GENE EXPRESSION ANALYSIS IN THE PRESENCE OF HETEROGENEITY

ABHA BELORKAR

B.E. (Hons) BITS Pilani

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

NATIONAL UNIVERSITY OF SINGAPORE

2018

Supervisor:

Professor Wong Lim Soon

Examiners:

Professor Sung Wing Kin

Associate Professor Leong Hon Wai

Professor Sun Kim, Seoul National University

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Abha Belorkar

January 30, 2018

Acknowledgments

While working towards my PhD, I have come to realize that the process is not only a pursuit of academic knowledge, but also a journey in self-discovery. Some people, with their presence, enthusiasm, and support have inspired and strengthened this conviction, and lent purpose and meaning to my quest. To all of them, I owe my most heartfelt acknowledgments.

I am immensely grateful to my supervisor, Prof. Wong Lim Soon, who motivated me with his enthusiasm for scientific research, demonstrated to me the power of logic in generating, filtering, and refining research ideas, encouraged me through the rough phases of this work, and led by example in many ways with his dedication and consideration. It is impossible for me to imagine a more nurturing PhD advisor.

I am grateful to Prof. Rajanikanth Vadigepalli for giving me the opportunity to intern in his lab at Thomas Jefferson University. His insights into research and writing, his sense of aesthetics in data visualization, and his support and guidance have contributed remarkably to my work.

I would like to thank my thesis committee members, Prof. Sung Wing Kin, and Prof. Leong Hon Wai, for their helpful suggestions and feedback on my reports and presentations.

I am grateful to my lab members at NUS for creating a cordial and conducive research environment. Special thanks are due to my senior, Dr. Kevin Lim, from whose insights and help, I benefited a lot during the earlier part of my program.

I would like to express my gratitude towards my friends at NUS and TJU – Phanita

Vemulapalli, Shruti Tople, Iana Pirogova, Lu Bingxin, and Babita Verma – for their warmth, friendship, and encouragement.

I am indebted to my teacher for the purpose, direction, and transformation which he brought to my life.

I wish to deeply thank my family – my mother for her extraordinary love and patience, sacrifices and courage; my husband for his sense and sensibility, his phenomenal friendship and positivity; my mother-in-law and father-in-law for their affection and support, heartfelt prayers and blessings; my grandfather, my uncles and aunts for their many contributions to my life and character; and my cousins for their love and inspiration.

Summary

Differential expression analysis is a popular approach for identifying genomic biomarkers that distinguish various phenotype conditions. Using the identified biomarkers, biological mechanisms responsible for the phenotype differences are inferred. While methods for such analysis have evolved significantly in the last two decades, they are unable to account for undeclared heterogeneity in the groups under comparison. On the other hand, heterogeneity, of either biological or non-biological origins, is observed to be invariably present in gene expression datasets. Delineating the basis of gene expression heterogeneity in relation to biological pathways is a difficult problem. Our work is aimed at addressing this challenge in three ways:

First, we propose a normalization technique based on rank-fuzzification – Gene Fuzzy Scores (GFS), which retains meaningful variation in gene expression and attenuates obscuring noise. This is important for two reasons: (a) the quality of preprocessing heavily impacts the reliability of downstream gene expression analysis; and (b) popular normalization methods are reported to seldom enhance the quality of expression data. Comparison of GFS with other popular techniques – mean-scaling, quantile normalization, z-score normalization – showed that output from our normalization approach is more consistent and biologically coherent.

Second, we present SPSNet – a method for differential expression analysis of samples with potential heterogeneity. SPSNet reports a list of significant subnetworks (smaller components of biological pathways) whose expression reveals undeclared sub-populations within the given sample phenotypes. Current approaches to study

heterogeneity perform comparisons of individual genes across phenotypes, and thus shroud a holistic view of the underlying biological mechanisms. In contrast, our approach reveals factors relevant to biological heterogeneity (e.g. disease subtypes, developmental stages) or non-biological heterogeneity (e.g. platform differences, batch effects) in the form of gene subnetworks, amplifies their effects in the data, and facilitates discrimination of subpopulations within phenotypes. Using publicly available gene expression datasets containing disease heterogeneity and batch effects, we show that SPSNet has low false-positive rate, high sensitivity, and high biological coherence in analyzing heterogeneous gene expression data.

Finally, with the help of an illustrative case-study, we demonstrate the potential of our methods for normalization and heterogeneity analysis – GFS and SPSNet – to analyze RNA-Seq datasets. We observe that data generated on RNA-Seq platforms, unlike microarray data, is subject to sampling stochasticity when sequencing depth is insufficient. This fact plays a critical role in the performance of methods which analyze RNA-Seq data. We present a Bernoulli trial-based model to explain sampling stochasticity, and propose the use of discretized-GFS (D-GFS) to attenuate the stochasticity effect. In our analyses, we also note that silhouette score fails to accurately represent the degree of clustering in data which is characterized by high dispersion. In response, we suggest a simple and effective alternative for clustering assessment, based on a metric we define as kNN score – the proportion of samples whose label matches the majority of its k nearest neighbors.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"A hundred instances of Hodgkin's disease, even though pathologically classified as the same entity, were a hundred variants around a common theme. Cancers possessed temperaments, personalities – behaviors. And biological heterogeneity demanded therapeutic heterogeneity; the same treatment could not indiscriminately be applied to all..."

– Dr. Siddhartha Mukherjee, "The Emperor of All Maladies: A Biography of Cancer"

Heterogeneity is a well-recognized phenomenon in many complex diseases – cancers [FPS13], diabetes [TSC+14], asthma [MB14], cystic fibrosis [DZD12], and schizophrenia [LCF+13] are some common examples. Understanding heterogeneity in such diseases requires a systematic analysis of their multiple subtypes which arise from the exploitation and breaking of diverse biological mechanisms, while maintaining similar overall characteristics. For this reason, diagnostic criteria, which are based on a combination of clinical and morphological features, are unable to fully capture the multi-faceted diversity amongst patients. A molecular understanding of heterogeneity is essential so that patients may be appropriately classified into more homogeneous subgroups, entailing more predictable and effective response to available treatment options.

Gene expression studies are increasingly recording the presence of heterogeneity in several contexts – heterogeneity is critical in the study of disease subtypes, developmental stages, time-series gene expression, nested experimental conditions, as well as technical variation due

to batch effects, platform differences in integrated meta-analyses, etc. We define heterogeneity in a given phenotype as any variation in the transcriptional pattern that gives rise to smaller subpopulations within the phenotype. Thus, the term is used in the sense of its literal English meaning (according to the Oxford dictionary, heterogeneity is "the quality or state of being diverse in character or content"). This way of defining heterogeneity is particularly significant in our work because our aim is to gain a modular understanding of *undeclared* heterogeneity in gene expression data. This means that we do not know in advance the source(s) of variation present in the given gene expression datasets. The methods proposed in this thesis help to delineate the basis of heterogeneity, irrespective of its source, in relation to biological pathways.

While methods for differential gene expression analysis have evolved considerably in the past two decades, accounting for the presence of undeclared heterogeneity remains a problem. Since a homogeneous sample is characterized by a set of common biological mechanisms, it is generally described with a set of features (e.g. expression of selected genes, aggregate expression of selected pathways, or aggregate expression of smaller components of pathways called subnetworks) following unimodal distributions. Comparing a well-defined set of such features across samples is relatively easier. In contrast, heterogeneous samples contain patients with a variety of biological mechanisms, which result in features with more complex statistical distributions. Thus, they are more difficult to compare.

Analyzing heterogeneous gene expression data and gaining a modular understanding of heterogeneity is a difficult task. Towards addressing this challenge, our work makes three contributions. First, we examine the role and effectiveness of normalization techniques in preprocessing heterogeneous data. Second, we propose a method to obtain sub-population-specific signatures in heterogeneous microarray expression data based on the activity of subnetworks in biological pathways. Finally, we present two case-studies demonstrating the application of our normalization and differential expression methods on RNA-Seq data, and discuss the specific considerations involved in doing so.

1.1 Motivation

1.1.1 Preprocessing

Preprocessing data with a suitable normalization method is essential to making expression values from multiple samples comparable, especially when they are heterogeneous, are from separate platforms, or in independent batches. Yet, it was reported [LSS+10] that popular normalization techniques are not very successful in discriminating between real and obscuring variation to produce quality input for downstream gene expression analysis. In fact, it was noted by Luo et. al [LSS+10] that preprocessing using standard normalization methods led to reduction in the quality of subsequent predictive models in up to 25% of the cases. On the other hand, certain differential analysis techniques (such as PFSNet [LW13]) which use a rank fuzzification transform on gene expression data as their preprocessing step, are known to be more reproducible across multiple data batches. This motivates us to examine the role and effectiveness of preprocessing in analyzing heterogeneous gene expression data.

1.1.2 Differential expression analysis

Heterogeneity, regardless of its origin, is often undeclared, as incomplete knowledge prevents the accurate identification of sub-populations in a phenotype. A systematic understanding of variation at the cellular and molecular level is key to deciphering such diversity. Many studies attempt to achieve this using unsupervised techniques to learn gene expression-based subtype-specific molecular signatures. Typically, gene expression data is subjected to hierarchical clustering or orthogonal transformation, and sub-populations in the sample are inferred using observations regarding variation patterns. However, such analysis is carried out at the individual-gene level, and leaves considerable room for subjective, and possibly incorrect, interpretation of the underlying biological mechanisms. It also prevents a systemic view of these mechanisms, and leads to a high false-positive rate, and low reproducibility [ZZZ+09]. The famous work

by Venet et al. [VDD11] even shows that such gene-based signatures are rarely better than random, at least in the context of breast cancer. In contrast, PFSNet [LW13] and other techniques (e.g. [LLCW15]) that report signatures based on subnetworks within pathways are shown to be highly reproducible and exhibit very low false-positive rate. However, these techniques are only designed to compare homogeneous phenotypes. This prompts us to take a network-based approach to understanding heterogeneity in gene expression.

1.1.3 Analysis of RNA-Seq expression data

Gene expression data obtained on microarray and next-generation sequencing platforms differ in a few important ways – range of precise measurements, effect of sampling stochasticity, ability to measure low expression transcripts, to name a few. Therefore, methods of analysis originally designed for and tested on microarray datasets, often need modifications for application on RNA-Seq data. Therefore, we examine the characteristics of RNA-Seq data specifically important for normalization and differential expression analysis, and evaluate the potential of GFS and SPSNet using a case-study on a HCC RNA-Seq dataset.

1.2 Focus and contributions of the thesis

This dissertation makes three important contributions towards understanding and analyzing heterogeneity.

1. We propose Gene Fuzzy Score (GFS), a simple rank-based preprocessing technique, that is able to largely reduce obscuring variation while retaining useful biological information. Using four sets of publicly available microarray datasets containing batch effects and heterogeneity, we compared GFS with three standard normalization techniques as well as raw gene expression. Each method was evaluated with respect to the quality, consistency, and biological coherence of its processed output. It was found that GFS outperforms

other transformation techniques in all three aspects.

A paper resulting from this work was published in BMC Bioinformatics:

Abha Belorkar, and Limsoon Wong. "GFS: fuzzy preprocessing for effective gene expression analysis." BMC Bioinformatics 17.17 (2016): 540.

2. We present a method – viz. Sub-Population-Specific Network-based analysis (SPSNet) – to analyze heterogeneity in gene expression data, and obtain subtype-specific signatures based on activity of subnetworks in biological pathways rather than individual genes. When heterogeneity is biological in nature, our approach identifies sub-populations within a sample and the underlying diverse biological mechanisms. In the presence of extrinsic or non-biological heterogeneity such as batch effects, it amplifies these effects, and helps to identify and eliminate factors irrelevant to the biology of the phenotypes being studied. Using publicly available microarray datasets containing disease heterogeneity and batch effects, we provide evidence for low false-positive rate, high sensitivity, and high biological coherence of our method in analyzing heterogeneous gene expression data.

   A paper resulting from this work was published in BMC Systems Biology: **Abha Belorkar, Rajanikanth Vadigepalli, and Limsoon Wong. "SPSNet: subpopulation-sensitive network-based analysis of heterogeneous gene expression data." BMC Systems Biology 12.2 (2018): 28.**

3. We perform a case-study of an RNA-Seq dataset, and observe that RNA-Seq data is imprecise when sequencing depth is insufficient. This is because, in RNA-Seq, transcripts compete with each other to be assigned the limited number of total reads, which may result in sampling a random subset of the original transcriptome. We propose a Bernoulli trial-based model to describe this stochasticity. Further, to attenuate the effects of this stochasticity on RNA-Seq data, we incorporate discretization into our normalization approach (GFS). We also show that SPSNet can be used to analyze heterogeneity on

RNA-Seq data. Moreover, we identify a context in which silhouette score fails to quantify the extent of clustering – when intra-cluster distance is very high due to large dispersion in data. We then define a metric called the kNN score, and propose a method to use this score for performing a quantitative assessment of clustering based on the nearest neighbors of a sample. We illustrate that the technique provides an intuitive and useful way to quantify the extent of clustering.

1.3 Thesis organization

This thesis is organized in 6 chapters. Chapter 2 gives the background on popular methods for differential expression analysis, and specifically illustrates that PFSNet is a method that – in comparison to other common methods – is particularly effective and consistent across multiple datasets. Chapter 3 discusses insights from our examination of the role of normalization techniques in preprocessing heterogeneous data, and how a rank fuzzification transform such as that used in PFSNet can be used as a tool to improve their effectiveness. Chapter 4 describes our proposed approach SPSNet, a generalization of PFSNet, to analyze heterogeneity and obtain subtype-specific signatures based on activity of subnetworks in biological pathways. Chapter 5 highlights the effect of sampling stochasticity on the precision of RNA-Seq data, and presents an illustrative case-study to demonstrate the potential of our normalization and differential expression analysis approach on RNA-Seq. Chapter 6 summarizes our work and presents directions for future work.

Chapter 2

Related Work

"There's two possible outcomes: if the result confirms the hypothesis, then you've made a discovery. If the result is contrary to the hypothesis, then you've made a discovery."

– Enrico Fermi

2.1 Introduction

Methods analyzing data generated by microarray technology and next-generation sequencing have substantially accelerated the rate of hypothesis generation in genomics. By observing patterns in gene-expression data obtained from these high-throughput technologies, it is possible to propose several interesting conjectures aimed at understanding the functioning and co-ordination of genes in biological mechanisms. Genes which are differentially expressed in two (or more) phenotypes of interest are likely to be instrumental in explaining biological phenomena that cause, or arise from the differences in phenotypes under consideration (normal and disease patients, patients before and after receiving a particular course of treatment, etc.) This forms the premise for differential expression analysis, and helps to discover causal genes and pathways, identify appropriate targets for drug treatments, etc.

Such analysis is impeded by various challenges, including biological noise, technological noise, and batch effects, all of which complicate the process of recognizing true signals in

gene-expression data. For example, Raser & O'Shea [RO05] discuss various sources of biological noise in gene-expression data, including the inherently stochastic nature of pertinent biochemical reactions. Moreover, technological noise continues to be an inevitable component of gene-expression data as even RNA-seq was shown to suffer from obscuring measurement variability similar to what was earlier observed in microarrays [HIZ12]. In another study [LSB+10], an analysis of second-generation sequencing data generated by the 1000 Genomes Project revealed that 32% of the features (features = experimental measurements such as expression levels of genes) were associated with the date of sequencing while merely 17% were associated with biological outcomes, demonstrating the extent to which batch effects influence measurement variability in genomic technologies. However, some issues in performance of gene-expression analysis techniques do not stem from the challenges mentioned above. Many concerns have been raised [GB07, KSB12, ACPS06] regarding the fundamental methodologies adopted and assumptions made in statistical procedures that are used in some of the most popular techniques.

2.2 Analysis at the individual-gene level

The earliest differential expression analysis methods relied on testing individual genes for significance. A common way to do this has been to use the fold-change observed in the average expression of each gene. Genes whose fold-change values lie on the tails of the distribution so obtained, are considered to be significant. Alternatively, a t-test is done to compare the means of expression value distributions corresponding to the two phenotypes of interest, and thereby identify genes that are significantly differentially expressed. A similar strategy is to use the Wilcoxon signed-rank test which compares population mean ranks instead of means themselves. Attempts to improve the performance of these simplistic methods were made in SAM [TTC01] and Rank Products [BAAH04], where, instead of evaluating individual genes, the ordering information of statistics over the entire set of genes was used.

There are multiple drawbacks associated with the above set of approaches. Most notably, it gives rise to the issue of multiple hypothesis testing, since thousands of genes are used to compare two phenotypes. When two classes (phenotypes) are compared based on multiple features (genes), it is likely that more and more features will distinguish the classes by chance alone, as the number of features increases. For example, if 20,000 genes are being tested with a significance threshold of 0.05 each, the expected number of genes identified as significant merely by chance would be a thousand. This leads to a high false-positive rate, and may only be corrected by either obtaining a remarkably large sample, or requiring a much higher threshold of significance. It was noted by Zhang et al [ZZZ+09], that the overlap between lists of differentially expressed genes from two studies (of the same disease) is often low, even in the presence of negligible technical noise. It was even found in some cases [VDD11], that the outcome of such methods may be no better than a list of random genes. Moreover, even if a reliable list of significant genes is obtained in this manner, explanatory biological themes are hardly apparent from such an end result. So, in the last one and a half decade, the focus of efforts in differential expression analysis has shifted from finding individual significant genes to identifying sets of related causal genes [Won11], where the information regarding groups of related genes (and their interactions) are generally obtained from Gene Ontology (GO) or pathway databases.

2.3 Incorporating biological pathway information

Testing gene sets for significance as opposed to individual genes would mitigate the issue of multiple hypothesis testing, provided the gene sets under consideration are relatively few in number. Indeed, if we were to examine all possible gene set formations, we would hugely exacerbate the multiple hypothesis testing issue. An ideal way to limit such candidate gene sets is to incorporate additional information, available in the form of biological pathways (or ontologies), so that biologically unreasonable gene sets are excluded. This idea has given rise to the next generation of differential expression analysis methods, which may be conceptually

classified under the following categories.

2.3.1 Over-representation analysis

In over-representation analysis (ORA), a list of differentially expressed genes (DEGs) is first obtained with a pre-determined significance threshold. Each pathway (or GO group) is then tested for over-representation or under-representation in this pre-computed list, using a test based on the hypergeometric, chi-squared, or binomial distribution [KD05]. However, this class of methods suffers from several shortcomings that arise from the procedure used to pre-compute DEGs, as well as from the statistical tests used thereafter to test for over/under-representation. Firstly, the list of DEGs is highly sensitive to the pre-specified significance threshold. Yet, the choice of this threshold is almost arbitrary, implying that genes with marginally less significance than the threshold could be missed easily. Secondly, the list of DEGs may be influenced by the sample due to sampling bias [WSG16]. Thirdly, genes relevant to the differences in phenotypes may not always be differentially expressed themselves but may cause other genes to significantly change their expression [LW13]. By focusing only on differentially expressed genes, ORA misses genes that may be biologically important. Fourthly, the null hypothesis of the hypergeometric test assumes that genes in a pathway are no different from a random set of genes. This is obviously false since genes in a pathway are known to be correlated, resulting in very frequent rejection of the null hypothesis, and consequently generating many false-positives. It was also pointed out in [GB07], that the sampling procedure in ORA is statistically invalid, leading to misleading interpretations of resultant p-values. Lastly, only a small portion of a pathway may be relevant to the differences in phenotypes. By testing entire pathways for over/under-representation, ORA risks missing important pathways where signals from a small part of a relevant pathway are diluted by noise from the large remainder of the pathway.

2.3.2 Direct group analysis

Direct group analysis methods assign a significance score to each gene in a pathway, based on the extent to which they are differentially expressed in the two phenotypes of interest and then aggregate it to obtain a pathway-level score (such as the Kolmogorov-Smirnov statistic in case of GSEA [STM+05], or mean of log-p values in FCS [GVDGDKVH04]). This score is then subjected to a statistical test. In FCS, the statistical test involves estimating significance of the pathway by comparing its score to that of repeatedly sampled sets of random genes of the same size. This test, like ORA, generates many false-positives, because the null hypothesis assumes independence of genes within a pathway. GSEA performs a permutation test which involves class-label swapping (each tissue in the sample is randomly assigned a phenotype class irrespective of its original membership) to generate its null distribution. Thus, the issue in FCS is corrected here by using the correct null hypothesis, namely, that the difference of the pathway-level statistic between the two phenotypes is irrelevant to the biological differences between the phenotypes. However, for small sample sizes, GSEA cannot generate such a null distribution, as the number of possible permutations is insufficient. Hence, in case of small sample sizes, it resorts to gene-label swapping, which again incorrectly assumes gene-gene independence in pathways. Direct group methods overcome an important problem in ORA – the need to specify a significance threshold at the level of individual genes is eliminated, as each gene is assigned a significance score based on its expression values in the two classes and is ultimately incorporated into the analysis. On the other hand, by considering entire pathways, they leave the noise from irrelevant parts of the pathways unaccounted for. Also, like ORA, this group of methods may miss biologically important genes that are not themselves differentially expressed.

2.3.3 Network-based analysis

Network-based analysis focuses on obtaining relevant *subnetworks* within pathways, which may be consequently subjected to significance testing so that true signals from affected parts of a pathway may be successfully discriminated from the noise in the remainder of a pathway. NEA [ALP+12] attempts to handle this issue by forming subgraphs based on immediate neighborhoods of each gene. Subsequent analysis is done by testing the subgraphs so formed, using a permutation test such as that used in GSEA. Although this enhances the power to recognize affected parts of a pathway, it provides little insight into the underlying biological mechanism as the subnetwork formation procedure is somewhat coarse, and provides no assistance in explaining the role of potential upstream influences. This is addressed by SNet [SDGW11], which relies on the gene-expression matrix itself to obtain relevant subnetworks, instead of adopting a generic pre-defined scheme for the task. SNet considers the top α% highly expressed genes that occur in at least β% of tissues in each phenotype sample (where the default values for α and β are set to 10 and 50 percent respectively). Genes obtained in this manner are used to fragment the pathway, and the remaining genes are removed from the following analysis. As SNet depends on gene ranks (which are relatively more stable) instead of absolute expression values, it results in higher consistency compared to most previous methods. However, this leads to the issue of determining appropriate values for the α and β parameters – an overtly relaxed α would lead to many false-positives, whereas a highly conservative value would incur the risk of missing important genes. PFSNet [LW13] improves on this deficiency by disposing the strict cut-offs in favor of a fuzzification process that assigns a score between 0 and 1 to genes that do not fall within the region defined by fixed thresholds. Further, it computes a paired t-statistic obtained by allowing genes in each subnetwork to vote for both phenotypes, which leads to more consistent results. A different method in this category is DEAP [HHS+13], which identifies the maximally differentially expressed path within each pathway and then performs an array rotation test to estimate its significance. In the rotation test, a null distribution is obtained by generating random expression profiles which preserve

the gene-gene correlations in the expression values. Since the number of such rotations is unlimited, reliable significance estimates can be obtained by this method even when the sample size is very small. In general, network-based methods perform much better in practice than ORA and direct group methods.

2.3.4 Model-based analysis

Model-based methods typically adopt a general framework that consists of three important steps. First, groups of closely related genes are formed by various methods. GGEA [GCK+11] does this by superimposing immediate neighborhoods of genes on a global gene regulatory network to obtain smaller fragments. SRI [ZLSA11] forms subnetworks by extracting groups of genes most correlated with each other, where the correlation threshold is defined by a rank parameter θ whose best value is estimated using Bayesian Information Criterion. Second, the gene-expression behavior of every subnetwork in one phenotype is modelled using a chosen machine learning or modeling technique (e.g. fuzzy petri nets in GGEA, and linear regression in SRI). Third, the resultant models are used to predict gene-expression values in another phenotype, and a comparison between observed and predicted values is performed. Inconsistency is interpretated to be indicative of perturbations responsible for biological differences within the two phenotypes. If reasonable models for subnetworks are known, model-based methods could be ideal for explaining biological mechanisms relevant to a disease, since they possess the capability to specifically learn the working of individual subnetworks in control sample and identify abnormalities in test sample based on this acquired understanding. This also implies that the genes being detected as significant need not necessarily be differentially expressed themselves. Despite the latent potential of model-based approaches derived from the benefits mentioned above, these methods have not gained much popularity in practice, since the parameter estimation involved in the learning process requires an enormous amount of quality data, which has been relatively hard to collect so far. For example, SRI uses as many as 1366 different microarray experiments from Gene Expression Omnibus (GEO) for

model training [ZLSA11]. However, it is possible that the single-cell sequencing technologies will help to mitigate this problem, since it is capable of generating data that is not only rich in biological information, but also adequate in quantity to overcome the classical issue of parameter estimation.

2.4 Performance analysis of network-based methods

We designed three experiments to assess the reliability of some of the latest generation of gene expression profile analysis methods – GGEA, DEAP, and PFSNet – based on their consistency, false discovery rate, and performance on small-sized samples. We observed that, in all three experiments, the performance of PFSNet is significantly better than the other two methods, with DEAP being marginally better than GGEA. Below, we explain the design of our experiments and present the results obtained from them.

In our setup, we used three pairs of datasets corresponding to three different diseases – Duchenne Muscular Dystrophy (DMD) [HSK+02, PBM+07], Leukemia [ASS+02, GST+99], and Childhood Acute Lymphoblastic Leukemia (ALL) [RMO+04, YRS+02] (described in table 2.1).

Table 2.1: Datasets used for performance analysis

Disease type	Source	Affy GeneChip	Dataset composition
DMD	Haslett et al. [HSK+02]	HG-U95Av2	12 DMD, 12 controls
	Pescatori et al. [PBM+07]	HG-U133A	22 DMD, 14 controls
Leukemia	Golub et al. [GST+99]	HU-6800	47 ALL, 25 AML
	Armstrong et al. [ASS+02]	HG-U95Av2	24 ALL, 24 AML
ALL	Yeoh et al. [YRS+02]	HG-U95Av2	15 BCR-ABL, 27 E2A-PBX1
	Ross et al. [RMO+04]	HG-U133A	15 BCR-ABL, 18 E2A-PBX1

We used pathway information from the database PathwayAPI [SDGW10] which contains the unification of human pathways from KEGG [OGS+99], Wikipathways [KvIH+12], and Ingenuity.

2.4.1 Consistency

To measure consistency of the methods, we use gene expression dataset pairs of the same disease phenotypes. Since each disease should have a common set of causes, a good method should consistently report the same causes and nothing else, when applied independently to two datasets which are sufficiently representative of a disease phenotype. Thus, the reliability of a differential gene expression analysis method can be assessed by evaluating how reproducible the set of causes reported by a method is when the method is independently applied on such a dataset pair.

We run GGEA, DEAP, and PFSNet independently on the two datasets in each of the three dataset pairs. Thus for each method and each dataset pair, we get a pair of results corresponding to the two datasets used. These results are in the form of a list of pathways (GGEA) or subnetworks (DEAP, PFSNet) along with their respective p-values. A significance threshold of $p = 0.05$ was used to extract pathways (subnetworks) which are reportedly significant in accordance to the respective methods. Further, jaccard similarity coefficient was used as a measure to calculate the consistency in the two results for each dataset pair. The performance of all methods is summarized in Figure 2.1.

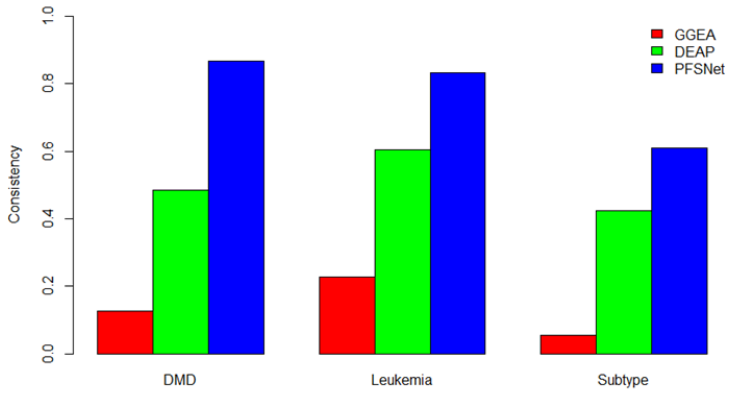


Figure 2.1: Results from experiment to evaluate pathway/subnetwork consistency

For PFSNet, the subnetwork consistency was calculated by using the approach mentioned

in [LW13] which uses jaccard-like agreement to calculate subnetwork consistency. However, since GGEA reports pathways, we simply calculated the jaccard coefficient for pathways. Even though DEAP detects subnetworks within pathways, we also calculated its consistency in the same manner as that of GGEA, as this does not adversely affect its performance profile. Since DEAP detects only one subnetwork (the maximally differentially expressed path) per pathway, pathway consistency would always prove to be an equally or more relaxed measure than subnetwork consistency. Thus, despite giving DEAP the advantage of a more lenient measure, we observe that PFSNet (with around 61% to 87% consistency) leads the other methods by a huge margin (DEAP – 42% to 61%, GGEA – 10% to 23%).

2.4.2 False-positive rate

In this experiment, we generated random expression matrices of the same size as our original three pairs of datasets. To generate these matrices, we computed the mean and standard deviation of the expression of each gene across all patients, and sampled expression values from the normal distribution of the same mean and standard deviation. For each dataset, 20 instances of such random matrices were generated. Since the matrices were randomly generated, we know that every pathway or subnetwork reported as significant is actually a false positive. We ran GGEA, DEAP, PFSNet on these random datasets, and calculated false-positive rate (FPR) as the ratio of the number of pathways (or subnetworks) reported as significant to the total number of input pathways. The average false-positive rate is shown in Figure 2.2.

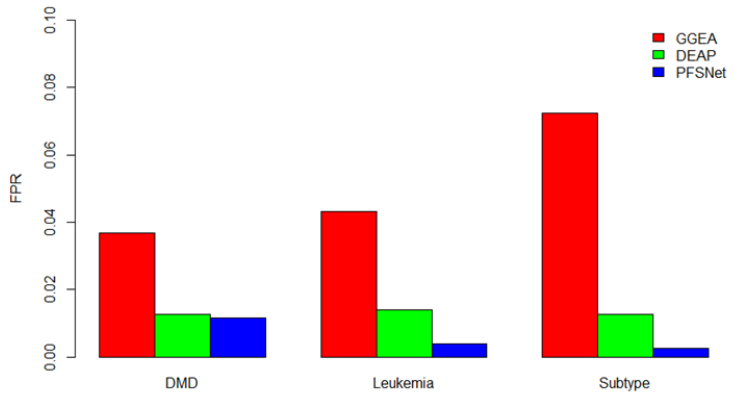


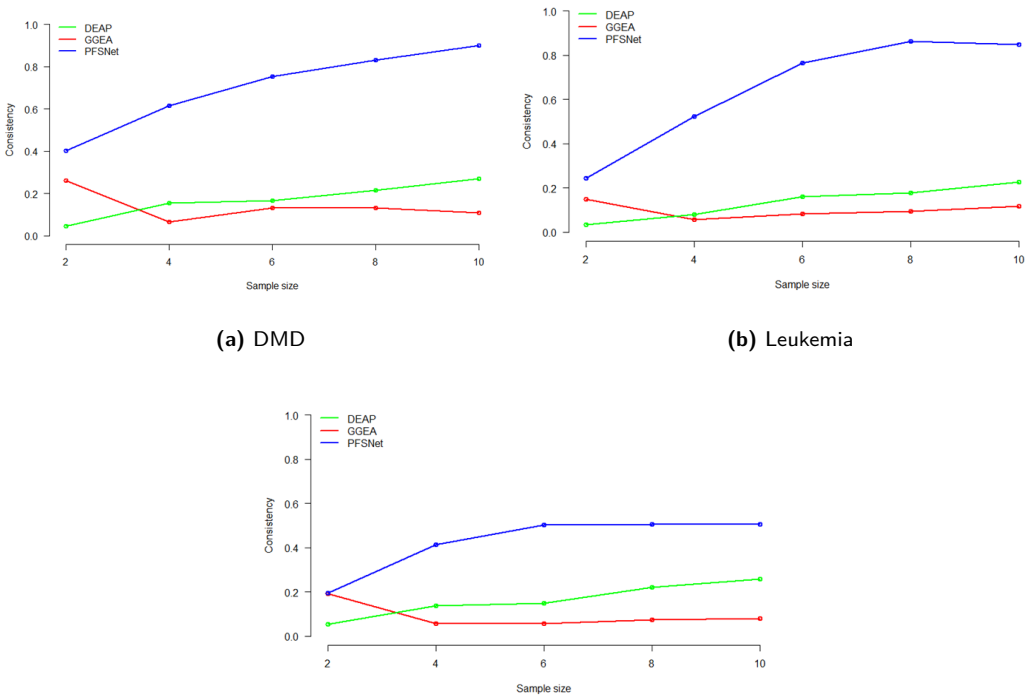
Figure 2.2: A comparison of the false-positive rates of GGEA, DEAP, and PFSNet obtained by using random expression matrices of same size as original datasets for DMD, Leukemia, and ALL subtype

It can be seen that PFSNet consistently gives the least FPR (around 0.002 to 0.01), whereas GGEA performs the worst (FPR around 0.04 to 0.07).

2.4.3 Performance on small-sized samples

To assess the effect of sample size on performance of the methods, we performed random sampling from the original datasets, and obtained new datasets of sample size 2, 4, 6, 8, 10 in each phenotype. We then calculated the consistency of GGEA, DEAP, and PFSNet based on these smaller datasets, using the same procedure as specified in section 2.4.1. This was repeated for 20 iterations and the average consistency was calculated for each method corresponding to each of the sample sizes.

In Figure 2.3, we observe that both GGEA and DEAP perform very poorly on small sample sizes, and are also slow to reach high consistency as sample size is increased. PFSNet is also seen not performing very well on extremely small-sized datasets, though it recovers quickly to give near perfect consistency as sample size is increased, except in the case of ALL subtype datasets.



(a) DMD

(b) Leukemia

(c) ALL subtype

Figure 2.3: Effect of sample size on performance of GGEA, DEAP, PFSNet

2.5 Conclusion

An ideal method for differential expression analysis should be tolerant to noise, leading to near-perfect consistency and very low false-positive rate, and should be powerful enough to show good consistency even on smaller datasets. Such a method can be trusted to draw accurate inferences regarding the biological phenomena underlying the reportedly significant pathways and subnetworks. From our experiments above, we illustrated that the performance of many popular current methods is not entirely satisfactory by these standards, and PFSNet is one of the rare exceptions in demonstrating high consistency and reliability. We conjecture that PFSNet partly derives its effectiveness from its normalization procedure – rank fuzzification. This motivated us to examine the different factors responsible for PFSNet's performance. We also proposed a generalization of PFSNet, which we called SPSNet, for identifying significant gene subnetworks in heterogeneous expression data, since PFSNet is designed to analyze pure phenotypes. Our approach and findings are discussed in the following chapters.

Chapter 3

GFS: Fuzzy Preprocessing for Effective Gene Expression Analysis

"Everybody believes in the normal approximation, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact."

– G. Lippman

3.1 Introduction

Gene expression profiling experiments and analysis are often designed with the objective of verifying one or more hypotheses that can help in building effective diagnostic or prognostic models in clinical settings. Typically, expression data are collected from groups manifesting differences in certain properties of interest, such as disease types or states, developmental stages, and response to specific treatments or interventions over time. The collected data are then mined for appropriate variation patterns relevant to the hypotheses under consideration. The underlying assumption in such studies is that the input gene expression values from different samples accurately reflect the amounts of RNA produced by the corresponding genes and, thus, are properly comparable. However, in practice, unless an effective normalization technique is applied to preprocess the expression data, a number of factors may lead to the violation of this assumption [SRJ+06], [LSS+10].

Firstly, the entire technical process of isolation and quantification of RNA leading up to the final measurements is unlikely to be completely error-free, as inaccuracies may insinuate any of the steps in the long procedure. Secondly, with change in time, place, and other variables in experimental settings, systematic biases of non-biological origins invariably enter during measurement experiments in the form of batch effects. When such biases are correlated with the biological properties under investigation, they can severely confound interesting variation [LSB+10]. Thirdly, differences in experimental settings may also introduce changes in local environments of cells, thus inducing fluctuations in gene expression that further contribute to noise in the measurement data [RO05].

All these factors together make it improbable for multiple samples to naturally have comparable expression values. Therefore, we rely heavily on the capabilities of a preprocessing method to recover meaningful biological information, and remove or account for noise in the form of obscuring variation. Yet, it was reported [LSS+10] that popular normalization techniques are not very successful in discriminating between real and obscuring variation to produce quality input for downstream gene expression analysis. In fact, it was noted by Luo et al. [LSS+10] that preprocessing using common methods led to reduction in the quality of subsequent predictive models in up to 25% of the cases.

To mitigate the performance issues commonly presented by preprocessing techniques, we propose Gene Fuzzy Score (GFS), a transformation method that uses fuzzy scores derived from rank values of gene expression within individual tissues in a sample. We chose four different sets of gene expression data containing substantial batch effects and heterogeneity for the analysis. On these datasets, we compared the performance of GFS and other popular preprocessing methods with respect to the quality, consistency, and biological coherence of their processed output.

3.2 Background

Preprocessing techniques typically attempt to make expression values from multiple samples comparable in two different ways:

1. by scaling expression values such that each sample has an equal value for a statistic such as mean or median; or
2. by adjusting expression values such that each sample has the same expression distribution across genes.

The first approach includes methods such as mean and median scaling, and is popular for Affymetrix genechips. For example, in the mean-scaling method, the mean gene expression value of each microarray in the sample is first calculated, and a grand mean is then computed as the mean of all means. Finally, expression value of each microarray in the sample is scaled such that the mean expression of each microarray is equal to the grand mean. Median scaling also follows the same procedure, with the mean statistic being replaced by median. While these methods are simple to implement, they assume that expression values of all microarrays in the samples share a linear relationship. They – especially mean-scaling – also suffer from a few other drawbacks such as sensitivity to outlier distortions [CVFB03].

The second approach includes more sophisticated methods such as z-score and quantile normalization. In z-score normalization, the expression values of genes in each microarray are transformed to fit the standard normal distribution with a mean of zero and 1 unit standard deviation. On the other hand, quantile normalization uses the rank values of gene expression within individual microarrays to make the distribution of all microarrays identical in statistical properties. Since ranks are known to be relatively more robust to batch effects than absolute expression values [SRJ+06], this is expected to lead to better performance on datasets with batch effects. In the quantile normalization procedure, the expression values of each microarray are first sorted in ascending order, and the mean expression corresponding to each rank across microarrays is stored separately. Following this, the original expression values in each microarray

are assigned ranks based on their relative quantitative order. Finally, a transformed matrix is obtained by replacing each gene's rank value by the mean expression value corresponding to that rank as stored earlier.

The z-score and quantile normalization methods are relatively more robust to outliers, provided that the number of microarrays in a dataset is sufficiently large. However, the actual distributions of underlying data are assumed to be identical in all microarrays, and specifically assumed to be Gaussian in case of z-score normalization. This assumption is especially likely to break down in datasets with disease-state samples where the regular functions of the genes and their synchronization with each other may be substantially disrupted. In such cases, the expression patterns within a microarray of the disease sample may not be identical to those in the normal phenotype sample. It also may not be identical to other microarrays in the disease sample if the disease is heterogeneous and is able to manifest itself through the exploitation and/or breaking of multiple mechanisms.

It is also commonly observed that low-expression genes and proteins exhibit a much greater coefficient of variance than highly expressed ones in their expression levels (see figure 2E in the work by Goh et al. [GGAW15]). Thus, the expression rank of low-expression genes is highly unstable. This may adversely affect the performance of a ranking-based normalization method such as quantile normalization.

Therefore, we are inspired to present GFS as a preprocessing technique for gene expression data. Like quantile normalization, GFS also makes use of gene expression ranks instead of absolute values, thus earning more robustness to batch effects. However, unlike the above techniques, we do not make any assumptions on the similarity of distribution or the equality of any mean-, median-like statistic across microarrays in samples. Moreover, in GFS, we fuzzify the expression ranks such that irrelevant fluctuations introduced by minor differences in ranks are alleviated, and noise from low-ranked genes is discarded.

The idea of fuzzification has also been used earlier in a few gene expression profile analysis methods [LW13] (PFSNet), [GCK+11] (GGEA) and also proteomic profile analysis methods

[GGAW15] (qPSP), [GW16b]. However, these works merely use it as a component of their respective methods, and do not study its role and effectiveness as a normalization procedure.

3.3 Material and Methods

3.3.1 Datasets

We collected datasets (see Table 3.1) from three different disease types – Duschenne Muscular Dystrophy (DMD), Leukemia, and Acute Lymphoblastic Leukemia (ALL).

Table 3.1: Datasets used for comparing preprocessing methods

Disease type	Source	Affy GeneChip	Dataset composition
DMD	Haslett et al. [HSK+02]	HG-U95Av2	12 DMD, 12 controls
	Pescatori et al. [PBM+07]	HG-U133A	22 DMD, 14 controls
Leukemia	Golub et al. [GST+99]	HU-6800	47 ALL, 25 AML
	Armstrong et al. [ASS+02]	HG-U95Av2	24 ALL, 24 AML
ALL	Yeoh et al. [YRS+02]	HG-U95Av2	15 BCR-ABL, 27 E2A-PBX1
	Ross et al. [RMO+04]	HG-U133A	15 BCR-ABL, 18 E2A-PBX1
ALL	Yeoh et al. [YRS+02]	HG-U95Av2	6 Normal, 26 TEL-AML1, 22 Hyperdip>50, 15 T-ALL, 10 Pseudodip, 6 BCR-ABL, 7 MLL, 8 Hyperdip47-50 9 E2A-PBX1, 3 Hypodip

A single gene expression matrix was produced by merging the two DMD datasets from Haslett et al. [HSK+02] and Pescatori et al. [PBM+07]. Similarly, data were merged from Armstrong et al. [ASS+02] and Golub et al. (Leukemia) [GST+99], as also from Yeoh et al. [YRS+02] and Ross et al. (ALL subtypes) [RMO+04].

Note that each of the first three pairs of the chosen datasets (as in Table 3.1) are independent and were produced on different microarray platforms. Thus, the merged gene expression matrices obtained from them contain batch effects by default. We consider only genes that

are common between the two samples in the dataset pair, and run all the four preprocessing techniques – GFS, mean-scaling, z-score normalization, and quantile normalization – on these input matrices, and evaluate their effectiveness in dealing with batch effects. To observe the effect of preprocessing on highly heterogenous data, we also use another more heterogeneous dataset from Yeoh et al. [YRS+02] that has 9 disease subtypes (ALL) and normal patients to compare the selected methods. Thus, in total, four sets of input gene expression matrices belonging to three different disease types are used in our analysis.

3.3.2 Approach

In GFS, we transform a raw gene expression matrix by making use of the rank values of genes within each microarray, rather than by using their absolute expression values. Further, we use two quantile thresholds – θ_1 and θ_2 – to assign a fuzzified score to each gene in each patient. Ranks below θ_2 in a microarray are all reduced to a score of zero, those above θ_1 are given a score of 1, and intermediate ranks are interpolated to obtain a score between 0 and 1. In particular, let $r(g_i, p_j)$ be the rank of gene expression of a gene g_i in patient p_j, and $q(p_j, \theta)$ be the rank corresponding to the upper θth quantile of gene expression in patient p_j. Then, the gene fuzzy score $s(g_i, p_j)$ assigned to a gene g_i in patient p_j is given by the following function:

$$s(g_i, p_j) = \begin{cases} 1, & \text{if } q(p_j, \theta_1) \leq r(g_i, p_j) \\ \frac{r(g_i, p_j) - q(p_j, \theta_2)}{q(p_j, \theta_1) - q(p_j, \theta_2)}, & \text{if } q(p_j, \theta_1) > r(g_i, p_j) \geq q(p_j, \theta_2) \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

Apart from the use of rank values in computing transformed scores, GFS also benefits from the fact that it allows for selection of quantile thresholds such that noise from low-ranked genes is safely removed by assigning a score of 0, while genes with very high expression are all treated equally with a score of 1. A graphic representation of GFS is shown in Fig 3.1. For the purpose

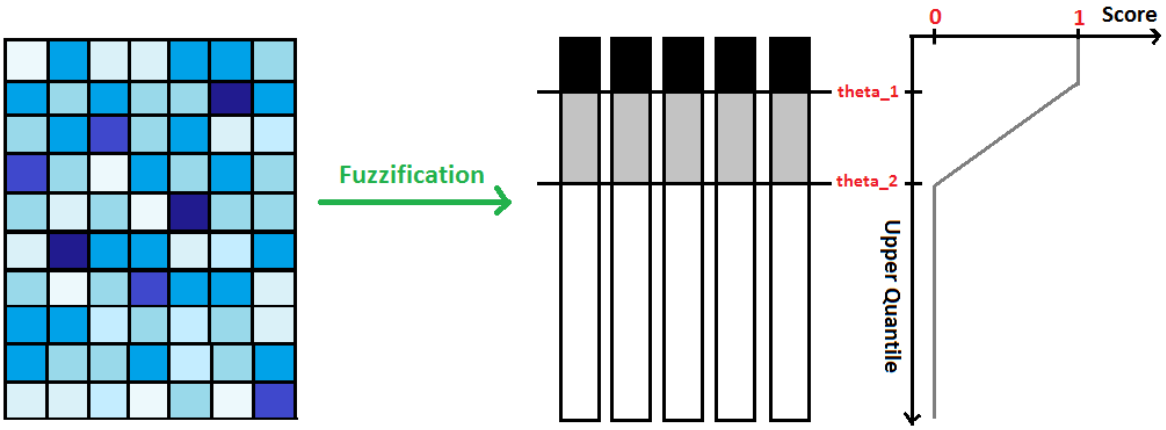


Figure 3.1: GFS normalization methodology

of uniformity in comparison, we fix θ_1 to 5% and θ_2 to 15% for all GFS runs mentioned in this chapter. However, using a θ_1 value between 5% to 10% and θ_2 value between 15% to 20% also leads to similar results.

In evaluating the proposed approach against other normalization techniques discussed earlier, we focus on three salient questions in this dissertation:

1. Does the preprocessing technique produce consistent results across different datasets, provided that they have the same composition phenotypes?
2. What is the quality of the output produced by the preprocessing technique? How well does the preprocessing retain useful information while mitigating obscuring effects?
3. Is the output produced by the technique biologically coherent?

We compared GFS with three standard normalization methods described in the previous section – mean-scaling, z-score normalization, and quantile normalization. The description of our design and approach to each experiment is given in the next section.

3.4 Results

3.4.1 Visualizing data after PCA transformation

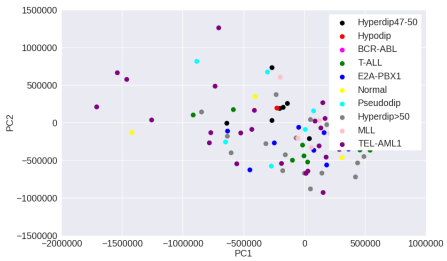
We preprocess the raw gene expression matrices with each of the four methods – mean-scaling, z-score normalization, quantile normalization and GFS. For each method, we select the top 15% genes with maximum variance in the processed matrix, as these are most likely to be the genes contributing to interesting variation. We then reduce the processed matrix to include only these high-variance genes, and apply PCA transformation on the reduced matrix. A scatter plot of the coordinates corresponding to the first two principal components (PC1 and PC2) corresponding to each tissue in the sample is visualized.

A good preprocessing method is expected to show a clear clustering of tissues of the same phenotype, and separation between tissues of different phenotypes. Moreover, the quality of clustering would ideally not be adversely affected by the presence of samples from multiple batches in the data.

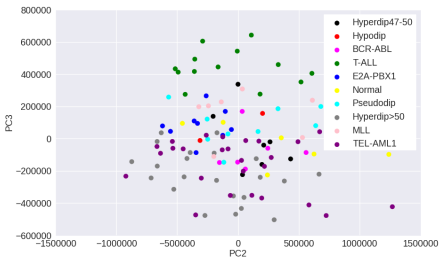
Observations: While in the Leukemia, DMD, and childhood ALL datasets, patients from different batches are clearly separated, GFS (Figure 3.6) shows the best phenotype-wise clustering of patients among all preprocessing techniques. Mean scaling (Figure 3.3) does not perform well on any of the datasets, and in some cases, obscures the separation seen even in raw gene expression (Figure 3.2). This degradation in performance is in line with previous findings [LSS+10]. Z-score normalization shows good performance on DMD and Leukemia (Figure 3.4) datasets, and quantile normalization performs well only on the DMD dataset (Figure 3.5).

In case of the more heterogeneous ALL dataset (9 disease subtypes and normal sample), GFS is the only method to discriminate between patients of the different ALL subtypes (Figures 3.2 - 3.6 (a)).

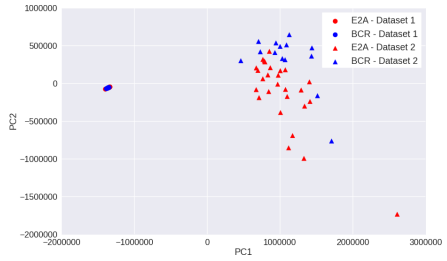
From the PCA scatterplots for all the three datasets with batch effects (Leukemia, DMD, and



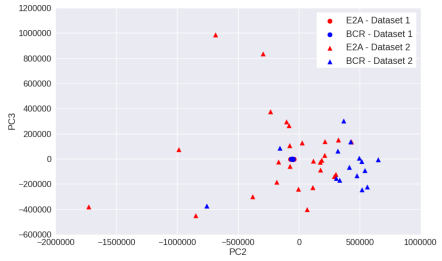
(a) ALL (9 subtypes) : PC1 vs. PC2



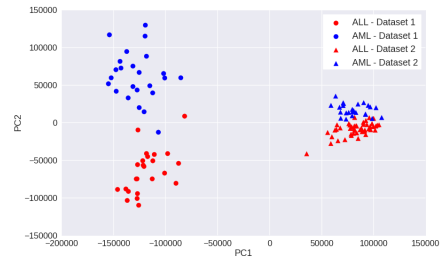
(b) ALL (9 Subtypes) : PC2 vs. PC3



(c) ALL (2 subtypes) : PC1 vs. PC2

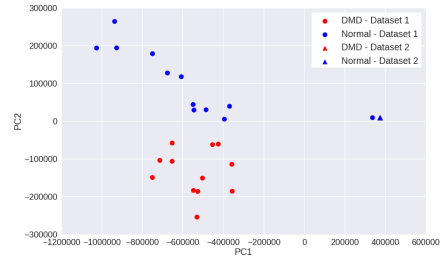


(d) ALL (2 Subtypes) : PC2 vs. PC3

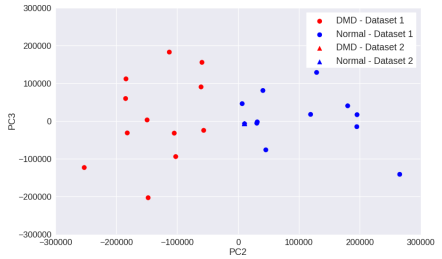


(e) Leukemia : PC1 vs. PC2

(f) Leukemia : PC2 vs. PC3

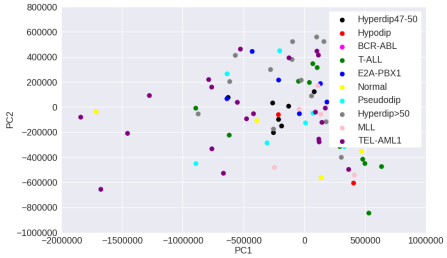


(g) DMD : PC1 vs. PC2

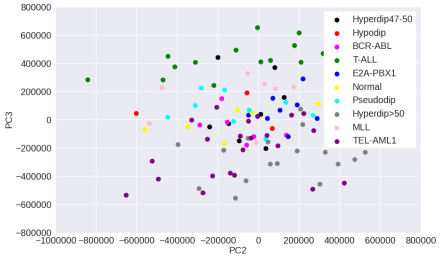


(h) DMD : PC2 vs. PC3

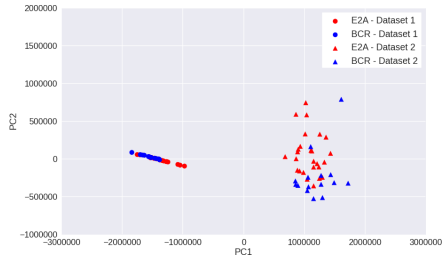
Figure 3.2: Visualisation with PCA scatter plots – Raw expression



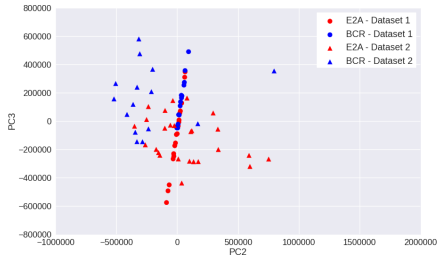
(a) ALL (9 subtypes) : PC1 vs. PC2



(b) ALL (9 Subtypes) : PC2 vs. PC3

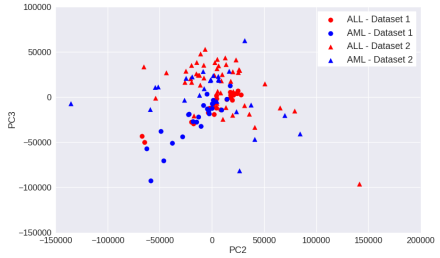


(c) ALL (2 subtypes) : PC1 vs. PC2

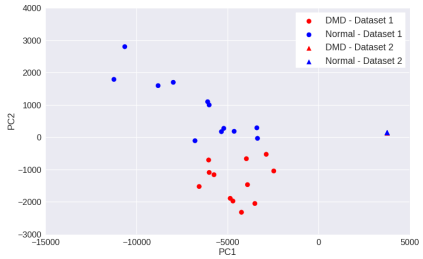


(d) ALL (2 Subtypes) : PC2 vs. PC3

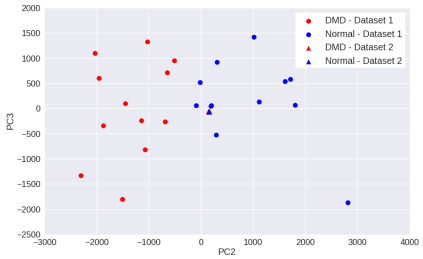
(e) Leukemia : PC1 vs. PC2



(f) Leukemia : PC2 vs. PC3

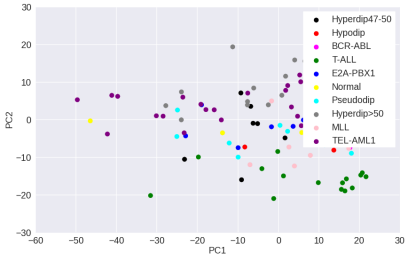


(g) DMD : PC1 vs. PC2

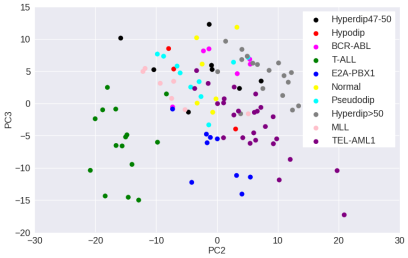


(h) DMD : PC2 vs. PC3

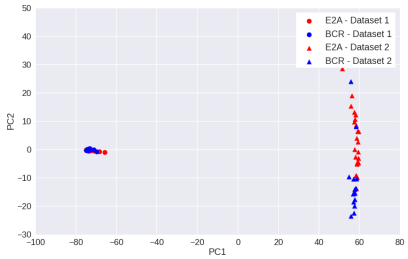
Figure 3.3: Visualisation with PCA scatter plots – Mean-scaled expression



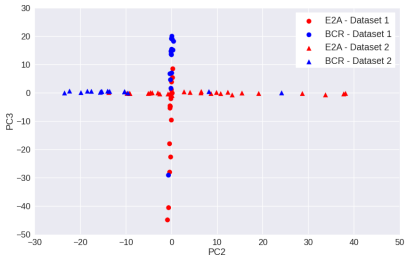
(a) ALL (9 subtypes) : PC1 vs. PC2



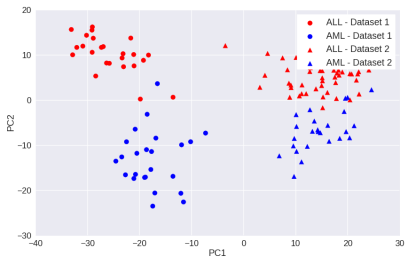
(b) ALL (9 Subtypes) : PC2 vs. PC3



(c) ALL (2 subtypes) : PC1 vs. PC2

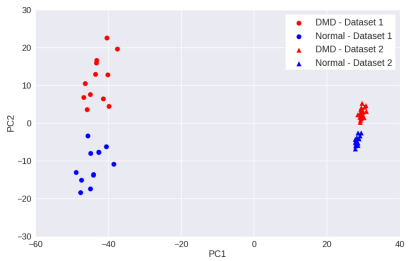


(d) ALL (2 Subtypes) : PC2 vs. PC3

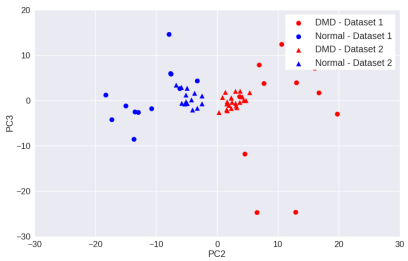


(e) Leukemia : PC1 vs. PC2

(f) Leukemia : PC2 vs. PC3

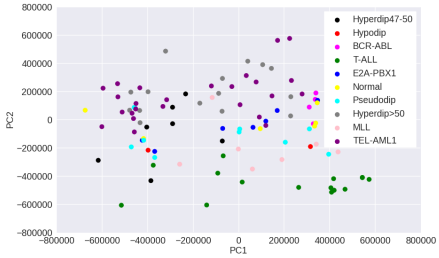


(g) DMD : PC1 vs. PC2

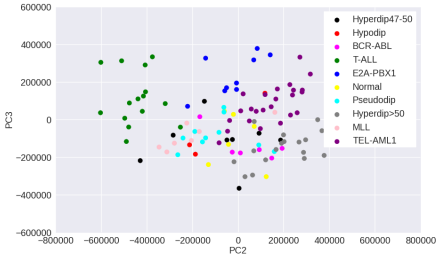


(h) DMD : PC2 vs. PC3

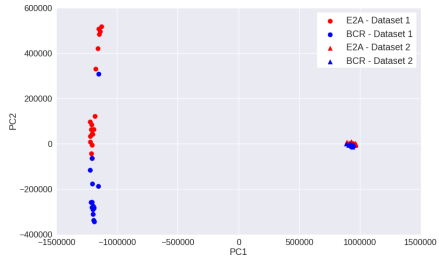
Figure 3.4: Visualisation with PCA scatter plots – Z-score normalized expression



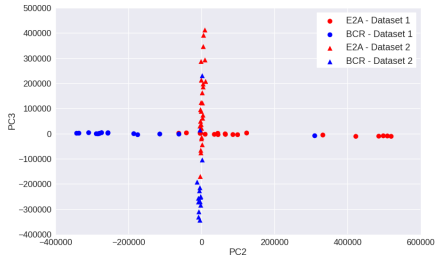
(a) ALL (9 subtypes) : PC1 vs. PC2



(b) ALL (9 Subtypes) : PC2 vs. PC3



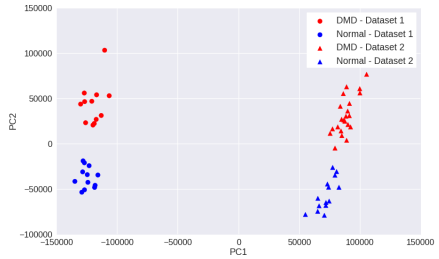
(c) ALL (2 subtypes) : PC1 vs. PC2



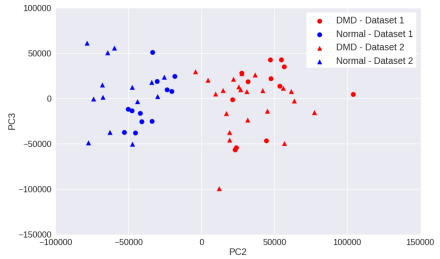
(d) ALL (2 Subtypes) : PC2 vs. PC3

(e) Leukemia : PC1 vs. PC2

(f) Leukemia : PC2 vs. PC3

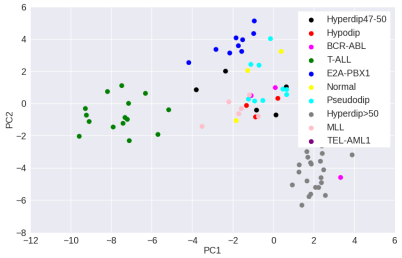


(g) DMD : PC1 vs. PC2

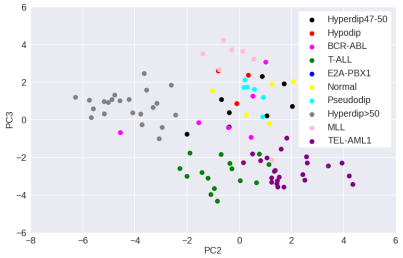


(h) DMD : PC2 vs. PC3

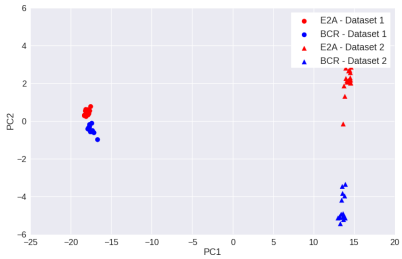
Figure 3.5: Visualisation with PCA scatter plots – Quantile normalized expression



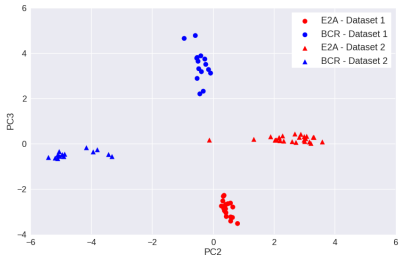
(a) ALL (9 subtypes) : PC1 vs. PC2



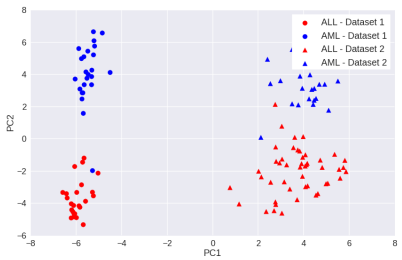
(b) ALL (9 Subtypes) : PC2 vs. PC3



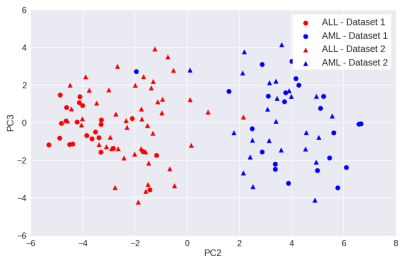
(c) ALL (2 subtypes) : PC1 vs. PC2



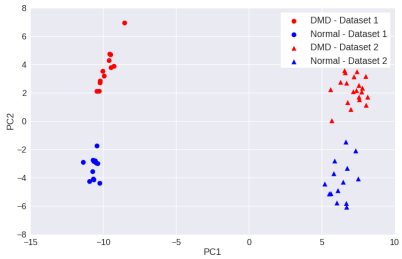
(d) ALL (2 Subtypes) : PC2 vs. PC3



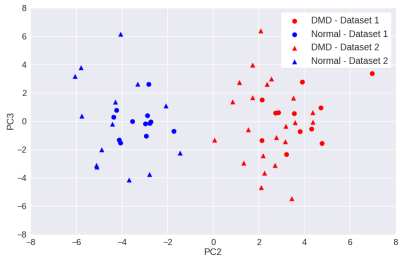
(e) Leukemia : PC1 vs. PC2



(f) Leukemia : PC2 vs. PC3



(g) DMD : PC1 vs. PC2



(h) DMD : PC2 vs. PC3

Figure 3.6: Visualisation with PCA scatter plots – GFS normalized expression

ALL with 2 subtypes), we observed that patients from two batches are always clearly separated along PC1. This implies that the first principal component is highly enriched in batch effects. Therefore, we exclude the first principal component (PC1), and draw scatterplots corresponding to the second and third principal component (PC2, PC3). In PC2 vs PC3 scatterplots, there is much less separation between patients from different batches but belonging to the same phenotype, as compared to that in PC1 vs PC2 scatterplots (Figures 3.2-3.6). This trend is consistent across all three datasets with batch effects. Thus, removing PC1 can be an effective technique to reduce batch effects in gene expression data to a great extent. However, for the more heterogeneous ALL dataset where batch effects are absent, removing PC1 results in loss of important variation information, and subsequently, less clear separation between different phenotypes.

3.4.2 Comparing processing quality

Quality of a preprocessing method is determined by its ability to separate interesting from obscuring variation. An inferior preprocessing method leads to an output in which expression variation across microarrays is confounded with irrelevant information. In contrast, expression variation across microarrays in the output of an ideal preprocessing method corresponds to interesting biological variation alone.

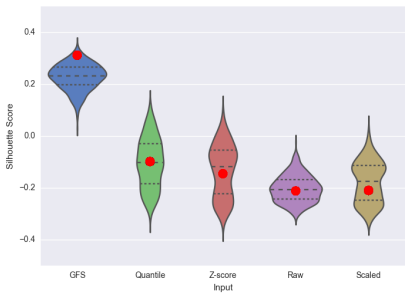
Experiment: We estimate the quality of preprocessing methods with respect to the capability of their transformed output to separate patients of different phenotypes. In particular, we randomly select 15% of the genes, reduce the processed matrix to include the selected genes, and apply PCA on the resultant matrix. The PCA co-ordinates of all patients are then used to compute a clustering performance metric called the silhouette score. The silhouette score is calculated based on the mean intra-cluster distance a and the mean nearest-cluster distance b for each patient, as $(b-a)/max(a,b)$ [Rou87]. The score ranges from -1 to 1. In general, a higher silhouette score indicates a better clustering.

For the ALL dataset with 9 subtypes, co-ordinates corresponding to the first three principal components are used, while for the other three datasets with batch effects, co-ordinates corresponding to only the second and third principal components are used. This is repeated over 1000 iterations, and the distribution of silhouette scores corresponding to each preprocessing method is used to infer the quality of clusters formed by its transformed output.

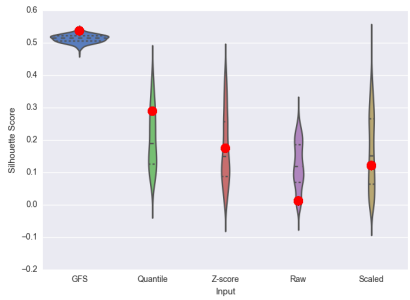
Observations: For all the four datasets, the distribution of silhouette scores obtained using randomly chosen 15% genes is stable at a higher value in case of GFS, in comparison to other preprocessing methods (see Figure 3.7). This shows that the assigned scores to each microarray-gene pair after GFS preprocessing are more relevant to the interesting variation in gene expression and thus, even randomly chosen features are better able to capture the phenotype-based clusters. Moreover, the reference silhouette scores obtained from the top 15% variance genes in GFS processed matrices are consistently higher than the 75th percentile score of its null distribution obtained from random 15% genes, across all datasets (Figure 3.7). For quantile normalization, while the silhouette scores obtained from its top 15% variance genes are also consistently higher than the 75th percentile score of the corresponding null distribution, these observed silhouette scores are consistently lower than those for GFS. On the other hand, the silhouette scores derived using the top 15% variance genes in z-score normalized and raw expression are lower than the 75th percentile score of their corresponding null distributions in the DMD dataset and ALL dataset with 2 subtypes. The silhouette score computed on top 15% variance genes in mean-scaled expression data is lower than the median score of its null distribution in all datasets. This shows GFS-processed expression values are more effective than the other methods.

Table 3.2: Silhouette Scores with respect to phenotype labels obtained using the transformed expression values from top 15% variance genes on applying different preprocessing techniques (using first three principal components)

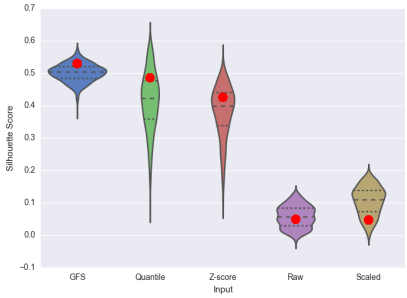
	Raw	Scaled	Z-Score	Quantile	GFS
ALL (9 subtypes)	-0.212	-0.209	-0.145	-0.099	**0.312**
ALL (2 subtypes)	0.009	0.027	0.043	0.070	**0.145**
DMD	0.025	0.044	0.096	0.202	**0.203**
Leukemia	0.153	0.128	0.177	0.227	**0.289**



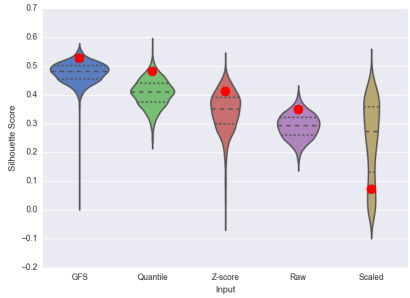
(a) ALL (9 Subtypes)



(b) ALL (2 Subtypes)



(c) DMD



(d) Leukemia

Figure 3.7: Null distributions of silhouette scores obtained with raw and processed expression matrices taking 15% random genes as features (the three dashed lines show 25th quartile, median and 75th quartile, while the red dot indicates the score obtained from top 15% variance genes)

Table 3.3: Silhouette Scores with respect to phenotype labels obtained using the transformed expression values from top 15% variance genes on applying different preprocessing techniques (using only PC2 and PC3, ignoring PC1)

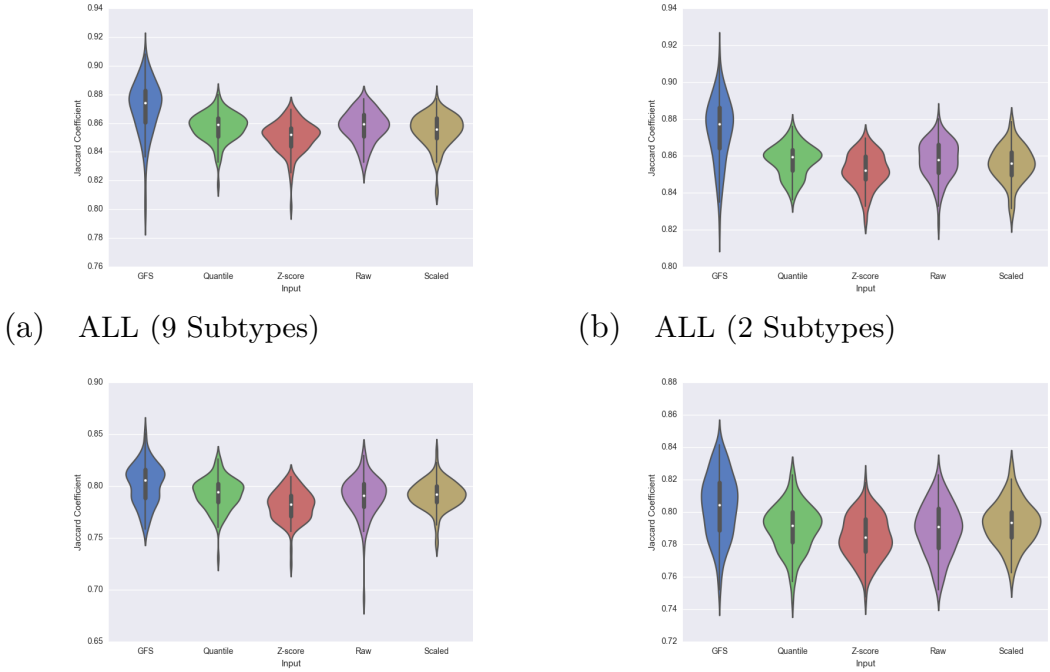
	Raw	Scaled	Z-Score	Quantile	GFS
ALL (9 subtypes)	-0.243	-0.186	0.017	0.027	**0.217**
ALL (2 subtypes)	0.012	0.121	0.176	0.289	**0.538**
DMD	0.049	0.047	0.426	0.486	**0.530**
Leukemia	0.349	0.072	0.412	0.482	**0.528**

The silhouette scores obtained from the PCA transformed co-ordinates of patients using the top 15% high-variance genes are recorded in Table 3.2 and 3.3. In all datasets, with and without the first principal component (which is often the richest in batch effects), GFS is seen to have a better score relative to other processing methods. Also, in the three datasets with batch effects, removing PC1 improves phenotype-wise clustering, while in the heterogeneous ALL dataset with no batch effects, removing PC1 leads to discarding important variation and thus a reduction in clustering performance.

3.4.3 Comparing consistency

It is important that a reliable preprocessing method produces an output that remains consistent in multiple runs over datasets of the same type. For instance, if two datasets of the same disease are independently transformed by a preprocessing method, and the genes indicated to have the highest contribution to interesting variation have very little overlap, it is natural to infer that the variation is confounded by noise and the genes are likely to be false positives. In contrast, consistency in such output lends confidence that the preprocessing method is indeed reliable, since similarity (in terms of sample phenotypes) in input ensures similarity (in terms of differentially expressed genes) in output. Thus, a preprocessing technique assigning meaningfully transformed expression values should indicate a consistent set of high-variance genes, when applied to different datasets with the same phenotype distribution.

Experiment: In order to evaluate the consistency of different preprocessing methods, we split each dataset into two datasets such that each contains the same number of patients of each phenotype, independently apply the preprocessing technique on the resultant split data, and obtain the two resulting lists of the top 15% high-variance genes from the splits. Further, we apply PCA to the normalized data, and remove genes that have a coefficient of zero in all of the first three principal components for the ALL dataset with 9 disease subtypes. For the other three batch effects-ridden datasets, we only remove genes that have a coefficient of zero in the second and third principal component. This process is repeated 100 times using different



(a) ALL (9 Subtypes) (b) ALL (2 Subtypes)

(c) DMD (d) Leukemia

Figure 3.8: Consistency of preprocessed output - Jaccard coefficient distribution of top variance-contributing genes on comparing 100 data splits

splits of each dataset. We then examine the distribution of similarity (measured in terms of the jaccard coefficient) between the two gene lists.

Observations: A consistent preprocessing technique is expected to demonstrate a high overlap in high-variance genes. It is seen that the distribution of jaccard coefficient when the split datasets are processed using GFS, is stable at an equal or higher value than the other methods in all the datasets (Figure 3.8). The other methods fluctuate in performance and, in some cases, show worse consistency than raw gene expression.

3.4.4 Comparing biological coherence

For a phenotype to manifest, the causal genes often co-ordinate with other genes, and seldom act alone. Therefore, genes contributing to interesting variation in data are more likely to

be connected to each other in biological pathways. Thus, we expect that a more biologically coherent preprocessing technique will result in high-variance genes that induce significantly more and/or bigger subnetworks on known biological pathways.

Experiment: We assess the biological coherence of the preprocessing methods by examining the subnetwork size distribution obtained when high-variance genes are used to induce subnetworks on pathways. The subnetwork size distribution for each processing method is obtained as follows:

1. Preprocess the gene expression matrix using the chosen technique.
2. Select top 15% genes with maximum variance across patients.
3. Reduce processed expression matrices to only include the selected genes.
4. Perform a PCA transformation on the reduced matrix, and list genes with non-zero coefficients in any of the first three principal components.
5. Using genes in step 4, induce subnetworks on known pathways from the PathwayAPI database [SDGW10] and store the subnetwork size distribution.

To generate the null model, step 2 is replaced with randomly selecting 15% of all genes, and steps 1-5 are repeated over 1000 iterations. Finally, for each subnetwork size, a p-value is calculated as the proportion of subnetwork frequencies in the null model found to be greater than the frequency from original distribution.

The same analysis is repeated for the three datasets with batch effects by modifying step 4 to include only those genes that have a non-zero coefficient in the second or third principal component.

Observations: The distribution of subnetwork sizes induced by the top 15% variance genes are shown in Figure 3.9 (using the first three principal components) and Figure 3.10 (using PC2 and PC3 only). The figures show the actual subnetwork count distribution across different subnetwork sizes, while the inset figures show the corresponding percentage frequencies. In the

Leukemia dataset and ALL dataset with 2 subtypes, GFS has the highest percentage frequency of subnetworks of size greater than or equal to 5 and, in most datasets, GFS induces more subnetworks overall.

Table 3.4: ALL (2 Subtypes) – Significance comparison of size of subnetworks induced by high-variance genes in preprocessed output; p_1 = p-value using first three PCs, p_2 = p-value using PC2, PC3 only

	Raw			**Scaled**			**Z-score**			**Quantile**			**GFS**		
size	**freq**	p_1	p_2	**freq**	p_1	p_2	**freq**	p_1	p_2	**freq**	p_1	p_2	**freq**	p_1	p_2
2	89	0.620	0.604	94	0.476	0.482	93	0.502	0.509	92	0.527	0.532	82	0.128	0.105
3	44	0.646	0.663	44	0.646	0.663	50	0.419	0.430	44	0.646	0.663	46	0.030	0.030
4	31	0.196	0.173	32	0.162	0.153	28	0.312	0.316	29	0.268	0.259	33	0.001	0.001
5	14	0.429	0.398	13	0.509	0.487	18	0.169	0.156	17	0.226	0.193	20	0.001	0.002
6	12	0.082	0.101	15	0.018	0.024	12	0.082	0.101	14	0.032	0.038	6	0.045	0.043
7	7	0.133	0.117	7	0.133	0.117	6	0.224	0.220	9	0.035	0.030	14	0.000	0.000
8	6	0.050	0.043	6	0.050	0.043	7	0.019	0.017	5	0.098	0.097	5	0.006	0.005
9	2	0.324	0.345	2	0.324	0.345	1	0.594	0.607	4	0.061	0.069	3	0.043	0.031
10	3	0.076	0.075	1	0.451	0.449	1	0.451	0.449	1	0.451	0.449	1	0.177	0.168
11	1	0.350	0.357	-	-	-	-	-	-	-	-	-	1	0.129	0.117
12	1	0.300	0.278	1	0.300	0.278	1	0.300	0.278	1	0.300	0.278	1	0.066	0.083
13	-	-	-	1	0.233	0.264	1	0.233	0.264	1	0.233	0.264	-	-	-
14	-	-	-	-	-	-	-	-	-	-	-	-	1	0.021	0.019
15	-	-	-	1	0.156	0.145	1	0.156	0.145	1	0.156	0.145	-	-	-
16	1	0.133	0.139	3	0.005	0.002	2	0.038	0.027	2	0.038	0.027	1	0.002	0.005
17	3	0.006	0.001	1	0.093	0.099	2	0.020	0.018	1	0.093	0.099	1	0.003	0.002
18	-	-	-	1	0.077	0.070	1	0.077	0.070	2	0.013	0.012	1	0.000	0.001
19	3	0.001	0.001	-	-	-	1	0.008	0.007	-	-	-	2	0.000	0.000
20	1	0.035	0.041	-	-	-	-	-	-	-	-	-	-	-	-
21	-	-	-	-	-	-	-	-	-	-	-	-	1	0.000	0.000
22	-	-	-	1	0.008	0.007	-	-	-	-	-	-	1	0.000	0.000

From the low p-values in Tables 3.4, 3.5, 3.6, 3.7, we observe that the significance of frequencies is high for subnetworks induced by GFS, regardless of their size. Further, comparison with other methods shows that the frequency of subnetworks induced by high-variance genes in GFS-processed datasets is much more significant than those induced on datasets processed with other methods and raw gene expression.

Hence, we infer that GFS-transformed output is highly biologically coherent. Moreover, we

Table 3.5: DMD – Significance comparison of size of subnetworks induced by high-variance genes in preprocessed output; p_1 = p-value using first three PCs, p_2 = p-value using PC2, PC3 only

	Raw			**Scaled**			**Z-score**			**Quantile**			**GFS**		
size	**freq**	p_1	p_2	**freq**	p_1	p_2	**freq**	p_1	p_2	**freq**	p_1	p_2	**freq**	p_1	p_2
2	74	0.901	0.903	970	0.429	0.415	57	0.995	0.995	104	0.298	0.278	85	0.015	0.009
3	83	0.004	0.007	44	0.649	0.644	23	0.999	0.999	40	0.794	0.777	81	0.000	0.000
4	19	0.817	0.799	22	0.660	0.643	17	0.894	0.894	18	0.861	0.861	28	0.002	0.004
5	15	0.337	0.324	11	0.692	0.665	12	0.588	0.586	13	0.499	0.485	18	0.001	0.000
6	7	0.536	0.521	11	0.147	0.145	7	0.536	0.521	10	0.213	0.206	11	0.000	0.000
7	8	0.084	0.106	12	0.005	0.005	4	0.521	0.519	10	0.021	0.022	9	0.000	0.000
8	7	0.025	0.018	6	0.053	0.045	3	0.379	0.392	6	0.053	0.045	3	0.019	0.011
9	1	0.598	0.615	5	0.029	0.031	3	0.182	0.148	7	0.004	0.008	4	0.000	0.002
10	2	0.209	0.229	1	0.449	0.467	3	0.089	0.084	2	0.209	0.229	2	0.007	0.007
11	2	0.134	0.140	5	0.001	0.001	1	0.372	0.372	2	0.134	0.140	1	0.012	0.006
12	4	0.006	0.003	3	0.021	0.027	-	-	-	3	0.021	0.027	1	0.005	0.003
13	3	0.017	0.016	2	0.078	0.077	3	0.017	0.016	2	0.078	0.077	2	0.000	0.001
14	3	0.011	0.012	1	0.200	0.189	3	0.011	0.012	-	-	-	1	0.000	0.002
15	2	0.054	0.039	3	0.012	0.009	1	0.181	0.164	1	0.181	0.164	2	0.000	0.000
16	3	0.004	0.002	-	-	-	-	-	-	1	0.133	0.142	2	0.000	0.000
17	1	0.104	0.091	-	-	-	-	-	-	2	0.016	0.019	1	0.000	0.000
18	1	0.097	0.072	-	-	-	-	-	-	1	0.097	0.072	-	-	-
19	1	0.058	0.073	-	-	-	-	-	-	-	-	-	1	0.000	0.000
20	1	0.041	0.040	1	0.041	0.040	-	-	-	-	-	-			
21	-	-	-	-	-	-	1	0.026	0.038	-	-	-	1	0.000	0.000
28	-	-	-	1	0.001	0.000	-	-	-	-	-	-	-	-	-

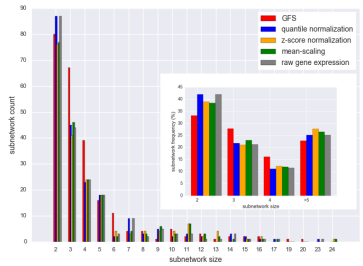
Table 3.6: ALL (9 subtypes) - Significance comparison of size of subnetworks induced by high-variance genes in preprocessed output; p = p-value of the frequency using first three principal components

	Raw		**Scaled**		**Z-score**		**Quantile**		**GFS**	
size	**freq**	p	**freq**	p	**freq**	p	**freq**	p	**freq**	p
2	87	0.672	77	0.861	76	0.876	87	0.672	80	0.071
3	44	0.621	46	0.545	41	0.722	45	0.577	67	0.000
4	24	0.483	24	0.483	24	0.483	23	0.546	39	0.000
5	18	0.105	18	0.105	18	0.105	18	0.105	16	0.001
6	3	0.890	2	0.958	4	0.804	2	0.958	11	0.000
7	9	0.025	4	0.408	3	0.588	9	0.025	4	0.029
8	2	0.492	3	0.289	4	0.144	3	0.289	4	0.013
9	5	0.017	6	0.004	4	0.057	5	0.017	1	0.170
10	3	0.062	3	0.062	4	0.021	2	0.165	5	0.000
11	3	0.038	7	0.001	7	0.001	3	0.038	2	0.015
12	1	0.289	3	0.021	2	0.092	2	0.092	3	0.001
13	1	0.230	2	0.059	4	0.007	-	-	1	0.011
14	3	0.005	1	0.203	-	-	3	0.005	2	0.000
15	1	0.193	1	0.193	1	0.193	2	0.047	2	0.002
16	1	0.124	1	0.124	2	0.031	1	0.124	2	0.000
17	1	0.122	1	0.122	-	-	1	0.122	-	-
19	-	-	-	-	-	-	-	-	1	0.000
20	-	-	-	-	-	-	-	-	1	0.000
23	1	0.006	-	-	-	-	1	0.006	-	-
24	-	-	1	0.003	1	0.003	-	-	-	-

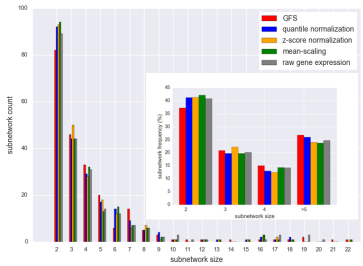
Table 3.7: Leukemia – Significance comparison of size of subnetworks induced by high-variance genes in preprocessed output; p_1 = p-value using first three PCs, p_2 = p-value using PC2, PC3 only

	Raw			**Scaled**			**Z-score**			**Quantile**			**GFS**		
size	**freq**	p_1	p_2	**freq**	p_1	p_2	**freq**	p_1	p_2	**freq**	p_1	p_2	**freq**	p_1	p_2
2	44	0.994	0.993	57	0.920	0.924	47	0.987	0.986	59	0.893	0.894	47	0.570	0.582
3	34	0.557	0.584	23	0.954	0.945	27	0.842	0.844	26	0.885	0.883	34	0.073	0.075
4	14	0.664	0.679	14	0.664	0.679	15	0.588	0.589	17	0.454	0.471	15	0.046	0.044
5	7	0.597	0.579	6	0.700	0.686	8	0.474	0.465	7	0.597	0.579	4	0.244	0.253
6	5	0.279	0.318	2	0.762	0.779	1	0.904	0.925	4	0.423	0.462	6	0.011	0.013
7	1	0.688	0.696	1	0.688	0.696	4	0.130	0.149	1	0.688	0.696	2	0.159	0.166
8	4	0.048	0.039	3	0.119	0.104	-	-	-	1	0.487	0.500	1	0.259	0.220
9	1	0.384	0.369	2	0.153	0.159	3	0.051	0.047	4	0.014	0.011	3	0.021	0.017
10	1	0.285	0.252	2	0.107	0.098	1	0.285	0.252	1	0.285	0.252	2	0.032	0.031
11	1	0.201	0.224	-	-	-	-	-	-	1	0.201	0.224	2	0.020	0.017
12	-	-	-	-	-	-	-	-	-	-	-	-	1	0.030	0.028
15	-	-	-	-	-	-	-	-	-	-	-	-	1	0.006	0.001

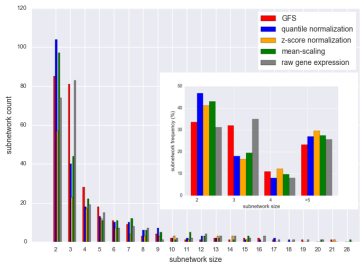
observe that on excluding the batch effects-enriched PC1 from the analysis, the p-values corresponding to larger subnetwork sizes are lower than those of smaller sizes, indicating higher significance, and hence greater biological coherence, of the large subnetwork sizes.



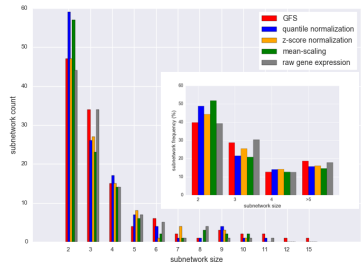
(a) ALL (9 Subtypes)



(b) ALL (2 Subtypes)

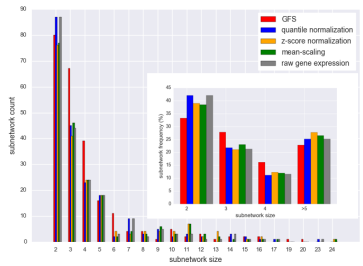


(c) DMD

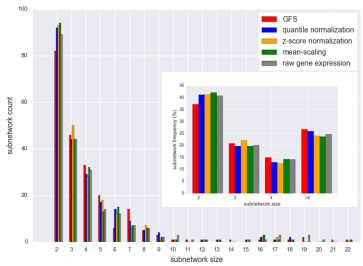


(d) Leukemia

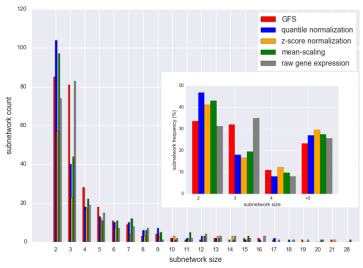
Figure 3.9: Distribution for size of subnetworks induced by high-variance genes in different preprocessed outputs (using first three components); Inset figure shows the same as percentage frequency



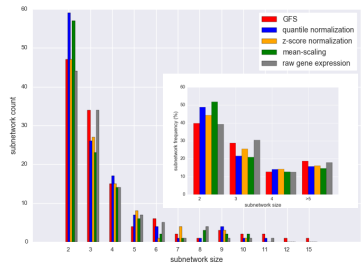
(a) ALL (9 Subtypes)



(b) ALL (2 Subtypes)



(c) DMD

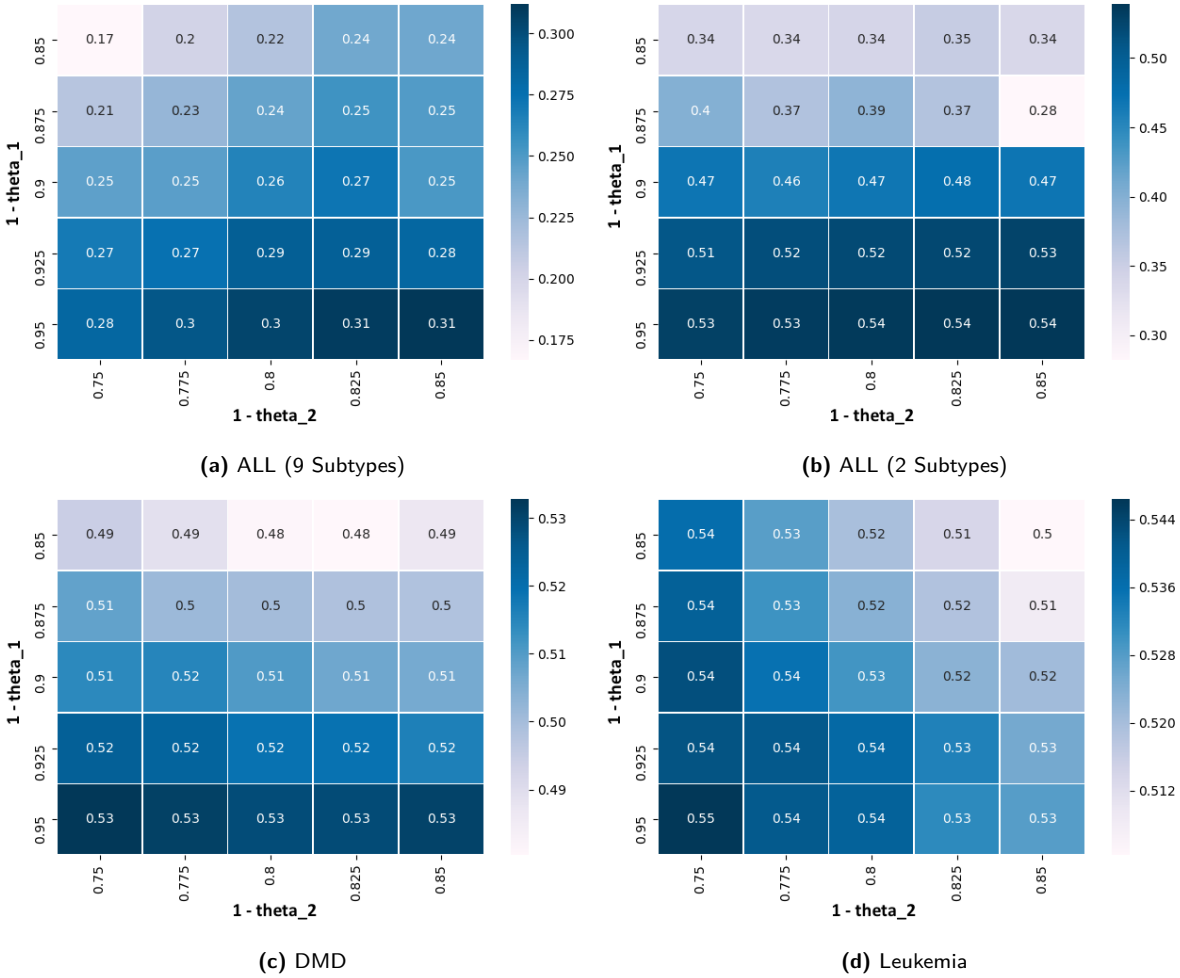


(d) Leukemia

Figure 3.10: Distribution for size of subnetworks induced by high-variance genes in different preprocessed outputs (using PC2, PC3 only, ignoring PC1 from analysis); Inset figure shows the same as percentage frequency

3.4.5 Effect of θ_1 and θ_2 thresholds on the performance of GFS

We examined the effect of variation in θ_1 and θ_2 on the performance of GFS by computing the silhouette scores corresponding to different θ_1, θ_2 combinations.



(a) ALL (9 Subtypes) **(b)** ALL (2 Subtypes)

(c) DMD **(d)** Leukemia

Figure 3.11: Heatmaps of silhouette scores on different datasets after normalization with GFS

Fig 3.11 shows the resultant heatmaps of these silhoeutte scores for each of our datasets. We make two observations from this:

1. The silhouette score is the highest for the values $\theta_1 = 5\%$, $\theta_2 = 15\%$ in all the three datasets other than Leukemia. However, the silhouette scores in the Leukemia dataset are high across all threshold combinations, and the difference between all the silhouette scores is negligible.

2. The extent to which the silhouette score degrades on using non-optimal θ_1, θ_2 thresholds depends on the particular dataset – possibly the underlying biology, and precision of the platform in measuring expression in the higher and lower ranges.

3.4.6 Selecting θ_1 and θ_2: Are we throwing away critical information?

As an example, say that θ_1 and θ_2 are set to 5% and 15%. Then, for a given microarray, 5% of the genes (with the highest expression values) are given a score of 1, 85% of the genes (with the lowest expression values) are given a score of 0, while only the intermediate genes (10%) are given scores between 0 and 1 by linear interpolation of their ranks. So are we throwing away critical information?

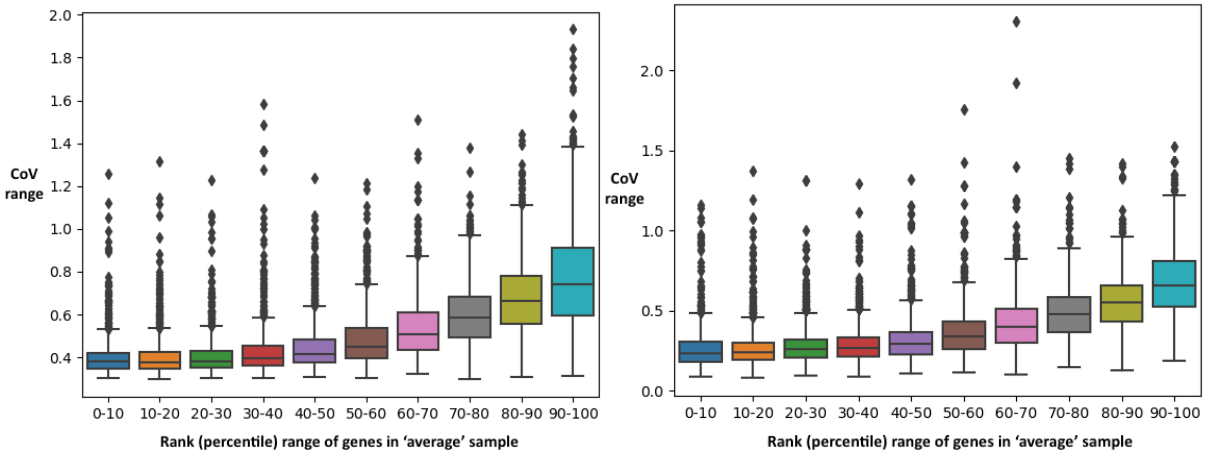
Firstly, notice that the genes which get a score of 0, will differ across microarrays, as the phenotype varies. This implies that for each gene, we have a vector of fuzzy scores - which may be 0, 1, or some other fractional number between 0 and 1 - corresponding to all the microarrays in which it has been measured.

Still, a large number of low expression genes are given a score of 0 in multiple microarrays. This is justified because low expression genes tend to have a very high Coefficient of Variation (CoV), and often introduce noise rather than true signals in the data. We illustrate this with the following procedure:

Duchenne Muscular Dystrophy (DMD) – Haslett dataset:

1. Measure the CoV of each gene across the disease sample, and across the normal sample.
2. Create an 'average' normal and 'average' disease sample, by averaging the expression of each gene across all microarrays in normal and disease phenotypes respectively. For the normal and disease 'average' samples, rank genes from the highest to lowest expression level.
3. For the normal and disease samples, plot the range of CoV observed in genes with the

top 10% highest expression, genes with ranks between top 10-20%, 20-30%, and so on. We then observe how the CoV varies across the different gene rank groups.



(a) Haslett dataset: Normal sample

(b) Haslett dataset: DMD sample

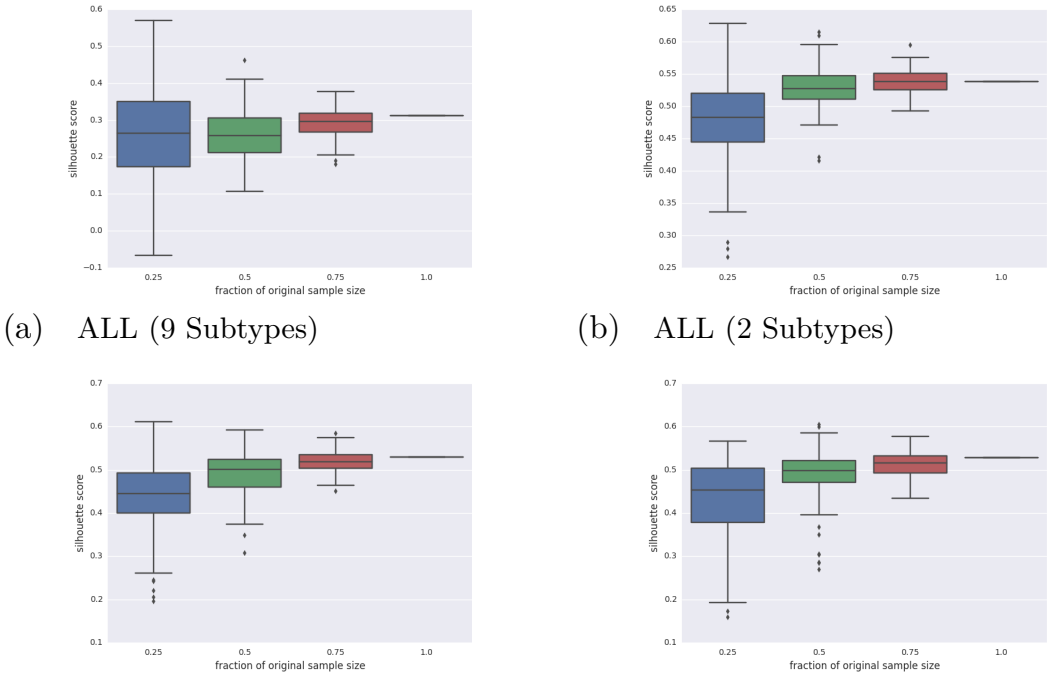
We see that ranks from 20-30% onwards show increasingly noticeable increase in CoV, indicating noise in expression measurements. This is particularly significant because DMD is caused by the absence of dystrophin, a protein that helps keep muscle cells intact, and hence a homogeneous disease with little expected variation.

This illustration serves two purposes: (a) It shows that a vast majority of lower expressed genes have very high CoV, and the expression of such genes often does not contribute very useful biological information. (b) A rank-wise boxplot of CoV across samples (such as the one above) may be a helpful indicator for defining suitable values for θ_1 and θ_2.

3.4.7 Effect of sample size on performance of GFS

To examine the effect of sample size on GFS, we randomly selected samples of the size of 0.25, 0.50, 0.75 times the original sample size over 100 iterations. We then noted the range of silhouette scores obtained from the iterations for each sample size. (For the heterogeneous ALL dataset, the first three PCs were used to calculate the silhouette scores, while for the other datasets, only the second and third PCs were used.) As expected, Figure 3.13 shows that

the clustering performance improves with increase in sample size. Interestingly, the boxplots in Figure 3.13, interpreted together with Tables 3.2 and 3.3, also indicate that the median performance of GFS when provided with even 0.25 times of the entire sample size is still comparable with, and often better than, that of other normalization methods when they are supplied with the entire sample size.



(a) ALL (9 Subtypes) (b) ALL (2 Subtypes)

(c) DMD (d) Leukemia

Figure 3.13: Effect of sample size on clustering performance of GFS

3.5 Conclusion

An effective preprocessing technique is expected to transform the gene expression matrix such that data of the same phenotype from different sources is made similar. This can be achieved by removing or accounting for obscuring noise in gene expression measurement, and retaining interesting variation relevant to properties of biological interest. Such a processing is essential to ensure reliable downstream analysis of gene expression data. However, popular normalization

techniques do not necessarily improve the quality of expression data, and sometimes even exacerbate the issue by mistaking real variation for noise and discarding it.

We discussed a new approach, Gene Fuzzy Score, to address this issue and compared it with other popular preprocessing methods with respect to three important criteria. First, we assessed the capability of the transformed output of each technique to resolve differences in phenotypes within the dataset. Secondly, we estimated the consistency of their output when presented with different datasets with the same phenotype distribution. Finally, we analysed the distributions of size of subnetworks induced by genes indicated to be sources of interesting variation in each processed expression matrix. In each of these aspects, GFS was successful in improving the transformation outcome, proving its applicability in datasets with batch effects and heterogeneity. Moreover, the performance of GFS improves with increase in sample size.

A recurring observation from our experiments is that in datasets with significant batch effects, the batch effects are generally captured by the first principal component in PCA. Thus, applying a PCA transformation and excluding the first principal component from subsequent analysis leads to significant reduction in batch effects in any dataset, and improves the performance of all preprocessing techniques. Further, we note that GFS outperforms other methods irrespective of whether this additional step is implemented.

Another merit of GFS is the interpretability of its transformed outcome. A biologist may quickly understand how highly the gene is ranked in a particular patient. For example, when a gene has a GFS score of 0.5 in a patient, it means the gene is in the top 10% most highly expressed genes in that patient (assuming $\theta_1$and θ_2 are set at 5% and 15% respectively). Thus, apart from being a robust and effective preprocessing technique, GFS is also easily interpretable.

Chapter 4

SPSNet: Sub-Population Sensitive Network-based Analysis of Heterogeneous Expression Data

"An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem."

– John Tuckey

4.1 Introduction

Diseases and biological processes are highly heterogeneous due to variation in the underlying mechanisms. Regardless of its origin, heterogeneity is often implicit and undeclared, as incomplete knowledge prevents the accurate identification of subpopulations in a phenotype. Undeclared heterogeneity in transcriptomic data can arise from biological variation such as diversity of disease subtypes, treatment subgroups, time-series gene expression, nested experimental conditions, as well as technical variation due to batch effects, platform differences in integrated meta-analyses, etc. Unless the underlying heterogeneity is appropriately considered, comprehensive analysis of disease mechanisms is hindered, potentially resulting in misleading conclusions. In general, a systematic understanding of the biological basis of heterogeneity is

critical in many practical contexts, e.g.:

- developing effective treatments by precise identification of dysregulated mechanisms in distinct disease subtypes.
- identifying differences in the molecular states of stem cells resulting in distinct lineage progression, to better understand organ development and regeneration; and
- detecting and eliminating the effects of intrinsic heterogeneity (e.g., cell cycle differences across cells, variation in cellular composition), which can hinder the discovery of physiologically relevant variation in the gene expression profiles.

A systematic analysis of non-biological and extrinsic heterogeneity is also useful in many cases, even when analyzing apparently homogeneous experimental conditions, for:

- extracting knowledge with greater confidence from a meta-analysis of independently generated datasets;
- discovering unsuspected anomalies or technical errors; and
- identifying and eliminating factors most influenced by extrinsic elements and/or batch effects.

Yet, handling heterogeneity in gene expression is a major problem with few and ineffective solutions.

4.2 Background

Previous studies have attempted to unravel heterogeneity using unsupervised techniques to identify gene expression-based, subtype-specific, molecular signatures [AED+00, SPT+01, MdRD+13, BSW+11]. In these approaches, gene expression data is typically subjected to hierarchical clustering or orthogonal transformation, and subpopulations in the sample are inferred using observations on the patterns of variation in gene expression. However, analysis

carried out at the individual-gene level prevents a systemic view of the underlying mechanisms, and leaves considerable room for subjective, and potentially incorrect, interpretation of the underlying biological mechanisms. It also leads to a high false-positive rate, and low reproducibility [ZZZ+09]. Notably, Venet et al. showed that, in the case of breast cancer, such gene-based signatures are no better than randomly chosen signatures [VDD11].

Several methods have been proposed for analyzing differential expression between homogeneous phenotypes at the level of biological pathways and subnetworks, including Over-Representation Analysis (ORA)[KDOK02], Gene Set Enrichment Analysis (GSEA)[STM+05], Gene Graph Enrichment Analysis (GGEA)[GCK+11], and Differential Expression Analysis in Pathways (DEAP)[HHS+13]. However, it has been demonstrated that, when analyzing independent datasets consisting of identical phenotypes, these methods produce results that considerably differ between the independent datasets, demonstrating lack of consistency. This issue arises mainly due to ineffective data normalization and/or the utilization of incorrect null hypothesis/distribution. Two recent methods overcome these issues to yield consistent results across data sets: SNet[SDGW11] and its refinement PFSNet[LW13]. However, these methods are designed to analyze only homogeneous phenotypes without subclasses.

We propose a generalized approach to analyze heterogeneity in gene expression data, and obtain subtype-specific signatures based on the differential gene expression of subnetworks in biological pathways rather than individual genes. Our generalization of PFSNet is termed SPSNet (SubPopulation-sensitive PFSNet). While PFSNet reports subnetworks that are differentially expressed between two samples representing homogeneous phenotypes, SPSNet makes no assumptions on the homogeneity of given phenotypes and automatically identifies subnetworks that are differentially expressed between the subpopulations within phenotypes. Thus, SPSNet serves a two-fold purpose: (i) when heterogeneity is biological in nature, it provides insights into how subpopulations within a sample set indicating diverse biological mechanisms manifest as sample subphenotypes; and (ii) in the presence of extrinsic or non-biological heterogeneity, it amplifies these effects, facilitating identification and elimination of

factors extraneous to biology of the phenotypes being studied. We demonstrate the utility and performance of SPSNet using publicly available gene expression datasets containing disease heterogeneity, batch effects, and varied experimental treatments.

4.3 Methods

4.3.1 Data

- Leukemia dataset by Yeoh et al. [YRS+02]: We use the normal class (12 training, 6 test patients) and two large ALL subtypes, TEL-AML1 (52 training, 25 test patients), T-ALL (29 training, 15 test patients) from this microarray dataset.
- Hepatocellular Carcinoma (HCC) dataset by Roessler et al. [RJB+10]: This microarray dataset consists of 247 tumor and 241 adjacent non-tumor samples.
- HCC dataset by Burchard et al. [BZL+10]: This microarray dataset consists of 268 tumor and adjacent 249 non-tumor samples.
- TCGA RCC dataset—[N+13]: This microarray dataset contains 30 normal and 30 clear cell Renal Cell Carcinoma (ccRCC) tumor samples.
- We obtained human pathway information from the PathwayAPI database which consists of 300 human pathways [SDGW10].

4.3.2 Notations and terminology

- G: the set of all genes g_i ($i \in \{1, 2, \ldots, n\}$) whose expression has been measured
- P_C, $P_{\neg C}$: set of patients in the control and test phenotypes respectively, where the phenotypes potentially contain undeclared sources of heterogeneity. The objective of SPSNet is to identify gene subnetworks that are significantly differentially expressed

between P_C and $P_{\neg C}$, while accounting for this potential heterogeneity.

- $E(g, p)$: expression value of gene g in patient p
- $F(g, p)$: the fuzzy score of gene g in patient p, as obtained by applying a GFS transform (as described in Chapter 3) on the gene expression matrix. Briefly, genes are ranked in each patient according to their raw expression, and a fuzzy score is obtained by using two thresholds θ_1 and θ_2; genes in the upper θ_1 quantile are assigned a score of 1, genes below the θ_2 quantile are assigned a score of 0, and those in between are assigned a score by linear interpolation. In Chapter 3, we demonstrated that this transformation leads to great improvement in the quality of downstream analysis, as compared to preprocessing by mean-scaling, z-score, and quantile normalization.
- $\beta(g, X)$: the relevance factor of gene g in a population represented by a set of patients X. The factor denotes how consistently g gets highly expressed in X, and is computed as the average fuzzy score of g over all patients in X:

$$\beta(g, X) = \sum_{p \in X} \frac{F(g, p)}{|X|} \tag{4.1}$$

- S: the set of all candidate subnetworks S_k ($k \in \{1, 2, \ldots, r\}$) generated from known biological pathways.

4.3.3 Approach

Generating candidate subnetworks

The primary goal of SPSNet is to identify biological factors that distinguish subpopulations within a sample. Therefore, pathways were chosen to generate subnetworks as they represent the biological processes in an organism, and differences in their functioning contribute to differences within phenotypes. SPSNet does not preclude generating subnetworks from high-quality PPI networks. Both PPI networks and biological pathways can be supplied, even simultaneously,

as input to SPSNet (and also to PFSNet [LW13]). However, in this dissertation, we do not investigate PPI networks since there are confounding issues when using PPI networks. For example, a PPI network is strictly speaking an artificial assembly of pairwise PPIs: While each individual PPI is a real biological interaction, the subnetwork itself is misleading because e.g. not all partners of a protein in the subnetwork actually simultaneously bind the protein. To ensure a straightforward interpretation and evaluation of SPSNet, we prefer to exclude PPI networks.

The standard PFSNet methodology uses highly expressed genes from each phenotype to induce subnetworks on known biological pathways. However, this technique for generating candidate subnetworks is not suitable for heterogeneous data, as the presence of multiple subpopulations in a phenotype is likely to dilute high expression in any specific subtype. Therefore, we generate subnetworks as in NEA[ALP+12]; i.e. we form a subnetwork from each gene and its immediate neighbors in a biological pathway. We filter out subnetworks with less than 5 genes. We generate a total of 5654 such subnetworks from 300 human pathways in PathwayAPI [SDGW10].

Computing subnetwork scores

A GFS tranform is first applied to the gene expression matrix, as described in Section 4.3.2. In PFSNet, all subnetworks are then assigned phenotype-wise scores for each patient as follows. A subnetwork S_k is scored in phenotype C by summing the fuzzy votes of all patients towards each member gene in S_k, weighted by the respective gene relevance factors in C. Similarly, a score corresponding to $\neg C$ is obtained by weighing the gene fuzzy votes with the respective relevance factors in $\neg C$. With the null hypothesis that subnetwork S_k is not relevant to the difference between phenotypes C and $\neg C$, we test whether distribution of the difference between their corresponding scores is centered around zero. In particular,

$$PScore(p, S_k, C) = \sum_{g \in S_k} F(g, p) \times \beta(g, C) \tag{4.2}$$

$$PScore(p, S_k, \neg C) = \sum_{g \in S_k} F(g, p) \times \beta(g, \neg C) \tag{4.3}$$

$$PFS\text{-}Score(p, S_k, C, \neg C) = PScore(p, S_k, \neg C) - PScore(p, S_k, C) \tag{4.4}$$

Since PFSNet assumes no underlying heterogeneity in the phenotypes, the two relevance factors $\beta(g, C)$ and $\beta(g, \neg C)$ are computed using the average of fuzzy votes in *all* patients in the respective phenotype. However, since SPSNet deals with heterogeneous data, we wish to compute subpopulation-specific relevance factors, rather than relevance factors over entire phenotypes. For this, we assume that each subpopulation in a phenotype has at least one subnetwork for which it has the highest expression among members of the phenotype. We then select *representative patients* for each subpopulation as the top x patients with highest expression of the subnetwork (supposing that the smallest subpopulation has at least x members), and use these to compute the subpopulation specific relevance factors. In our analysis, we set the value of x to 10, unless specified otherwise. The effect of variation in x on the performance of SPSNet is discussed later in Section 4.4.6.

For each subnetwork S_k, we compute the sum of gene fuzzy votes in patients belonging to both phenotypes C and $\neg C$. Thus, two vectors $V(S_k, C)$ and $V(S_k, \neg C)$ are generated as:

$$V(S_k, C) = [\sum_{g \in S_k} F(g, p_1), \sum_{g \in S_k} F(g, p_2), ..., \sum_{g \in S_k} F(g, p_{|C|})] \tag{4.5}$$

$$V(S_k, \neg C) = [\sum_{g \in S_k} F(g, p'_1), \sum_{g \in S_k} F(g, p'_2), ..., \sum_{g \in S_k} F(g, p'_{|\neg C|})] \tag{4.6}$$

The top x patients each with the highest values in $V(S_k, C)$ and $V(S_k, \neg C)$ are then selected as the *representative patients*. Let the set of these patients be denoted as $Q(S_k, C)$ and $Q(S_k, \neg C)$ respectively. Then, we compute the final scores for each subnetwork as:

$$SScore(p, S_k, C) = \sum_{g \in S_k} F(g, p) \times \beta(g, Q(S_k, C)) \tag{4.7}$$

$$SScore(p, S_k, \neg C) = \sum_{g \in S_k} F(g, p) \times \beta(g, Q(S_k, \neg C)) \tag{4.8}$$

$$SPS\text{-}Score(p, S_k, C, \neg C) = SScore(p, S_k, \neg C) - SScore(p, S_k, C) \tag{4.9}$$

Similar to PFSNet, the null hypothesis in SPSNet is that subnetwork S_k is not relevant to the difference between phenotypes C and $\neg C$. Therefore, it is tested whether the distribution of $SPS\text{-}Score(p, S_k, C, \neg C)$ (as mentioned in Equation 4.9) across all patients is centered around zero. However, before testing the subnetworks for statistical significance, we eliminate candidate subnetworks which do not contain at least five genes with a phenotype-specific (subpopulation-specific) relevance factor greater than or equal to 0.5 in PFSNet (SPSNet). Setting this cutoff ensures that genes in each candidate subnetwork are highly expressed in at least half of the patients of that phenotype/subpopulation, and thus helps to reduce false positives.

Determining statistical significance

In the standard PFSNet methodology, a null score distribution for each phenotype is generated by randomly swapping class-labels between patients in the control and test samples, and computing subnetwork scores using the permuted labels. However, we use the theoretical t-distribution as our null distribution, as a class-label permutation approach is not practical for SPSNet. This is because the number of representative patients (recall $x = 10$) is insufficient for generating the necessary number of class-label permutations. We test how distant the mean $SPS\text{-}Score$ of each subnetwork is from zero (on either side), and thereby estimate the corresponding statistical significance. All subnetworks with p-value below a given threshold are reported as significant. In here, we use the customary significance threshold of 0.05.

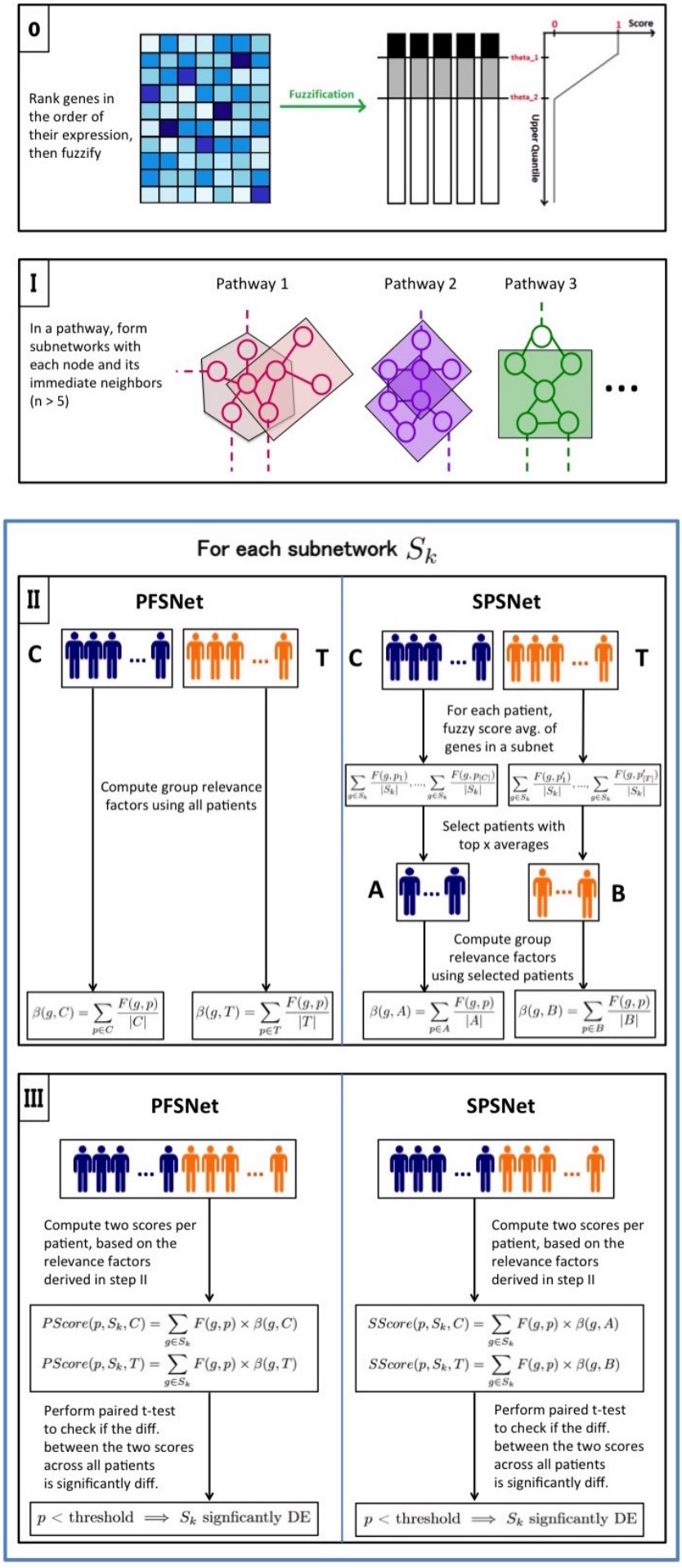


Figure 4.1: Flowchart illustrating the SPSNet methodology (in comparison to PFSNet)

SPSNet as the generalization of PFSNet

As stated earlier in Section 2.1, SPSNet is a generalization of PFSNet. When a 'subpopulation' expands to accommodate the entire phenotype, and all patients in the phenotype can be considered *representative* of it, SPSNet is equivalent to PFSNet:

$$SPS\text{-}Score(p, S_k, C, \neg C) = \sum_{g \in S_k} F(g, p) \times \beta(g, Q(S_k, \neg C)) - \sum_{g \in S_k} F(g, p) \times \beta(g, Q(S_k, C)) \tag{4.10}$$

$$= PFS\text{-}Score(p, S_k, Q(S_k, C), Q(S_k, \neg C)) \tag{4.11}$$

An overview of the PFSNet and SPSNet methodology is presented in Figure 4.1.

4.4 Results

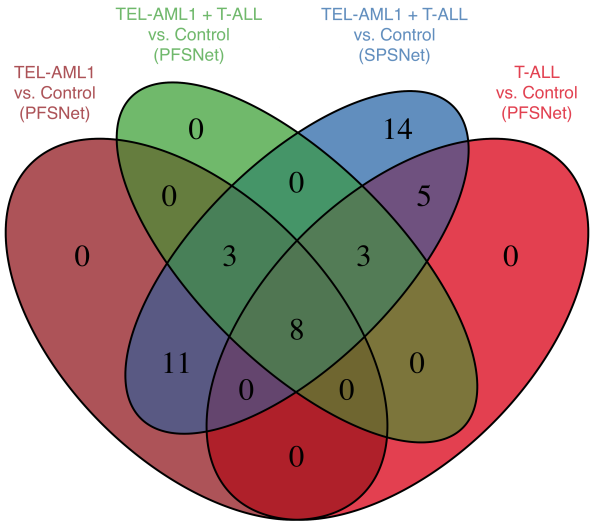
In analyzing the performance of SPSNet, we take a four-fold approach: (i) First, we merge samples with known experimental conditions; and test whether SPSNet is able to discover subnetworks known to be differentially expressed in the individual subpopulations in the merged dataset. We also quantitatively assess the discriminatory power of SPSNet by transforming the subnetwork scores into feature matrices, and computing silhouette scores on their PCA transform. (ii) To analyze the sensitivity and specificity of the method, we simulate test datasets with induced heterogeneity, and evaluate if SPSNet correctly identifies the differentially expressed subnetworks as such. (iii) To validate the reproducibility of SPSNet, we examine the overlap between subnetworks reported significantly differentially expressed on independent datasets with the same phenotype composition. (iv) Finally, we investigate the utility of SPSNet scores in separating different sub-populations within the heterogeneous phenotypes under comparison.

4.4.1 Comparison using homogeneous phenotypes

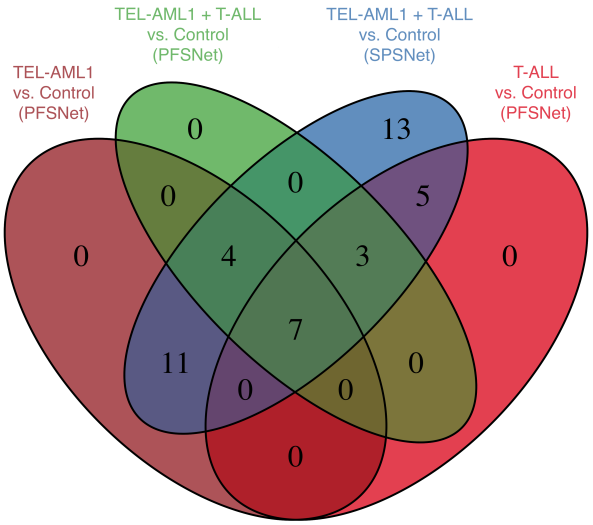
Since PFSNet performs well on homogeneous phenotypes [LW13], it is reasonable to assume that subnetworks reported by it when comparing two homogeneous classes are truly differentially expressed. Therefore, we compare the subnetworks reported significant from PFSNet runs on homogeneous classes (e.g. A vs. C and B vs. C), with those reported by SPSNet and PFSNet on heterogeneous classes obtained by merging multiple homogeneous phenotypes (e.g. A + B vs. C).

Acute Lymphoblastic Leukemia

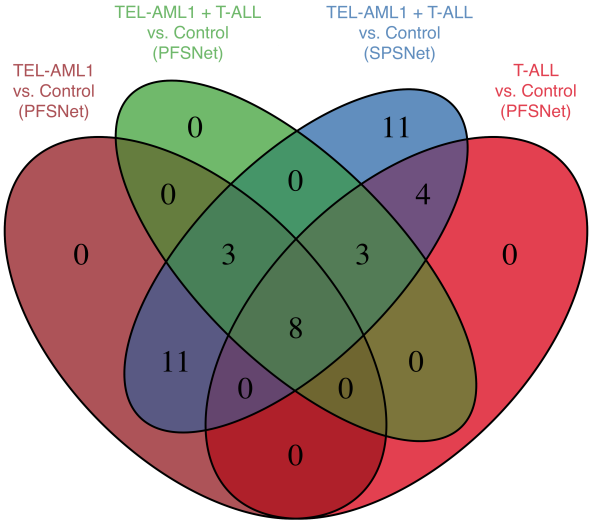
We obtain subnetworks highly expressed in the TEL-AML1 subtype and are reported by PFSNet as significantly differentially expressed with respect to the normal class, and a similar set of subnetworks highly expressed in the T-ALL subtype. To simulate the heterogeneous case, we combine patients from both disease subtypes into a single "heterogeneous" disease class, and then obtain subnetworks highly expressed in it that are reported by PFSNet and SPSNet as significantly differentially expressed with respect to the normal class. Finally, we perform a pathway-level comparison of the subnetworks reported significant in the homogeneous and heterogeneous cases. Figure 4.2 records three sets of observations corresponding to datasets of increasing heterogeneity (where the disease sample is created by incrementally merging 10, 20, and 29 patients of the T-ALL subtype respectively, with 30 TEL-AML1 patients in each case). From the figure, we observe that both PFSNet and SPSNet are successful in identifying pathways common to the TEL-AML1 and T-ALL subtypes. However, SPSNet is more sensitive in detecting pathways that are specific to either of the disease subtypes.



(a) 30 TEL-AML1 + 29 T-ALL



(b) 30 TEL-AML1 + 20 T-ALL



(c) 30 TEL-AML1 + 10 T-ALL

Figure 4.2: Acute Lymphoblastic Leukemia (ALL) – pathways containing differentially expressed subnetworks

Hepatocellular Carcinoma

We conduct a similar experiment on the two batches of HCC data, whereby subnetworks highly expressed in HCC and differentially expressed with respect to the normal sample are obtained for each batch separately, and after merging the two batches. Pathway-level comparison of

these subnetworks is recorded in Figure 4.3. We observe that PFSNet and SPSNet are able to discover pathways that have subnetworks differentially expressed in both HCC batches. However, SPSNet is able to better identify pathways differentially expressed only in one of the two batches, indicating its sensitivity to heterogeneity in samples.

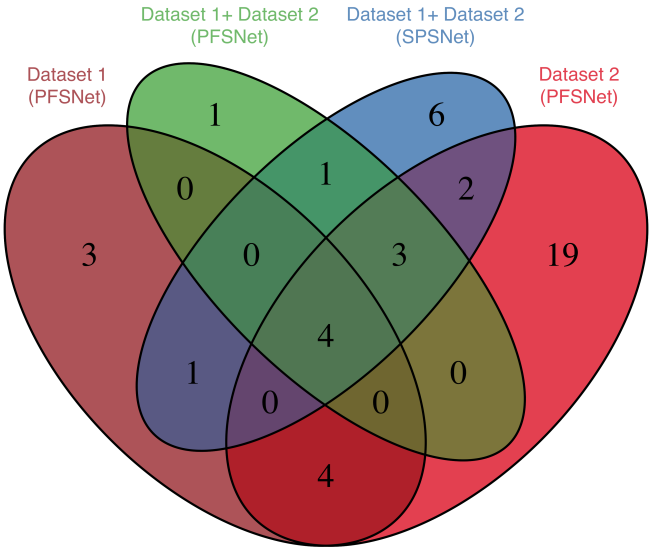


Figure 4.3: Hepatocellular Carcinoma (HCC) – pathways containing differentially expressed subnetworks that are highly expressed in HCC

4.4.2 Estimating sensitivity and specificity from simulation

Simulation experiments, when carefully designed, have the advantage that 'correct' outcomes from the application of a method can be known in advance. Thus, they can be powerful tools for objective performance evaluation.

We simulate test samples with injected heterogeneity, pair them with homogeneous control samples, and compare subnetworks that are known to be differentially expressed between the two sample groups with those reported significant by SPSNet to estimate the sensitivity and specificity of SPSNet. The detailed procedure is described below, and illustrated in Figure 4.4:

We choose a homogeneous normal sample, which is unlikely to contain any significantly

differentially expressed genes at the outset. The normal sample is randomly split into two equal halves, N_1 and N_2, and one of these parts (N_2) is allocated for injecting differential expression. To induce heterogeneity, N_2 is further divided into two subtypes, N_{21} and N_{22}, with α% and $(100-\alpha)$% of its patients respectively. We sub-sample 10% of the total number of genes and induce differential expression in patients in N_{21} for these selected genes, in a manner similar to the description from Langley et al [LM15]. i.e. we multiply the expression of patients in N_{21} by a factor of r, where r is chosen randomly from the set {1.2, 1.5, 1.8, 2.0, 3.0}, for each gene in the sub-sample. Another independent sub-sample of 10% genes is chosen, and differential expression corresponding to genes in this sub-sample is induced in patients belonging to the set N_{22}.

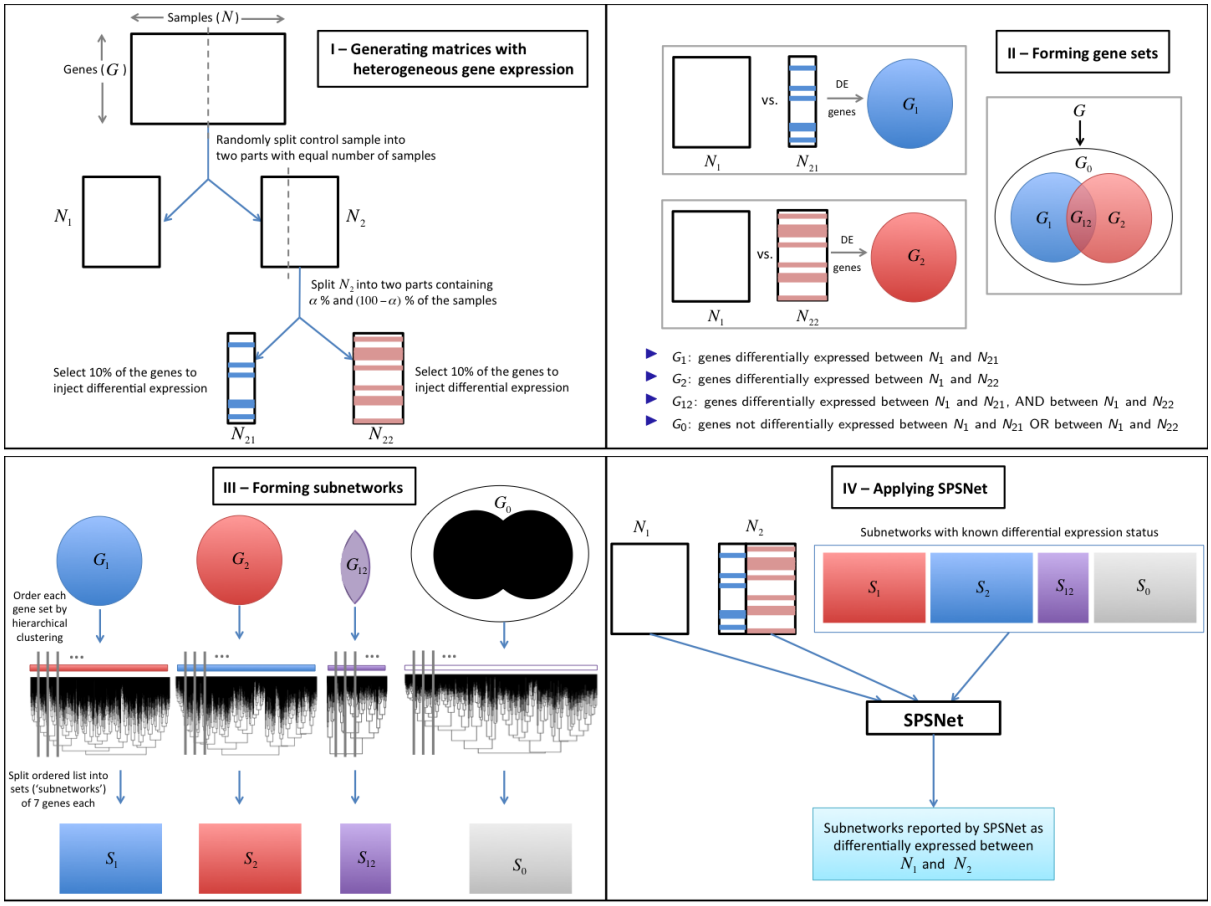


Figure 4.4: Flowchart illustrating the simulation methodology for estimating sensitivity and specificity of SPSNet

Thus, we obtain four sets of genes, which we use to generate four sets of subnetworks:

- G_1: genes differentially expressed between N_1 and N_{21}
- G_2: genes differentially expressed between N_1 and N_{22}
- G_{12}: genes differentially expressed between N_1 and N_{21}, AND between N_1 and N_{22}
- G_0: genes not differentially expressed between N_1 and N_{21} and between N_1 and N_{22}

To generate subnetworks from these genes, we adopt the procedure used by Goh and Wong [GW16a], emulating the feature of real biological subnetworks that genes in a subnetwork tend to have correlated expression patterns. In particular, we perform a hierarchical clustering of genes in G_1, and reposition them within their clusters such that the most similar genes are next to each other. Subnetworks are then generated by splitting the resulting ordered list into sets of seven genes each. As discussed in the work by Goh and Wong [GW16a], this procedure is a sample prototype to generate pseudo-subnetworks in the absence of any gold-standard for simulation purposes, and not a fool-proof to form groups of genes that approximate real subnetworks. However, we avoid forming subnetworks with a very small number (e.g. < 5) of genes, since it is likely to lead to a high fluctuation in the test statistic.

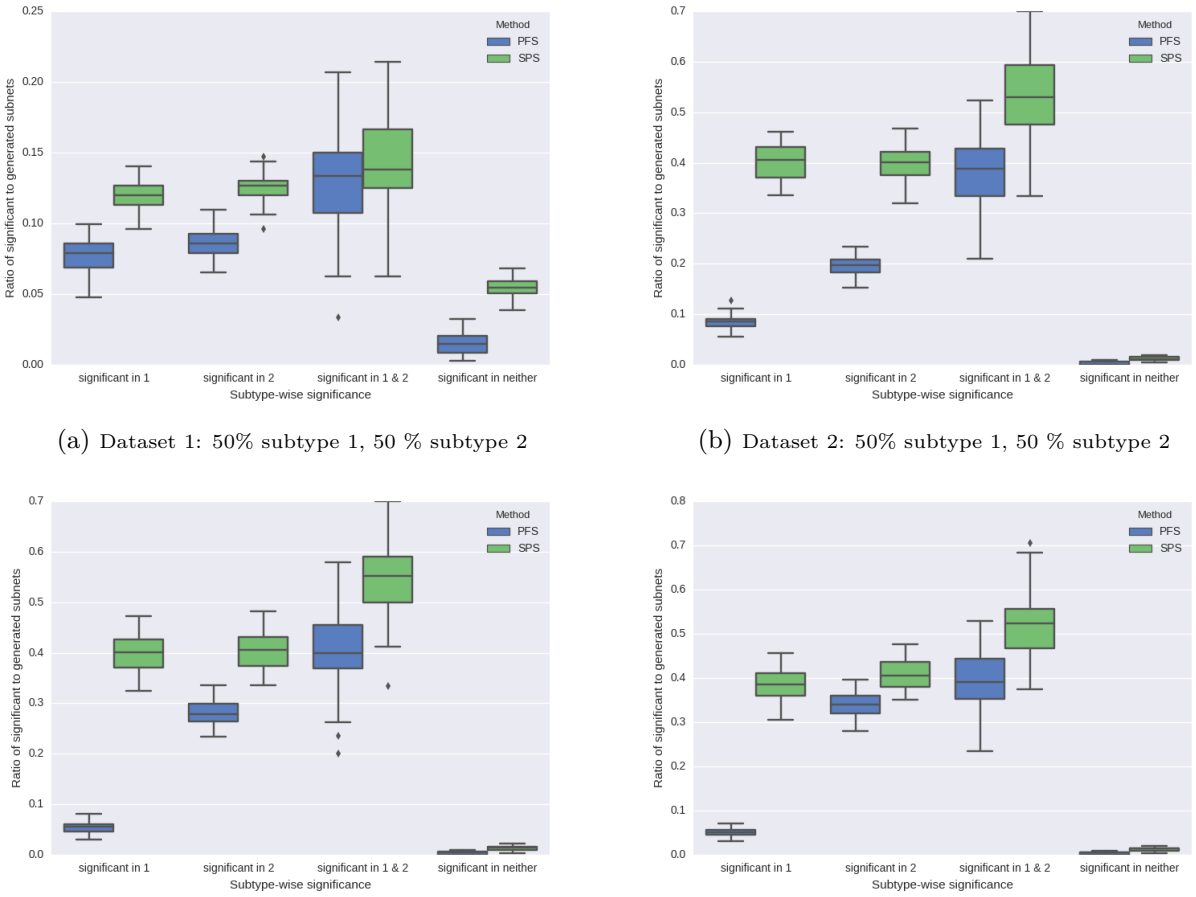
A similar ordering after hierarchical clustering is obtained separately for G_2, G_{12}, and G_0. However, for G_0, we do not use all the non-differentially expressed genes to form subnetworks, but only four times the number of genes in G_1. This emulates the effect of incompleteness in biological pathway databases, and also saves computation time required to generate a vast number of negative control subnetworks.

The entire simulation process is repeated for 100 iterations. In each iteration, PFSNet and SPSNet are run on newly simulated data, and subnetworks generated from G_1, G_2, G_{12}, and G_0 in the corresponding iteration are tested for significance.

Estimating sensitivity

We use two datasets for simulation, normal kidney and normal liver tissue expression data from TCGA [N+13] (Dataset 1) and Roessler et al. [RJB+10] (Dataset 2), which profile 20,502

and 13,801 genes respectively. The number of subnetworks generated in each iteration from Dataset 1 using G_1, G_2, G_{12}, G_0 are 292, 292, 30, 1168 respectively; while 197, 197, 20, 788 subnetworks are generated from Dataset 2. To understand the effect of different levels of heterogeneity within the data on the performance of PFSNet and SPSNet, we vary the parameter α in our simulations. For Dataset 1, α is set to 50% (the test sample is divided into two subtypes with 50% of its patients each), while for the larger Dataset 2, separate simulations are performed with α set to 20% (subtype 1 – 20%, subtype 2 – 80%), 40% (subtype 1 – 40%, subtype 2 – 60%), and 50% (subtype 1 – 50%, subtype 2 – 50%).



(a) Dataset 1: 50% subtype 1, 50 % subtype 2

(b) Dataset 2: 50% subtype 1, 50 % subtype 2

(c) Dataset 2: 40% subtype 1, 60 % subtype 2

(d) Dataset 2: 20% subtype 1, 80 % subtype 2

Figure 4.5: Proportion of significant subnetworks reported by PFSNet and SPSNet on test samples injected with different levels of heterogeneity

Figure 4.5 (a) shows four boxplots for Dataset 1 corresponding to the fraction of subnetworks reported significant by PFSNet and SPSNet from subnetworks that are simulated to be

significant in subtype 1, significant in subtype 2, simulated to be significant in both, and non-significant in both subtypes. Figures 4.5 (b) to (d) show similar boxplots for Dataset 2, with varying levels of heterogeneity (different values of α).

As expected, both PFSNet and SPSNet show higher sensitivity for subnetworks significant in both subtypes, when compared with those significant in only one of the subtypes. In all three subnetwork categories—significant in subtype 1, subtype 2, and both—the sensitivity of SPSNet is higher than PFSNet (SPSNet improves the median sensitivity by about 10% in case of subnetworks significant in both subtypes, and by a larger margin in the subtype-specific subnetworks). The subnetworks not significant in either subtypes are rarely reported significant by PFSNet and SPSNet (high specificity); the false-positive rate, although a little higher in SPSNet than PFSNet, is within or around the 5% bound in all cases.

It is also interesting to note the impact of varying heterogeneity on the sensitivity of the two methods for simulations on Dataset 2. We notice that the output of PFSNet is strongly dominated by the majority subtype, while SPSNet is relatively insensitive to the level of heterogeneity. Thus, when α is set to 50%, the median sensitivity of PFSNet for subnetworks significant in subtype 1 and 2 is about 10% and 20% respectively. When α is decreased to 40%, the median sensitivity for subnetworks significant in subtype 1 (minority) drops to below 5% and median sensitivity for subnetworks significant in subtype 2 (majority) rises to about 25%. At an even lower α of 20%, the recall for subnetworks significant in subtype 1 remains almost the same, while the median sensitivity for subtype 2 rises to about 35%. On the other hand, SPSNet performs relatively better at all levels of heterogeneity; irrespective of the value of α, it consistently shows a median sensitivity of about 40%.

Estimating false-positive rate

To assess whether the false-positive rate in SPSNet is well-controlled, we use the same simulation setup as that in the previous subsection 4.4.2, and explained in Fig 4.4. We generate 1000 subnetworks using G_0. Since the genes in G_0 are differentially expressed

between neither N_1 and N_{21}, nor N_1 and N_{22}, no subnetworks generated from G_0 are expected to be differentially expressed. We run SPSNet and test whether the subnetworks are reported to be differentially expressed. For this experiment, we used the normal tissues from one of the HCC datasets [RJB+10]. To observe whether sample size affects false-positive rate, we randomly selected subsamples of size 240, 210, 180, 150, 120, 90, 60, and 30, fifty times each.

Figure 4.6 shows boxplots depicting the range of false-positive rates corresponding to subsamples of each size. In samples of all sizes, the false positives were seen to be well-controlled: less than 50 of 1000 subnetworks are reported significant (FP rate < 0.05).

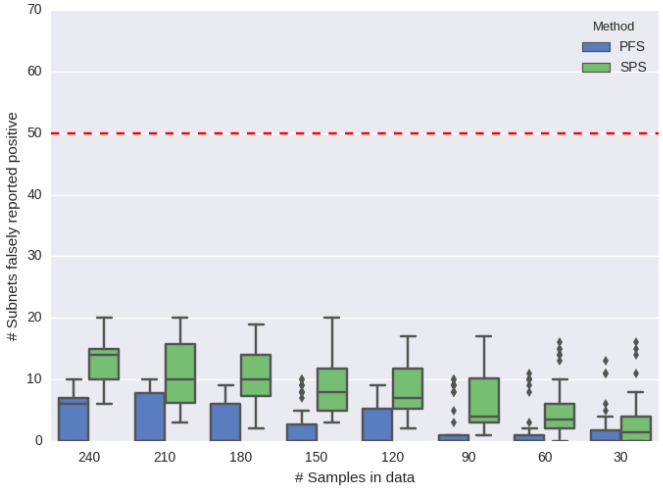


Figure 4.6: False-positive rate of SPSNet on simulated data with varying sample size

4.4.3 Reproducibility on independent datasets

A reliable method would produce significant subnetworks that agree highly when run on independent datasets with the same phenotypical composition. Therefore, we run PFSNet and SPSNet to obtain significantly differentially expressed subnetworks between normal sample and the heterogeneous ALL sample (with all patients from subtypes TEL-AML1 and T-ALL combined). This is done separately for the training and test data, and the agreement (in the form of jaccard coefficient) between significant subnetworks obtained on the two sets of data is recorded in Table 4.1. We observe that SPSNet shows much higher reproducibility on the

heterogeneous dataset, as compared to PFSNet.

	Training	Test	Training ∩ Test	Training ∪ Test	Jaccard Coefficient
PFSNet	27	24	11	40	**0.28**
SPSNet	87	77	62	102	**0.61**

Table 4.1: Jaccard coefficients showing agreement between significant subnetworks obtained by PFSNet and SPSNet on training and test data;

4.4.4 Quality of feature selection

A good method for network-based differential expression analysis of heterogeneous data would report significant subnetworks that can serve as relevant features in distinguishing the classes being compared, as well as their component subpopulations. Therefore, we use the scores of significant subnetworks in PFSNet and SPSNet as features, and visualize scatter plots based on PCA transformation of the resulting feature matrices. Further, we quantitatively assess the ability of these features to distinguish between subpopulations, with silhouette scores computed using the feature matrices and known labels corresponding to patient subtype and/or subpopulation.

Acute Lymphoblastic Leukemia

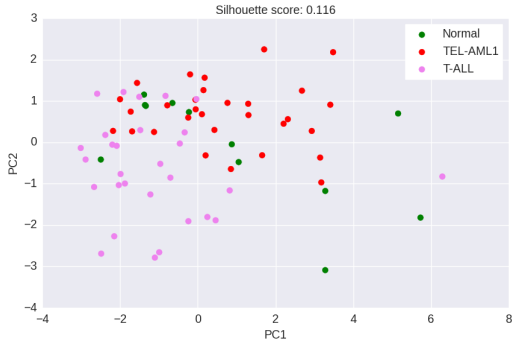
We use the same samples as mentioned in previous sections with experiments on the ALL dataset [YRS+02] – normal class against datasets of increasing heterogeneity (where the disease sample is created by incrementally merging 10, 20, and 29 patients of the T-ALL subtype respectively, with 30 TEL-AML1 patients in each case). We draw PCA scatter plots corresponding to subnetworks reported as differentially expressed between normal and each heterogeneous disease sample (Figure 4.7).

Table 4.2 shows three sets of silhouette scores corresponding to feature matrices obtained from scores of significantly differentially expressed subnetworks reported on comparing normal sample with disease samples of increasing heterogeneity. From the silhouette scores, as well as

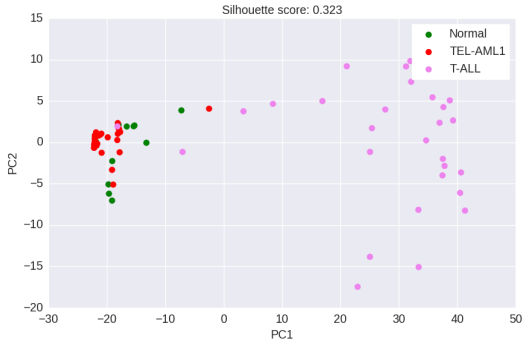
PCA scatter plots of subnetwork scores, we observe that SPSNet is able to better discriminate between different disease subtypes within the ALL sample, across varying levels of heterogeneity.

	30 TEL-AML1 + 10 T-ALL	30 TEL-AML1 + 20 T-ALL	30 TEL-AML1 + 29 T-ALL
PFSNet	0.079	0.12	0.116
SPSnet	**0.288**	**0.342**	**0.323**

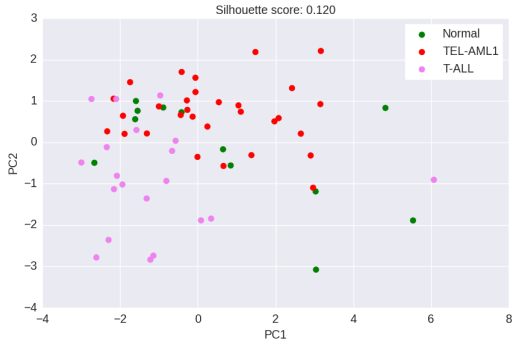
Table 4.2: ALL – Silhouette scores based on the first 3 PCs of feature matrices built using scores significant subnetworks in PFSNet and SPSnet;



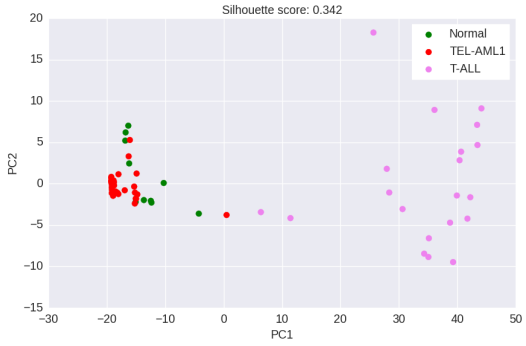
(a) PFSNet – (Normal vs. 30 TEL-AML1 + 29 T-ALL)



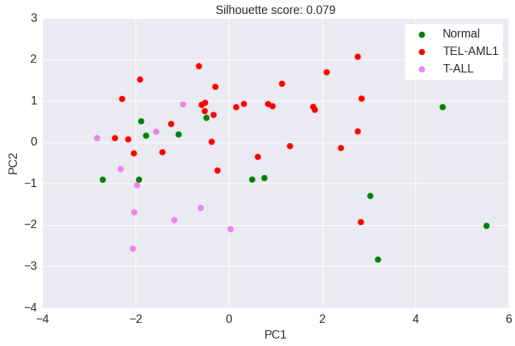
(b) SPSNet – (Normal vs. 30 TEL-AML1 + 29 T-ALL)



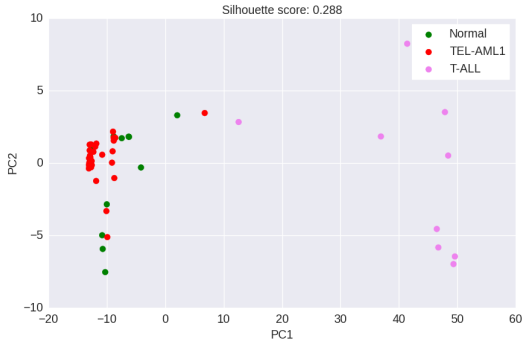
(c) PFSNet – (Normal vs. 30 TEL-AML1 + 20 T-ALL)



(d) SPSNet – (Normal vs. 30 TEL-AML1 + 20 T-ALL)



(e) PFSNet – (Normal vs. 30 TEL-AML1 + 10 T-ALL)

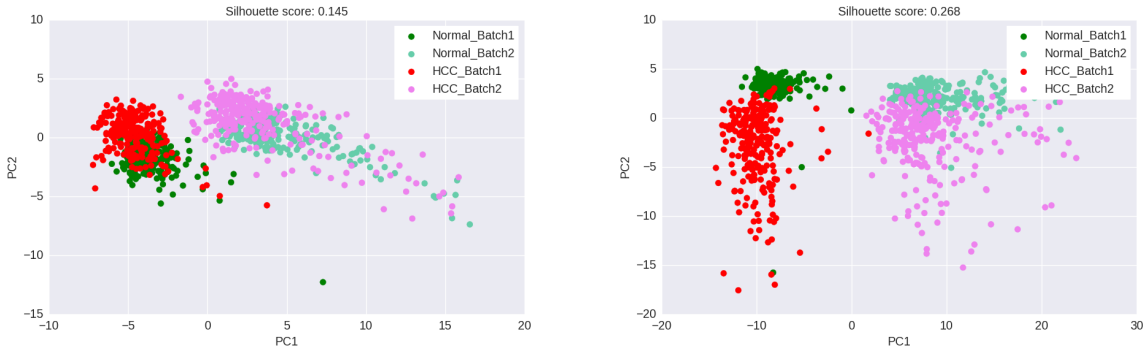


(f) SPSNet – (Normal vs. 30 TEL-AML1 + 10 T-ALL)

Figure 4.7: Normal vs heterogeneous ALL disease sample – PCA scatter plots based on scores of significant subnetworks in PFSNet and SPSnet

Hepatocellular Carcinoma

We use the two HCC datasets from [RJB+10] and [BZL+10], and create a new normal and HCC sample by merging the normal and disease samples respectively from both batches. PCA scatter plots drawn using scores of significant subnetworks are shown in Figure 4.8 (a) and (c).



(a) PFSNet – PC1 and PC2

(b) SPSNet – PC1 and PC2

Figure 4.8: Normal vs HCC sample combined from Dataset 1 ([RJB+10]) and Dataset 2 ([BZL+10]) – PCA scatter plots based on scores of significant subnetworks in PFSNet and SPSnet

We observe that in the scatter plot corresponding to SPSNet features, patients appear better separated with respect to their batch as well as phenotype labels. Further, PC1 is able to capture and isolate almost all of the batch effects in the SPSNet scatter plot, whereas the batch effects spill over to the lower PCs in the case of PFSNet. This is despite the fact that PC1 in SPSNet covers only 66% of the total variance while PC1 in PFSNet covers 72% of its total variance. Thus, SPSNet proves to be effective at identifying the heterogeneity induced by batch effects.

	Normal vs HCC (first 3 PCs, with batch labels)	Normal vs HCC (2^{nd}, 3^{rd} PC, without batch labels)
PFSNet	0.145	0.117
SPSnet	**0.268**	**0.298**

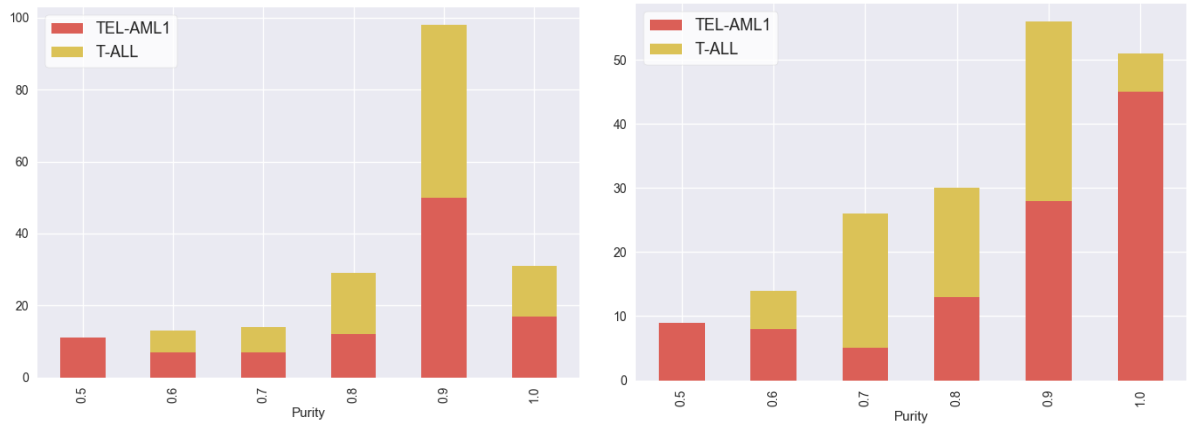
Table 4.3: HCC – Silhouette scores based on PCA transform applied to scores of subnetworks reported as significantly DE by PFSNet and SPSNet;

Next, we eliminate PC1 to see if the normal and HCC samples (combined from two batches) can be clearly separated by the remaining PCs based on their phenotypes alone. From the

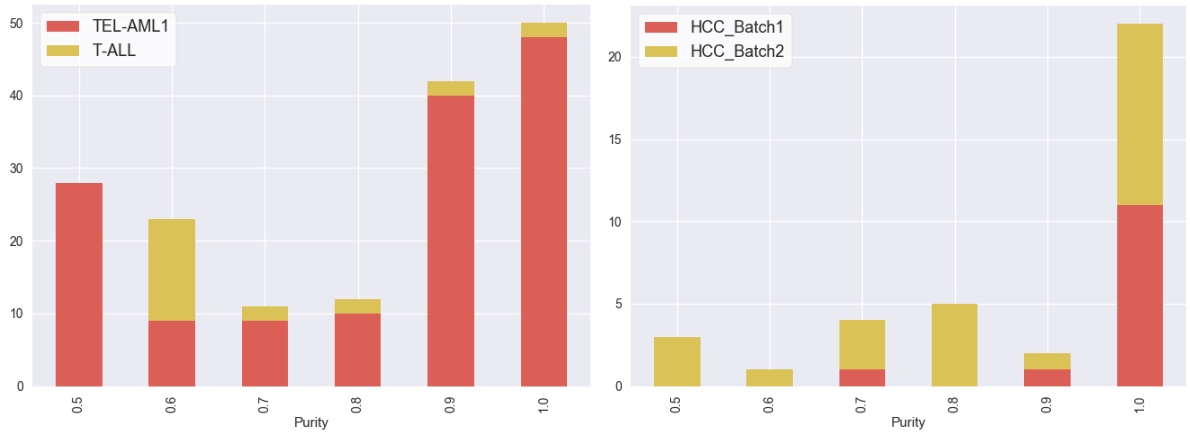
silhouette scores in Table 4.3, it is seen that PC2 and PC3 from SPSNet features are able to better distinguish between normal and HCC samples, as compared to their counterparts from PFSNet features. These observations are in line with the remarks from Chapter 3 that eliminating PC1 often leads to removal of batch effects and a clearer separation based on phenotypes.

4.4.5 Are *representative patients* of significant subnetworks enriched in specific subpopulations?

Since SPSNet utilises a subset of patients for each subnetwork to represent potential subpopulations in the phenotype, we study a) whether such subsets are enriched in one of the constituent subpopulations, and b) how such enrichment is affected by the relative proportions of the constituent subpopulations in the data.



(a) 30 TEL-AML1 + 29 T-ALL (p-val: 1.1×10^{-3}) (b) 30 TEL-AML1 + 20 T-ALL (p-val: 1.1×10^{-10})



(c) 30 TEL-AML1 + 10 T-ALL (p-val: 1.02×10^{-17}) (d) HCC merged dataset (p-val: 5.5×10^{-4})

Figure 4.9: Number of subnetworks reported significant by SPSNet corresponding to different purity levels. A chi-squared test is performed to see if the number of significant subnetworks with high purity (purity > 0.75) is larger than those with low purity (purity ≤ 0.75); p-values are reported in brackets.

To assess this, we once again use the ALL [YRS+02] and HCC datasets [RJB+10, BZL+10], and define a measure 'purity' as the proportion of patients belonging to the majority subpopulation (subtype/batch) in the *representative patients* subset for a given significant subnetwork. Figure 4.9 records the number of significant subnetworks with purity levels between 0.5 to 1.0 and the colors indicate the majority subpopulation which resulted in the purity value.

We observe that a large proportion of significant subnetworks are enriched in one of the constituent subpopulations (high purity); such subnetworks help distinguish the subpopulations from each other. There are also a few significant subnetworks which have low purity (almost equal proportion of subpopulations); these indicate common biological characteristics shared by

the subpopulations. Also, in the ALL dataset, when SPSNet is used to compare control sample with a heterogeneous disease sample containing 30 TEL-AML1 patients and 29 T-ALL patients, the contribution of the two disease subtypes to high purity levels (purity > 0.75) is similar; i.e. the number of significant subnetworks with representative patients having TEL-AML1 and T-ALL patients in majority is similar. This phenomenon persists even when the number of T-ALL patients is reduced to 20. However, when only 10 T-ALL patients are included in the heterogeneous sample, there are very few significant subnetworks with representative patients having a T-ALL majority. This suggests that SPSNet is able to recover minority subpopulations unless the size of the smaller subpopulations drops below a certain threshold (viz. x).

4.4.6 Effect of varying number of representative patients on the performance of SPSNet

For each subnetwork, representative patients are chosen by SPSNet to ensure representation of a potential subpopulation in which the subnetwork is highly expressed. Ideally, the number of representative patients, say x, is lower than or equal to the number of patients in the smallest subpopulation within the phenotype. Thus, when top x patients with the highest expression of a given subnetwork are chosen, the selected patients are likely belonging to the same subpopulation.

We study the effect of varying the parameter x on the performance of SPSNet, in terms of its ability to distinguish between subpopulations based on subnetworks reported to be differentially expressed. A PCA transform was applied to the SPSNet scores of differentially expressed subnetworks, and a silhouette score was computed using the first three principal components as features, and sample labels – phenotype x batch. Figure 4.10 shows the trend in this silhouette score with gradual increase in the value of x.

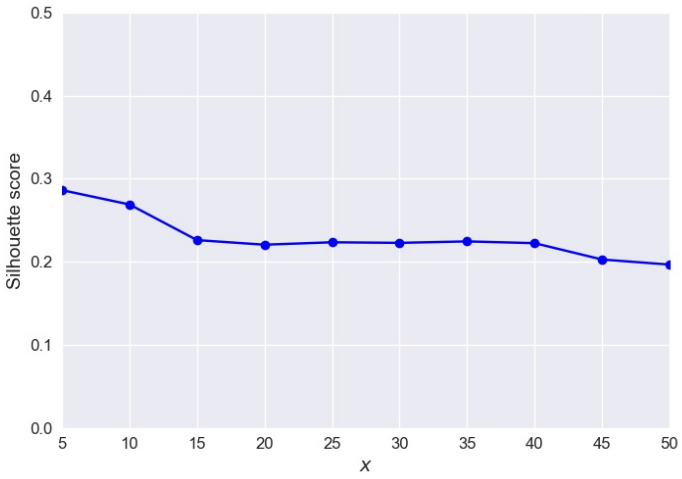


Figure 4.10: HCC merged dataset: Effect of varying x (number of representative patients) in SPSNet on silhouette scores

From Figure 4.10, we see that silhouette score drops as the value of x increases from 5 to 15, and remains stable thereafter. The dataset is likely to contain at least one small subpopulation of 5–10 patients who are most similar to each other in terms of their biological mechanisms, than with other patients. The subsequent stability in silhouette scores uptil $x = 40$ suggests that size of the next smallest subpopulation is around 40 patients. It is interesting to note that SPSNet is relatively robust to minor changes in the parameter x – therefore, the silhouette score also remains fairly stable between the range 15-40.

4.5 Conclusion

Presence of undeclared heterogeneity in gene expression data hinders identification of subpopulations present in the phenotype sample and the specific biological factors associated with them. We presented a method, SPSNet, which discovers and analyzes such heterogeneity. As opposed to previous approaches that derived gene-based signatures to identify potential subpopulations within specific diseases, our method is a generic tool which provides subnetwork-based signatures for subpopulations in any phenotype.

While many methods are available for differential expression analysis on homogeneous pheno-

types, only a few produce consistent results over independent datasets containing the same phenotypes, and none are designed to deal with potential heterogeneity in the data. PFSNet is one method among the rare exceptions which results in consistent outcomes, but it is designed to analyze only homogeneous phenotypes. We proposed SPSNet, a generalization of PFSNet, which is able to solve an important problem – handling undeclared heterogeneity in gene expression samples by identifying subnetworks associated with hidden subpopulations within phenotypes. The approach also helps recognize and eliminate extrinsic heterogeneity such as batch effects. We demonstrated that SPSNet has high sensitivity, low false-positive rate, high reproducibility, and high biological coherence when analyzing gene expression data with heterogeneity.

However, there is room for improvement in the design and performance of SPSNet. For example, SPSNet could benefit from a better subnetwork generation scheme. Although the current procedure for generating candidate subnetworks—selecting each gene and its immediate neighbors in a pathway—is a simple way to account for connections between genes in biological pathways, it is relatively naive and results in fragmented components of pathways. Complementing the information in pathways with that extracted from gene expression datasets could possibly lead to generation of subnetworks that are more cohesive and biologically meaningful. Research is also necessary to further improve the sensitivity of SPSNet.

Chapter 5

Analyzing heterogeneity in RNA-Seq data

"The price to sequence a base [of the human genome] has fallen 100 million times. That's the equivalent of you filling up your car with gas in 1998, waiting until 2011, and now you can drive to Jupiter and back twice."

– Richard Resnick

5.1 Background

With dramatic decline in the price of sequencing, RNA-Seq is becoming increasingly popular as a means to discover novel transcripts and transcript-phenotype associations. Recent RNA-Seq studies have been able to report genes which were previously unidentified in microarray datasets, to have critical roles in disease mechanisms [JSFDK12, TGJ+12, SFUdR+15].Due to the promise that the RNA-Seq technology holds in making new biological discoveries, normalization and analysis of RNA-Seq data are topics of great research interest to computational biologists. In this chapter, we describe some important differences between data generated on the microarray and RNA-Seq platforms, which suggest that the methods for normalization and analysis of microarray datasets may not necessarily generalize to RNA-Seq data. We then present an illustrative case-study on RNA-Seq datasets which demonstrates the application of our normalization and heterogeneity approaches in this context.

5.1.1 Differences between microarray and RNA-Seq data

While microarrays and RNA-Seq share a common purpose of quantifying the transcriptome, each technology has its own merits and drawbacks.

In RNA-Seq, the expression level of each RNA unit is quantified as the number of sequenced fragments mapping to each transcript species [RKL+13]. Depending on the sequencing depth, a target number of bases/reads is sequenced. When the sequencing depth is sufficient, RNA-Seq has the advantage of capturing a broad dynamic range of expression measurements, and is not limited to the discovery of a fixed set of transcript species. Weakly expressed genes and rare transcript species can also be detected by increasing the sequencing depth.

However, a typical human RNA-Seq experiment sequences about 40-50 million paired-end reads. This implies that on an average, every gene is covered approximately 100 times. Since bulk RNA-seq sequences many cells, this can cover only a small fraction of the transcripts. When the sequencing depth is insufficient, the sequenced fragments may be mapped to a random subset of transcripts, distorting the estimation of the actual gene expression. The subsequent observed variations in RNA-Seq measurements are more likely due to stochastic under-sampling than to biological variations.

It is also important to note that, when read counts in RNA-Seq are mapped to genes, they contain an inherent length bias. In particular, more reads are mapped to longer genes, resulting in higher read counts for these genes. Thus for within-sample comparison (i.e. whether a given gene is more highly expressed than another gene in the same sample), raw count data from RNA-Seq cannot be used, unless normalized with respect to transcript length. However, for inter-sample comparison (i.e. whether this same given gene is more highly expressed in this given tissue than another tissue), such normalization with respect to gene length is not needed and can even be harmful [CMT+16].

In contrast, microarrays use a fixed set of probes, which are designed to have specific target transcripts. Due to this specificity, probes do not generally compete with each other to bind

to their target transcripts. Thus, the probability of a probe binding to its target transcript is primarily determined by the abundance of the transcript, and the resulting probe intensities are independent of each other. Since the transcripts are tagged with molecular markers, transcript abundance can be precisely captured with imaging techniques that quantify the hybridized probes.

Due to limited number of probes in microarrays, the range of transcripts whose expression can be obtained is limited. However, unlike RNA-Seq platforms, microarrays are not affected by fluctuations due to sampling stochasticity – the transcripts of different genes do not compete with each other for binding to a microarray as the microarray has dedicated probes for each transcript species.

These and other technical differences necessitate that when generalizing methods developed for analysing microarray datasets to RNA-Seq, the datasets as well as the methods be examined properly for their specific characteristics.

5.1.2 Normalization of RNA-Seq data

RNA-Seq data is complex – it contains large differences in the number of reads produced between different sequencing runs, technical variation arising due to nucleotide compositions, sequencing platforms, library preparation protocols, and so on [RKL+13]. Normalization of such data is essential to make sample expression measurements comparable.

Mortazavi et al. [MWM+08] proposed a simple normalization technique in 2008, RPKM (Reads Per Kilobase of transcript per Million mapped reads), in which gene counts are normalized by the transcript length and the total number of mapped reads in each library. In 2010, Trapnell et al. [TWP+10] presented a variation of RPKM to accommodate paired end reads, FPKM (Fragments Per Kilobase of transcript per Million mapped reads). However, in both RPKM and FPKM, changes in the expression levels of genes affect all others. Consequently, these normalization methods are suitable only for within-sample comparison, but not for inter-sample

comparison.

To illustrate this point, we consider the following example. Suppose in sample A, there are 5,000 expressed genes X_i, each of length 2,000 bases, and each is expressed at a level of 2,000 transcripts. Suppose the RNA-seq budget is 10,000,000 reads (for simplicity, assume also that each read covers an entire transcript). Then the RPKM is 100 for each gene X_i in sample A. Suppose in sample B, the same 5,000 genes X_i are also expressed at a level of 2,000 transcripts, and an additional 5,000 genes Y_j (each of length 2,000 bases) are expressed at a level of 2,000 transcripts as well. Again, suppose the RNA-seq budget is 10,000,000 reads (and each read covers an entire transcript). Then, since the budget is insufficient to sequence every transcript in sample B, in the best-case scenario, exactly 1,000 transcripts of each X_i and Y_i get sampled in the RNA-seq. Then the RPKM is 50 for each gene X_i and Y_i in sample B. If sample A and B are compared for differentially expressed genes, each X_i will be declared differential, along with each Y_i. However, we know that each X_i has exactly the same expression level in sample A and B, viz. 2,000 transcripts in both samples. Thus, the reported differential expression of X_i is an artefact of wrong use of RPKM normalization.

A straightforward way to avoid this problem is to spike samples A and B with a constant number of transcripts of a positive-control gene. By design, this gene is known to have exactly the same expression level in samples A and B. Thus if it is reported at say RPKM $= 100$ in sample A but at RPKM $= 50$ in sample B, we know that all RPKM levels reported for sample B must be multiplied by a factor of 2 before we can compare samples A and B. Alternatively, we need to pick some genes that are known to have similar actual expression levels in samples A and B, and use them to determine the multiplicative factor needed. Housekeeping genes might be good candidates for this purpose for comparison of non-cancer samples; however, they might not suitable for cancer samples since many housekeeping genes are aberrantly expressed in cancer cells.

Sometimes an RNA-Seq dataset is available only in RPKM- or FPKM- normalized form. In such a situation, GFS may be suitable for rectifying the distortions induced by RPKM and

FPKM normalization. In raw RNA-Seq data, the rank of genes based on read counts do not correspond to the rank based on their actual gene expression level, since the former is influenced by gene length. RPKM and FPKM normalize a sample with respect to transcript length and sequencing depth. Thus, despite distorting actual gene expression values in a sample, they also restore the relative rank of the genes in the sample with respect to actual expression level. As GFS replaces the expression values in a sample by fuzzified ranks of these values in that sample, it discards the RPKM- and FPKM-induced distortions and makes the relative rank information more robust.

5.1.3 Concordance between RNA-Seq and microarray datasets

Concordance between datasets of the same phenotype generated on the microarray and RNA-Seq platforms is a topic of popular research interest. The study by Wang et al. [WGB+14a] reports that concordance between microarray and RNA-Seq depends on the degree to which the underlying biological mechanisms in the given phenotypes are perturbed, after adjusting for chance. RNA-Seq has greater sensitivity in detecting lowly expressed genes when the sequencing depth is reasonable, and expression of genes which are only in RNA-Seq data and not microarrays, show strong correlation with the degree of perturbation. Thus, the performance of RNA-Seq is significantly better than that of microarrays when comparing groups with similar phenotypes, such as progressive stages of a disease. Microarray and RNA-Seq show similar performance when comparing groups with very different phenotypes, such as cancer and normal tissues [WGB+14b].

5.2 A case-study on Hepatocellular Carcinoma

In our case-study, we use the Hepatocellular Carcinoma (HCC) RNA-Seq dataset from The Cancer Genome Atlas GDC portal [GHF+16]. The dataset contains 367 tumor samples and 48 adjacent non-tumor samples (without pairing information between the tumor and non-tumor

samples). We used count and FPKM values (as generated by HTSeq [APH15]) for our experiments reported below for this dataset.

In conjunction, we also use two HCC datasets generated on microarrays:

- Dataset from Roessler et al. [RJB+10] (GSE14520) consisting of 246 tumor (predominantly HBV-positive) and 240 adjacent non-tumor samples
- Dataset from Burchard et al. [BZL+10] consisting of 268 tumor (predominantly HBV-positive) and adjacent 249 non-tumor samples.

We perform three types of analysis with the three datasets. First, we examine whether GFS is a suitable technique to normalize RNA-Seq count data, and explore its robustness across different θ parameters (quantile thresholds). Second, we test the concordance between the microarray and RNA-Seq data by comparing subnetworks reported significantly differentially expressed between HCC and control samples in both the platforms. Third, we merge RNA-Seq data with the two microarray batches, and apply SPSNet to see if the underlying heterogeneity of batches and phenotypes is discovered. In our second and third analysis, we use FPKM values of our RNA-Seq dataset.

Note that, as we described in Section 5.1.2, it is potentially harmful to use FPKM values when comparing gene expression across samples. However, since read count-based ranks in RNA-Seq data do not correspond to ranks based on actual gene expression level, GFS should not be applied to raw RNA-Seq data directly when analysing RNA-Seq and microarray data together. Applying GFS to raw RNA-Seq data causes GFS to produce fuzzified ranks based on read counts; these ranks are incompatible with ranks produced by applying GFS to microarray gene expression data. Fortunately, normalizing RNA-Seq data by FPKM, followed by GFS makes the processed data compatible with microarray data normalized by GFS. Thus, GFS can be used for combining data produced by these two platforms.

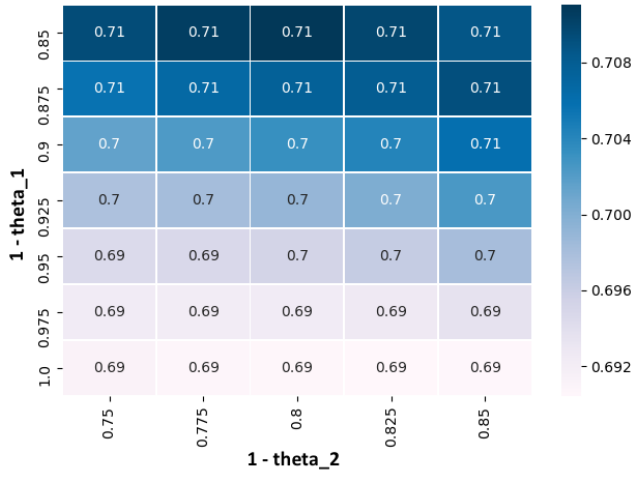
5.2.1 Normalization

Recall that, GFS involves two quantile thresholds – θ_1 and θ_2 – to assign a fuzzified score to each gene in each patient. Ranks below θ_2 are all reduced to a score of zero, those above θ_1 are given a score of 1, and intermediate ranks are interpolated to obtain a score between 0 and 1. In the case of the microarray datasets that we analysed (presented in Chapter 3), we found that the performance of GFS is robust against changes in the upper and lower quantile thresholds, when they are varied within specific ranges – i.e. silhouette scores with respect to the underlying phenotypes remain stable across θ_1 and θ_2 values. In the following experiment, we similarly test the robustness of GFS on our HCC RNA-Seq dataset against changes in the quantile thresholds θ_1, θ_2 and discuss the findings in the context of sampling stochasticity.

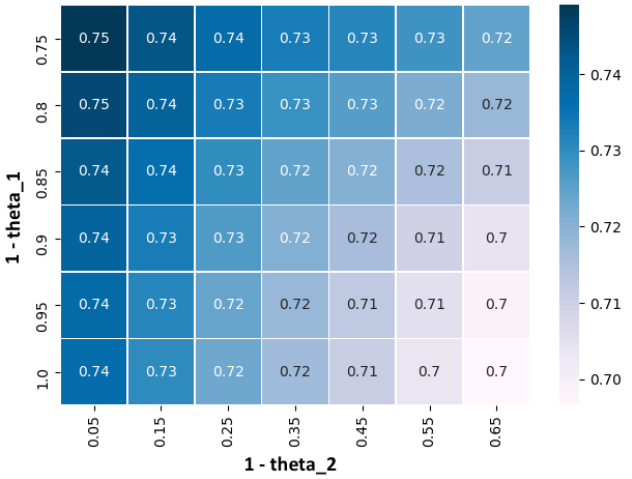
For a given θ_1, θ_2 combination,

1. Apply GFS to the RNA-Seq count matrix, and microarray gene expression matrices
2. Apply PCA transform to the GFS-transformed matrices
3. Compute silhouette scores for each of the first 5 principle components, given the sample phenotype labels.
4. Select the maximum of the silhouette scores obtained in 3.

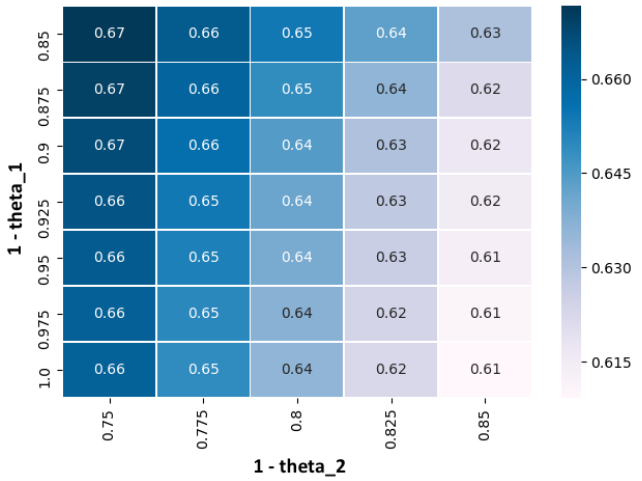
Fig 5.1 shows heatmaps of the resultant silhouette scores for different θ_1, θ_2 combinations, for the RNA-Seq dataset, and the two microarray batches. Separate heatmaps are generated for higher and lower θ_2 ranges.



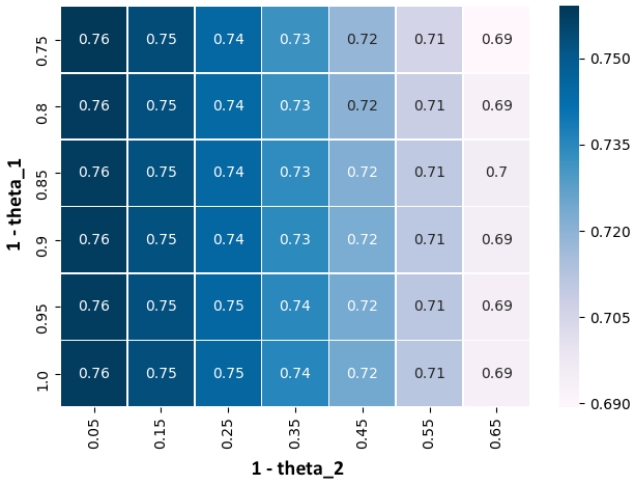
(a) HCC Microarray (Roessler et al.): High θ_2 cutoff range



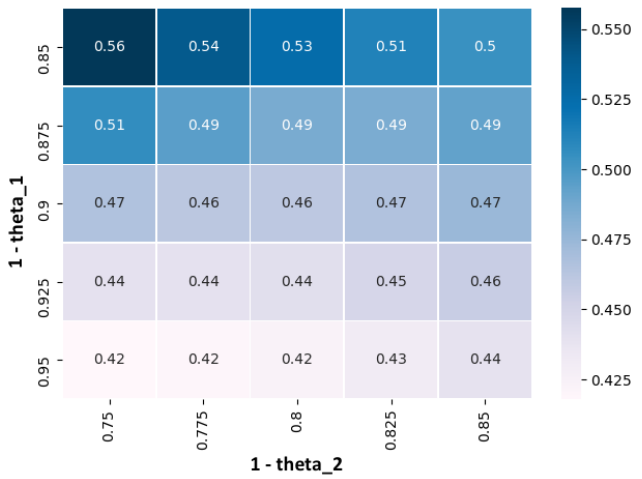
(b) HCC Microarray (Roessler et al.): Low θ_2 cutoff range



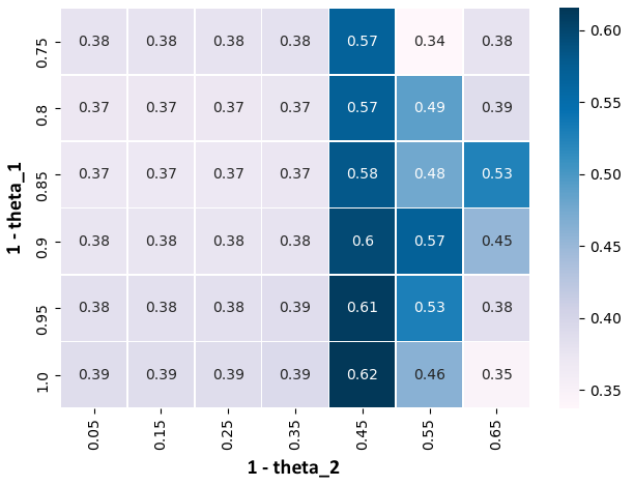
(c) HCC Microarray (Burchard et al.): High θ_2 cutoff range



(d) HCC Microarray (Burchard et al.): Low θ_2 cutoff range



(e) HCC RNA-Seq (counts): High θ_2 cutoff range



(f) HCC RNA-Seq (counts): Low θ_2 cutoff range

Figure 5.1: Silhouette scores on applying GFS with varying θ_1, θ_2 thresholds on HCC datasets – RNA-Seq and Microarray

The following observations can be made with regard to the heatmaps:

- For microarrays, the silhouette scores show only a small difference across the different θ ranges (Fig. 5.1 (a), (b), (c), (d)), and are thus relatively robust against changes in θ_1 and θ_2.
- For RNA-Seq expression data, we found that accounting for the highest and lowest expressed genes leads to deterioration in the silhouette score. We suspect that this is due to the effect of sampling stochasticity.

Note that the number of genes whose expression is measured in the RNA-Seq and microarray datasets is not the same. Therefore, we repeat the above experiment, restricting the analysis to genes common between RNA-Seq and microarray (Figure 5.2). This results in lower silhouette scores than accounting for all the expressed genes in the data, but the scores are more robust against changes in θ_1 and θ_2. This reaffirms the understanding that the broader dynamic range of expression in RNA-Seq data comes at the cost of increased stochasticity.

(a) HCC RNA-Seq: High θ_2 cutoff range

(b) HCC RNA-Seq: Low θ_2 cutoff range

Figure 5.2: Silhouette scores obtained by applying GFS with varying θ_1, θ_2 thresholds on RNA-Seq HCC data restricted to genes common between microarrays and RNA-Seq

A model for sampling stochasticity in RNA-Seq data

To understand the nature of this observed stochasticity in RNA-Seq data, we appeal to the following Bernoulli trial model. This model is based on the assumption that RNA-Seq only samples a subset of transcripts, and that the transcripts have to (fairly) compete with each

other to get sequenced. To explain our model, we first describe a simpler setting that represents a fixed number of sequenced reads being mapped to two competing transcripts.

An analogous setting to emulate sampling stochasticity – 'the red-ball model'

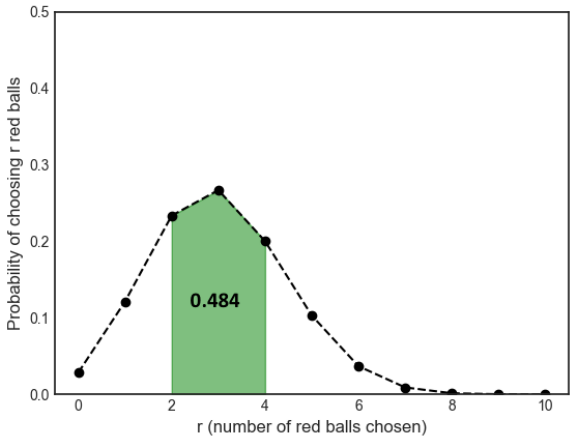
A box (B_1) contains 300 red balls and 700 blue balls (300 and 700 transcripts of competing transcript species). If a ball is drawn out (transcript is sequenced), there is a probability $p = 0.3$ that the ball is red (read is mapped to the less abundant transcript species), and a probability $q = 0.7$ that the ball is blue (read is mapped to the more abundant transcript species). If 10 balls are drawn (only a limited total number of transcripts can be sequenced), the expected number of red and blue balls are 3 and 7 respectively. This is true if the number of times we make the 10-ball draw is very large, and if the numbers of red and blue balls observed in each draw are averaged. However, if only a single draw of 10 balls is made, what is the probability that r red balls are drawn? Equivalently, given a fixed number of total reads, what is the probability of a given count value being assigned to one of the transcripts?

Computing probabilities of events in the sample space

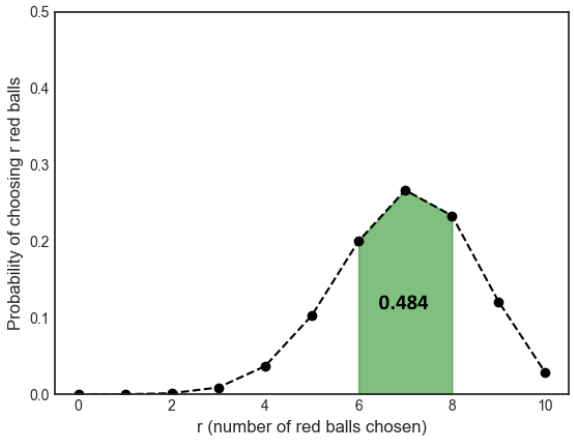
When 10 balls are drawn, the possible outcomes include: 0 red balls (10 blue balls), 1 red ball (9 blue balls), ... , 10 red balls (and 0 blue balls). The probability of the 10-ball draw containing r red balls is given by:

$$P[num\_red\_balls = r] = \binom{10}{r} \times 0.3^r \times 0.7^{10-r} \tag{5.1}$$

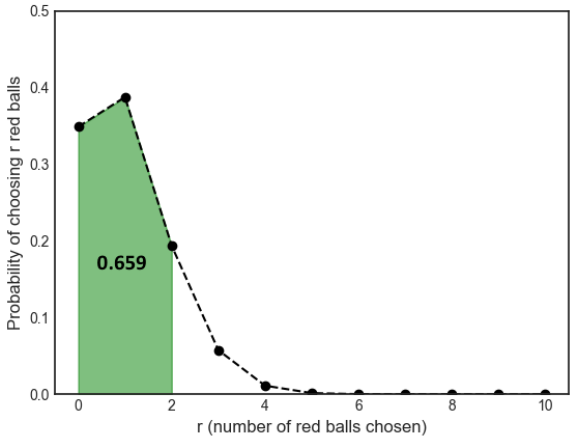
Figure 5.3 (a) shows the probabilities of all possible events when 10 balls are drawn from the box B_1. Note that the expected number of red balls is 3, and the probability that the number of red balls in the 10-ball draw is 3 or 3 ± 1 is given by the region colored green in Fig. 5.3 (a).



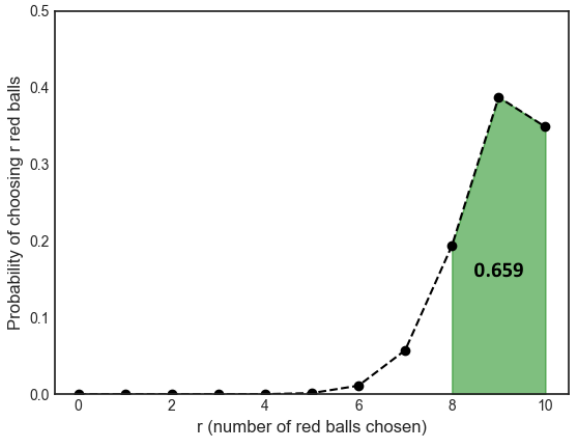
(a) A draw of 10 balls from 300 red and 700 blue balls



(b) A draw of 10 balls from 700 red and 300 blue balls



(c) A draw of 10 balls from 100 red and 900 blue balls



(d) A draw of 10 balls from 900 red and 100 blue balls

Figure 5.3: Probability distribution of choosing r red balls in a 10-ball draw taken from a set of balls in a box

In the transcript mapping scenario, the area of this region indicates the probability of a transcript species being assigned a read count value, which precisely (± 1) represents its actual abundance. The greater the area of the shaded region, the more likely it is to get precise measurement of the transcript species' abundance.

On the other hand, if a box B_1' contains 700 red balls and 300 blue balls, then the probabilities of possible events when 10 balls are drawn from B_1' are shown in Figure 5.3 (b). The area of the green shaded region (area = 0.484 units) is equal in 5.3 (a) and (b). This is expected due to a symmetry in the binomial distribution used in the red-ball model;, the effect of sampling stochasticity is the same on the high and low expressed transcript species.

Next, we change the red-to-blue ball ratio (transcript abundance ratio), and plot the probabilities of getting r red balls in the 10-ball draw (Figure 5.3(c),(d)). The area of the green shaded region in both the cases is 0.659 units. This leads us to another interesting observation: the more skewed the abundance ratio of competing transcripts, the more likely it is that the assigned counts to the transcript species are precise.

Thus, the red-ball model provides important and interesting insights regarding the nature of sampling stochasticity. However, it is a rough representation of the actual scenario, since it overlooks that a transcript is not sequenced in one shot but in pieces. For a more accurate model, transcript species can be divided into segments of 1,000 bases each, where each segment h has a read count n_h that follows the same red-ball model above. If m is the read count total over all segments over all transcripts, then each n_h/m is an estimate for p (the relative abundance of the transcript species among all transcript species). By the central limit theorem, these n_h/m are normally distributed, and taking the average of these results in the expected value for p, with a variance proportional to the inverse of the number of segments that the transcript species has. Thus, the longer the gene, the more segments there are, and the more accurate this expected value for p is.

Note that $m \times p/10^6$ equals the RPKM for this transcript species, from which the TPM [WKL12] can be computed (TPM is the RPKM for this transcript species divided by the sum of RPKM of all transcript species in the sample, multiplied by a million). Due to this equivalence, the TPM measure also has a variance proportional to the inverse of the number of segments the transcript species has (i.e. its length).

In short, shorter transcript species more strongly exhibit the stochasticity described by the red-ball model, while measures like RPKM, FPKM, TPM mitigate long transcript species against this stochasticity.

Discretized-GFS

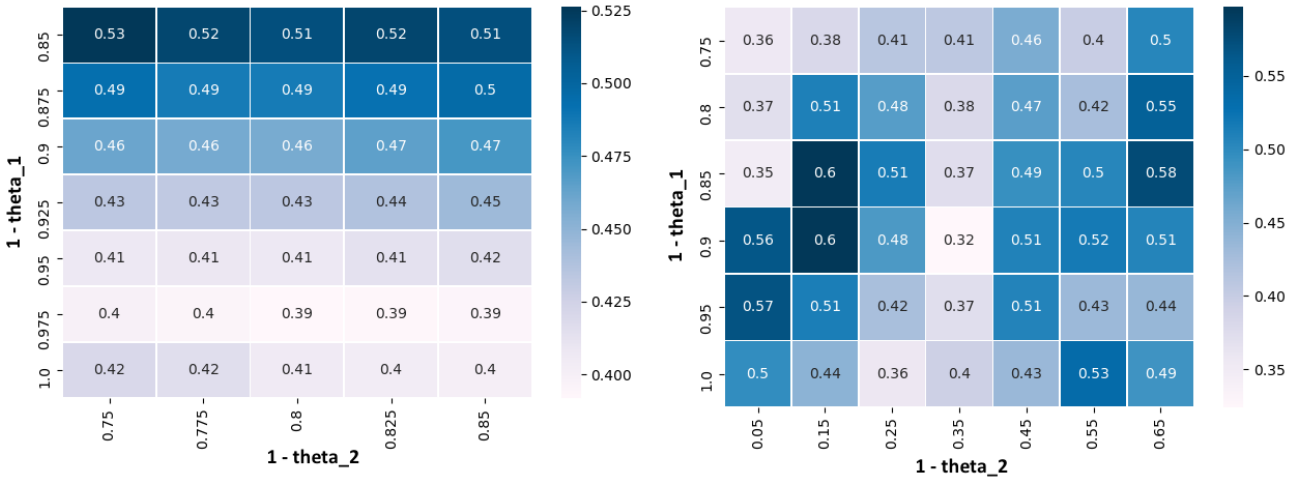
To attenuate the effect of sampling stochasticity on RNA-Seq gene expression, we propose the use of discretized-GFS (D-GFS), as introduced in the work of Goh and Wong [GW16a] for normalizing proteomics data. In D-GFS, gene ranks are interpolated into discrete bins, instead of a continuous interval, thus, irrelevant variation is reduced. A formal description of D-GFS (with 4 bins) is as follows:

Let $r(g_i, p_j)$ be the rank of gene expression of a gene g_i in tissue p_j, and $q(p_j, \theta)$ be the rank corresponding to the upper θth quantile of gene expression in tissue p_j. Also, let θ_1 and θ_2 be the upper and lower quantile thresholds respectively (as in GFS), and $\Delta\theta = \theta_2 - \theta_1$. Then, the D-GFS score $ds(g_i, p_j)$ assigned to a gene g_i in tissue p_j is given by the following function:

$$ds(g_i, p_j) = \begin{cases} 1, & \text{if } q(p_j, \theta_1) \leq r(g_i, p_j) \\ 0.8, & \text{if } q(p_j, \theta_1) > r(g_i, p_j) \geq q(p_j, \theta_2 - \frac{3\times\Delta\theta}{4}) \\ 0.6, & \text{if } q(p_j, \theta_2 - \frac{3\times\Delta\theta}{4}) > r(g_i, p_j) \geq q(p_j, \theta_2 - \frac{2\times\Delta\theta}{4}) \\ 0.4, & \text{if } q(p_j, \theta_2 - \frac{2\times\Delta\theta}{4}) > r(g_i, p_j) \geq q(p_j, \theta_2 - \frac{\Delta\theta}{4}) \\ 0.2, & \text{if } q(p_j, \theta_2 - \frac{\Delta\theta}{4}) > r(g_i, p_j) \geq q(p_j, \theta_2) \\ 0, & \text{otherwise} \end{cases} \tag{5.2}$$

Silhouette scores obtained after applying D-GFS normalization on the RNA-Seq dataset are shown in the heatmaps in Fig. 5.4.

We observe that D-GFS captures more meaningful variation in the data, and is more robust than GFS against variation in the quantile thresholds – the silhouette scores with respect to the normal and HCC phenotypes are both higher and more stable (Fig 5.1 (f), 5.4 (b)). This is more easily observed in the heatmaps depicting lower expression ranges in which the $1 - \theta_2$ value was varied across a wide range (0.05 – 0.65). This suggests that introducing discretization in the GFS methodology helps in reversing the adverse impact of sampling stochasticity.



(a) HCC RNA-Seq: High θ_2 cutoff range (with discretization)

(b) HCC RNA-Seq: Low θ_2 cutoff range (with discretization)

Figure 5.4: Silhouette scores on applying discretized GFS with varying θ_1, θ_2 thresholds on HCC datasets – RNA-Seq

We note that even after applying D-GFS, the silhouette scores obtained on RNA-Seq data do not compare favorably with those obtained on the microarray data. This difference could be attributed to three reasons. First, the RNA-Seq dataset is potentially more heterogeneous than the microarray datasets, which are reported to be predominantly HBV-positive. Second, in the case of RNA-Seq, the ratio of tumor and non-tumor samples is more imbalanced, which likely weakens the differential signal between the normal and HCC phenotypes. Third, the difference could be caused by some anomaly in the silhouette score (described in the next section).

5.2.2 Concordance across platforms, and the curious case of silhouette scores

It is also interesting to study the concordance between the HCC microarray datasets and RNA-Seq dataset. For this, we verify whether the subnetworks significantly differentially expressed (as reported by PFSNet and SPSNet) in microarrays are also able to differentiate between the HCC tumor and adjacent non-tumor samples in RNA-Seq data. In the analysis below, we present our encounter of a particular scenario, that is an instructive example on the meaning and interpretation of silhouette scores.

The procedure for our analysis is described below:

1. Merge the two microarray datasets, and run PFSNet/SPSNet on the new dataset comparing HCC tumor and adjacent normal samples. (Although we use both PFSNet and SPSNet for this analysis, we expect PFSNet to perform better as it should be unable to detect and report subnetworks that capture heterogeneity arising from batch effects due to merging of the two microarray datasets.).

   PFSNet: 12 subnets consisting of 124 genes and belonging to 4 different pathways (*Jak Stat Signaling*, *Wnt Signaling*, *Pentose Phosphate Pathway*, *Proteosome Degradation*) were reported significant by PFSNet on this microarray data.

   SPSNet: 31 subnets consisting of 245 genes and belonging to 13 different pathways were reported significant by SPSNet on the microarray data. The 12 subnets reported by PFSNet are a subset of the 31 subnets reported by SPSNet.

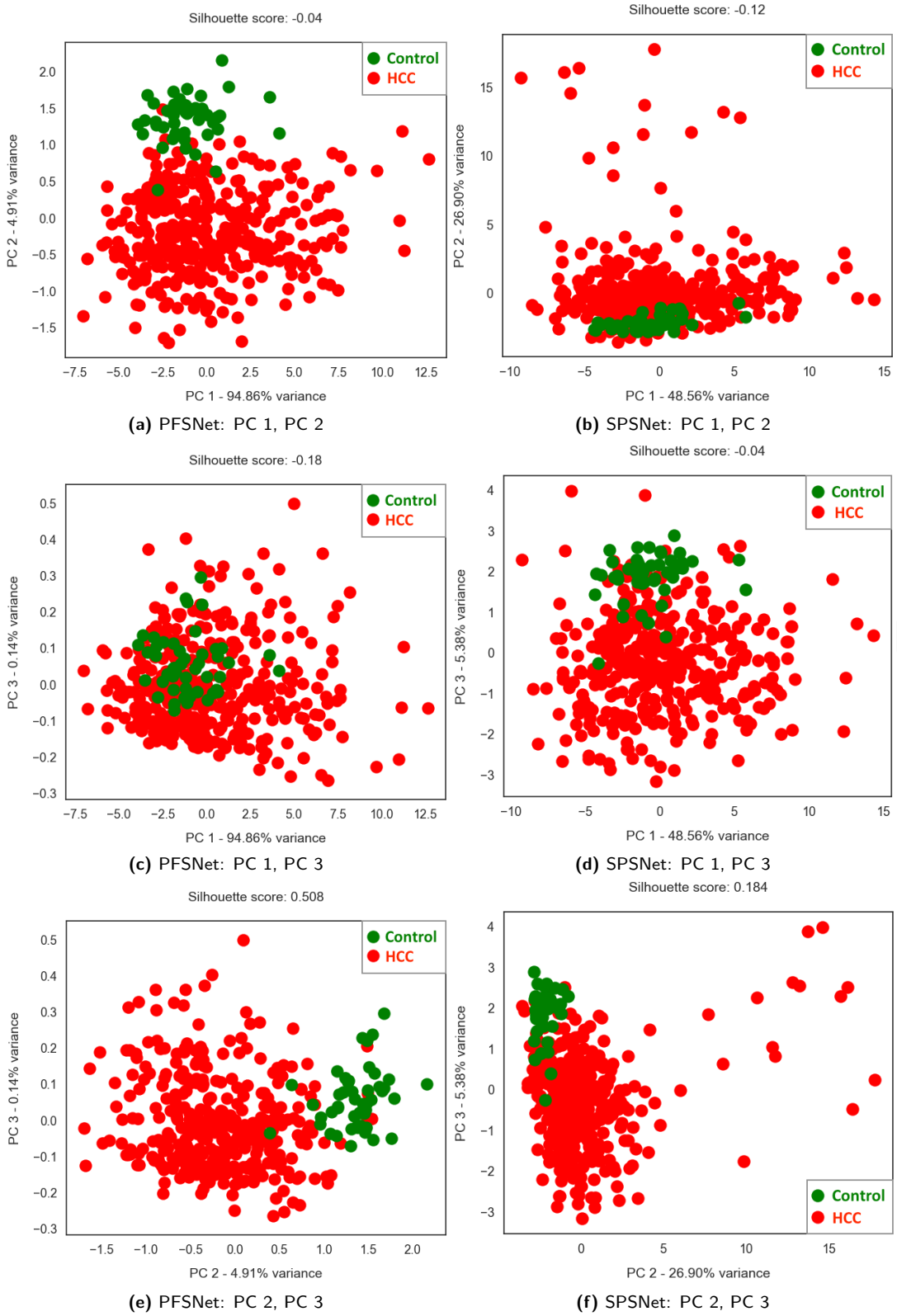
2. Run PFSNet/SPSNet comparing HCC tumor and adjacent normal samples in the RNA-Seq dataset. In this run, use subnetworks reported significant by PFSNet/SPSNet on microarray as input subnetworks to be tested for significance on the RNA-Seq dataset.

   PFSNet: Of the 12 subnets provided, PFSNet reported 4 sigificant subnets in this run. These subnetworks consisted of 86 genes and belonged to 2 pathways – *Jak Stat Signaling* and *Proteosome Degradation*.

   SPSNet: Of the 31 subnets provided, SPSNet reported 22 sigificant subnets in this run. The significant subnetworks consisted of 220 genes, and belonged to 11 pathways. The 4 subnets reported by PFSNet are included within the 22 subnets reported by SPSNet.

3. Apply PCA transformation to PFSNet/SPSNet scores of subnetworks reported significant in run 2, and generate scatter plots to visualize PC scores.

4. Compute silhouette scores to quantify the extent to which same phenotype samples have been clustered together.



(a) PFSNet: PC 1, PC 2

(b) SPSNet: PC 1, PC 2

(c) PFSNet: PC 1, PC 3

(d) SPSNet: PC 1, PC 3

(e) PFSNet: PC 2, PC 3

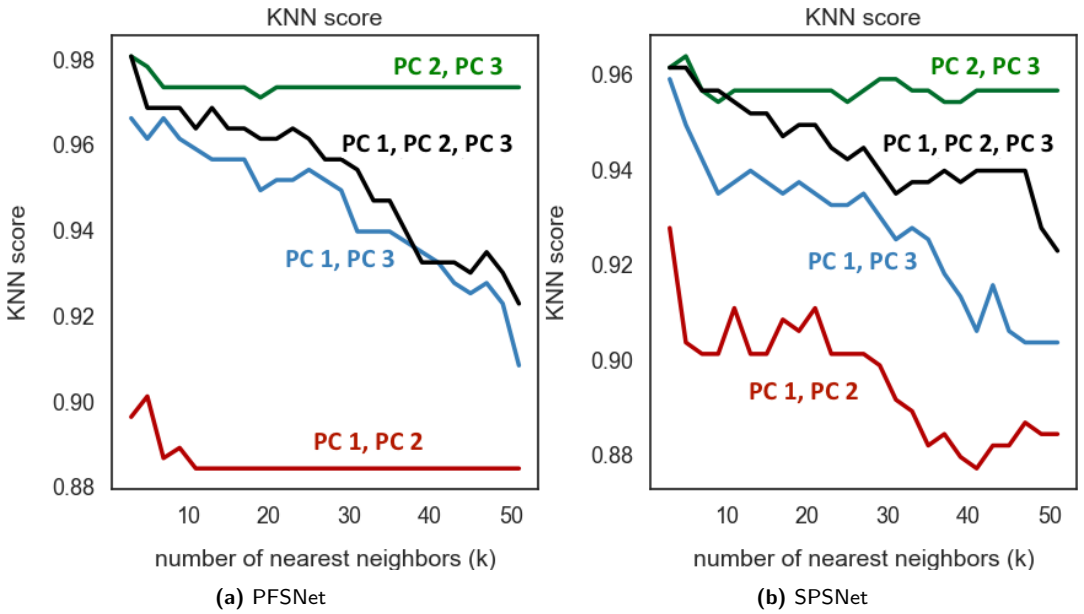
(f) SPSNet: PC 2, PC 3

Figure 5.5: Using subnetworks reported significant by PFSNet on microarray to compare control and HCC samples generated on RNA-Seq platform

Fig 5.5 shows that the 4 subnetworks selected by PFSNet are successful in separating the HCC

and control phenotypes. However, it is interesting to note that the scatter plots show that even though the plots corresponding to (PC 1, PC 2) and (PC 2, PC 3) both show good separation between the control and HCC samples, the silhouette score corresponding to (PC 1, PC 2) is extremely low: -0.04. This shows an anomaly associated with silhouette scores, possibly arising due to high and unbalanced dispersion. The silhouette score is calculated based on the mean intra-cluster distance a and the mean nearest-cluster distance b for each patient, as $(b-a)/max(a,b)$ [Rou87]. Here, b is the distance between a sample and the nearest cluster (amongst clusters excluding the one to which the sample belongs). Thus, for data with high dispersion, the mean intra-cluster distance (a) would be a large number, often greater than the mean nearest cluster distance (b), resulting in an overall negative value.

The above analysis was carried out in the setting $\theta_1 = 5\%$ and $\theta_2 = 15\%$ in all runs of PFSNet and SPSNet. While selecting other values of θ_1 and θ_2 for PFSNet/SPSNet runs on microarray and RNA-Seq datasets may result in a better separation between the tumor and non-tumor samples, we intentionally present the above scenario, because it serves as a cautionary example against the misinterpretation of silhouette scores.



(a) PFSNet **(b)** SPSNet

Figure 5.6: Variation in kNN score based on first three principle components with varying k

To address the anomaly, we propose a new way of assessing the separation of samples into

clusters. Our method relies on kNN score, a metric we define as the proportion of samples whose actual label matches the labels of a majority ($> 50\%$) of its k nearest neighbors. Figure 5.6 shows the kNN scores obtained using different combinations of the first three principle components based on scores of significant subnets reported by PFSNet. Only odd k's were chosen to perform this experiment in order to avoid the issue of having no majority label.

From the pattern, it can be inferred that the combinations (PC 1, PC 2) and (PC 2, PC 3) result in good clustering of the phenotypes, as they remain stable at a high kNN score even with increasing k, whereas the kNN score corresponding to the combination (PC 1, PC 3) continues to decrease with increasing k, although it starts off at a high score. Interestingly, this matches with the observation from the PCA scatter plot for (PC 1, PC 3) – although the control samples are clustered together, they are surrounded by HCC samples and not separated from them. This example demonstrates the potential of the kNN score metric for analyzing clustering effectiveness where silhouette score is not a reasonable metric due to high data dispersion.

Further, it is possible to report kNN scores adjusted for chance. The adjusted scores can be computed by permuting the class labels to get a null distribution of kNN scores, so that we could obtain (a) a p-value for the kNN score; and/or (b) an expected value of the kNN score which can be used to compute adjusted kNN scores as:

$$adjusted\_kNN\_score = \frac{(observed\_kNN\_score - E[kNN\_score])}{(1 - E[kNN\_score])}. \tag{5.3}$$

5.2.3 Integrative analysis of multi-platform data

In our final analysis in the case-study, we perform a meta-analysis of our multi-platform HCC dataset using SPSNet. The procedure for our analysis is as below:

1. Normalize the microarray HCC datasets independently using GFS, and the RNA-Seq dataset using D-GFS with appropriate quantile thresholds. Based on Figure 5.4, we

chose thresholds around which the performance of GFS does not change considerably: for microarrays, $\theta_1 = 5\%, \theta_2 = 15\%$; for RNA-Seq, $\theta_1 = 15\%, \theta_2 = 35\%$.

2. Merge the RNA-Seq dataset [GHF+16] and two microarray datasets [RJB+10, BZL+10] (on common genes) to prepare the multi-platform dataset.

3. Run SPSNet, sans its GFS preprocessing step, to compare HCC tumor and adjacent control samples in the merged dataset.

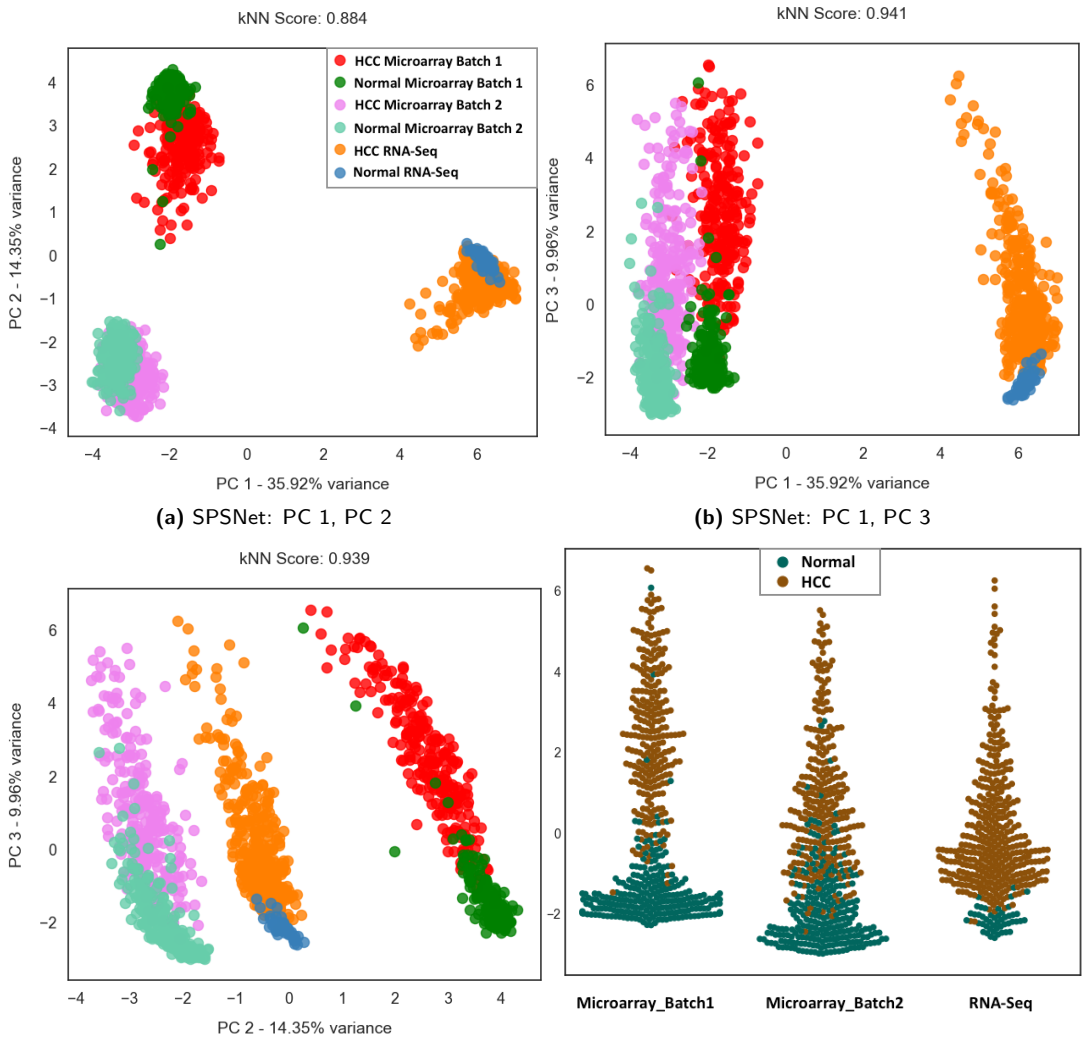
4. List the genes belonging to the subnetworks reported significant by SPSNet in step 3.

5. Create a feature matrix of GFS values (from step 1) of genes listed in step 4.

6. Apply a PCA transform and generate scatter plots of the principle component scores.

7. Compute kNN scores to quantify the extent to which same phenotype, same batch samples have been clustered together.

From Fig. 5.7 (a-c), it can be seen that the genes from significant subnetworks in SPSNet are able to identify three levels of heterogeneity in the data – platforms, microarray batches, and phenotypes. Particularly, PC 1 and PC 2 are both able to separate platforms and microarray batches, and the sign of PC 1 indicates whether the platform is microarray (negative) or RNA-Seq (positive). From Fig. 5.7 (d), the disease effect is concentrated in PC 3. Also, PC 3 is relatively free from platform/batch effects. This illustrates the potential of SPSNet in systematically uncovering heterogeneity in a meta-analysis of multi-platform datasets.



(a) SPSNet: PC 1, PC 2

(b) SPSNet: PC 1, PC 3

(c) SPSNet: PC 2, PC 3

(d) SPSNet: PC 3

Figure 5.7: (a-c) PCA scatter plots of GFS transformed expression of genes belonging to significant subnetworks reported by SPSNet on a multi-platform gene expression dataset on containing control and HCC samples generated with 2 batches of microarray data and 1 batch of RNA-Seq data (d) Swarmplot of PC 3 showing normal and HCC samples across all datasets

5.3 Conclusion

Data generated from RNA-Seq platforms and microarrays have fundamental differences. RNA-Seq measures expression across a broader range than microarrays. However, when sequencing depth is insufficient, RNA-Seq is subject to sampling stochasticity. This means that the abundance of some transcripts is inaccurately represented in RNA-Seq, because the sequencer generates a limited number of total reads that sometimes result in a non-uniform sampling of transcripts across the transcriptome. Thus, when analyzing RNA-Seq data using methods that

are originally designed for and tested on microarray samples, consideration of these differences becomes critical.

We also note that RNA-Seq data that is normalized by RPKM or FPKM should be interpreted with caution. These methods normalize read counts by million mapped reads before normalizing by gene length. Therefore, while they are suitable for comparing genes within a given sample, they also tend to induce distortions in the relative expression levels of a gene across samples.

In this chapter, we presented a case-study on a Hepatocellular Carcinoma RNA-Seq dataset and two HCC microarray datasets. In our discussion of the analysis of these datasets, we (a) demonstrated the potential of our methods for normalization and heterogeneity analysis – GFS and SPSNet – to analyze RNA-Seq datasets, and (b) highlighted some critical points regarding the application of these methods on RNA-Seq data.

First, we observed that the stability of RNA-Seq data is adversely affected by sampling stochasticity. We proposed a model to explain this stochasticity based on the assumption that only a limited portion of the transcriptome is sequenced by RNA-Seq, and there is competition amongst transcripts to be sequenced. To attenuate the effect of this imprecision in RNA-Seq data, we propose the use of discretized-GFS (D-GFS). In D-GFS, gene ranks within a certain range are interpolated into discrete bins, instead of a continuous interval, and thus, confounding effects due to stochasticity are reduced.

Second, we found that the concordance between the HCC microarray and RNA-Seq datasets is high, with respect to the features distinguishing tumor samples from adjacent normal samples, as reported by PFSNet/SPSNet. Interestingly, in our analysis, we noticed an anomalous behavior of silhouette scores – their quantification of the extent of clustering is inaccurate when data dispersion is high. Therefore, we described a method using the kNN score metric (the number of samples whose label matches with majority of their k nearest neighbors) for a more accurate quantitative assessment of clustering. Visualizing kNN scores across a series of odd k values in an increasing order provides an easy-to-compute, effective substitute for silhouette score in scenarios where data-dispersion is high.

Finally, we observed that SPSNet is able to discover and isolate heterogeneity at the level of platform, batch, and phenotype, in a multi-platform analysis of merged dataset containing microarray and RNA-Seq HCC datasets. This is suggestive of its potential use in performing multi-platform meta-analysis of gene expression datasets.

Chapter 6

Conclusion

"It's more fun to arrive at the conclusion than to justify it."

– Malcolm Forbes

In this thesis, we discussed our research work in three parts – preprocessing expression data, analyzing expression to uncover undeclared heterogeneity, and generalizing the heterogeneity analysis across platforms. Below, we provide a summary of our contributions, and discuss potential directions for future work.

6.1 Summary

6.1.1 Role of preprocessing in gene expression analysis

Despite the critical impact of normalization on downstream gene expression analysis, popular techniques often fail to enhance the quality of expression data. We proposed a technique – Gene Fuzzy Scores (GFS) – a simple rank-based normalization technique that effectively retains important sources of variation while removing obscuring noise. Using publicly available datasets, we compared our approach against three other popular normalization methods – mean-scaling, quantile normalization, and z-score normalization – with respect to the quality, consistency, and biological coherence of the normalized output. The performance of GFS is

equal to or better than the selected methods in all the three respects. Moreover, we illustrated that as sample size increases, the performance of GFS improves further.

6.1.2 Differential gene expression analysis of heterogeneous phenotypes

From batch effects to disease subtypes, heterogeneity has been gathering a lot of attention in many different contexts, in the past decade. However, this heterogeneity in gene expression datasets is seldom understood or classified in advance. To our knowledge, there are no network-based approaches at present to systematically analyze such undeclared heterogeneity. Therefore, we proposed SPSNet, a generalization of PFSNet, which handles undeclared heterogeneity in gene expression samples by identifying subnetworks associated with hidden subpopulations within phenotypes. We demonstrated that SPSNet shows low false-positive rate, high reproducibility, and high sensitivity in analyzing expression data with undeclared heterogeneity.

6.1.3 Analysis of RNA-Seq datasets

RNA-Seq measures expression across a broader range than microarrays, but is subject to sampling stochasticity when sequencing depth is insufficient. Consideration of this factor is important to ensure reliable analysis of RNA-Seq datasets. We presented a case-study demonstrating the potential of our methods for normalization and heterogeneity analysis – GFS and SPSNet – to analyze RNA-Seq datasets. We presented a Bernoulli trial-based model to explain sampling stochasticity and proposed the use of discretized-GFS (D-GFS) to attenuate the stochasticity effect. We also proposed a method using a kNN-score metric, which we define as the number of samples whose label matches with majority of their k nearest neighbors as an effective substitute for silhouette scores, which show anomalous behavior when dispersion in the underlying data is high. We also demonstrated the potential of SPSNet in performing

multi-platform meta-analysis of gene expression datasets.

6.2 Future work

6.2.1 Improving the design and performance of SPSNet

It would be interesting to explore ways to improve the design and performance of SPSNet. A few important questions in this direction are:

- The current scheme of generating candidate subnetworks for SPSNet is simplistic, and is independent of the gene expression data being analyzed. How can it be improved?
- Can we improve the way SPSNet picks reference subsamples to represent potential subpopulations? Given a reference subsample, is it possible to identify the entire subpopulation?
- The SPSNet scores of candidate subnetworks are tested for significance using the theoretical t-distribution, which may not be a universally appropriate null distribution. How can better null distributions be designed to replace this?
- How can the homogeneity of reference subsamples picked by SPSNet be evaluated in the absence of actual subtype labels? Is it possible to e.g. infer this homogeneity by drawing a PCA scatterplot of the SPSNet scores of these subsamples and assessing the degree of their clustering?

In addition, SPSNet features could be used to perform clustering for subpopulation prediction within phenotypes, and to build classifiers for predicting subpopulations in new data.

6.2.2 Heterogeneity analysis of paired gene expression data

There are many scenarios where samples in heterogeneous datasets are paired – tumor tissues before and after treatment, developmental stages of cells through different time points, analysis of the same samples on independent platforms, etc. Currently, SPSNet analyzes the heterogeneity within two potentially heterogeneous samples, and does not consider any pairing information between tissues. Incorporating the ability to analyze paired data in the SPSNet methodology would further increase its utility.

6.2.3 Generalized model of sampling stochasticity in RNA-Seq data

We proposed a Bernoulli trial model using two transcripts competing to be sequenced, and show interesting insights regarding the effect of sampling stochasticity in RNA-Seq data. However, it is possible to extend the current model using the generalized Bernoulli distribution to account for multiple competing transcripts. It would be interesting to see if the extension captures sampling stochasticity in a more realistic manner.

Bibliography

[ACPS06] David B Allison, Xiangqin Cui, Grier P Page, and Mahyar Sabripour. Microarray data analysis: from disarray to consolidation and consensus. Nature Reviews Genetics, 7(1):55–65, 2006.

[AED+00] Ash A Alizadeh, Michael B Eisen, R Eric Davis, Chi Ma, Izidore S Lossos, Andreas Rosenwald, Jennifer C Boldrick, Hajeer Sabet, Truc Tran, Xin Yu, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 403(6769):503–511, 2000.

[ALP+12] Andrey Alexeyenko, Woojoo Lee, Maria Pernemalm, Justin Guegan, Philippe Dessen, Vladimir Lazar, Janne Lehtiö, and Yudi Pawitan. Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. BMC Bioinformatics, 13(1):226, 2012.

[APH15] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics, 31(2):166–169, 2015.

[ASS+02] Scott A Armstrong, Jane E Staunton, Lewis B Silverman, Rob Pieters, Monique L den Boer, Mark D Minden, Stephen E Sallan, Eric S Lander, Todd R Golub, and Stanley J Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature Genetics, 30(1):41–47, 2002.

[BAAH04] Rainer Breitling, Patrick Armengaud, Anna Amtmann, and Pawel Herzyk. Rank products: a simple, yet powerful, new method to detect differentially

regulated genes in replicated microarray experiments. FEBS letters, 573(1-3):83–92, 2004.

[BSW+11] Katherine J Baines, Jodie L Simpson, Lisa G Wood, Rodney J Scott, and Peter G Gibson. Transcriptional phenotypes of asthma defined by gene expression profiling of induced sputum samples. Journal of Allergy and Clinical Immunology, 127(1):153–160, 2011.

[BZL+10] Julja Burchard, Chunsheng Zhang, Angela M Liu, Ronnie TP Poon, Nikki PY Lee, Kwong-Fai Wong, Pak C Sham, Brian Y Lam, Mark D Ferguson, George Tokiwa, et al. MicroRNA-122 as a regulator of mitochondrial metabolic gene network in hepatocellular carcinoma. Molecular Systems Biology, 6(1):402, 2010.

[CMT+16] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, et al. A survey of best practices for RNA-seq data analysis. Genome Biology, 17(1):13, 2016.

[CVFB03] Chris Cheadle, Marquis P Vawter, William J Freed, and Kevin G Becker. Analysis of microarray data using z-score transformation. Journal of Molecular Diagnostics, 5(2):73–81, 2003.

[DZD12] Mitchell L Drumm, Assem G Ziady, and Pamela B Davis. Genetic variation and clinical heterogeneity in cystic fibrosis. Annual review of pathology, 7:267, 2012.

[FPS13] R Fisher, L Pusztai, and C Swanton. Cancer heterogeneity: implications for targeted therapeutics. British journal of cancer, 108(3):479–485, 2013.

[GB07] Jelle J Goeman and Peter Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics, 23(8):980–987, 2007.

[GCK+11] Ludwig Geistlinger, Gergely Csaba, Robert Küffner, Nicola Mulder, and Ralf Zimmer. From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems. Bioinformatics, 27(13):i366–i373, 2011.

[GGAW15] Wilson Wen Bin Goh, Tiannan Guo, Ruedi Aebersold, and Limsoon Wong. Quantitative proteomics signature profiling based on network contextualization. Biology Direct, 10:71, 2015.

[GHF+16] Robert L Grossman, Allison P Heath, Vincent Ferretti, Harold E Varmus, Douglas R Lowy, Warren A Kibbe, and Louis M Staudt. Toward a shared vision for cancer genomic data. New England Journal of Medicine, 375(12):1109–1112, 2016.

[GST+99] Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286(5439):531–537, 1999.

[GVDGDKVH04] Jelle J Goeman, Sara A Van De Geer, Floor De Kort, and Hans C Van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics, 20(1):93–99, 2004.

[GW16a] Wilson Wen Bin Goh and Limsoon Wong. Advancing clinical proteomics via analysis based on biological complexes: A tale of five paradigms. Journal of Proteome Research, 15(9):3167–3179, 2016.

[GW16b] Wilson Wen Bin Goh and Limsoon Wong. Evaluating feature-selection stability in next-generation proteomics. Journal of Bioinformatics and Computational Biology, 14(05):1650029, 2016.

[HHS+13] Winston A Haynes, Roger Higdon, Larissa Stanberry, Dwayne Collins, and Eugene Kolker. Differential expression analysis for pathways. PLoS Comput

Biol, 9(3):e1002967, 2013.

[HIZ12] Kasper D Hansen, Rafael A Irizarry, and WU Zhijin. Removing technical variability in rna-seq data using conditional quantile normalization. Biostatistics, 13(2):204–216, 2012.

[HSK+02] Judith N Haslett, Despina Sanoudou, Alvin T Kho, Richard R Bennett, Steven A Greenberg, Isaac S Kohane, Alan H Beggs, and Louis M Kunkel. Gene expression comparison of biopsies from duchenne muscular dystrophy (DMD) and normal skeletal muscle. Proceedings of the National Academy of Sciences USA, 99(23):15000–15005, 2002.

[JSFDK12] Ali Jabbari, Mayte Suárez-Fariñas, Scott Dewell, and James G Krueger. Transcriptional profiling of psoriasis using RNA-seq reveals previously unidentified differentially expressed genes. The Journal of investigative dermatology, 132(1):246, 2012.

[KD05] Purvesh Khatri and Sorin Drăghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics, 21(18):3587–3595, 2005.

[KDOK02] Purvesh Khatri, Sorin Draghici, G Charles Ostermeier, and Stephen A Krawetz. Profiling gene expression using onto-express. Genomics, 79(2):266–270, 2002.

[KSB12] Purvesh Khatri, Marina Sirota, and Atul J Butte. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol, 8(2):e1002375, 2012.

[KvIH+12] Thomas Kelder, Martijn P van Iersel, Kristina Hanspers, Martina Kutmon, Bruce R Conklin, Chris T Evelo, and Alexander R Pico. WikiPathways: building research communities on biological pathways. Nucleic Acids Research, 40(D1):D1301–D1307, 2012.

[LCF+13] Tristram A Lett, Mallar M Chakavarty, Daniel Felsky, Eva J Brandl, Arun K Tiwari, Vanessa F Gonçalves, Tarek K Rajji, Z Jeffery Daskalakis, Herbert Y Meltzer, Jeffery A Lieberman, et al. The genome-wide supported microRNA-137 variant predicts phenotypic heterogeneity within schizophrenia. Molecular Psychiatry, 18(4):443–450, 2013.

[LLCW15] Kevin Lim, Zhenhua Li, Kwok Pui Choi, and Limsoon Wong. A quantum leap in the reproducibility, precision, and sensitivity of gene expression profile analysis even when sample size is extremely small. Journal of Bioinformatics and Computational Biology, 13(04):1550018, 2015.

[LM15] Sarah R Langley and Manuel Mayr. Comparative analysis of statistical methods used for detecting differential expression in label-free mass spectrometry proteomics. Journal of Proteomics, 129:83–92, 2015.

[LSB+10] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics, 11(10):733–739, 2010.

[LSS+10] J Luo, M Schumacher, A Scherer, Despoina Sanoudou, D Megherbi, T Davison, T Shi, W Tong, L Shi, H Hong, et al. A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. Pharmacogenomics Journal, 10(4):278–291, 2010.

[LW13] Kevin Lim and Limsoon Wong. Finding consistent disease subnetworks using PFSNet. Bioinformatics, 30(2):189–196, 2013.

[MB14] Wendy C Moore and Eugene R Bleecker. Asthma heterogeneity and severity–why is comprehensive phenotyping important? The lancet. Respiratory

medicine, 2(1):10, 2014.

[MdRD+13] Laetitia Marisa, Aurélien de Reyniès, Alex Duval, Janick Selves, Marie Pierre Gaub, Laure Vescovo, Marie-Christine Etienne-Grimaldi, Renaud Schiappa, Dominique Guenot, Mira Ayadi, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med, 10(5):e1001453, 2013.

[MWM+08] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods, 5(7):621–628, 2008.

[N+13] Cancer Genome Atlas Research Network et al. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature, 499(7456):43–49, 2013.

[OGS+99] Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research, 27(1):29–34, 1999.

[PBM+07] Mario Pescatori, Aldobrando Broccolini, Carlo Minetti, Enrico Bertini, Claudio Bruno, Adele D'amico, Camilla Bernardini, Massimiliano Mirabella, Gabriella Silvestri, Vincenzo Giglio, et al. Gene expression profiling in the early phases of DMD: a constant molecular signature characterizes dmd muscle from early postnatal life throughout disease progression. FASEB Journal, 21(4):1210–1226, 2007.

[RJB+10] Stephanie Roessler, Hu-Liang Jia, Anuradha Budhu, Marshonna Forgues, Qing-Hai Ye, Ju-Seog Lee, Snorri S Thorgeirsson, Zhongtang Sun, Zhao-You Tang, Lun-Xiu Qin, et al. A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. Cancer Research, 70(24):10202–10212, 2010.

[RKL+13] Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E Mason, Nicholas D Socci, and Doron Betel. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biology, 14(9):3158, 2013.

[RMO+04] Mary E Ross, Rami Mahfouz, Mihaela Onciu, Hsi-Che Liu, Xiaodong Zhou, Guangchun Song, Sheila A Shurtleff, Stanley Pounds, Cheng Cheng, Jing Ma, et al. Gene expression profiling of pediatric acute myelogenous leukemia. Blood, 104(12):3679–3687, 2004.

[RO05] Jonathan M Raser and Erin K O'Shea. Noise in gene expression: origins, consequences, and control. Science, 309(5743):2010–2013, 2005.

[Rou87] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20:53–65, 1987.

[SDGW10] Donny Soh, Difeng Dong, Yike Guo, and Limsoon Wong. Consistency, comprehensiveness, and compatibility of pathway databases. BMC Bioinformatics, 11(1):449, 2010.

[SDGW11] Donny Soh, Difeng Dong, Yike Guo, and Limsoon Wong. Finding consistent disease subnetworks across microarray datasets. BMC Bioinformatics, 12(13):1, 2011.

[SFUdR+15] Mayte Suárez-Fariñas, Benjamin Ungar, Joel Correa da Rosa, David A Ewald, Mariya Rozenblit, Juana Gonzalez, Hui Xu, Xiuzhong Zheng, Xiangyu Peng, Yeriel D Estrada, et al. RNA sequencing atopic dermatitis transcriptome profiling provides insights into novel disease mechanisms with potential therapeutic implications. Journal of Allergy and Clinical Immunology, 135(5):1218–1227, 2015.

[SPT+01] Therese Sørlie, Charles M Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, Michael B Eisen, Matt Van De Rijn, Stefanie S Jeffrey, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences, 98(19):10869–10874, 2001.

[SRJ+06] Leming Shi, Laura H Reid, Wendell D Jones, Richard Shippy, Janet A Warrington, Shawn C Baker, Patrick J Collins, Francoise De Longueville, Ernest S Kawasaki, Kathleen Y Lee, et al. The microarray quality control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. Nature Biotechnology, 24(9):1151–1161, 2006.

[STM+05] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43):15545–15550, 2005.

[TGJ+12] M Tandon, A Gallo, S-I Jang, GG Illei, and I Alevizos. Deep sequencing of short RNAs reveals novel microRNAs in minor salivary glands of patients with sjögren's syndrome. Oral diseases, 18(2):127–131, 2012.

[TSC+14] Tiinamaija Tuomi, Nicola Santoro, Sonia Caprio, Mengyin Cai, Jianping Weng, and Leif Groop. The many faces of diabetes: a disease with increasing heterogeneity. The Lancet, 383(9922):1084–1094, 2014.

[TTC01] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences, 98(9):5116–5121, 2001.

[TWP+10] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon

Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology, 28(5):511–515, 2010.

[VDD11] David Venet, Jacques E Dumont, and Vincent Detours. Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Comput Biol, 7(10):e1002240, 2011.

[WGB+14a] Charles Wang, Binsheng Gong, Pierre R Bushel, Jean Thierry-Mieg, Danielle Thierry-Mieg, Joshua Xu, Hong Fang, Huixiao Hong, Jie Shen, Zhenqiang Su, et al. A comprehensive study design reveals treatment-and transcript abundance–dependent concordance between rna-seq and microarray data. Nature Biotechnology, 32(9):926, 2014.

[WGB+14b] Charles Wang, Binsheng Gong, Pierre R Bushel, Jean Thierry-Mieg, Danielle Thierry-Mieg, Joshua Xu, Hong Fang, Huixiao Hong, Jie Shen, Zhenqiang Su, et al. The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nature Biotechnology, 32(9):926–932, 2014.

[WKL12] Günter P Wagner, Koryu Kin, and Vincent J Lynch. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory in Biosciences, 131(4):281–285, 2012.

[Won11] Limsoon Wong. Using biological networks in protein function prediction and gene expression analysis. Internet Mathematics, 7(4):274–298, 2011.

[WSG16] Wei Wang, Andrew CH Sue, and Wilson WB Goh. Feature selection in clinical proteomics: with great power comes great reproducibility. Drug Discovery Today, 2016.

[YRS+02] Eng-Juh Yeoh, Mary E Ross, Sheila A Shurtleff, W Kent Williams, Divyen Patel, Rami Mahfouz, Fred G Behm, Susana C Raimondi, Mary V Relling, Anami Patel, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer cell, 1(2):133–143, 2002.

[ZLSA11] Mattia Zampieri, Giuseppe Legname, Daniel Segrè, and Claudio Altafini. A system-level approach for deciphering the transcriptional response to prion infection. Bioinformatics, 27(24):3407–3414, 2011.

[ZZZ+09] Min Zhang, Lin Zhang, Jinfeng Zou, Chen Yao, Hui Xiao, Qing Liu, Jing Wang, Dong Wang, Chenguang Wang, and Zheng Guo. Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. Bioinformatics, 25(13):1662–1668, 2009.