S$ greater than 3 were set to 3, and values less than -3 were set to -3.

Algorithm description

To refine a gene set $G$ containing $n$ genes, we extracted those rows from $S$ that represented genes in the given set, giving us the sub-matrix $S_G$ of dimension $(n \times 158)$. We then used the R statistical software package [7] to run k-means clustering on the rows of $S_G$. The non-deterministic k-means algorithm partitioned the $n$ genes into $k$ clusters, for our desired value of $k$, based on $k$ centroids that were computed from the data starting with a random initialization. Because we were comparing the success of k-means using different numbers of clusters, we performed this procedure for $k = 2$ through 5. In each case, we discarded any cluster containing less than 10 genes, since such small gene sets generally do not score highly in GSEA.
2.3.2 NMF consensus clustering using all GNF tissues

The use of the non-negative matrix factorization (NMF) consensus clustering method was inspired by research showing it to be effective for extracting meaningful biological patterns from microarray data [8]. However, the use of this method for current research is significantly different due to the type of input data. NMF was described in its modern form by Lee and Seung [9] as a parts-based method of decomposing a matrix into its additive components, and has since been used in a variety of applications. The parts-based additive NMF algorithm allows the decomposition results to be interpreted more easily than non-additive algorithms such as PCA.

Data pre-processing

As in the k-means clustering method, we created $S_G$ from $S$ based on the genes in the given gene set and used the NMF consensus algorithm to cluster the rows of $S_G$. However, we had to process our expression data further in order to run NMF consensus clustering on it, because the algorithm requires non-negative inputs.

Another property of the NMF consensus clustering algorithm is that high values in the input matrix tend to contribute more to the clustering result than do values close to 0. Since the pre-processed data in $S$ has high values for upregulated genes and low values for downregulated genes, any clustering done on some non-negative version of this data (perhaps simply shifted up) would be based strictly on upregulated genes while downregulated genes would be ignored. We wanted up- and downregulated genes to be represented equally in $S_G$. For each column $j$, we created a negative version $-j$ which represented a reverse view of differential expression for that phenotype, and appended this negative column to $S_G$. $S_G$ was now of dimension $(n \times 316)$, where each value was in the range $(-3, 3)$.

The further processing of data was performed in order to meet the non-negativity constraint. This was done in two different ways, which we treat as two separate clustering techniques:

- Add 3 to each value in $S_G$ so that all values are in the range $(0, 6)$. 

31
• For all values of $S_G$ that are negative, set them to 0 ("zero-out").

These two types of data pre-processing result in different outputs of the NMF consensus clustering algorithm, and we therefore treat them as two separate algorithms in our analysis. Each approach has certain advantages and disadvantages, addressed in the Discussion chapter.

**Algorithm description**

NMF takes as input a matrix $S_G$ and returns a factorization of $S_G$ into two matrices $W$ and $H$, where $W$ contains a description of $k$ "meta-samples", or centroids, as described in [8]. Each gene is then assigned to a cluster by choosing the centroid with the highest value for that gene, which corresponds to assigning it to the meta-sample for which it has the highest differential expression level. After this algorithm is run 20 times, consensus clustering is applied to the outputs in order to generate the final partition of $S_G$ [10]. As in the case of k-means clustering, we discarded any cluster containing less than 10 genes.

### 2.3.3 Leading edge clustering

Another way to extract differential expression information from GNF was through the use of GSEA to compare samples and generate leading edges for those gene sets that were significantly enriched for either of the two phenotypes. From the 79 human tissues in GNF, we generated 79 one-vs-all GSEA comparison analyses where for each experiment, one of the tissues represented one phenotype and all other tissues were pooled into the other phenotype. Such a separation of tissues into phenotypes allowed us to obtain analyses of gene sets that were up-regulated in the "one" tissue.

**Data pre-processing**

For each gene set $G$, we created a leading-edge matrix $B_G$ as follows:

1. Evaluate the significance of the enrichment of $G$ in either of the two tissue phenotypes in each of the 79 one-vs-all GSEA analyses.
2. For each experiment in which \( G \) is significantly enriched \((p < 0.05, \text{nominal})\), note its leading edge \( G_L \), the subset of genes in \( G \) that contributed to its enrichment score in the given experiment.

3. Using \( G_L \), construct binary vector \( V \) of length \( k = |G| \) where \( V[i] = 1 \) if gene \( i \) is in \( G_L \), and 0 otherwise.

4. Append \( V \) as a column to \( B_G \).

5. Remove rows containing only 0s, since these represent genes that never show up in the leading edge and therefore never contribute positively to \( G \)'s GSEA score.

\( B_G \) is therefore a binary matrix of dimension \((k \times s)\), where \( s \) is the number of one-vs-all GSEA analyses where \( G \) was significantly enriched \((p < .05, \text{nominal})\). Thus, \(|s|\) is restricted to be between 1 and 79. By this scheme, if \( G \) is not significantly enriched in either phenotype in any of the 79 comparison experiments, we do not refine \( G \). However, for those gene sets \( G \) that are identified at least several times in the GNF one-vs-all experiments, \( B_G \) reflects strong differential expression of genes because only strongly differentially expressed genes tend to contribute to the leading edge.

Algorithm description

NMF consensus clustering takes as input the binary matrix \( B_G \) and clusters the rows, as described in section 2.3.2. As in the case of k-means clustering, we discarded any cluster containing less than 10 genes.

2.4 Refinement method validation

Choosing a clustering algorithm for microarray data is difficult: there exist many algorithms but few guidelines for choosing among them. Therefore, to derive any sort of pattern in the successes of different algorithms, we chose to perform "systematic
spotchecks” using several published GSEA analyses [1], each one comparing samples represented on the Affymetrix HG-U95 microarray chip. The general idea was to look at a gene set in context of one of the experiments, refine it independently (using the GNF compendium), and then use GSEA to test it in the experiment from which it came to see if it improved.

We tested different refinement algorithms as follows. For each gene set, we created many possible clusterings using these algorithms (each clustering consisting of several children gene sets) and examined the overall improvement patterns for each algorithm. Ultimately, we wanted to test our refinement of gene sets that were either already enriched in the given GSEA experiment \( (p < .05, \text{ nominal}) \), or had the potential to be enriched (at least one of the child gene sets was significant at nominal \( p < .05 \)). We chose our figure of merit to be the NES ratio of the child gene set and the original gene set.

We used a visualization technique to aid us in choosing an algorithm. To study the performance of a clustering algorithm producing \( k \) clusters, we plotted the NES ratios produced by each of the \( k \) clusters for every gene set, and tracked the best NES ratio per gene set for the algorithm of interest. As long as one of the \( k \) resulting clusters had an NES ratio greater than 1 (the cluster had a higher NES than the parent), we considered that gene set to be successfully refined, even if the other clusters had NES ratios lower than 1. We used these plots to get an approximation of the improvement trends for different algorithms.

The analysis of the next chapter further describes the process of arriving at a final methodology for refining gene sets.

2.5 Overlap significance testing

Once the manually curated gene sets from MSigDB were refined and split into clusters, each of the resulting clusters was annotated with the name of the parent gene set as well as a list of original gene sets from the entire C2 collection that were the most similar to the cluster. Similarity of two gene sets was measured as the overlap between
them.

The significance of the overlap between two gene sets was computed using the hypergeometric cumulative distribution, where the total population size equaled 13,154, the number of genes represented on the HG-U133A microarray chip. A flat Bonferroni correction was applied to the p-values to account for multiple-hypothesis testing.
Chapter 3

Results

In this chapter, we describe the preliminary analysis of a refinement methodology tested in a particular gene set and use that example to argue that all clusters resulting from the refinement split of a gene set should be considered valid co-regulated children gene sets. We then describe the choice of clustering algorithms appropriate for refining gene sets of different sizes in the final methodology, and detail the outcomes of the procedure we used to make a decision between clustering algorithms.

3.1 Clustering identifies a co-regulated subset of P53_SIGNALING that is downregulated in mutant p53 cells

P53_SIGNALING, shown in context of an experiment in Fig. 1-2, is a manually curated gene set obtained from the initial collection of gene sets described in Mootha et al. [3]. This gene set is comprised of 101 genes that are involved in the p53 transcription factor signaling pathway. An experiment in which one would expect such a gene set to be differentially expressed is the comparison of cells with a normal p53 gene and a mutant p53 gene. Therefore, we examined the expression of the P53_SIGNALING gene set in the wild-type and mutant p53 cells from the NCI-60 collection of cancer cell lines [4]. In this experiment, P53_SIGNALING was significantly enriched in cells
with the wild-type p53 gene ($p = .033$, nominal), with the normalized enrichment score 1.36. This score served as a confirmation that while overall the gene set is not coherent, P53_SIGNALING contains genes that are at least somewhat correlated in the p53 wild-type phenotype, and coordinately down-regulated in the mutant cells.

Fig. 1-2 suggests that there may be several more strongly co-regulated subsets of P53_SIGNALING. Because the gene set contains genes that are both up- and down-regulated in the wild-type p53 phenotype, we hypothesized that a coordinately regulated subset of the gene might score better in the same GSEA analysis. In a preliminary analysis, in order to isolate that subset, we performed a cluster analysis using the NMF consensus clustering algorithm with 5 clusters on GNF data that had undergone the “zero-out” transformation as described in Methods. Before feeding the resulting clusters as input to the p53+/- GSEA analysis, we wanted to make sure that the clustering algorithm produced clusters that were coherent within the GNF data. A visualization of these clusters is displayed in Fig. 3-1.

Based on Fig. 3-1, most of the clusters were coherent. We then evaluated the enrichment of each cluster as a separate gene set in the context of the p53+/p53- experiment, using GSEA. One of the resulting clusters scored higher than its parent gene set, getting an NES of 1.73 and a nominal $p$-value of .015. This small though important improvement suggested that the cluster was more coordinately down-regulated than the entire P53_SIGNALING set.

### 3.2 Evidence for three distinct co-regulated subsets of P53_SIGNALING

Since our method produced one cluster from the P53_SIGNALING gene set that was more coordinately differentially expressed in the P53 experiment than the original gene set, we sought to determine whether the other clusters also contained genes that were co-regulated and functionally related in a way that made the clusters significantly enriched in some other related comparison. We examined two other expression
Figure 3-1: A heatmap of the P53 SIGNALING gene set representing rows of the standardized clipped GNF expression matrix. The rows/genes are ordered according to the 5-clustering using "zero-out" NMF consensus clustering method, and the colors to the left of the heatmap mark the cluster boundaries.
datasets that were used in validating the current GSEA method: acute lymphoid leukemia (ALL) vs. acute myeloid leukemia (AML), and data from a lung adenocarcinoma study relating to “good” or “poor” clinical outcome [1]. Since the p53 pathway is activated by DNA damage in ALL cells [11], we would expect some of the signal to show through the P53,SIGNALING gene set. In addition, Subramanian et al. [1] discovered that gene sets related to hypoxia are enriched in the poor outcome samples from lung adenocarcinoma data, which could be explained by the overcrowding of cells in tumor environments [12]. Since the p53 transcription factor is regulated by hypoxia [13], it seemed reasonable that if P53,SIGNALING contained a hypoxia-related subset, the subset could cause the whole gene set to score well in context of the lung adenocarcinoma data.

Using the increase in NES and the nominal p-value to measure success, we found that two different clusters performed well in the leukemia and lung adenocarcinoma experiments, and that neither of these was the cluster scoring highly in the p53 experiment. The results are summarized in Tables 3.1, 3.2 and 3.3. In each case, the bold font cluster is the best performing cluster.

Table 3.1: P53,SIGNALING cluster performance in context of NCI-60 p53 wild-type and mutant cells. The original gene set and most clusters are enriched in the wild-type phenotype, shown as positive NES scores. Negative values for NES indicate enrichment score in the opposite phenotype. The cluster numbers correspond to those from Fig. 3-1.

<table>
<thead>
<tr>
<th></th>
<th>Nominal p-value</th>
<th>NES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>0.033</td>
<td>1.36</td>
</tr>
<tr>
<td>Cluster 1</td>
<td>0.493</td>
<td>0.96</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>0.140</td>
<td>1.36</td>
</tr>
<tr>
<td>Cluster 3</td>
<td><strong>0.015</strong></td>
<td><strong>1.73</strong></td>
</tr>
<tr>
<td>Cluster 4</td>
<td>0.780</td>
<td>0.79</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>0.290</td>
<td>-1.22</td>
</tr>
</tbody>
</table>

Out of 5 clusters, a different one had the highest NES and the most significant nominal p-value in each of the 3 contexts. The emergence of the different clusters as significant in these different experimental contexts suggests that the several co-
Table 3.2: P53.SIGNALING cluster performance in context of ALL/AML data. The original gene set and most clusters are enriched in the ALL phenotype, shown as positive NES scores.

<table>
<thead>
<tr>
<th></th>
<th>Nominal p-value</th>
<th>NES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>0.143</td>
<td>1.26</td>
</tr>
<tr>
<td>Cluster 1</td>
<td>0.193</td>
<td>-1.28</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>0.340</td>
<td>1.10</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>0.351</td>
<td>-1.10</td>
</tr>
<tr>
<td><strong>Cluster 4</strong></td>
<td><strong>0.037</strong></td>
<td><strong>1.60</strong></td>
</tr>
<tr>
<td>Cluster 5</td>
<td>0.161</td>
<td>1.37</td>
</tr>
</tbody>
</table>

Table 3.3: P53.SIGNALING cluster performance in context of lung adenocarcinoma data. The original gene set and most clusters are enriched in the poor outcome phenotype, shown as positive NES scores.

<table>
<thead>
<tr>
<th></th>
<th>Nominal p-value</th>
<th>NES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>0.131</td>
<td>1.26</td>
</tr>
<tr>
<td><strong>Cluster 1</strong></td>
<td><strong>0.006</strong></td>
<td><strong>1.81</strong></td>
</tr>
<tr>
<td>Cluster 2</td>
<td>0.931</td>
<td>-0.61</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>0.547</td>
<td>-0.97</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>0.443</td>
<td>1.02</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>0.813</td>
<td>0.72</td>
</tr>
</tbody>
</table>
regulated subsets of a biologically significant gene set may themselves be biologically significant, perhaps in contexts as yet unseen.

We note that significant enrichment of a gene set using the GSEA statistic is usually quantified by the relationship of the gene set’s score to the scores of other gene sets, in order to correct for multiple hypothesis testing. In our analysis, we did not compare the performance of the clusters generated from P53_SIGNALING to that of all other C2 gene sets, but we did obtain a better nominal score for the clusters in context of the above 3 experiments than the original gene set had received. The clusters are therefore more coherent versions of P53_SIGNALING that each maintain a different part of the original gene set’s biological significance.

It is also interesting to note that the noisiest of the resulting clusters, Cluster 3, had the smallest relative improvement, while the other two high-scoring clusters presented more significant improvements from the original and were much more coherent in GNF. These observations further support the main idea driving the work described in this thesis: that gene sets that are strongly coherent across many conditions will score higher in a gene set-level statistical analysis than gene sets that are less coherent.

3.2.1 Biological significance of P53_SIGNALING clusters

The fact that the clusters had significant scores when the original P53_SIGNALING gene set did not, in 2 of 3 experiments, implied that these clusters may represent subprocesses of the p53 signaling pathway. Based on our ideas about which processes are relevant to these experiments, we decided to see if the clusters were similar to any gene sets already in our database that contained the genes that participate in these processes. We used the MSigDB annotation tool [1] to find the top 5 most significant overlaps between the P53_SIGNALING clusters and all other gene sets in the manually curated component of MSigDB [1]. Cluster 1, enriched in the poor outcome in lung adenocarcinoma data, overlaps highly significantly with the MSigDB gene set P53HYPOXIAPATHWAY [1] \((p < 6.12 \times 10^{-6}, \text{adjusted})\). Cluster 4, enriched in ALL cells, overlaps highly significantly with another MSigDB set, DNA_DAMAGE_SIGNALING [1] \((p < 2.97 \times 10^{-8}, \text{adjusted})\). Cluster 3, enriched in
the wild-type p53 cells in the NCI-60 cancer cell line data, overlaps significantly with many gene sets describing apoptosis. All of these results are biologically consistent, which suggests that these clusters contain genes that are functionally related. Since we showed that more than one of the P53_SIGNALING clusters is functionally significant, and since we had no a priori way of recognizing such clusters, we decided to keep all clusters generated from a gene set using some clustering algorithm that we would settle on as the result of further analysis.

### 3.3 Choosing a clustering algorithm

Based only on the P53_SIGNALING clustering, it is difficult to discern whether the success of the NMF consensus clustering algorithm on “zero-out” data is typical or unusual. Therefore, we tried several clustering approaches on many manually curated gene sets in order to choose a final refinement algorithm. We tested NMF consensus clustering on leading edge data and processed GNF expression data, as well as k-means clustering on GNF data. In our choice of clustering algorithms to attempt, we rejected hierarchical clustering, used to refine OXPHOS successfully in the original GSEA research [3], because specifying a cluster number for that algorithm leads to unstable clustering results, dependent on the precise location of the cut at a certain branch.

We studied the improvement patterns of several clustering algorithms using the visualization technique described in the Methods, where we plotted the NES improvement ratio of all clusters for each gene set. We considered a gene set to be successfully refined by some algorithm if at least one cluster generated by that algorithm had a higher NES than did the original gene set, using the reasoning from the P53_SIGNALING clusters example. We tracked the success of each algorithm across the gene sets in order to choose the algorithm producing the largest and most consistent refinements. We chose the GSEA analyses examining p53 wild-type/mutant, ALL/AML and lung adenocarcinoma good/poor outcomes as the sample contexts in which to test the refinement of gene sets. We assumed that the success of the re-
finement algorithms in these three contexts is representative of the overall refinement success. From seeing representative patterns of refinement success of the algorithms in these contexts, we became convinced that these methods could be applied to the entire manually curated component of MSigDB.

3.3.1 Choosing optimal number of clusters

For the clustering algorithms that we are considering, the number of clusters $k$ must be specified. Therefore, we also explored ways to choose the optimal number of clusters. We hypothesized that co-regulated subsets of genes within a gene set were within a particular size range, and consequently that the optimal number of clusters for a clustering algorithm would scale with gene set size. We tested this idea by examining the relative success rates of NMF consensus clustering with shifted data in relatively large gene sets, using different values of $k$, in the context of the three GSEA analyses. Since we were interested in the success of the refinement of gene sets which we believed to be biologically relevant in each experiment, we only looked at sets of genes that were either already enriched in either of the two phenotypes or resulted in at least one cluster from at least one algorithm that was enriched in either of the two phenotypes (significant at nominal $p < .05$). For each such gene set, we plotted the NES ratio of its clusters generated by the algorithms of interest. The results are shown in Figs. 3-2, 3-3, and 3-4.

Overall, we observed the same pattern from these three experiments: for large gene sets, the larger the choice of cluster number, the better the performance of the algorithm. The fact that increasing $k$ leads to incrementally better performance is especially evident in Fig. 3-4. This suggests that large gene sets often represent complex biological processes which may be broken down into more and more distinctly-regulated subparts, yielding more and more primitive sub-processes.

In Figs. 3-3 and 3-4, we see that NMF consensus clustering with 5 clusters consistently outperforms the same algorithm with smaller $k$. A similar pattern is observed in Fig. 3-2, though not as unambiguously. Considering the evidence from all three experiments, we conclude that the optimal $k$ for the NMF consensus clustering al-
Figure 3-2: Pattern of NES improvement ratios in the context of the p53+/p53- GSEA analysis for the NMF consensus clustering algorithm using shifted row-standardized GNF data with different values of $k$, where $k$ is the desired number of clusters. The columns represent gene sets, and the color-coded tick marks represent the NES improvement ratios for all child gene sets for each $k$. The colored lines track the progress of the associated $k$-cluster algorithm, following the maximum NES ratio cluster for each gene set. Columns that contain fewer than $k$ tick marks for a $k$-cluster algorithm signify gene sets for which one or more resulting clusters contained less than 10 genes and was discarded. The gray line at 1 is a reference point to help distinguish successful refinements from failures. For better visibility, the columns are ordered by increasing maximum NES improvement ratio of any single cluster generated from that gene set by any of the displayed algorithms. Each gene set shown is manually curated and contains between 100 and 500 genes.
Figure 3-3: Pattern of NES improvement ratios in the context of the ALL/AML GSEA analysis for the NMF consensus clustering algorithm using shifted standardized GNF data with different values of $k$. 
Figure 3-4: Pattern of NES improvement ratios in the context of the GSEA analysis comparing good vs. poor clinical outcome of lung adenocarcinoma, for the NMF consensus clustering algorithm using shifted standardized GNF data with different values of $k$. 

Lung cancer, good/poor outcome: sizes 100–500
algorithm is 5 in the case of large gene sets, containing between 100 and 500 genes. These findings support the hypothesis that there is some notion of a good subset size for the end result of a clustering, which relates to the size of the original gene set, at least for use with GSEA.

3.3.2 Choosing best algorithm overall

For large gene sets, K-means, NMF consensus clustering on the “zero-out” standardized GNF expression data and NMF consensus clustering on the leading-edge matrices followed patterns similar to that of NMF consensus clustering on the shifted data: the best cluster number size appeared to be 5 for each algorithm. In this fashion, we chose the best \( k \) for each kind of algorithm. We then compared these best performers against each other in order to decide on the best overall algorithm for large gene sets. Figs. 3-5, 3-6 and 3-7 display the NES ratio plots for the clustering algorithms we tested.

It is difficult to choose a clear winner among the clustering algorithms for refining large gene sets, as seen in the plots in Figs. 3-5, 3-6 and 3-7. We reasoned that the algorithm we ultimately choose should refine gene sets as consistently as possible and should produce the fewest clustering results in which no child gene set performed at least as well as the parent. In the lung adenocarcinoma and leukemia data sets, seen in Figs. 3-6 and 3-7, NMF consensus clustering on shifted row-standardized GNF data almost always produces clusters in which a child has a better NES than the parent gene set. In these data sets, each of the other algorithms tested has worse performance. In the P53 +/- data set, it is very difficult to see which algorithm has the best performance, though each one produces refinements that can be considered successful. We then made a subjective judgment based on these observations that for large gene sets, i.e. those containing between 100 and 500 genes, we would use NMF consensus clustering on shifted row-standardized GNF data with \( k = 5 \) for refinement.

We repeated this procedure for slightly smaller gene sets, containing between 85 and 99 genes, inclusive. We first chose the best \( k \) for each kind of algorithm and found
Figure 3-5: NES ratio plot for gene sets containing between 100 and 500 genes, in the context of the p53+/- GSEA analysis. The results for all algorithms using 5 clusters are shown.
Figure 3-6: NES ratio plot for gene sets containing between 100 and 500 genes, in the context of the ALL/AML GSEA analysis. The results for all algorithms with 5 clusters are shown.
Figure 3-7: NES ratio plot for gene sets containing between 100 and 500 genes, in the context of the lung adenocarcinoma good vs. poor outcome GSEA analysis. The results for all algorithms with 5 clusters are shown.
that k-means and NMF consensus clustering on leading edge data perform best with $k = 5$, while NMF consensus clustering on both kinds of row-standardized GNF data performs best with $k = 4$. We compared the 4 algorithms against each other, using the best $k$ for each. The NES ratio plot results for these comparisons are shown in Figs. 3-8, 3-9 and 3-10.

As in the case of the largest gene sets, NMF consensus clustering on shifted row-standardized GNF data appears to have the best performance of the four algorithms that we compared. This is especially evident in Figs. 3-8 and 3-9. In Fig. 3-10, there is no clear winner. Focusing on the two contexts in which a winning algorithm was evident, we decided to use NMF consensus clustering on shifted data using 4 clusters to refine gene sets containing between 85 and 99 genes.
Figure 3-9: NES ratio plot for gene sets containing between 85 and 99 genes, in context of the leukemia ALL/AML GSEA analysis. The results for each algorithm with its associated best $k$ are shown.
Figure 3-10: NES ratio plot for gene sets containing between 85 and 99 genes, in context of the GSEA analysis involving good vs. poor outcomes of lung adenocarcinoma. The results for each algorithm with its associated best $k$ are shown.
We used similar reasoning for other size ranges in order to choose the best refinement algorithm: we first chose an optimal $k$ for each algorithm based on its NES ratio plot comparing several values of $k$, and then we compared the performance of the algorithms to each other using their associated best $k$. In many cases, there was no unambiguous winning algorithm. At the same time, most of the algorithms we considered for the different size ranges produced good refinements in the majority of cases.

We were also concerned with possible overfitting of the data. In most cases, we dealt with 5-10 gene sets per size range, in each of the three GSEA analyses used as contexts for refinement method validation, thus having limited data to work with. Choosing the best performing algorithm for each gene set size range independently would produce an overall set of refinement algorithms that was too varied, which seemed to be a consequence of overfitting our choice of algorithm. We were concerned that the conclusions we drew about which clustering algorithm and $k$ are best suited for refinement did not generalize well because they were only based on 3 expression data sets. To address these concerns, we assumed that there were one or two algorithms that produced good refinement results, and strived to choose a set of clustering algorithms that performed consistently well but were not strictly the best in their associated gene set size ranges. We developed a heuristic strategy for refining MSigDB manually curated gene sets, specified in Table 3.4.

Table 3.4: Guidelines for refining manually curated MSigDB gene sets, broken down by gene set size range.

<table>
<thead>
<tr>
<th>Gene set size range</th>
<th>25-49</th>
<th>50-84</th>
<th>85-99</th>
<th>100+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td>NMF, “zero-out”</td>
<td>NMF, “zero-out”</td>
<td>NMF, shifted</td>
<td>NMF, shifted</td>
</tr>
<tr>
<td>$k$</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Refining gene sets containing fewer than 25 genes proved largely unsuccessful, and we chose to omit such small gene sets from the refinement procedures.
3.3.3 Refining all manually curated gene sets of a certain size

On the basis of the refinement guidelines explained in Section 3.3.2, we applied these algorithms to every gene set in the manually curated component of MSigDB, for every gene set containing at least 25 genes. There were 175 such gene sets. We then created a new version of this subcollection which contained all clusters generated by the appropriate algorithms for every gene set in the allowed size range, where each resulting cluster contained at least 10 genes. We annotated each new gene set by listing its top overlaps with current C2 manually curated gene sets.

From 175 original gene sets, 436 new gene sets were created. Each of the original gene sets was replaced by at least one new set. Most gene sets were replaced by 2 clusters. The number of gene sets replaced by \( k \) clusters, for \( k \) between 1 and 5, is summarized in Table 3.5.

Table 3.5: Counts of the number of gene sets replaced by \( k \) clusters for the new version of the manually curated subcollection, for \( k \) between 1 and 5.

<table>
<thead>
<tr>
<th>( k )</th>
<th>Number of gene sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
</tr>
</tbody>
</table>

For the 175 gene sets that were refined, the average gene set size prior to refinement is approximately 64 genes, and the median size is 51 genes. The average size of the resulting new gene sets is approximately 23 genes, and the median size is 18 genes.
Chapter 4

Discussion

In this work, we examined manually curated gene sets in the context of a human normal tissue expression compendium compiled by the Genomics Institute of the Novartis Research Foundation (GNF). We implemented and tested several clustering algorithms for refining sets of genes by splitting them into disjoint co-regulated subsets. On the basis of these tests, we chose a set of clustering algorithms that we applied to gene sets from the manually curated sub-collection of the Molecular Signatures Database (MSigDB). We annotated each of the new gene sets resulting from the refinement by noting the parent gene set and a list of the most similar gene sets from the original manually curated collection in MSigDB. We then compiled the new gene sets into an improved version of that subcollection of gene sets.

4.1 Gene expression compendium for refinement

We used the GNF human normal tissue expression atlas for creating a context in which to examine the coherence of expert-curated gene sets. The choice of compendium was appropriate because functionally related genes may often be co-expressed across tissues. The C2 collection in MSigDB contains manually curated gene sets that are associated with particular functions, but we hypothesize that most of these gene sets are actually composed of distinct co-regulated subsets. Subsets within a single functionally related gene set will be coherent, but may also have divergent behavior.
across the tissues. Moreover, we trust that coherence patterns observed across such
a wide and varied expression compendium represent a fairly unbiased view of general
coop-regulation of genes within functionally related sets.

However, the GNF human tissue atlas is still only a subset of possible expression
data that would be useful for our goals. Many of the manually curated gene sets
may be associated with disease states. For such gene sets, even though the patterns
of co-regulated subsets may be deduced from observing the coherence of the sets in
normal tissue, these patterns may be more apparent in the context of expression
profiles of cells from disease samples such as tumors or cancer cell lines. If we had
more expression data on which to base our refinements, we could be more confident
of finding the co-regulation patterns in the MSigDB gene sets.

The DNA microarray chips that we used in our refinement procedure do not cover
all the known genes. Therefore, some genes will not be represented in the clustering
results. For that reason, using multiple expression compendia from multiple kinds of
microarrays can be beneficial.

4.2 Data transformation

In search of the right approach to refining the MSigDB gene sets, we explored the use
of expression data from GNF in several forms. Our first attempt was inspired by the
suggestion in Subramanian et al. [1] to learn about the core subset of genes within a
set by examining leading edge overlaps across several GSEA analyses. In our work,
we used GSEA to analyze comparisons between the tissues in GNF and extracted
leading edge information from these analyses as one basis for gene set refinement. In
subsequent attempts, we used row-standardized data from the GNF compendium. We
have found that working with row-standardized data often yields better refinement
results. However, both approaches have benefits and drawbacks.
4.2.1 GSEA leading edge data

Since our goal in this project is to refine gene sets for use with GSEA, and since the leading edge drives the GSEA score of a gene set, it would seem that studying a gene set's leading edges across many contexts can be especially useful for understanding its behavior and learning how to improve it for GSEA. Specifically, the leading edge partitions the genes in a set into those that are strongly differentially expressed in the given context, and those that are not. This partitioning allows us to extract genes that tend to drive the score across many biological contexts. Intuitively, any new gene set created from these genes should score better in GSEA. Examining which genes often occur together in the leading edge of the gene set may also illuminate its co-regulated subsets.

However, the binary nature of the leading edge throws away useful information about the expression patterns of the gene set by effectively flattening the expression data. The leading edge approach highlights genes that often drive the GSEA enrichment score, and allows us to see patterns of co-expression in the case where the genes are highly differentially expressed. However, we may also be interested in weaker cases of differential expression, because this information may help us in creating a successful refinement. Ultimately, co-regulated subsets can be observed from co-expression of genes in the middle portion of the ranked gene list as well as at the extremes. Studying the leading edge of a gene set therefore obscures the co-regulation of genes that are not strongly differentially expressed. In addition, if a gene set is not significantly enriched in many tissues, there will also be limited leading edge information on which to base the refinement.

4.2.2 Row-standardized GNF data

This approach does not suffer from some of the problems of the leading edge approach. Firstly, rather than selecting significant phenotype comparisons as in the leading edge approach, all samples are taken into account. Secondly, the data spans a much wider range of possible values, which allows us to examine the co-expression of genes within...
a set at a finer level.

One of the major issues that must be dealt with in this approach is the problem of outliers. There are cases where a gene is unusually highly expressed in a tissue. Such outliers present a computational problem because they may skew the refinement results. We have dealt with this problem by clipping the expression data in the atlas at the range (-3, 3). However, it is unclear whether this solution is justified biologically. On one hand, we may not be interested in just how strongly a gene is upregulated: there may not be a meaningful difference, for our purposes, between cases where a gene has a five-fold and a ten-fold change in expression. We may think that in both cases, the gene is clearly upregulated, and that is all the information we need to determine its patterns of differential expression. On the other hand, we may in fact want to distinguish ten-fold differential expression from five-fold, especially when the gene is upregulated most of the time. It is therefore unclear whether removing outliers from the refinement data is biologically justified. At the same time, clipping the data makes the computational refinements more reasonable by removing the extreme dominance of the outliers and allowing most genes to contribute to the clustering. Another issue to address is the choice of the value at which to clip the data. The loss of data that results from clipping has to be balanced against the gain in stability of the refinement algorithm.

4.3 Clustering algorithms

In Mootha et al. [3], the authors used hierarchical clustering to extract the co-regulated subsets of the OXPHOS gene set. However, hierarchical clustering does not behave well systematically and is notoriously unstable. Since the desired number of clusters is obtained by cutting the hierarchy tree at a particular branch height, the resulting partition may be counterintuitive and not biologically meaningful. We have therefore explored other clustering algorithms for splitting a gene set into its co-regulated components. We found that in most cases, the performance of several of the algorithms is comparable, and it is difficult to choose the best. Most importantly,
the algorithms produce largely consistent refinements, which suggests that manually curated gene sets are indeed composed of co-regulated subsets which can be teased apart by looking at patterns of co-expression.

4.3.1 K-means vs. NMF consensus

We found that k-means clustering generally works well, but NMF consensus clustering outperforms k-means in our evaluation scheme. There are several possible reasons for this difference in performance. Firstly, the NMF clustering output may be more interpretable biologically since the algorithm is parts-based. Secondly, the difference in performance may be explained by the fact that we used consensus clustering in our NMF approach. The distance metric between two genes in consensus clustering is related to the frequency with which they are clustered together in multiple runs of another clustering algorithm. In our implementation, the genes were hierarchically clustered based on the distance matrix created from multiple runs of NMF. We speculate that in our case, consensus clustering may be performing outlier detection. Genes that are far from all cluster centroids will be joined to a cluster late in the hierarchy. Then, when we cut the tree at some branch height, all the outliers may be grouped in the same noisy cluster, as we see in the case of Cluster 3 from P53_SIGNALING. We note that more extensive analysis is required to evaluate this claim.

4.3.2 NMF consensus on different types of data

In both forms of row-standardized GNF expression data that we used, we represented up- and down-regulated genes equally by appending a negative version of each column in the gene set expression sub-matrix. In the first method, we set all negative values in the resulting matrix to 0; in the second, we simply shifted the data up to make all values nonnegative. We found that these two methods result in good gene set refinements in different situations: the shifting method results in good refinements for large gene sets, while the “zero-out” method works well for smaller gene sets. The breakpoint between the two algorithms occurs for genes containing around 50 genes,
where it is unclear which algorithm works better. We do not know why this happens, but there is a clear observable difference between the performances of these two algorithms for small and large gene sets, and an observable switch in best performance from one algorithm to another as gene set size increases.

We also used NMF clustering on binary leading edge data. For all gene set sizes, the refinement results for leading-edge data were fairly comparable to the results for the best performing algorithm in that gene set size range. However, refinements using the leading edge data were usually somewhat worse than the best refinements from other algorithms. This difference is likely due to the data loss in the leading edge approach. There were gene set ranges where one could have decided that leading-edge data refinements were the best of all algorithms presented (e.g. gene set size range 40-49). However, we decided that this result may be due to “overfitting”, since in most other cases the winning algorithm did not involve leading edge data. Another consideration is that working with leading edge data is computationally intensive because of the many GSEA analyses involved in creating the leading edge binary matrix. Given that these methods have comparable performance and that clustering row-standardized GNF data is less computationally intensive, we chose to use NMF consensus clustering on row-standardized data for refining gene sets.

4.4 Method validation

The visual procedure that we developed for evaluating an algorithm’s success in gene set refinement is very helpful for quickly viewing a summary of the algorithm’s performance. In many cases, it is useful for comparing the performance of several algorithms in order to choose the best one. When there is a clear winning algorithm for refinement, it can often be detected using our visualization. However, there are cases where the visualization is confusing and makes it difficult to see the best algorithm for some gene set size range. In addition, the visualization procedure as a validation technique is essentially subjective. This subjectivity may be appropriate, given the lack of a complete agreement in literature on existing solutions to the problem of clustering.
expression data. However, a more quantitative approach might have been useful as well. A combination of our approach and a more quantitative method validation procedure might have produced better results.

4.5 Future directions

There are many possible improvements on our methodology as well as follow-up analyses that may elucidate the reasons for the relative success of some of the algorithms that we tested for gene set refinement. One question to be explored is whether consensus clustering accounts for most of the success of the NMF consensus technique. This may be answered by using consensus clustering with other algorithms, such as k-means. It is possible that k-means consensus clustering would perform as well as NMF consensus clustering did in our analyses.

While consensus clustering may have accounted for the relative success of NMF, NMF as a standalone algorithm also has an advantage over other methods: it is possible to interpret the clustering results and to obtain insight into how exactly genes are assigned to clusters. The algorithm by its additive nature produces interpretable clustering results. In our work, we only used the resulting clusters from the method and did not examine the specifics of cluster assignments. An in-depth examination of the details of cluster assignment may be used to create more sophisticated approaches, or to choose a more appropriate data transformation. It is therefore worthwhile to explore these details of the NMF algorithm.

Another approach that should be explored is an algorithm that explicitly performs outlier elimination or detection. Such an approach would be biologically appropriate, because most functionally related sets of genes likely have several co-regulated subsets and also some genes that do not follow the expression patterns of any of the subsets. One could therefore benefit from an algorithm that does not assign genes to a cluster if they are far from every cluster centroid.

One of the most important suggestions for future gene set refinements is to include more expression data in assembling a set of biological contexts for the refinement.
Seeing the behavior of functionally related sets of genes in more and more contexts will yield an increasingly better partition of a gene set into co-regulated subsets.

The primary goal of this work was to create biologically meaningful gene sets that would score highly in GSEA on the basis of existing expert-curated sets of genes. However, we can also use the technique of examining a priori biologically significant sets of genes in the context of real expression data to refine our knowledge about biological processes. Discovering co-regulated subsets of sets of genes that are already known to be biologically significant will illuminate the real functionally related sets of genes and yield clues about their specific functions. Such knowledge will improve our ability to interpret microarray data, which will help us gain greater understanding of human biology.
Appendix A

NES Enrichment Plots For
Different Gene Set Size Ranges

On the basis of the following plots, we made subjective choices of the best clustering algorithm for refining gene sets within the given size range. Each plot is shown for all algorithms attempted with the same number of clusters, where we chose the number of clusters based on the gene set size for reasons explained in the Results chapter. In this section, we include detailed plots, for most gene set size ranges, of each algorithm with all possible cluster numbers $k$ as well as comparative plots of different algorithms.
Figure A-1: In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for k-means clustering, leading edge NMF consensus clustering, and consensus clustering with 2 different types of processed GNF data. Shown for $k = 2$, where the original gene sets contained between 25 and 29 genes.
Figure A-2: In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for k-means clustering, leading edge NMF consensus clustering, and consensus clustering with 2 different types of processed GNF data. Shown for $k = 2$, where the original gene sets contained between 30 and 34 genes.
Figure A-3: In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for k-means clustering, shown for $k = 2, 3, 4, 5$. The original gene sets contained between 35 and 39 genes.
Figure A-4: In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on leading edge data, shown for $k = 2, 3, 4, 5$. The original gene sets contained between 35 and 39 genes.
Figure A-5: In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on shifted row-standardized GNF data, shown for $k = 2, 3, 4, 5$. The original gene sets contained between 35 and 39 genes.
Figure A-6: In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on "zero-out" row-standardized GNF data, shown for $k = 2, 3, 4, 5$. The original gene sets contained between 35 and 39 genes.
Figure A-7: In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for k-means clustering, leading edge NMF consensus clustering, and consensus clustering with 2 different types of processed GNF data. Shown for $k = 2$, which was the optimal cluster number for each algorithm shown. The original gene sets contained between 35 and 39 genes.
Figure A-8: In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for k-means clustering, shown for $k = 2, 3, 4, 5$. The original gene sets contained between 40 and 49 genes.
Figure A-9: In the context of GSEA analyses of P53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on leading edge data, shown for $k = 2, 3, 4, 5$. The original gene sets contained between 40 and 49 genes.
Figure A-10: In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on shifted row-standardized GNF data, shown for $k = 2, 3, 4, 5$. The original gene sets contained between 40 and 49 genes.
Figure A-11: In the context of GSEA analyses of p53+/−, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on “zero-out” row-standardized GNF data, shown for k = 2, 3, 4, 5. The original gene sets contained between 40 and 49 genes.
Figure A-12: In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for k-means clustering, leading edge NMF consensus clustering, and consensus clustering with 2 different types of processed GNF data. NES ratios shown for $k = 2$, which was the optimal cluster number for each algorithm shown. The original gene sets contained between 40 and 49 genes.
Figure A-13: In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for k-means clustering, shown for k = 2, 3, 4, 5. The original gene sets contained between 50 and 84 genes.
Figure A-14: In the context of GSEA analyses of p53+/−, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on leading edge data, shown for $k = 2, 3, 4, 5$. The original gene sets contained between 50 and 84 genes.
Figure A-15: In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on shifted row-standardized GNF data, shown for $k = 2, 3, 4, 5$. The original gene sets contained between 50 and 84 genes.
Figure A-16: In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on “zero-out” row-standardized GNF data, shown for $k = 2, 3, 4, 5$. The original gene sets contained between 50 and 84 genes.
Figure A-17: In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for k-means clustering, leading edge NMF consensus clustering, and consensus clustering with 2 different types of processed GNF data. NES ratios shown for $k = 3$ in the cases of NMF consensus clustering on “zero-out” GNF data and on leading edge data, and for $k = 4$ in the cases of k-means clustering on raw data and NMF consensus clustering on shifted GNF data. These values of $k$ were the optimal cluster numbers for their associated algorithms. The original gene sets contained between 50 and 84 genes.
Figure A-18: In the context of GSEA analyses of p53+/−, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for k-means clustering, shown for $k = 2, 3, 4, 5$. The original gene sets contained between 85 and 99 genes.
Figure A-19: In the context of GSEA analyses of p53+/−, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on leading edge data, shown for $k = 2, 3, 4, 5$. The original gene sets contained between 85 and 99 genes.
Figure A-20: In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on shifted row-standardized GNF data, shown for $k = 2, 3, 4, 5$. The original gene sets contained between 85 and 99 genes.
Figure A-21: In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for NMF consensus clustering on “zero-out” row-standardized GNF data, shown for $k = 2, 3, 4, 5$. The original gene sets contained between 85 and 99 genes.
Figure A-22: In the context of GSEA analyses of p53+/-, ALL/AML and good vs. poor outcomes of lung adenocarcinoma, patterns of NES improvement ratios for k-means clustering, leading edge NMF consensus clustering, and consensus clustering with 2 different types of processed GNF data. NES ratios shown for $k = 4$ in the cases of NMF consensus clustering on “zero-out” GNF data and shifted GNF data, and for $k = 5$ in the cases of k-means clustering on raw data and NMF consensus clustering on leading edge data. These values of $k$ were the optimal cluster numbers for their associated algorithms. The original gene sets contained between 85 and 99 genes.
Bibliography


