ABSTRACT

Title of dissertation: KNOWLEDGE DISCOVERY FROM GENE EXPRESSION DATA: NOVEL METHODS FOR SIMILARITY SEARCH, SIGNATURE DETECTION, AND CONFOUNDER CORRECTION

Louis Licamele, Doctor of Philosophy, 2012

Dissertation directed by: Professor Lise Getoor
Department of Computer Science

Gene expression microarray data is used to answer a variety of scientific questions. For example, it can be used for gaining a better understanding of a drug, segmenting a disease, and predicting an optimal therapeutic response. The amount of gene expression data publicly available is extremely large and continues to grow at an increasing rate. However, this rapid growth of gene expression data from laboratories across the world has not fully achieved its potential impact on the scientific community. This shortcoming is due to the fact that the majority of the data has been gathered under varying conditions, and there is no principled way for combining and fully utilizing related data. Even within a closely controlled gene expression experiment, there are confounding factors that may mask the true signatures when analyzed with current methods. Therefore, we are interested in three core tasks that we believe are important for improving the utilization of gene array data: similarity search, signature detection, and confounder correction. We have developed novel methods that address each of these tasks.

In this work, we first address the similarity search problem. More specifically, we propose methods which overcome experimental barriers in pariwise gene expression similarity calculations. We introduce a method, which we refer to as *indirect similarity*, which, unlike previous approaches, uses all of the information in a database to better inform the similarity calculation of a pair of gene expression profiles. We demonstrate that our method is more robust and better able to cope with experimental barriers such as vehicle and batch effects. We evaluate the ability of our method to retrieve compounds with similar therapeutic effects in two independent datasets. We evaluate the recall ability of our approach and show that our method results in an improvement of 97.03% and 49.44% respectively in the two datasets over existing state of the art approaches.

The second problem we focus on is signature detection. Gene expression experiments are performed to test a specific hypothesis. Generally, this hypothesis is that there is some genetic signature common in a group of samples. Current methods try to find the differentially expressed genes within a group of samples using a variety of methods, however, they all are parametric. We introduce a nonparametric approach to group profile creation which we refer to as the *Weighted Influence Model - Rank of Ranks* method. For every probe on the microarray, the average rank is calculated across all members of a group. These average ranks are then re-ranked to form the group profile. We demonstrate the ability of our group profile method to better understand a disease and the underlying mechanism common to its treatments. Additionally, we demonstrate the predictive power of this group profile to detect novel drugs that could treat a particular disease. This method leads the detection of robust group signatures even with unknown confounding effects.

The final problem that we address is the challenge of removing known (annotated) confounding effects from gene expression profiles. We propose an extension to our non-parametric gene expression profile method to correct for observed confounding effects. This correction is performed on ranked lists directly, and it provides a robust alternative to parametric batch profile correction methods. We evaluate our novel profile subtraction method on two real world datasets, comparing against several state-of-the-art parametric methods. We demonstrate an improvement in group signature detection using our method to remove confounding effects. Additionally, we show that in a dataset with the true group assignments removed and only the confounding effects labelled, our profile subtraction method allows for the discovery of the true groups. We evaluate the robustness of our methods using a gene expression profile generator that we developed.

KNOWLEDGE DISCOVERY FROM GENE EXPRESSION DATA: NOVEL METHODS FOR SIMILARITY SEARCH, SIGNATURE DETECTION, AND CONFOUNDER CORRECTION

by

Louis Licamele

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2012

Advisory Committee:
Professor Lise Getoor, Chair/Advisor
Professor Héctor Corrada Bravo
Professor Carl Kingsford
Professor Stephen M. Mount
Professor Mihai Pop

Foreword

Portions of this dissertation are derived from research and papers co-authored by the candidate and published elsewhere. Chapter 2 is based on *Indirect two-sided relative ranking: a robust similarity measure for gene expression data* [32]. Chapter 3 is based on *A method for the detection of meaningful and reproducible group signatures from gene expression profiles* [33] and is the method used in *Common effect of antipsychotics on the biosynthesis and regulation of fatty acids and cholesterol supports a key role of lipid homeostasis in schizophrenia* [47]. The profile subtraction method presented in Chapter 4 will be submitted for publication separately.

Dedication

For my wife Raquel and my children Michael, Nadia and James, who have given up so much to let me work on this thesis and yet remained supportive throughout the whole process.

Acknowledgments

I would like to thank my advisor, Dr. Lise Getoor, for all the guidance and support that she has provided. I also would like to thank my other committee members: Dr. Héctor Corrada Bravo, Dr. Carl Kingsford, Dr. Stephen M. Mount, and Dr. Mihai Pop for serving on my committee and providing useful feedback throughout this process.

I want to acknowledge my family and friends for the role that everyone has played in my life. I thank my parents for their love and support in addition to the education that I have received. I am especially grateful for my amazing wife Raquel who has sacrificed the most throughout the years while I have completed this research and who has continued to remain incredibly supportive. I also need to recognize and thank our children Michael, Nadia and James for being so understanding of the times that I was busy with completing this work.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ABA	Abscisic Acid
ACC	Aminocyclopropane Carboxylic Acid
ALL	Acute Lymphoblastic Leukemia
AML	Acute Myeloid Leukemia
ATC	Anatomical Therapeutic Chemical
AUC	Area Under Curve
BDNF	Brain-Derived Neurotrophic Factor
BS	Brassinolide
CMAP	Connectivity Map
CNS	Central nervous system
CYT	Cytokinins
DEGS	Differentially Expressed Genes
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
EtOH	Ethanol
MeOH	Methanol
EV	Expression Value (raw)
GO	Gene Ontology
GPS	Global Positioning System
HT	High Throughput
IAA	Indoleacetic Acid
ID	Identifier
INN	International Nonproprietary Name
KS	Kolmogorov-Smirnov (modified)
KS_D	Kolmogorov-Smirnov DownTags Score
KS_U	Kolmogorov-Smirnov UpTags Score
LIMMA	Linear Models for Microarray Data
MAS5	Microarray Suite - software from Affymetrix
MEP	Megakaryocyte-Erythrocyte Progenitors
MJ	Methyl Jasmonate
mRNA	messenger RNA
NG	Number of group profiles selected
NSE	Number of confounding (side effect) profiles selected
PDR	Physicians Desktop Reference
RNA	Ribonucleic acid
ROC	Receiver Operating Characteristic
SVA	Surrogate Variable Analysis
WIMRR	Weighted Influence Model - Rank of Ranks

Chapter 1

Introduction

Gene expression microarrays are used to answer a variety of scientific questions. However, there still exist many challenges that current methods do not overcome. This may lead to incorrect or missed scientific discoveries. In this thesis, I address three core problems that I believe are important:

1. Similarity Search

2. Signature Detection

3. Confounder Correction

We develop new algorithms to address each of these: 1) *Indirected Similarity*, 2) *Weighted Influence Model - Rank of Ranks*, and 3) *Profile Subtraction for Ranked Lists*. My hypothesis is that by using as much information from the database while focusing on the development of nonparametric methods, we will overcome the variability and noise introduced by confounding factors. This will lead to an improvement in methods for similarity search, signature detection and confounder correction.

1.1 Background

There is a large amount of gene expression data, generated from microarray experiments, that exists in the public domain. Gene expression microarrays attempt to measure the amount of mRNA transcribed. This gives an estimate of the amount of protein that is translated from this mRNA. Proteins are responsible for most of the work in the cell, whether it is breaking down biomolecules or compounds , signaling other cells or pathways, or making up the infrastructure and machinery to continue to transcribe DNA into mRNA. Gene expression profiling is often used to understand the underlying mechanism of biological processes and pathways [12, 29], to explain diseases and segment patients into subtypes of a disease [16, 44], and to predict cancer prognosis [48, 60]. In addition, because it captures how the cell is responding to each compound, gene expression data may be a excellent source for investigating whether two drugs could have a similar therapeutic effect.

Unfortunately, gene expression data are inherently complex and difficult to analyze and compare. There are many factors that complicate the process including post-transcriptional modification (e.g., splicing), degradation of the mRNA, changes in the translation rates from mRNA to polypeptide chains, as well as post-translational modification (e.g., phosphorylation). In addition, the existing data have been generated by many different laboratories across the world under a variety of experimental conditions. These experiments can be testing many different hypotheses, such as the effect of a drug, i.e., pathways and genes affected by the drug, or the cause of a disease, i.e., pathways and genes differentiated in affected individuals. Furthermore, the gene expression profiles represent a complex response to many unobserved factors in tandem beyond those being explicitly controlled in the experiments. Lastly, the microarray technology itself introduces noise into the results. All of these factors in combination result in confounding effects on the gene expression profiles.

1.1.1 Similarity Search

Historically, when researchers compare gene expression profiles, they limit themselves to data generated under similar experimental conditions. The ability to compare gene expression profiles across experiments would substantially increase both the questions that could be answered as well as the reliability of the results. Recently, researchers at the Broad institute developed a new approach for detecting gene expression similarity. Their tool, the Connectivity-Map (CMAP) [28], tackles the problem of comparing gene expression profiles generated under diverse experimental conditions. Unlike previous methods, they use a distribution statistic to compare the ranked lists of expression probes. They show that this method is able to overcome some of the experimental noise that can affect the gene expression profile. This noise can arise from a wide range of confounding factors such as the vehicle used to deliver the compound and the cell line used for the experiment. This method is a substantial improvement over simple, hierarchical clustering which was one of the previously preferred methods [21, 62]. We believe that the robustness of this non-parametric method is largely driven by its dynamic use of ranked lists.

1.1.2 Signature Detection

The goal of gene expression experiments can vary widely but at the core they have a common goal: the detection of a genetic signature in common within a group of interest. This group may be a given treatment compared to controls, or a subtype of a disease versus healthy individuals, or a profile of a patient who is more likely to respond to a given treatment, or even relating to the prediction of the toxicity of a compound versus safer alternatives to guide promotion into future human studies. A gene expression experiment could be designed to evaluate any of these scientific questions, but a method that generates more reliable and reproducible results (e.g., lists of DEGs) from this gene expression data is needed. Reproducibility has remained low among these types of experiments, calling into doubt the validity of the detected signatures. For example, using an identical set of RNA samples across several different commercial platforms, Tan et al. [58] found only four common (differentially expression genes) DEGs. Both Ramalho-Santos [50] and Ivanova [22] independently found only six DEGs in common among approximately 200 identified in each study (even though they had a similar study design using the same platform). In another study by Miller et al. [41], who compared the effect of varying platforms on the same samples, there were only 11 DEGs in common of 425 DEGs found by CodeLink and 138 DEGs found by the Affymetrix platform. These are all examples of studies that exhibit how current methods are producing irreproducible signatures. This lack of reproducible findings indicates the presence of false positives, and that these methods may be overfitting the data. Furthermore, many methods are complex and only explain a group in a piecewise fashion (e.g., a decision tree-type model). We believe that the ideal method does not require such strict filtering that scientists in the field are used to and yet also remains simple and robust.

1.1.3 Confounder Correction

Realizing that sometimes there are explicit, labelled confounding effects in a gene expression dataset, methods for removing these confounded effects are of high utility. This will lead to group profiles with increased robustness in the downstream analysis. There have been a number of methods developed for dealing with the problem of detecting and removing confounding effects within gene expression profiles [10, 25, 31, 35, 57]. Linear models are a commonly used approach and Limma [57] is a popular package for R[49]. In addition to this method, other common methods include Combat[25] and SVA[31]. Additional methods include Geometric ratio-based methods[38], Mean-centering methods[56], and Distance-weighted discrimination[5] based on SVMs.

1.2 Our Proposed Solutions

We have identified the three main challenges that we see as being the biggest barriers to knowledge discovery from gene expression data. Overall, we have focused on developing methods that are robust to confounding effects. We believe that by developing methods that use as much data as possible and by using this data in a way that remains nonparametric, our methods will not be as influenced by confounding effects. These methods will therefore be able to make use of the large amount of gene expression data in the public domain. We begin in Chapter 2 by introducing a similarity measure that can be used to compare two gene expression profiles. We have shown that this method is more robust to vehicle and batch effects. Next, in Chapter 3, we focus on signature detection among a group of samples: one of the core tasks of gene expression experiments. We focus on a nonparametric approach that we believe leads to a method that is robust to experimental noise and unlabeled confounding factors. We introduce our last method in Chapter 4 to allow for confounder correction to be done efficiently when there are known, labelled confounding effects. In order to remove these confounding effects we have created a profile subtraction method that works on ranked lists.

1.2.1 Similarity Search

In Chapter 2, we describe the problem of overcoming experimental barriers in pariwise gene expression similarity calculations. We introduce our new similarity measure for comparing ranked lists. Current methods for this problem consider only the information contained in the pair of gene expression profiles being compared. Our approach is novel because it incorporates information in the rest of the database to further refine our similarity calculation. Unlike the CMAP and other current approaches, which performs a direct comparison of the gene expression profiles, our approach captures the correlations in rankings between the target pair and the rest of the database. This results in a method which has empirically proven to be more robust and. We show that our new method is able to better cope with experimental barriers such as vehicle and batch effects. We evaluate the ability of this method to retrieve compounds with similar therapeutic effects in two different datasets.

1.2.2 Signature Detection

Chapter 3 introduces and describes our rank of ranks method for group profile creation. Our method is novel because it uses a nonparametric approach to detect a robust but consistent signature of a group. This is different from the commonly used parametric methods for detection signatures and finding differentially expressed genes (DEGs). We evaluate the utility of this group analysis method using a pilot study. Our evaluation consists of meta analysis methods for both understanding the group profiles biologically as well as for demonstrating the ability to use a signature from these profiles as a predictive model of therapeutic use. We conclude with a full analysis of the newer, and larger CMAP 2.0 dataset, including a sensitivity evaluation of each group as well as the validation of the most robust profiles within an independent dataset. In addition, another contribution of this work is the independent validation of the published expression signature of antipsychotic drugs.

1.2.3 Confounder Correction

In Chapter 4 we provide an extension to our non-parametric gene expression profile method to correct for observed confounding effects. This correction is performed on ranked lists directly and provides a robust alternative to parametric batch profile correction methods. Our model is novel because it is non-parametric; most other methods are parametric. We show that this method is more robust than all of the parametric based methods that we have evaluated, which includes Limma, Combat and SVA. We evaluate our method on two independent datasets and show the improvement over alternative methods.

1.3 Contributions

The contributions of this thesis include a set of methods to improve knowledge discovery from gene expression data. We have identified three main challenges that are encountered when analyzing gene expression experiments: similarity search, signature detection, and confounder correction. We have introduced novel methods to solve each of these challenges. For the task of similarity search, we have developed an indirect similarity method, demonstrating the methods ability to overcome experimental noise in finding the most similar gene expression profiles in a database. To solve the signature detection problem we developed a robust group profile method, referred to as our Weighted Influence Model, Rank of Ranks method. Again, we demonstrated the ability of this method to overcome experimental noise and create robust signatures of a group of profiles. We demonstrated the utility of these group profiles for two key tasks: gaining biological insight into the underlying function of a class of compounds, e.g., leading to a new hypothesis into the etiology of a disease, and performing similarity search and classification to predict new members of a class. An additional contribution of this thesis is the analysis of over 200 therapeutic classes of compounds and the release of their profiles to our website, GEPedia.org. This includes the validation of our previously published finding of the common effect of antipsychotics on lipid homeostasis in schizophrenia. The last task deals with known, labelled confounding factors and being able to successfully correct for them. To solve this task we have developed a profile subtraction method which is novel since it works on ranked lists. We have shown how this method can lead to much more robust group profile detection. Another contribution of this thesis is the creation of a group profile generator which has been used to more closely control and evaluate the robustness of our methods.

1.4 Outline of Thesis

The roadmap for the chapter on similarity search is as follows. In Section 2.1 we define the problem of gene expression similarity detection as a comparison of ranked profile lists. Next, we discuss the CMAP method and formalize how it works (Section 2.2). In Section 2.3.1, we introduce and motivate a novel indirect two-sided relative ranking method. The evaluation methodology is explained in Section 2.4, results are presented in Section 2.5, and a brief discussion appears in Section 2.7. A brief overview of related work for this area is presented in Section 2.6.

We then introduce our method to detect the signature of a group of samples in Chapter 3. The motivation and description of the proposed group profile creation method are explained in Section 3.2. We evaluate the utility of this group analyses method in Section 3.3 where we focus on the antipsychotic group from the Broad dataset (Section 3.3.1). The evaluation consists of both understanding the group profiles biologically (Section 3.4) as well as demonstrating the ability to use a signature of these profiles as a predictive model in Section 3.5. We analyze the large dataset consisting of over 200 therapeutic classes (Section 3.3.1 and provide a sensitivity analysis and independent validation in Section 3.6.2.

In Chapter 4 we propose an extension to our non-parametric gene expression profile method to correct for observed confounding effects. This correction is performed on ranked lists directly and provides a robust alternative to parametric batch profile correction methods. Our profile subtraction method is described in Section 4. We evaluate this method on two gene expression datasets (Section 4.3.1 and Section 4.3.2). Additionally, we develop a gene expression profile generator which is described in Section 4.4. Simulations from this gene expression profile generator are allow us to make general assumptions on how our method is robust to varying the tuning parameters of these newly introduced methods.

We conclude in Chapter 5 and discuss the contributions of this work. We have introduced three core problems in working with gene expression data and have presented our solutions to each of these problems. We have developed methods that allow for better similarity search, signature detection, and confounder correction in the presence of noise in the data, which is know to be a large issue. We have demonstrated how new scientific hypothesis can be generated from using these improved methods, including among other topics, a new a hypothesis of how antipsychotics work. We finish with a brief discussion on future research and examine the importance of continuing research in this field.

Chapter 2

Similarity Search in Gene Expression Profiles

In this chapter, we introduce and motivate the problem of detecting pairwise similarity among gene expression profiles. We formalize our definition of the problem and present the current state of the art method. We then explain our novel indirect similarity method and empirically demonstrate how it is more robust to experimental noise that is a known issue in gene expression analysis. Specifically, we evaluate our indirect method and show how it can achieve an improvement of 49.44% and 97.03% in two independent datasets.

2.1 Problem Definition

Given a database D of treatments, i.e., drugs or other compounds, $D = t_1, \ldots, t_n$, suppose we are interested in querying the database with a selected query treatment and returning other similar treatments. Typically we know the therapeutic use or indication for the query, but may not have complete therapeutic information for all the entries in the database. We may be trying to discover other drugs or treatments, perhaps originally developed for a different therapeutic purpose, that are likely to also share the same therapeutic properties as the query. These drugs then are good candidates for further evaluation of a new use.

More specifically, for each treatment instance t in the database, there is both general information about the experimental conditions of the sample as well as the actual experiment data from the microarray itself. The microarray data consists of a collection of probe sets, $probes(t)$. Each probe $p \in probes(t)$ measures the match to a particular genomic sequence. For each probe p, there is a raw expression value $EV(p)$ (calculated using MAS 5 algorithm [20]), as well as an amplitude $A(p)$ (the difference compared to control). The control is a reference baseline which is the average expression value calculated from multiple untreated samples run within the same vehicle and batch. Information specific to the treatment, i.e., the name of the drug, the therapeutic class (class) and subclass (subclass) as defined by the Physicians Desktop Reference (PDR) is also represented. There is also information that describes the experimental conditions of the sample, specifically the molar amount of substance (mol), the vehicle used for delivery of the drug (one of water, EtOH, MeOH, DMSO) and the batch or round in which the sample was run.

We are interested in retrieving treatments t that are similar in some way to a query treatment q. We measure similarity based on the probes of t and q. Rather than measuring the absolute similarity in expression levels, we compare the *ranking* of the probes. Using the ranks allows for a nonparametric comparison of the gene expression profiles. Nonparametric methods have been shown to work well for detecting differentially expressed genes in microarray data [59, 39, 40]. As mentioned above, probes have both a raw expression value and an amplitude. This ranking can be done based on either the raw value or the amplitude. The amplitude represents the change in expression as compared to the control. We utilize the amplitude because it measures the treatment effect. Amplitude a is defined as $(t \times c)/((t+c)/2)$, where t is the thresholded scaled average difference value (treatment) and c is the thresholded average difference value (control). Control average difference values were set to the arithmetic mean of the values from all matched controls. Average difference values less than a given threshold value of 50 were set to that threshold value. Any probes that yielded an amplitude change of 1 were re-evaluated with a lower threshold of 5 and then these probes were sub-sorted within the overall ranked list. These calculations follow what was done by Lamb et al. [28].

We use $rank(p, probes(t))$ to denote the rank of p in $probes(t)$; i.e., if the probes are sorted in order of their amplitude, then the rank is the position of p in that ordering. We also introduce the uptags of t, $Up(t)$ and the downtags of t, $Down(t)$. $Up(t)$ is the set of k highest ranked probes in $probes(t)$, i.e., the most upexpressed as compared to control, and $Down(t)$ is the set of k lowest ranked probes in $probes(t)$, i.e., the most downexpressed as compared to control.

2.2 Comparing Rankings

We are interested in finding drugs with similar therapeutic effect by comparing the rankings of probes in gene expression profiles of the drugs. The most straightforward approaches to compare these ranked lists, for example calculating the intersection of $Up(q)$ and $Up(t)$, quickly fail when there is any experimental noise. More robust methods are needed to be able to combine and draw conclusions from the large amount of gene expression data that has been created across many laboratories.

2.3 A Two-sided Approach

A more sophisticated approach to comparing the similarity of two rankings is to compare both the uptags and downtags, and rather than looking simply at the overlap in the sets of tags, take into account the relative ranking of the probe. We will refer to this approach as the *two-sided relative ranking* approach. This type of approach may be able to correctly weight both ends of the ranking and overcome noise in the experimental data.

The CMAP approach [28] is a recently introduced treatment retrieval method that is an example of a two-sided relative ranking approach. Here we formalize the CMAP method and ground it in our example domain. The following equations are adapted from [28]. The CMAP method is based on a similarity measure which uses a truncated Kolmogrov-Smirnov (KS) statistic applied to the up and down probes of the treatments. The KS statistic measures the similarity between two distributions; the truncated KS statistic focuses on the tail end of the distributions. Given a query treatment q and target treatment t, the KS score is high if a) the probes in $Up(q)$ tend to also be highly ranked in t, b) the probes in $Down(q)$ tend to have low ranks in t, and finally c) the probes in $Up(q)$ tend to be more highly ranked in t than the probes in $Down(q)$. This is similar to the truncated statistical approach seen in [2] in the whole genome association study in search of genetic markers for continuous traits.

The KS statistic of treatment instance t, given a query instance q, $KS(t, q)$, is computed using the uptags and downtags of the two treatments. $KS(t, q)$, in turn, is computed from two separate statistics, $KS_u(t,q)$ and $KS_d(t,q)$, which are calculated on the uptags and downtags respectively.

$KS_u(t,q)$ measures where the uptags of the query are located within the distribution of probes in a treatment instance t. It is a number between -1 and 1. If it is close to 1, it tells us that the uptags of q are also highly ranked in t, or more specifically that the probes that are most upexpressed in the query instance also tend to be upexpressed in the treatment instance.

In order to compute KS_u, based on the selected set of probes, $Up(q)$, we define $Up_t(q)$ to be the probes in $Up(q)$ sorted according to their rank in t, $rank(p, probes(t))$. Next we define the rank of p in this new sorted set of probes:

$$rank(p, Up_t(q)) = \text{the position of } p \text{ in } Up_t(q) \tag{2.1}$$

We introduce shorthand $p_i = rank(p, probes(t))$ and $p_j = rank(p, Up_t(q))$

Now we have the required information to compare the probe distributions between the query and each treatment. Let

$$a = \max_{p \in Up(q)} \left[\frac{p_j}{k} - \frac{p_i}{n} \right] \tag{2.2}$$

and

$$b = \max_{p \in Up(q)} \left[\frac{p_i}{n} - \frac{p_j - 1}{k} \right] \tag{2.3}$$

KS_u is computed as follows:

$$KS_u(t,q) = \begin{cases} a & \text{if } a > b \\ -b & \text{otherwise} \end{cases} \tag{2.4}$$

KS_d is calculated analogously using $Down(q)$.

Finally we can calculate the truncated KS statistic using the KS_u and KS_d as follows:

$$KS(t,q) = \begin{cases} KS_u(t,q) - KS_d(t,q) & \text{if } sgn(KS_u(t,q)) \neq sgn(KS_d(t,q)) \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

Referring back to our original description of the properties that we were looking for in the KS statistic, we see that when the sign of KS_u and KS_d are the same, whether both positive or both negative, then the KS score is set to zero. This indicates that there is a significant overlap between the two distributions. No clear separation means that the two distributions are randomly dispersed, and that this ranked list is not statistically similar to the query sequence. In the case where the sign of the two values is different then the final KS score represents the separation between the two distributions. This is done by calculating the difference between KS_u and KS_d.

The CMAP approach was developed as a query system that directly compares the query to each treatment in the database. It does not take into account any further information about how the treatment instances in the database relate to each other. We refer to this as a *direct* approach.

2.3.1 Indirect Two-sided Similarity

Next, we introduce an *indirect two-sided relative ranking* method which compares the similarity between the query and treatment instance by comparing their corresponding similarity to *all* the other instances in the database. Ideally, by combining hundreds or even thousands of pairwise distances, a more robust similarity measure can be obtained. This is similar in spirit to a vantage point method for computing similarity in metric spaces, where the distance between a pair of points is computed based on their distance to a collection of vantage points [9].

Our indirect two-sided relative ranking is calculated by comparing the correlation between how two treatments compare to the rest of the database. There are many correlation measures, including parametric statistics such as Pearson coefficient and nonparametric statistics such as Spearman rank correlation coefficient. Since we do not know ahead of time if the gene expression data is normally distributed, it is safer to use a nonparametric correlation measure. While there are a number of nonparametric correlation measures which could be used, we chose the Spearman rank correlation coefficient because of its widespread acceptance and ease of use.

We compute the Spearman rank correlation coefficient by measuring the difference between the KS statistics for the query q and target treatment t, for all the treatments in the database D.

Let $KS_D(q) = \{KS(q, t_1), KS(q, t_2), \ldots, KS(q, t_n)\}$ and let $KS_D(t) = \{KS(t, t_1), KS(t, t_2), \ldots, KS(t, t_n)\}$. Then we define the indirect two-sided relative ranking of a query q and a treatment t, $I2R(t, q)$ to be:

$$I2R(t, q) = \text{Spearman}(KS_D(q), KS_D(t)) \tag{2.6}$$

where *Spearman* is the Spearman correlation statistic, which we will formally define here.

If there are no tied ranks, then the Spearman correlation is calculated as follows.

$$1 - \frac{6\sum_{i=1}^{n}(KS(q,t_i) - KS(t,t_i))^2}{n(n^2-1)} \tag{2.7}$$

Where n is the number of instances in the database. This score is calculated using all of the pairwise KS scores from q and t to each other instance t_i. This is equivalent to taking the Pearson's correlation over the ranks. In the case where there are tied ranks, the full Pearson's correlation over ranks must be calculated.

The indirect similarity score is therefore a calculation of how two instances individually compare to the rest of the database. If they tend to be similar, or dissimilar, to the same instances then they are more likely to be similar to each other.

An advantage of this method is that it can build on any individual pair-wise similarity available. Here we have taken what we believe to be the current best method, the KS statistic from the CMAP approach, and used this as our source of pairwise similarities. If other, possibly better, direct similarity measures for the treatments become available, we can easily incorporate them. Another advantage of this method is that as more treatments are added to the database, additional evidence is available, which can further increase the accuracy of our indirect similarity calculation.

2.4 Evaluation

As mentioned at the outset, we are interested in finding similar treatments by comparing the gene expression profiles of drugs. Specifically, our goal is to improve the ability to detect similarity in the presence of experimental noise. We focus our evaluation on the case where we have known experimental noise, e.g., when 1) the samples are delivered in different vehicles, 2) they belong to different batches, or 3) they differ in both vehicle and batch (which corresponds to the most experimental noise). Though vehicle and batch are not the only sources of experimental noise they can easily be evaluated as they are both annotated.

The ideal outcome of such a discovery program is the in vivo validation of a drug predicted by gene expression similarity to be useful for an unknown, alternative indication. To simulate this goal, we propose calculating the average recall at rank k of drugs of the same PDR classification. We measure this recall of drugs which are known to be used for the same indication across vehicles, across batches, and across both vehicles and batches. We focus our analysis on the most populated PDR classifications, where 10 or more drugs from each group have been profiled, which leaves us with 14 different groups. This filtering of groups is done to avoid unrepresentative results caused by a small sample size. For the evaluation, we select recall at rank $k = 10$, but we also demonstrate that these results are not greatly affected by variations in k.

We begin with a simple example of our evaluation method for the PDR group *Histamine Antagonists*.

Given the group *Histamine Antagonists*:

1. For each histamine antagonist, determine the 10 most similar compounds using each method (direct vs indirect)

2. Count the number of compounds of the same class, i.e., Histamine Antagonist, that are screened in a different vehicle and different batch

3. Improvement of indirect over direct is represented as:

$$([indirect] - [direct])/[direct]$$

   where [indirect] and [direct] are the number of recalled treatments from the top 10 of each method respectively.

4. When [direct] = 0, and [indirect] ≥ 0 then we make note of this improvement as a special case. [1]

Table 2.1: Histamine Antagonists

Number of results returned by direct	2
Number of results returned by indirect	3
Improvement of indirect over direct	50%

The results of our example of comparing the two similarity methods for recall at rank 10 across both different vehicles and different batches for the Histamine

[1] As reporting percent improvement does not make sense when the baseline is 0, so we do not include these in our overall improvement calculation, but we note them as they are important special cases

Antagonists are shown in Table 2.1. In this case, the direct method finds two results while indirect finds three, which is an improvement of 50%. This example analysis compares the ability to recall other histamine antagonists across both vehicles and batches.

2.5 Results

Using the evaluation criteria presented above, we compare the ability of the two methods (direct and indirect) to overcome experimental noise. We have the ability to evaluate how these methods work on two completely separate datasets. The first is a large, proprietary dataset (GEPedia) from Vanda Pharmaceuticals. This dataset contains a large number of drugs that have been profiled. The second is a public dataset from the Broad Institute[2] which contains 453 profiles. It is important to note that this second dataset includes a substantial amount of replicates for many of the compounds.

2.5.1 Vanda GEPedia Dataset

We start by comparing the two methods, direct and indirect, using the Vanda GEPedia dataset. The average recall at rank 10 for the 14 PDR groups is presented in Table 2.2. The indirect method improves over the direct method and is able to recall 71.44% more true positives when searching across different vehicles. This improvement is increased when searching across batches (94.93%). When attempt-

[2] http://www.broad.mit.edu/cmap

ing to detect similarity across both vehicles and batches, which represents the most experimental noise in our setup, the indirect method has an improvement in recall of 97.03% as compared to the direct method. The indirect similarity method recalls almost twice the amount of true positives (similar drugs) as the direct method. This level of improvement brings the potential for important scientific discovery and impact of such a system.

As mentioned earlier, the average percentage improvement does not capture the important special case that occurs if one of the methods does not retrieve *any* treatments. These special cases are further examples of the ability of the indirect similarity method to detect similarity when the direct method cannot. These cases are listed below for those found across different vehicles , across different batches, and lastly across both different vehicles and different batches (Table 2.3).

Table 2.2: Percent improvement of indirect similarity recall over direct similarity recall in different conditions

Across Different Vehicles	71.44%
Across Different Batches	94.93%
Across Different Vehicles & Batches	97.03%

We have shown how overall, the indirect method which uses the Spearman rank correlation has a higher recall at rank 10 than the direct KS method.

Table 2.3: Novel Improvements - Instances found in conditions where the direct method had found none

PDR Classification	Vehicle	Batch	Vehicle and Batch
Antibiotic	12	11	8
Anesthetic	1	1	1
Antihypertensive	1	1	0
Anticonvulsant	0	2	2

2.5.2 In Depth Analysis

Next we study the largest groups (the groups with the most compounds profiled) in more detail (Antibiotic, Histamine Antagonist, and Analgesic), in order to a) verify that this is not an artifact of using $k = 10$ and b) to further inspect the differences in the results returned by each of the methods.

2.5.2.1 Antibiotics

The antibiotics group has the largest amount of compounds in the database ($n = 58$). This group is in the PDR class *Antiinfective* and the PDR subclass *Antibiotic*. An antibiotic drug is one that inhibits the growth of micro-organisms. The indirect method is able to recall eight antibiotics when searching across both vehicle and batch, compared to 0 recalled by the direct method. This result is not driven by any single treatment, i.e., each of these 8 recalled treatments is not only unique, but they are recalled by distinct query treatments.

Next, we demonstrate that these results are not biased by our selection of $k = 10$. Figure 2.1(a) shows the recall of the two methods across both different vehicles and batches with values of k ranging from 10 to 100. The indirect method is able to recall more true positives independent of k. We can also evaluate how the methods compare when searching over vehicle or batch separately. Figure 2.1(b) shows that when searching across different vehicles only, the same trend is seen as in Figure 2.1(a). Similarly, evaluating the two methods when searching across different batches (Figure 2.1(c)), a similar trend is seen in which the indirect method outperforms the direct method regardless of k.

2.5.2.2 Histamine Antagonist

The Histamine Antagonist group contains the second largest number of compounds profiled ($n = 24$). This group is made up of drugs in the PDR class *Respiratory Agent* and PDR subclass *Histamine Antagonist*. A histamine antagonist inhibits the release or minimizes the action of histamine. There are several subtypes of histamine antagonist based on their binding affinity to the different histamine receptors. The H_1 receptor antagonist, sometimes referred to as antihistamines, are clinically used to treat allergies. The other common subtype, the H_2 receptor antagonist, are commonly used to control the secretion of gastric acid. There are other subtypes, namely H_3 and H_4; however, they are not often used clinically. Once again we do not distinguish between these subtypes for our analysis; we use the PDR classification.

For the histamine antagonists, the direct method is able to recall two antihistamines at or below rank 10 while the indirect method is able to recall three. This corresponds to an increase of 50%. More specifically, the indirect method recalls the same two treatments as the direct method in addition to a third novel treatment. Figure 2.2(a) shows the ability of each method to recall histamine antagonists across both different vehicles and batches. The recall at rank 20 and at rank 30 is the same for the two methods, and then as k increases the indirect method improves in its ability to recall histamine antagonists as compared to the direct method. In splitting up the vehicle (Figure 2.2(b)) and batch (Figure 2.2(c)) analysis, we see that the direct method is outperforming the indirect method in the across vehicle analysis for smaller k, while underperforming against the indirect method in the across batch analysis, which contains more instances.

2.5.2.3 Analgesic

The last group that we individually analyze is the Analgesic group ($n = 23$). This group is defined as drugs belonging to the PDR class *Central Nervous System Agent* and PDR subclass *Analgesic*. An analgesic, more commonly known as a painkiller, acts in various ways on the peripheral and central nervous system in order to reduce pain. Searching across both vehicles and batches the direct method is able to recall one analgesic. The indirect method recalls the same treatment in addition to two other treatments. The indirect method is able to recall three analgesics in total which corresponds to a 200% percent increase.

Figure 2.3(a) shows that the indirect method has a higher recall rate at every level of k when searching across both vehicle and batch. The same is true when searching across just a different batch (see Figure 2.3(c)). We see in Figure 2.3(b) that the indirect method also does better for low k across different vehicles. It is more important for a method to do better for low k because in a drug discovery system you will start validation on the most promising hits first. It quickly becomes cost prohibitive to explore a large set of leads.

2.5.3 Broad Dataset

Next, we replicate our findings using the publicly available gene expression dataset from the Broad Institute. This dataset consists of 453 samples and was released with the Connectivity Map tool. To allow for easier reproducibility, we make use of the annotations provided on the CMAP website as opposed to custom matching to the PDR annotations. In terms of PDR indications we instead use what is described as *Therapeutic Uses* in the ChemBank [54] record linked to each CMAP instance. There is no information provided about the vehicles used for each sample. However, each instance is associated with a batch, and so we will use this information to segment our data. To remain consistent and in order to have more confidence in the results, only groups (Therapeutic Uses) with 10 or more instances are used. The groups, along with the number of instances in each group, are listed in Table 2.4.

Similar to what was observed before, the indirect similarity measure recalls more compounds of the same class than using the direct similarity measure alone. The improvement of the indirect method over the direct method is 49.44% on the Broad dataset. Once again, the indirect method allows for the ability to recall more true positives, and the improvement is substantial.

We now analyze the three largest therapeutic groups from the Broad dataset. Note that for this set of data we explore a smaller size for k. Given that this is a smaller database we want to guarantee that we are only evaluating the top pairs. We begin our analysis with the Antiinflammatory group.

2.5.3.1 Antiinflammatory

To illustrate this improvement, let us look at the group with the most compounds: the antiinflammatory group. An antiinflammatory drug is a substance that reduces inflammation. Many analgesics are antiinflammatory agents, alleviating pain by reducing inflammation. The direct approach is able to recall 16 compounds labeled as antiinflammatory that have been profiled in a different batch ($k = 10$). The indirect approach, however, is able to recall 28 compounds that also are classified with a therapeutic use of antiinflammatory. The improvement of the indirect method over the direct method can be seen in Figure 2.4(a). We see that the indirect method is always better than the direct method, and this holds even more true with lower k values.

Table 2.4: Broad Dataset

Therapeutic Use	Number of instances
antiinflammatory	28
anticonvulsant	21
antipsychotic (neuroleptic)	19
antiproliferative	16
tranquilizer	16
antineoplastic	15
analgesic	13
immunosuppressive agent	13
cardiovascular agent	10

2.5.3.2 Anticonvulsant

The second group that we will focus on in the Broad validation sample is the anticonvulsant group. Anticonvulsant drugs are used in the prevention and treatment of epileptic shock. The mechanism by which these drugs work is by suppressing the rapid firing of neurons. At $k = 10$ the direct method has a recall of 17 while the indirect method is able to recall 20 other anticonvulsants (from different batches). Figure 2.4(b) shows that this trend generally holds for this group as well, with a slight dip at $k = 20$.

2.5.3.3 Antipsychotic

The last group that we will evaluate is the antipsychotic group, which has the third highest number of instances in the Broad dataset. Antipsychotic drugs are used to treat psychosis. Neither method in this group is able to recall as many instances as in the previous two groups. The direct method recalls 9 antipsychotics while the indirect method recalls 6 ($k = 10$). The indirect method is able to gain an advantage at $k = 30$, however, then the methods switch back again (Figure 2.4(c)). The recall of both methods is much lower for this antipsychotic group (in the Broad dataset) than previous groups and this is a possible explanation for why we do not see improvement. Other potential explanations include the possibility that the antipsychotics could have different, and unique, biological mechanisms of action and therefore similarity is not detected by gene expression similarity, or alternatively that the common mechanism is below the threshold used for similarity, namely that we used the 500 top and bottom probes.

2.5.4 Evaluating the Statistical Significance of Improvement

We have demonstrated that our indirect method results in a large improvement in the recall of similar compounds over the direct method in the face of vehicle and batch effects. Specifically, we have shown an improvement of known true positives at rank 10 across a number of therapeutic groups. We evaluate if the difference in ranks of these true positives is statistically significant. Our indirect method does statistically better than the direct (CMAP) approach in 33% of the groups while remaining as accurate as the direct method on the remaining groups. This analysis is performed for the groups listed in Table 2.4 as follows. For each set of instances belonging to a particular therapeutic group we use both methods (direct and indirect) to determine the top 100 similar instances per method. We avoid cases where neither method recalls the true positives within the top 100 instances, as differences this far down the result listing is of limited practical interest. However, if one method recalls an instance within the top 100, we include the rank for the other method even if it is outside of the top 100, because we want to give either method credit for these cases. We perform a paired t-test between the two methods for each of the therapeutic groups. The indirect method is statistically better in 3/9 of the therapeutic groups when evaluating the top 100 results and is statistically equivalent (no statistical difference between the methods) in the other 6 groups. In selecting 100 as the threshold we evaluate at several other thresholds as well. The indirect method is never statistically worse than the direct method across all 900 evaluations (9 groups X 10 rank thresholds) and is statistically better in 22 cases, including being statistically better at all thresholds 20-100 for the anti-inflammatory group, which is the biggest group with 28 instances. The results for this analysis at a search threshold of 100 are show in Table 2.5 and the followup sensitivity analysis is shown in Table 2.6.

2.5.5 Evaluation For Classification on Additional Datasets

We have demonstrated how this novel method can work in a large, gene expression based, drug discovery framework which has been our motivating problem and focus. We now analyze our indirect method on three smaller (public) datasets. We evaluate how our indirect method performs in distinguishing cancer types (acute myeloid leukemia versus acute lymphoblastic leukemia) and in predicting drug sensitivity/resistance. Additionally, we demonstrate the ability to use our indirect method to distinguish three very similar and related cell types in a third dataset.

2.5.5.1 Molecular Classification of Cancer

Golub et al. [16] evaluated the use of gene expression signatures to classify acute leukemias. They created a database of expression profiles of both acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) samples and demonstrated how gene signatures can help to classify these subtypes of acute leukemia. This is an important task, as the appropriate treatment for an individual depends on understanding the tumor type. Maximizing efficacy and minimizing adverse events and toxicity is the goal, and this is best achieved by prescribing chemotherapies that target the correct pathogenetically distinct tumor types.

This dataset consists of 52 samples (24 ALL and 28 AML). Analogous to searching for similar drugs we can search for samples of the same cancer class, e.g., searching with an ALL sample should yield other ALL samples. In order to perform the classification we use the majority vote of the top k results. In this case, the majority vote of the top 11 results (as opposed to 10 to avoid ties) recalled by a given sample is used to classify the sample. In this example, the direct method does extremely well, accurately classifying every sample correctly. The indirect method also correctly classifies every sample correctly. However, this is a high level comparison and we can understand and evaluate the results in more detail by looking at the individual rankings upon which the voting relies.

We describe the average rank (of samples of the same class) across the two methods. The average rank for ALL by the indirect method is 12.2 compared to 14.5 for the direct method. The AML class also demonstrates an improvement where the average rank for the indirect method is 20.5 versus 21.6 for the direct method. While this improvement is consistent across both groups, the small number of groups in this dataset does not readily allow us to evaluate the statistical significance. For this we instead evaluate the underlying ranks for each sample. The full results are listed in Table 2.7 and Figure 2.5 shows the corresponding ROC curve. The AUC for the 0.829 for the indirect method and 0.727 for the direct method. This difference is statistically significant using the method described by DeLong et al [11] (p=1.13e-

32). Note that k is set to 250 to compensate for there being roughly half of the probes as used in the previous datasets $(12,564)$. The average rank improvement of recalling similar samples is 1.7 when using the indirect method as compared to the direct method.

2.5.5.2 Predicting Drug Sensitivity/Resistance

The next dataset (from Wei et al. [63]) that we evaluate consists of ALL expression profiles of individuals that are known to be sensitive or resistant to glucocorticoid treatment, specifically in regards to childhood ALL. This is an important task because a poor prognosis is linked to resistance to glucocorticoid-induced apoptosis of primary lymphoblastic leukemia cells in vitro [27, 45, 19, 26]. There are 13 glucocorticoid sensitive samples and 16 that are glucocorticoid resistant (total n=29). As before, we use a paired t-test comparing the average rank of recalling samples from the same class (i.e., sensitive or resistant). The indirect method improves upon the direct method by 0.45 on average across all samples.There are $22,283$ probes used in this dataset and k is set to 500. The full results are listed in Table 2.8 and the ROC curve is shown in Figure 2.6. The AUC is 0.635 for the indirect method and 0.608 for the direct method (p=1.90e-06).

2.5.5.3 Classifying (Related) Cell Types

The final dataset that we use to evaluate our indirect method is from Lu et al. [36]. It consists of megakaryocyte-erythrocyte progenitors (MEP) as well as the two cell types that MEPs can differentiate into, namely megakaryocytes and erythrocytes. Megakaryocytes are bone marrow cells that are responsible for the production of platelets while erythrocytes are red blood cells. We refer to this dataset as the MEP dataset. The original focus of Lu et al. [36] was to better understand the differentiation process and was not evaluating the classification of these three cell types. There are 320 probes, and we set $k = 10$ in order to maintain roughly the same ratio as before.

This is another example of how the indirect method can improve over the direct method even with a small dataset. There are only 3 classes and 27 total samples of which 9 are erythrocytes, 10 are megakaryocytes and 8 are MEPs. Analyzing the results in the same fashion as before we find that the indirect method statistically improves upon the direct method once again with an average improvement of 1.1. The ROC curve is shown in Figure 2.7 with the full listing of results in Table 2.9. The AUC for the indirect method is 0.670 compared with 0.633 for the direct method (p=4.13e-13).

2.5.6 Computational Complexity

We have presented a comparison of our novel indirect similarity method to the normal direct similarity method in terms of increased true positives recalled at a given threshold. We have thus far ignored the practical challenges that may occur in implementing either of these methods in a production system. The transition to using an indirect method has implications on both the time needed for calculating the similarity, as well as the space needed to store the information required by the system to work efficiently.

One important thing to note is that we are evaluating the difference between the indirect and direct similarity methods in the context of a knowledge discovery framework. In this situation, the system already performs all pairwise similarity comparisons to detect novel insights. This is quite different than the initial goal of the CMAP method, which was first developed as a real-time, query based tool for the web. The actual running time for the direct analysis was $\sim$ 24 hours to calculate the KS scores versus $\sim$ 3.6 hours for the indirect similarity calculation. A fair amount of effort was spent optimizing the indirect calculation and the indirect calculation additionally benefited from being run on a highly parallel SAS server optimized for such statistical calculations. We feel that it is more important to evaluate the theoretical computational complexity which follows below. Additionally, this theoretical analysis applies to any new direct method.

2.5.6.1 Time Complexity

We assume that all the data has been preprocessed and is stored as a rank-ordered list, reflecting the difference as compared to control. For the purpose of this analysis, we are interested in the relative complexity of our indirect method as compared to the complexity of the underlying direct method. As our indirect method can use any underlying direct method as its base, we focus on the relative complexity for a general comparison. Let us assume that the computational cost of one individual pairwise comparison using a given direct method is c_1. The complexity of performing all pairwise comparisons using the direct method is shown in Eq. (2.8).

$$T(n) = \frac{c_1 n(n-1)}{2} = O(n^2) \tag{2.8}$$

The implementation of the indirect method similarity takes advantage of reusing this full matrix of direct comparisons and does not naively recalculate any direct similarities. Once again, assuming a small constant, c_2 ,to calculate the Spearman correlation for a given pair, the time complexity of our indirect similarity method is given in Eq. (2.9). It should be noted that in our current experimental setup $c_2 < c_1$, as the computational complexity of the KS statistic far outweighs the complexity of the Spearman correlation.

$$T(n) = \frac{c_1 n(n-1)}{2} + \frac{c_2 n(n-1)}{2} = O(n^2) \tag{2.9}$$

We have managed to keep the time complexity of our indirect similarity method in the same order of magnitude as what is required by the underlying direct similarity method. We next determine the impact that we have on the space complexity of moving to an indirect approach.

2.5.6.2 Space Complexity

The data that needs to be stored is that of the individual ranked lists representing the treatments as compared to their respective control. For our given application, this works out to be $m \cdot n$, where m is the size of each ranked list (in our case $22,283$) and n is the number of treatment instances. We ignore the negligible space requirement needed for one individual direct comparison since this intermediary is not retained. The space complexity is then once again $O(n^2)$ and the direct similarities are stored as a $n \times n$ matrix. Analogously, the indirect similarities are also maintained in an $n \times n$ matrix requiring $O(n^2)$ space as well. In the current implementation, both of these space requirements are overshadowed by the large space requirements of the initial dataset itself.

2.6 Related Work

The underlying idea of indirect similarity can be observed within other areas of computer science research. For instance, examples of related approaches of indirect similarity appear within the domains of collaborative filtering and entity resolution. Work on item-based collaborative filtering, i.e. selecting items to recommend based on similarity to other items, has been represented as an item-item correlation [52, 34].

Additionally, the general task of predicting the similarity of two drugs can be seen as an extension of the entity resolution problem. The goal of entity resolution is to reconcile database references corresponding to the same real-world entities. However, as opposed to attempting to find identical items, we are interested in a less strict threshold of similarity. Recent research in the area of entity resolution has focused on the use of additional relational information between database references to improve resolution accuracy [7, 6]. This improvement is made possible by resolving related references or records jointly rather than independently. One example of resolving similar entities by their joint similarity to the rest of the entities can be seen in D-Dupe, a visual inspection tool for entity resolution, which has been shown to work well for the entity resolution problem[8].

Many specifics of our method distinguish it from this other work, the most obvious being that our similarity method is applied to complete, rather than partially, ranked lists. Additionally, our indirect similarity does not require labeled relationships but rather treats pair-wise similarities as the relationship.

2.7 Discussion

We have proposed a method for similarity search in gene expression data and an evaluation method based on recall at rank k. Additionally, we have focused on the ability to detect similarity despite known experimental biases, e.g., different vehicles, different batches, or both different vehicles and batches. The importance of improvements in recall are not confined to solely help to overcome such explained experimental effects, but they are in fact representative of the larger set of unknown environmental effects. It can thus be assumed that the indirect similarity measure will be able to overcome unknown changes in gene expression experiments, and is therefore well suited for comparing gene expression data from vastly different data sources, as these data sources can be viewed as being separate batches.

Furthermore, we have shown that in a large, proprietary dataset, this indirect method is able to overcome the experimental noise better and is able to recall a larger number of drugs that are similar to the query drug. More specifically, the indirect method was able to increase the amount of known similar drugs recalled by 97.03% over the direct method. These results have also been validated on a public dataset (Broad), for which the improvement in recall was 49.44%. The difference in improvement is representative of the fact that the Broad dataset is both smaller and less complex, i.e., contains more replicates. The benefit of the indirect method comes from the information in the rest of the database and therefore the improvement is expected to increase as the size and complexity of the database grows.

The ability to recall 50%-100% more compounds that are similar (similar based on the annotations), gives a researcher an advantage in his/her analysis. In one scenario, using the indirect method may decrease the amount of candidates that might have to be followed up in an experiment or drug discovery system, thereby saving time and effort by not chasing false negatives. This in turns allows more confidence in the results generated by such a system. More importantly, it also brings the community an additional step closer to being able to pull together and learn from the large amount of data that exists in the public domain, which continues to grow every day.

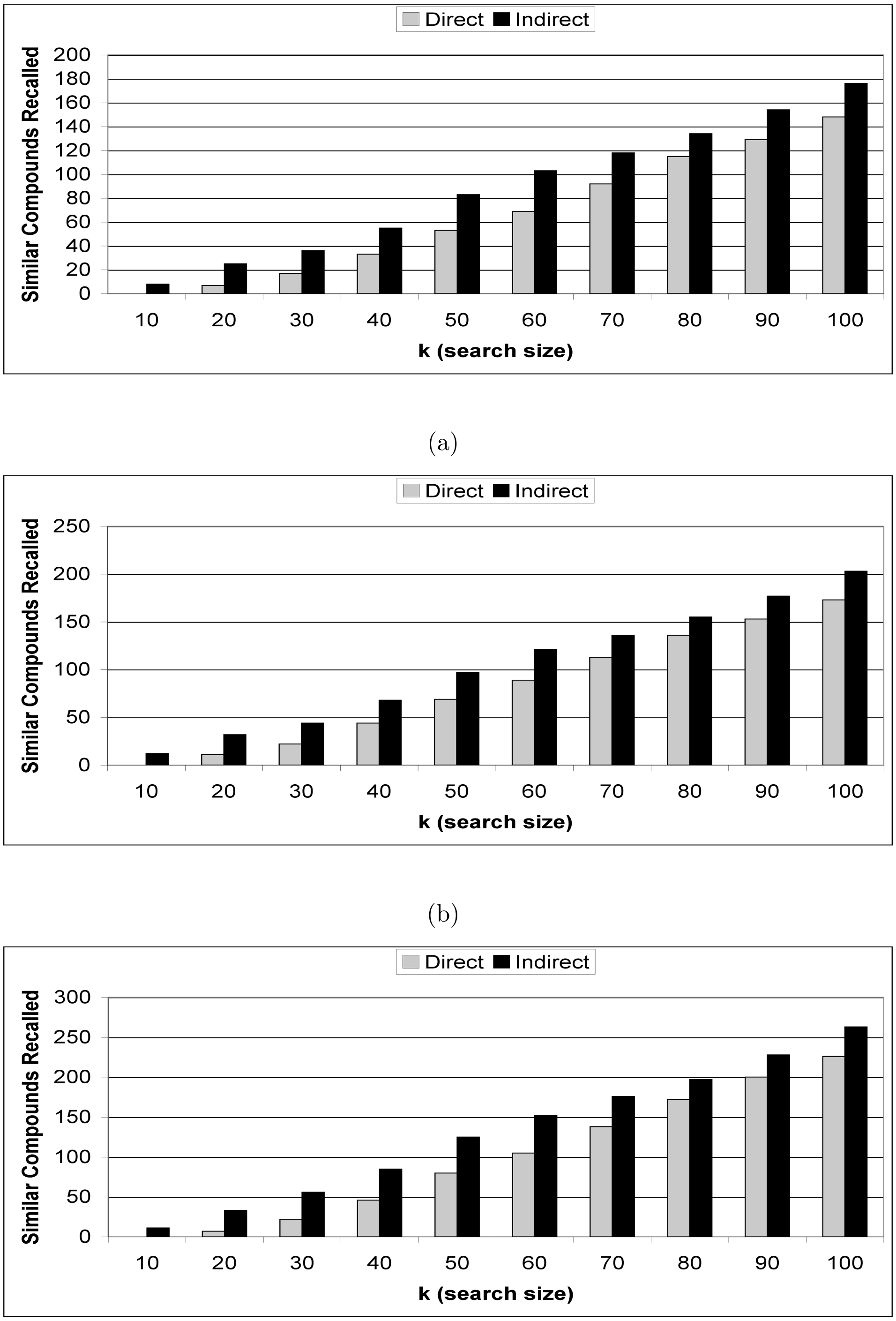
However, it is important to acknowledge that there are many potentially avenues for improvement and further research for this problem. One of the nice properties of the indirect approach is that it can be built upon any direct similarity method. If better direct methods are developed in the future then the indirect method can be adapted to use such methods. The work presented in this section has dealt with the task of adapting to and overcoming experimental bias in gene expression data, specifically for the task of comparing two samples directly. Other potential paths of future research could include more complex comparisons, e.g., analyzing groups of compounds together, as well as a more thorough evaluation of selecting the optimal size of k for a given dataset.

Table 2.5: Statistical Analysis of the Improvement in Rank

Therapeutic Use	N	Mean Improvement	StdDev	tValue	P
analgesic	29	90.43	117.55	4.14	0.0003
anti-inflammatory	160	30.12	115.04	3.31	0.0011
anti-psychotic (neurole	69	25.57	73.36	2.90	0.0051
immunosuppressive agent	67	7.96	40.42	1.61	0.1120
tranquilizer	68	14.43	73.95	1.61	0.1124
cardiovascular agent	22	10.82	69.37	0.73	0.4726
anti-neoplastic	54	10.33	114.65	0.66	0.5106
anti-convulsant	144	-4.50	92.09	-0.59	0.5585
anti-proliferative	86	-1.86	77.86	-0.22	0.8252

Table 2.6: Statistical Analysis of the Improvement in Rank - Sensitivity Analysis

Search Size:	10	20	30	40	50	60	70	80	90
Therapuetic Use									
analgesic	0.136	0.087	0.041	0.020	0.001	2.0E-04	0.004	0.001	0.002
anti-convulsant	0.811	0.529	0.932	0.636	0.918	0.856	0.478	0.481	0.512
anti-inflammatory	0.151	0.040	0.034	0.033	0.002	0.001	0.003	0.002	0.004
anti-neoplastic	0.412	0.842	0.943	0.928	0.577	0.804	0.919	0.909	0.779
anti-proliferative	0.148	0.553	0.778	0.659	0.987	0.881	0.942	0.837	0.837
anti-psychotic (neurole	0.905	0.852	0.686	0.674	0.529	0.129	0.063	0.024	0.006
cardiovascular agent	0.647	0.267	0.496	0.496	0.540	0.795	0.984	0.801	0.687
immunosuppressive agent	0.047	0.664	0.642	0.550	0.525	0.525	0.554	0.323	0.323
tranquilizer	1.000	0.580	0.579	0.921	0.665	0.619	0.413	0.290	0.159

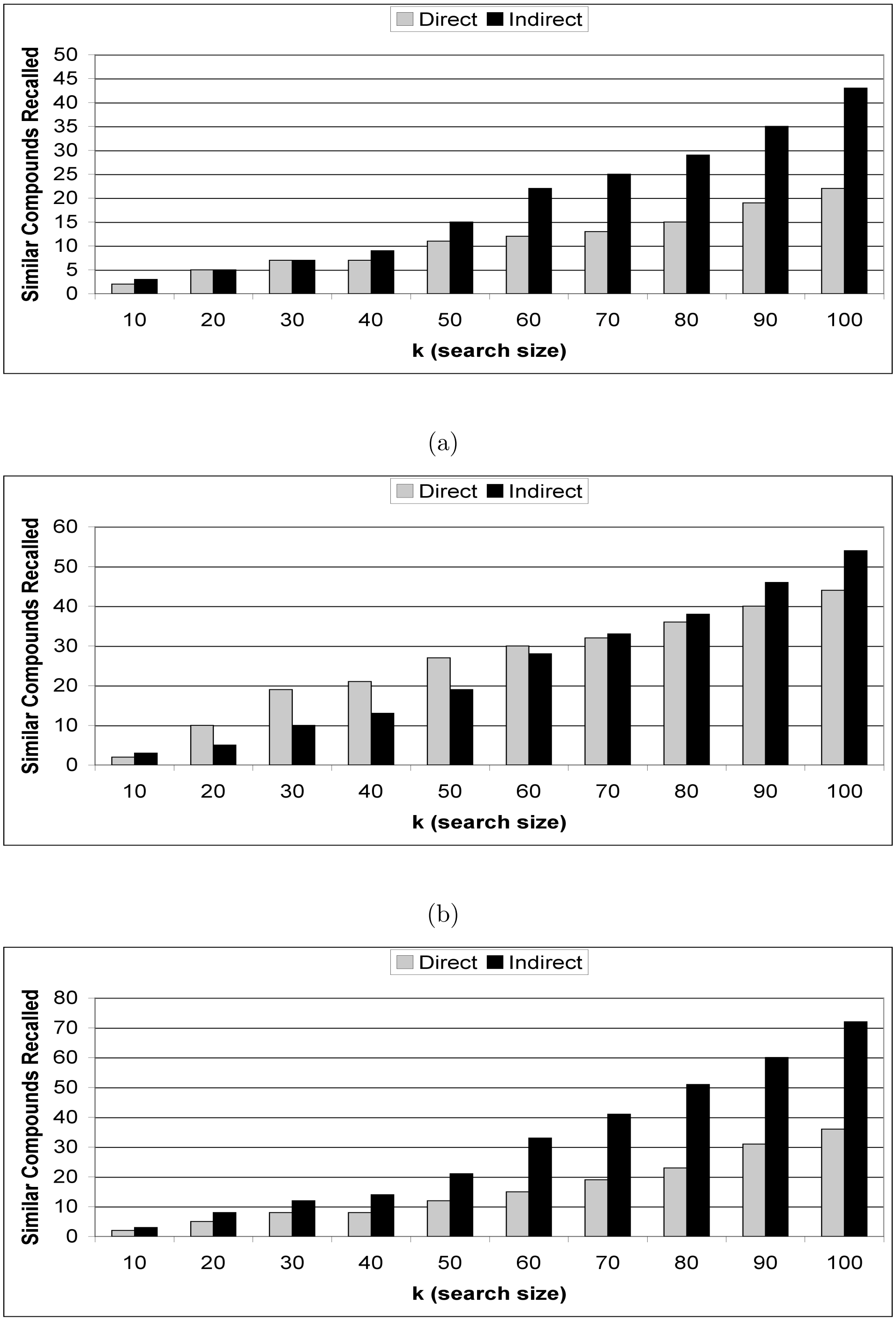


(a)

(b)

(c)

Figure 2.1: Antibiotic recall at rank k (a) across both different vehicles and batches, (b) different vehicles, and (c) different batches. There are $n = 56$ antibiotic instances in this dataset.

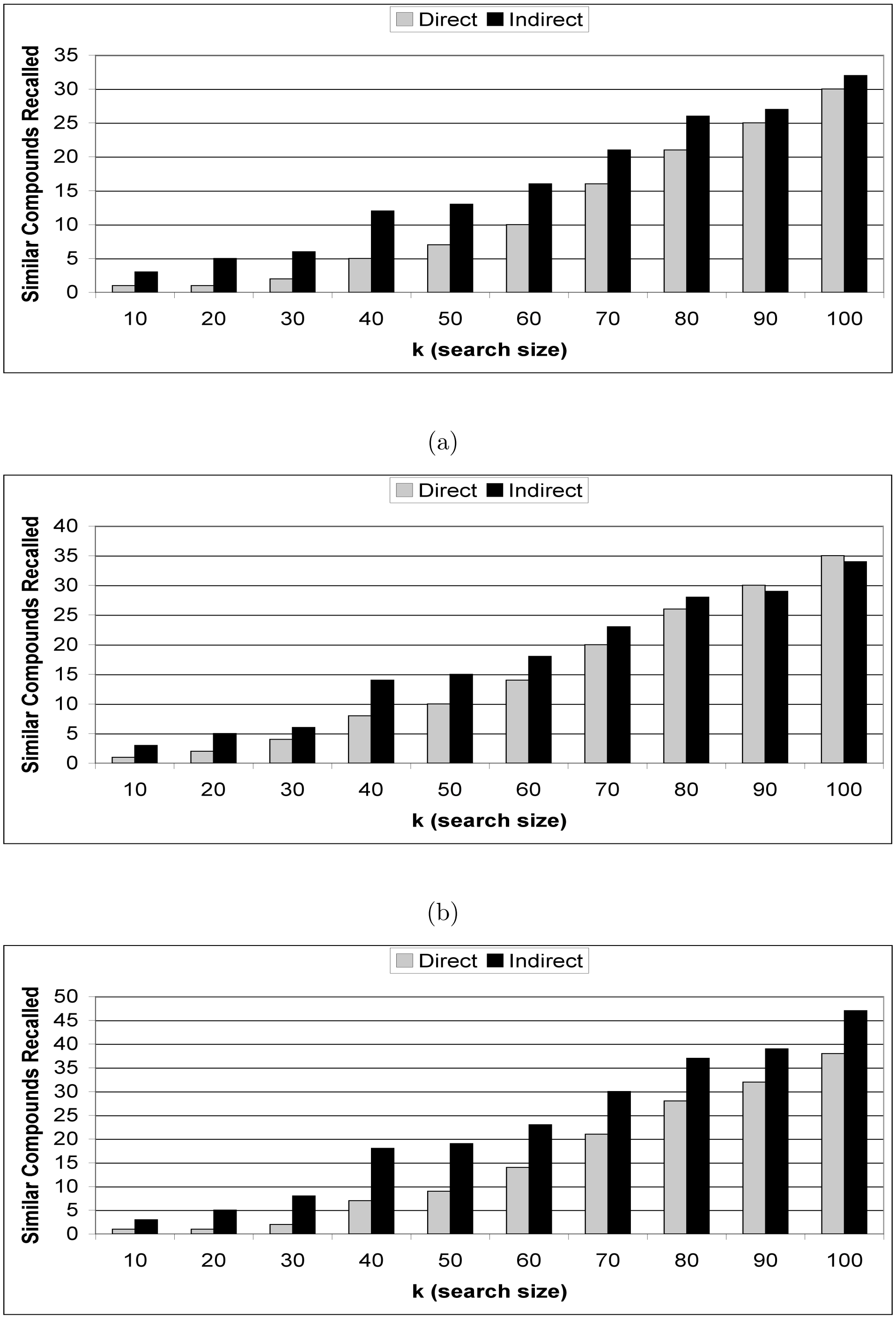


(a)

(b)

(c)

Figure 2.2: Histamine Antagonist recall at rank k across (a) both different vehicles and batches, (b) different vehicles, and (c) different batches. There are $n = 24$ Histamine Antagonist instances in this dataset.

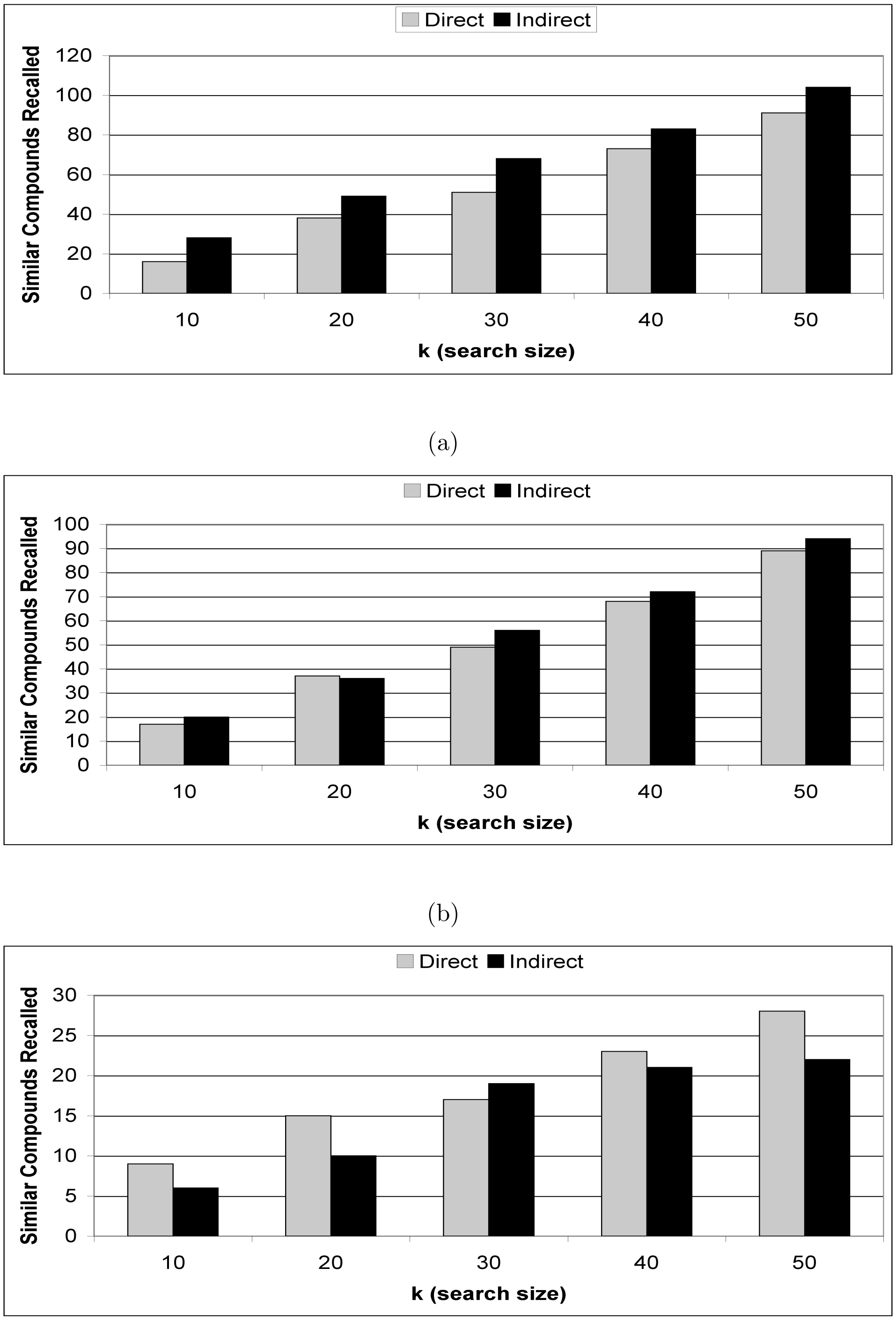


(a)

(b)

(c)

Figure 2.3: Analgesic recall at rank k across (a) both different vehicles and batches, (b) different vehicles, and (c) different batches. There are $n = 23$ antibiotic instances in this dataset.



(a)

(b)

(c)

Figure 2.4: Broad validation results with varying k on top 3 groups: a) AntiInflammatory ($n = 28$) b) Anticonvulsant ($n = 21$), and c) Antipsychotic ($n = 19$).

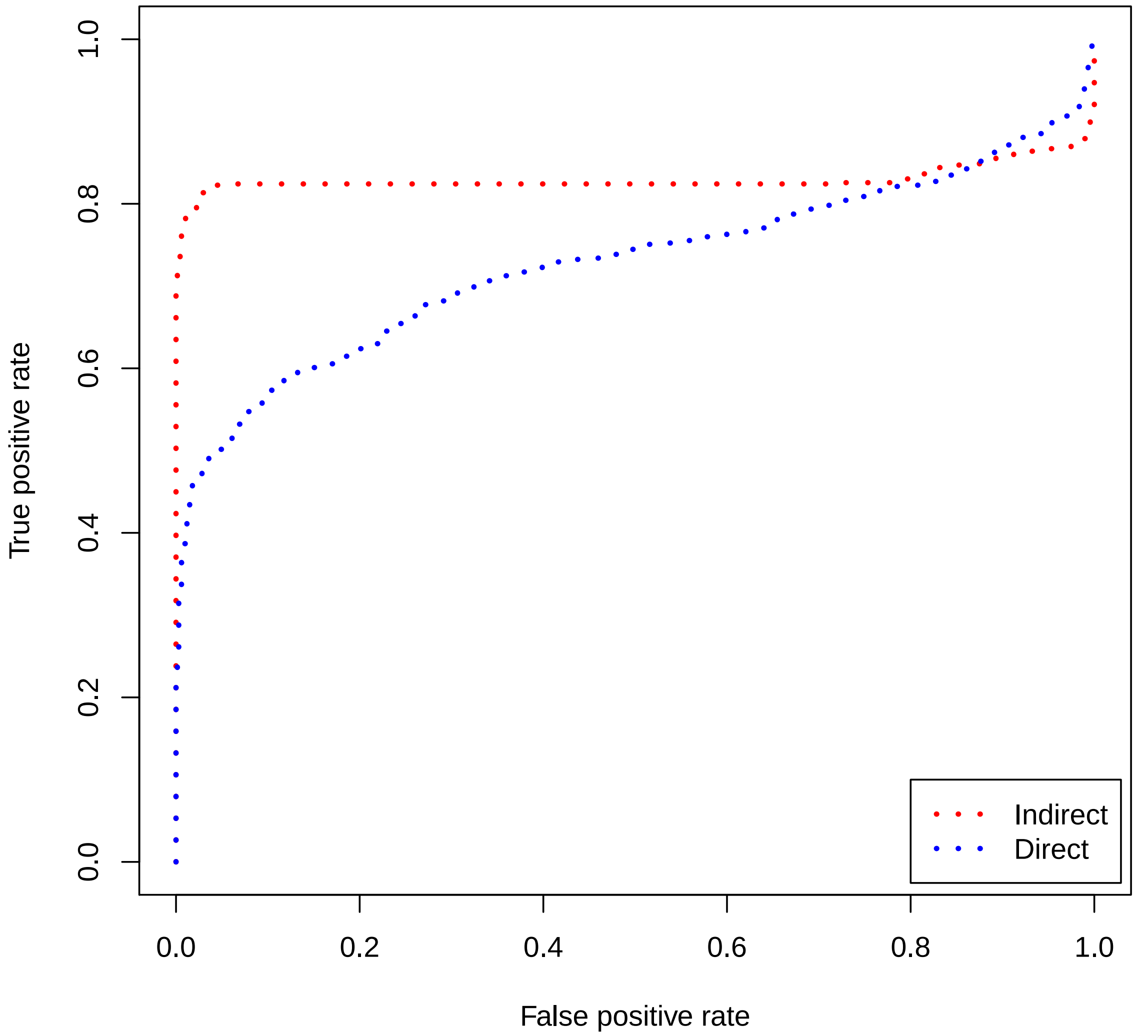


Figure 2.5: ROC curve showing the difference in sensitivity and specificity between the direct and indirect method for the task of predicting subtypes of cancer (AML versus ALL). The indirect method (shown in blue) performs better than the direct method (shown in red). The difference is statistically significant (p=0.0001 for the paired t-test on the underlying ranks).

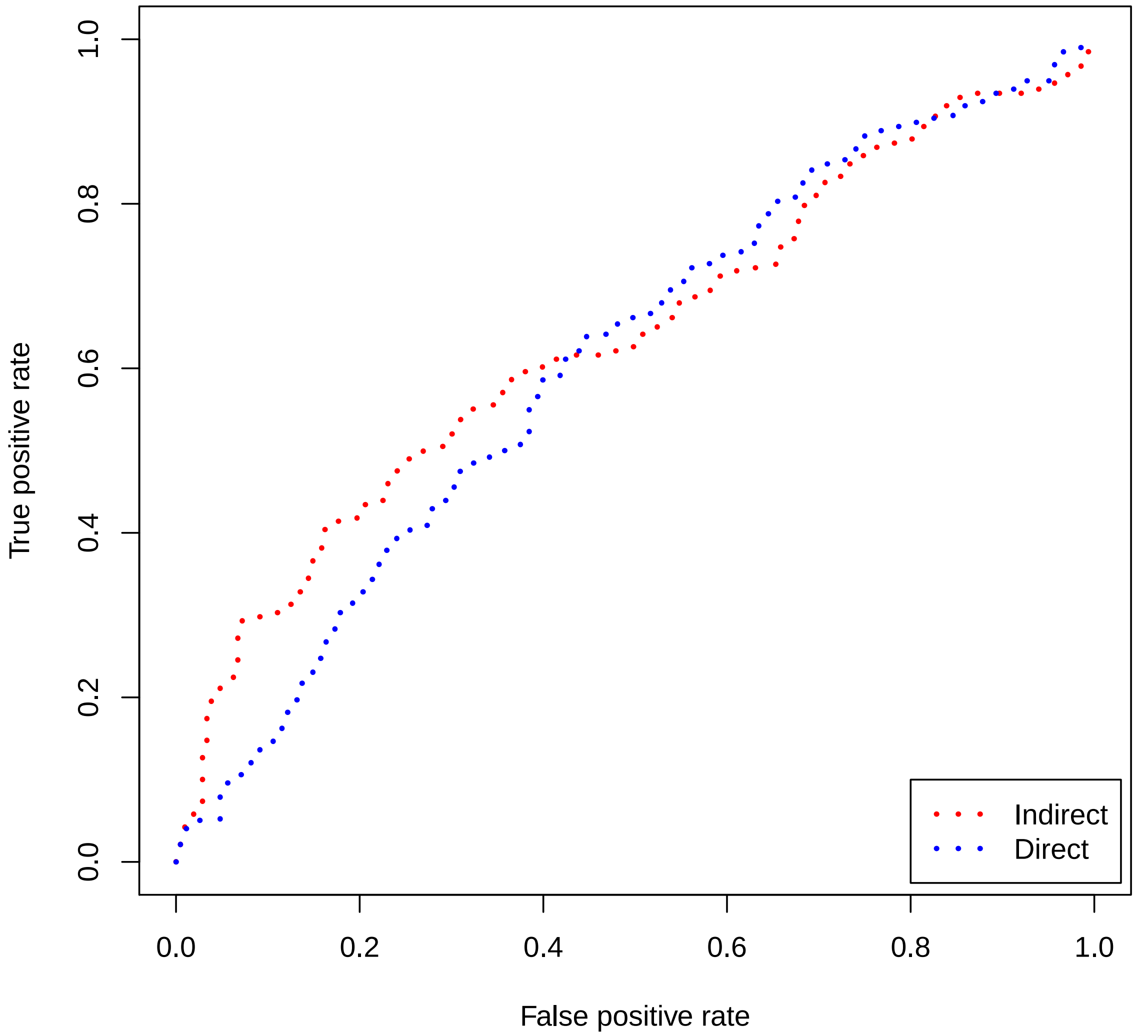


Figure 2.6: ROC curve showing the difference in sensitivity and specificity between the direct and indirect method for the task of predicting sensitivity or resistance to glucocorticoids. The indirect method (shown in blue) improves upon the direct method (shown in red). The difference is statistically significant (p=0.0156 for the paired t-test on the underlying ranks).

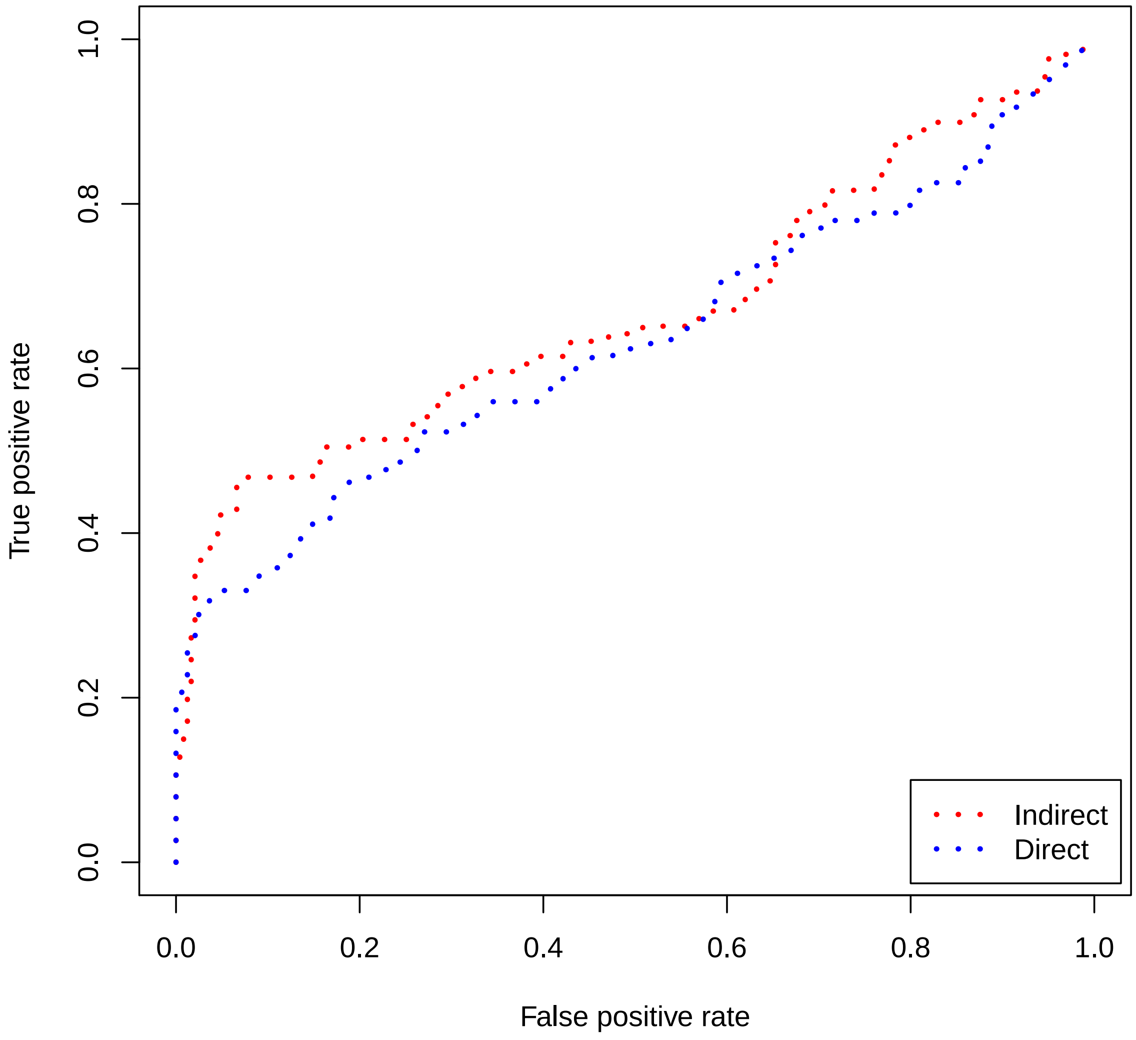


Figure 2.7: ROC curve showing the difference in sensitivity and specificity between the direct and indirect method for the task of classifying cell types (megakaryocyte, erythrocyte, and their corresponding progenitors, i.e., megakaryocyte-erythrocyte progenitors). The indirect method (shown in blue) outperforms the direct method (shown in red). The difference is statistically significant (p=0.0035 for the paired t-test on the underlying ranks).

Table 2.7: AML vs ALL Average Rank

Sample	Class	Indirect	Direct
ALL 1	ALL	12.0	13.3
ALL 2	ALL	12.0	14.1
ALL 3	ALL	12.1	19.2
ALL 4	ALL	12.0	14.3
ALL 5	ALL	12.0	15.4
ALL 6	ALL	12.0	12.7
ALL 7	ALL	12.0	12.1
ALL 8	ALL	12.0	16.0
ALL 9	ALL	12.0	12.3
ALL 10	ALL	12.0	12.3
ALL 11	ALL	12.0	14.1
ALL 12	ALL	12.0	12.0
ALL 13	ALL	12.2	16.4
ALL 14	ALL	12.0	13.0
ALL 15	ALL	12.0	13.0
ALL 16	ALL	12.0	16.9
ALL 17	ALL	12.1	14.8
ALL 18	ALL	12.0	13.1
ALL 19	ALL	12.0	14.1
ALL 20	ALL	17.0	21.3
ALL 21	ALL	12.0	13.5
ALL 22	ALL	12.0	13.3
ALL 23	ALL	12.0	14.1
ALL 24	ALL	12.0	17.4
AML 1	AML	18.5	20.7
AML 2	AML	18.3	19.5
AML 3	AML	18.3	19.9
AML 4	AML	18.3	19.2
AML 5	AML	18.6	21.7
AML 6	AML	18.4	20.3
AML 7	AML	18.3	19.3
AML 8	AML	18.4	20.7
AML 9	AML	18.3	19.6
AML 10	AML	18.4	21.4
AML 11	AML	20.4	25.9
AML 12	AML	18.3	19.7
AML 13	AML	18.4	20.7
AML 14	AML	18.2	19.1
AML 15	AML	18.5	20.8
AML 16	AML	18.2	19.4
AML 17	AML	18.5	20.7
AML 18	AML	18.4	20.3
AML 19	AML	18.2	19.9
AML 20	AML	18.3	19.9
AML 21	AML	18.3	19.8
AML 22	AML	18.4	20.3
AML 23	AML	18.3	19.6
AML 24	AML	34.3	28.1
AML 25	AML	34.3	28.5
AML 26	AML	33.9	28.3
AML 27	AML	34.3	30.4
AML 28	AML	14.0	22.4

Table 2.8: Glucocorticoid Sensitivity / Resistance Average Rank

Sample	Class	Indirect	Direct
DT2004021428-738	S	10.1	12.6
DT2004021429-976	S	9.8	10.5
DT2004021430-1047	S	10.0	10.6
DT2004021431-1219	S	13.6	12.8
DT2004021432-1241	S	13.3	12.1
DT2004021433-1299	S	9.0	9.6
DT2004021434-1307	S	7.9	8.7
DT2004021435-1477	S	7.3	7.8
DT2004021436-1533	S	12.5	12.2
DT2004021437-1553	S	12.4	13.1
DT2004021438-1657	S	14.7	14.0
DT2004021439-1684	S	10.2	9.5
DT2004021440-1696	S	12.1	13.2
DT2004021441-329	R	11.0	12.1
DT2004021442-557	R	9.8	10.7
DT2004021443-685	R	16.8	15.7
DT2004021444-789	R	12.1	12.7
DT2004021446-865	R	20.3	19.2
DT2004021448-1466	R	13.3	13.9
DT2004021449-1652	R	13.3	14.8
DT2004021451-1755	R	11.9	12.7
DT2004021452-2078	R	11.9	13.3
DT2004021453-2200	R	16.5	15.7
DT2004021454-2209	R	14.1	14.5
DT2004021455-vu8978	R	12.0	13.3
DT2004021456-vu9023	R	10.2	11.8
DT2004021457-vu9573	R	10.5	10.7
DT2004021458-vu9728	R	12.0	13.3
DT2004021459-vu9951	R	13.9	14.6

Table 2.9: MEPs (Megakaryocyte-Erythrocyte Progenitors Average Rank

Sample	Class	Indirect	Direct
ERY1-1	ERY	10.3	13.1
ERY1-2	ERY	9.5	8.8
ERY1-3	ERY	14.0	16.4
ERY1-4	ERY	10.3	14.9
ERY2-1	ERY	16.4	16.4
ERY2-2	ERY	15.8	18.1
ERY2-3	ERY	12.8	13.6
ERY3-1	ERY	18.6	18.4
ERY3-2	ERY	18.6	17.5
MEGA1-1	MEGA	7.3	7.3
MEGA1-2	MEGA	6.9	5.8
MEGA1-3	MEGA	13.7	11.3
MEGA1-4	MEGA	12.1	14.9
MEGA2-1	MEGA	8.2	7.7
MEGA2-2	MEGA	7.0	8.6
MEGA2-3	MEGA	7.3	7.1
MEGA2-4	MEGA	7.4	7.1
MEGA2-5	MEGA	8.7	8.7
MEGA2-6	MEGA	7.4	7.6
MEP-1	MEP	6.1	9.1
MEP-2	MEP	6.0	7.0
MEP-3	MEP	6.3	8.7
MEP-4	MEP	6.1	7.0
MEP-5	MEP	6.1	8.4
MEP-6	MEP	6.3	8.7
MEP-7	MEP	4.7	8.1
MEP-8	MEP	6.1	8.1

Chapter 3

Signature Detection

In the previous chapter, we tackled the problem of detecting pairwise similarity in gene expression data. We showed how the use of additional knowledge in the database can lead to more informed decisions by making use of our proposed indirect similarity method. However, there are many limitations of a pairwise similarity. The most important such limitation is understanding what is driving the similarity. In our example domain of detecting drugs with the same therapeutic use, let us assume that we detect a high similarity between two drugs, an antipsychotic drug and an antibiotic drug. We do not know why they are similar, or more formally: is the antibiotic drug acting like an antipsychotic drug, or vice versa? In terms of the hypothesis of a drug discovery system: should the antipsychotic drug be tested as an antibiotic drug? Or should the antibiotic drug be tested as an antipsychotic drug? Could it even be something completely different, maybe a side effect that they share in common? These are all questions that we cannot begin to answer when restricted to using solely a pairwise similarity.

Microarray experiments, whether they set out to discover biomarkers for a particular disease or to characterize a group of similar tissue samples, tend to have the same outcome: the signature detection of a list of differentially expressed genes (DEGs). We propose the creation of a group profile that will serve as the representative profile for a given group of interest. A gene expression profile is the representation of the activity of thousands of genes at once for a given sample. In our motivating examples these profiles each correspond to one microarray experiment, but the method is general and can extend to other input data types. A group profile represents the shared activity of these thousands of genes across all of the member samples belonging to the group. For example, we can create a group profile consisting of all available antipsychotic drugs; we refer to this as an antipsychotic profile. Traditionally, researchers attempt to find probes or genes that form the signature for a group by evaluating probes above a certain fold-change threshold. Fold-change refers to the ratio of change between Treated, t, and Control, c, such that the fold-change of treatment compared to control would be t/c. These methods will detect the signature common to the group in the rare case that the shared effect is incredibly strong (and there are no large experimental biases between the expression profiles). However, the majority of the time, the true signal is missed because it is not significantly up- or down-expressed in every one of the instances that make up a group (we refer to this as the full group). These methods preferentially detect very big changes within a subgroup of samples and then merge all of these differentially expressed genes with a combination function. Unfortunately, this approach does not find true signatures common to the full group and allows the method to overfit the data. Our method differs from most other methods by focusing on detecting signatures common to the full group, signatures that are normally overlooked by other methods, e.g., decision trees and support vector machines[46], linear models[57], etc., which can explain a group as a combination of rules defining unknown subgroups.

The representation of a group profile is a ranked list of all probesets on the microarray. A benefit of our approach is that this is the same representation as a single profile. This representation allows any current and future methods for non-parametric gene expression data to be used with our group profiles. We can focus on the most up- and down-expressed probesets from the profile, which we refer to as the signature of the group (separately they are the up and down signatures respectively). For example, we can make use of methods developed by others (e.g., Connectivity Map (CMAP) [28]) to use this antipsychotic group profile to search a database for drugs sharing the same signature. Alternatively, we can use still other methods (e.g., the L2L Microarray Analysis Tool [43]) to evaluate if any particular biological process is overrepresented within this signature, an approach that would provide additional insight into the common mechanism of antipsychotic therapies.

In this chapter, we introduce and describe our rank of ranks method for group profile creation. We evaluate the utility of this group analysis method using a pilot study in which we focus on the antipsychotic group from the original CMAP build 01 dataset. Our evaluation consists of both understanding the group profiles biologically and demonstrating the ability to use a signature from these profiles as a predictive model of therapeutic use. We conclude with a full analysis of the newer, and larger CMAP build 02 dataset, including a sensitivity evaluation of each group as well as the validation of the most robust profiles within an independent dataset. In addition, another contribution of this work is the independent validation of the published expression signature of antipsychotic drugs. All the results are available at GEPedia.org.

3.1 Problem Definition

Given a database D of treatments (i.e., drugs or other compounds), D = X_1,..., X_n, we are interested in creating a set of group profiles. A group can be defined as a set of instances (e.g., cells treated with a particular drug) that share something of interest in common (e.g., the same therapeutic use, mechanism of action, side effect, chemical structure). We are interested in understanding what is biologically common for a given group profile as well as evaluating the ability to query the database with the group profile to predict new members of the group. Our goal is to discover other drugs or treatments, perhaps originally developed for a different therapeutic purpose, which are likely to also share the same therapeutic properties as the query group. These therapeutic agents are thus good candidates for which new uses can then be evaluated.

For each treatment instance X in the database, there is both general information about the experimental conditions of the sample as well as the actual experiment data from the microarray itself. The gene expression profile is represented as a ranked list (amplitude of the treatment as compared to the control). Amplitude a is defined as follows: $a = (t - c)/((t + c)/2)$[28]. Information specific to the treatment (i.e., the name of the drug, the therapeutic class [class] and subclass [subclass] as defined by the chemicals Anatomical Therapeutic Chemical [ATC] code) is represented. There is also information that describes the experimental conditions of the sample, specifically the molar amount of substance (mol), the vehicle used for delivery of the drug (e.g., water, EtOH, MeOH, DMSO), and the batch or round in which the sample was run. A group, and therefore a group profile, can be created from any of these meta-labels associated with the samples.

3.2 Group Profile Creation (Weighted Influence Model - Rank of Ranks Method)

Previous methods have demonstrated that weighted distribution-based statistics can be more robust in detecting similarity in the pairwise comparison of gene expression data [28]; therefore, we propose a method for determining what is common among a group by also using a weighted method. This dynamic weighting of probes allows us to avoid strictly filtering any probes as is done with a fold-change threshold approach. We calculate the average rank of each probe across the members of the group and then re-rank the probes based on this average rank. We refer to this as the Weighted Influence Model, Rank of Ranks (WIMRR) method. The rank of each probe within each treatment X is known: rank(p, X). Let us assume we have a binary membership function, member(X, G), that returns 1 if treatment instance X is a member of group G and returns 0 otherwise. The size of the group is equal to the number of treatment instances that are members of the group (Equation 3.1).

$$size(G) = \sum_{i=1}^{n} member(X, G) \tag{3.1}$$

The average rank across all of the members of a given group for each probe is then calculated as described in Equation 3.2.

$$avgRank(p, G) = \frac{\sum_{i=1}^{n} rank(p, X_i) \times member(X_i, G)}{sizeG} \tag{3.2}$$

Given this set of average ranks across the members of a particular group, the probes are now re-ranked according to how consistently they are up- or down-expressed across the group. We define Profile(G) as the probes in probes(G) sorted by their average rank across all members of the group (Equation 3.3).

$$Profile(G) = \{(p_1, avgRank(p_1, G)), (p_2, avgRank(p_2, G)), \ldots, (p_m, avgRank(p_m, G))\} \tag{3.3}$$

3.3 Group Profile Evaluation - A Pilot Study

We make use of the original CMAP dataset (build 01) from the Broad Institute to evaluate our group profile method as part of a pilot study. We refer to this as the CMAP 1.0 dataset. We use this smaller, simpler dataset to characterize our method. Later, we analyze the newer CMAP build 02 dataset (CMAP 2.0), which contains many more treatments. For each treatment instance in the CMAP dataset, probe sets are first ranked based on their level of expression relative to the vehicle control in a fashion similar to the method described by Lamb et al. [28]. A group profile is then created for each therapeutic use according to the ChemBank annotation for the instances using our novel WIMRR method. The signature of each group profile is created by selecting the top and bottom k probes. For this evaluation, we set k

= 50.

3.3.1 Antipsychotics from Pilot Study

We focus on the antipsychotic profile from the CMAP 1.0 dataset as an example by which to analyze the WIMRR group profile creation method. The antipsychotic group is selected as the example because it includes a large number of unique drugs. The instances from the CMAP 1.0 dataset that are labeled as antipsychotic agents according to ChemBank are used to create this group. The antipsychotics profiled in this dataset include chlorpromazine, clozapine, haloperidol, thioridazine, and trifluoperazine. There are 19 profiles total for this group, consisting of replicates across different concentrations. The group profile is created and the top and bottom 50 probes are selected to serve as the signature for this group (shown in Table 1).

The top and bottom probes can both provide valuable insight. We focus on the top 50 probes, but the same analysis can be performed with the bottom 50 probes in an analogous way. The amplitude value for the top 50 probes across all antipsychotics is shown in Figure 1. The amplitude value for the top probe (Affymetrix probe id 201170 s at) is shown in Fig. 2A. This probe, which corresponds to the basic helix-loop-helix domain containing, class B, 2 (BHLHB2) gene, is almost exclusively up-expressed in all of the antipsychotic instances. We evaluate the specificity of this probe by determining how this probe behaves across the whole database (Fig. 3). All but one of the antipsychotic instances (pink dots in first column) show a clear increase in expression levels. The next set of groups all contain

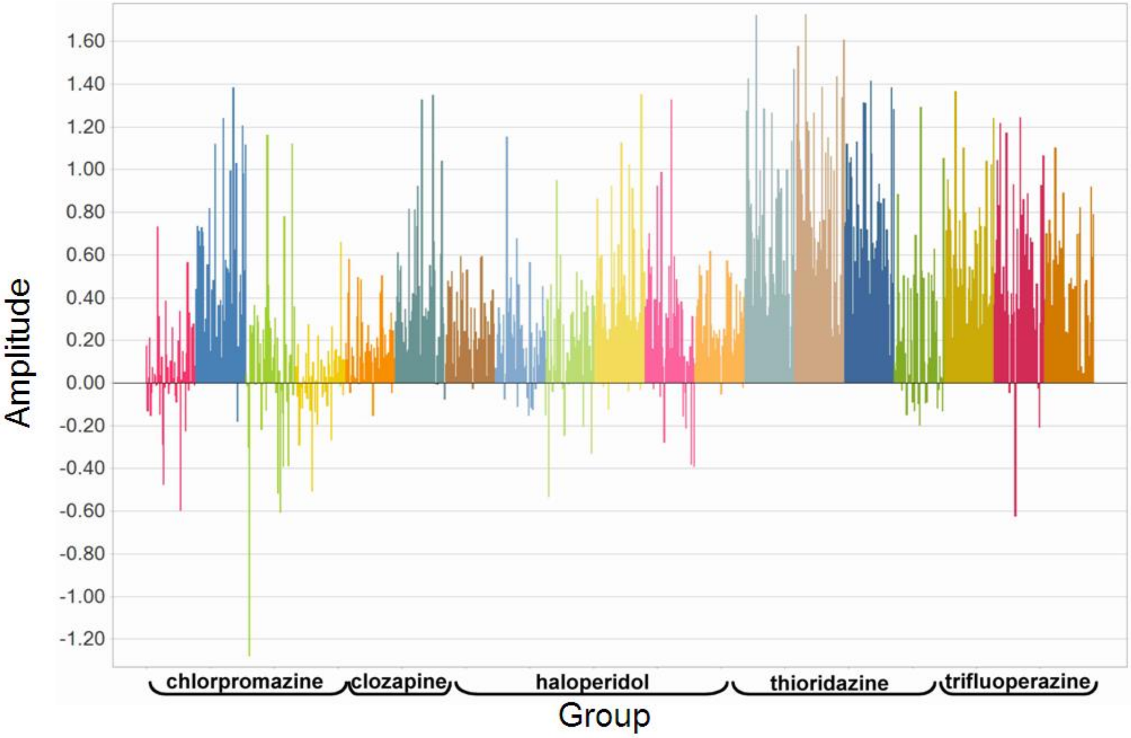


Figure 3.1: Amplitude values for top 50 probes of antipsychotic profile within each of the antipsychotic instances within the Broad dataset. Replicates are designated by different colors.

drugs that are known to also act as antipsychotics; this is expected if this probe is predictive of antipsychotic activity. The second group is the tranquilizers (includes prochlorperazine, fluphenazine, and trifluoperazine), the third group is antiemetics (includes prochlorpromazine and trifluoperazine), and the fourth group is the antineoplastics (includes prochlorpromazine). There is a clear pattern of antipsychotic activity related to the up-expression of this probe across the database.

We now compare what we have seen with the top probe from our method with a probe selected using more conventional methods. A potential alternative method for selecting probes (and genes) of interest that has been used extensively in the

Table 3.1: The top 50 probes of the up- and down-expressed signature from group profile created from the antipsychotic instances in the CMAP 1.0 dataset.

Up_Rank	Probe	Gene	Avg_Rank
1	201170_s_at	BHLHB2	2192.3684
2	212276_at	LPIN1	2431.2105
3	201627_s_at	INSIG1	2681
4	221577_x_at	GDF15	2710.3684
5	202672_s_at	ATF3	2833.3158
6	202769_at	CCNG2	2844.9474
7	208962_s_at	FADS1	3100.9474
8	209146_at	SC4MOL	3405.4211
9	208647_at	FDFT1	3427.7895
10	214326_x_at	JUND	3495
11	208933_s_at	—	3504.7895
12	210512_s_at	VEGFA	3556.7895
13	33304_at	ISG20	3594.3158
14	201626_at	INSIG1	3609.1053
15	208786_s_at	MAP1LC3B	3751.5789
16	209218_at	SQLE	3795.4211
17	207156_at	HIST1H2AG	3804.6842
18	202842_s_at	DNAJB9	3849.2105
19	204014_at	DUSP4	3896.2632
20	200779_at	ATF4	3970.2105
21	203751_x_at	JUND	4034.4737
22	216038_x_at	DAXX	4034.7895
23	212286_at	ANKRD12	4061.3158
24	201625_s_at	INSIG1	4081.8421
25	211559_s_at	CCNG2	4088.3158
26	202540_s_at	HMGCR	4104.4211
27	201631_s_at	IER3	4128
28	201465_s_at	JUN	4231.1579
29	211162_x_at	SCD	4292.3684
30	211979_at	GPR107	4308.3158
31	213877_x_at	TCEB2	4366.0526
32	221750_at	HMGCS1	4420.6842
33	200831_s_at	SCD	4496.8947
34	217996_at	PHLDA1	4505
35	203752_s_at	JUND	4525.9474
36	218041_x_at	SLC38A2	4548.8421
37	202419_at	FVT1	4575.5263
38	206648_at	ZNF571	4587.5789
39	202820_at	AHR	4610
40	202558_s_at	STCH	4610.2105
41	203665_at	HMOX1	4635.5789
42	203726_s_at	LAMA3	4637.1053
43	218412_s_at	GTF2IRD1	4658.7368
44	208961_s_at	KLF6	4673.8947
45	205047_s_at	ASNS	4698.8947
46	217310_s_at	FOXJ3	4708.1579
47	207601_at	SULT1B1	4708.5263
48	219527_at	MOSC2	4753.2632
49	220219_s_at	LRRC37A	4753.5263
50	212274_at	LPIN1	4773.6842

Down_Rank	Probe	Gene	Avg_Rank
1	218918_at	MAN1C1	19454.9474
2	204039_at	CEBPA	18929.1579
3	202613_at	CTPS	18600.4737
4	203699_s_at	DIO2	18516.0526
5	200768_s_at	MAT2A	18492.2105
6	219800_s_at	—	18461.3158
7	220771_at	LOC51152	18347.7895
8	208502_s_at	PITX1	18346.7895
9	214266_s_at	PDLIM7	18293.9474
10	201667_at	GJA1	18242.5263
11	204553_x_at	INPP4A	18220.6842
12	218944_at	PYCRL	18143.7895
13	90265_at	CENTA1	18125.8947
14	217759_at	TRIM44	18059.8947
15	203122_at	TTC15	17967
16	208080_at	AURKA	17862.4211
17	205613_at	SYT17	17858.2105
18	204307_at	KIAA0329	17840
19	219200_at	FASTKD3	17798.6316
20	212797_at	SORT1	17795
21	222028_at	ZNF45	17711.5789
22	201565_s_at	ID2	17695.0526
23	221552_at	ABHD6	17648.7895
24	205136_s_at	NUFIP1	17615.3158
25	218653_at	SLC25A15	17614.1053
26	221440_s_at	RBBP9	17561.8947
27	205966_at	TAF13	17544.5789
28	208885_at	LCP1	17536
29	206832_s_at	SEMA3F	17512.8421
30	215629_s_at	DLEU2L	17499.7368
31	204544_at	HPS5	17472.5789
32	204284_at	PPP1R3C	17458.5789
33	209515_s_at	RAB27A	17443.5263
34	203078_at	CUL2	17418.6316
35	218544_s_at	RCL1	17400.5263
36	218489_s_at	ALAD	17367.1053
37	205652_s_at	TTLL1	17365.1579
38	207458_at	C8orf51	17354.8421
39	205034_at	CCNE2	17353.4211
40	202818_s_at	TCEB3	17327.5789
41	209187_at	DR1	17305.5789
42	201436_at	EIF4E	17284.4211
43	214113_s_at	RBM8A	17263.8947
44	219031_s_at	NIP7	17259.8421
45	210007_s_at	GPD2	17250.2105
46	212753_at	PCGF3	17241.3684
47	205185_at	SPINK5	17224.3158
48	218707_at	ZNF444	17217.6316
49	213132_s_at	MCAT	17192.7895
50	210932_s_at	RNF6	17191.8421

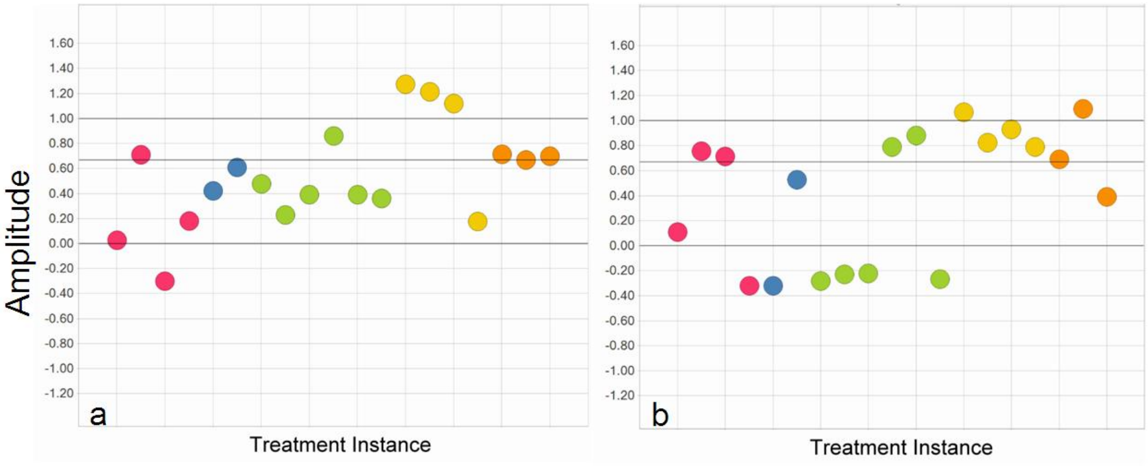


Figure 3.2: The amplitude values for a) the top probe found by the group profile method is from the BHLHB2 gene and b) the top probe by the fold-change method that is greater than 2. The lines correspond to a fold-change of 2 and 3, respectively. Colors represent compounds such that the first four pink dots correspond to the 4 chlorpromazine replicates, the next 2 are the clozapine replicates, etc. See Fig. 1 for order of compounds.

field has been to select probes that are commonly up-, or down-, expressed above a particular threshold. The most common thresholds used in the literature are fold-changes greater than or equal to either 2 or 3, which correspond to amplitude values of 0.67 and 1.0, respectively. We select the best probe from this alternative method, determining the probe that exhibits a fold-change greater than 2 in the most antipsychotic instances. The best probe found by this method was for the SEMA3B gene. The amplitude values across all of the antipsychotics for this probe are shown in Fig. 2B. Note that even though some of the individual instances have a very high amplitude value, roughly one-third of the instances have the opposite effect. Again, we determine the specificity of this probe to the antipsychotics by evaluating how it behaves across the rest of the database (Fig. 4). Visually, we can see that this probe is not specific to the antipsychotics at all, i.e., it up/down regulation is randomly distributed among all groups and not just those related to the antipsychotics.

As validation of our group profile method, we examine BDNF. BDNF (Brain-Derived Neurotrophic Factor) has long been a candidate gene for both schizophrenia and bipolar disorder [17, 14, 64]. Jiang et al. demonstrate that BHLHB2 regulates the BDNF transcription [23]. BHLHB2 has also recently been associated with bipolar disorder susceptibility [55]. These publications demonstrates how this method can give insight into the etiology of the disease that these drugs treat. While in this case the research community has already discovered and published this biological connection, there also will be other novel connections/signatures that will be found. It also demonstrates how the method extends beyond solely learning about the mechanism of action of drugs. More specifically, we have used the gene expression profile of antipsychotic drugs to learn of a mechanism that they share in common (regulating BHLHB2), which in turn has already been to be shown play a role in the underlying disease that these drugs are used to treat. Turning back to the best result from the alternative (fold-change threshold) method, there is no known link between SEMA3B and antipsychotics, schizophrenia, bipolar disorder or other topics expected to be related to antipsychotic agents and so we would treat this as a false positive without further evidence.

3.4 Understanding Group Signatures

As mentioned earlier, one of the major benefits of our group profile method is that we can easily plug our group profile results into many algorithms and tools developed to analyze (individual) gene expression data. The probe sets in the group profile signatures can be evaluated for significant overrepresentation of gene ontology (GO) terms, e.g., GO Biological Processes, using the L2L analysis tool [43]. Given a list of probe sets, e.g., DEGS, and a list to match them to, e.g. GO:BiolProc, L2L calculates the expected number of matches given the probes found on the microarray. From the actual and expected matches, an enrichment score and the corresponding P value for each GO term is then calculated [3]. Additional lists of published probe sets are also evaluated, including GO Cellular Component, GO Molecular Function, reactome protein-protein interactions [61], predicted human MicroRNA targets [24], and cancer gene expression modules [53].

We use the L2L method to evaluate the example group profile of the antipsychotics. The top 50 probes are evaluated for significant overrepresentation of GO Biological Process terms. The most significant terms are all related to lipid homeostasis (Table 2). There are five genes involved in the sterol biosynthetic process (GO:0016126) within the top 50 probes. Out of over 22,000 probes, only 41 are annotated as belonging to this GO term, so 0.11 probes for this term are expected by chance. This GO term, along with the next three in Table 2, pass Bonferroni correction for multiple testing ($p \leq 1.11E\text{-}05$ after correction for all four GO terms). The amplitude values for the five genes that are involved in this pathway are shown in

Table 3.2: The most significantly overrepresented GO Biological Process terms from the up-expressed antipsychotic signature.

GO Term	GO ID	Probes	Expected	Actual	Enrichment	P Value
sterol biosynthetic process	GO:0016126	41	0.11	5	44.73	1.04E-07*
steroid biosynthetic process	GO:0006694	88	0.24	5	20.84	4.89E-06*
alcohol metabolic process	GO:0006066	371	1.01	8	7.91	1.05E-05*
sterol metabolic process	GO:0016125	104	0.28	5	17.63	1.11E-05*
steroid metabolic process	GO:0008202	211	0.58	6	10.43	2.91E-05
cholesterol biosynthetic process	GO:0006695	31	0.08	3	35.50	8.60E-05
lipid biosynthetic process	GO:0008610	281	0.77	6	7.83	1.40E-04

Fig. 5. There is an obvious trend that the expression of these probes is increased in almost every antipsychotic instance in our database. However, even though they are always up-expressed, the amplitude value is normally below the common threshold used by other researchers (fold-change of 2 or 3). This is a good example of how the group profile method is able to detect consistent, and therefore more robust, signals in gene expression data; signals that are normally overlooked by current methods.

Support for these GO Biological Process findings comes from the work of other researchers aimed at understanding the molecular origin of the known metabolic side effects of antipsychotics that include increased weight gain and propensity to adiposity and insulin resistance [42]. Our observation is consistent with literature reports of an antipsychotic drug effect on the same or overlapping sets of genes involved in lipid homeostasis. Interestingly, a genome-wide screen of Saccharomyces cerevisiae heterozygotes had previously revealed that the antipsychotics haloperidol, chlorpromazine, and trifluoperazine had a strong effect on genes involved in yeast fatty acid biosynthesis (OLE1, the ortholog of the human SCD), sterol biosynthesis or phospholipid transport [37].

3.5 Querying with Group Signatures

The WIMRR method is able to create a specific representative profile for a group of gene expression profiles. We have demonstrated the ability to gain insight into the mechanism of action of a drug class (as well as the disease that it is used to treat) using WIMRR group profiles. Now we utilize the strength of a group profile to detect and predict the therapeutic use of a drug based on an individual gene expression profile.

We use the truncated KS statistics described previously for pairwise (instance-to-instance) similarity calculations [28] to detect instances that are similar to a group profile of interest (instance-to-group). Using the same antipsychotic group profile, we query the database of instances using k = 50 (i.e., the signature shown in Table 1). The instances most similar to this group profile are shown in Table 3, along with their KS score. The last column in Table 3 represents membership in the group of interest, i.e., if a given treatment is a member of the antipsychotic group used in creating the profile. Scanning the list, we see that prochlorperazine (Instance ID = 995) is the most similar non-antipsychotic drug. It turns out that prochlorperazine is in fact a phenothiazine antipsychotic; however, it is more commonly used for the treatment of nausea and vertigo. Prochlorperazine is a highly potent neuroleptic, which is considered a typical antipsychotic. The next non-antipsychotic is fluphenazine, for which two replicates show up as extremely similar to the antipsychotic profile. Fluphenazine is a typical antipsychotic drug used for the treatment of psychosis, e.g., schizophrenia and bipolar disorder. Fluphenazine is also an extremely potent phenothiazine. The next novel compound is calmidazolium, which is a calmodulin inhibitor. Though it is not used as an antipsychotic, it is validated because many of the antipsychotic drugs are potent inhibitors of calmodulin [13].

In fact, it turns out that many of the most significant results are already used as an antipsychotic agent even though they are not labeled in ChemBank as such. These examples are a validation of our method and increase the confidence in the other results that are not already supported by the literature, as these are potentially the important and still unknown alternative uses for these therapeutic agents.

3.6 Analysis of CMAP V2.0

We have introduced our method for creating group profiles from gene expression data. For this, we have used the original version of the CMAP dataset as our motivating example. We have seen how we can gain biological insight from these profiles as well as how to predict new members by querying the group signature. Here we present our analysis of the newly released CMAP 2.0 dataset with our method and describe the results. Groups are defined according to the compounds ATC code. We have analyzed all the groups at ATC level 3 and level 4. ATC level 3 defines the therapeutic/pharmacological subgroup, e.g., N05A = Antipsy-

Table 3.3: The database was queried with the antipsychotic signature (up and down together) and the most similar

Rank	Instance ID	Name	KS Score	Antipsychotic Member
1	1010	thioridazine[INN]	1.58	X
2	1068	thioridazine[INN]	1.483	X
3	1004	trifluoperazine[INN]	1.469	X
4	995	prochlorperazine[INN]	1.435	
5	910	trifluoperazine[INN]	1.408	X
6	417	thioridazine[INN]	1.387	X
7	983	haloperidol[INN]	1.352	X
8	1024	haloperidol[INN]	1.346	X
9	1017	fluphenazine[INN]	1.317	
10	1075	fluphenazine[INN]	1.293	
11	421	trifluoperazine[INN]	1.256	X
12	906	calmidazolium	1.223	
13	870	pyrvinium	1.209	
14	1053	prochlorperazine[INN]	1.201	
15	418	haloperidol[INN]	1.167	X
16	1009	clozapine[INN]	1.162	X
17	419	chlorpromazine[INN]	1.138	X
18	1003	nordihydroguaiareticacid	1.1	
19	416	clozapine[INN]	1.09	X
20	1105	monensin[INN]	1.077	
21	978	pyrvinium	1.065	
22	893	pararosaniline	1.051	
23	882	ionomycin	1.027	
24	941	rottlerin	1.023	
25	1012	troglitazone[INN]	1.018	
26	1082	haloperidol[INN]	1.009	X
27	1055	chlorpromazine[INN]	0.997	X
28	1041	haloperidol[INN]	0.992	X
29	997	chlorpromazine[INN]	0.99	X

chotics. ATC level 4 further defines a subgroup based on chemical properties, e.g., N05AE = Indole Derivative Antipsychotics. We focus on groups with three or more compounds, resulting in 117 ATC level 3 groups and 148 level 4 groups.

3.6.1 GEPedia.org

We have compiled all of the results from our analysis of CMAP 2.0 and have made them available online at GEPedia.org. In this chapter, we focus on evaluating our group profile method and only highlight a few interesting results from this analysis. We assume that there are many undiscovered biological insights within this dataset. We are releasing all of the data allowing researchers to examine the results for further discoveries and to compare with their own datasets.

Currently, the organization of GEPedia.org is based around the analysis presented in this chapter. We include the output of the complete analysis of all groups. For every group, i.e., for all ATC groups, we have made available a) the profile itself, including the up- and down-expressed signatures, b) the analysis of the profile according to the L2L tool, c) the sensitivity analysis of the profile, and d) the results of searching across the database with the signature. In the future, we plan to modify the website to allow more interactive analysis of the data in addition to allowing scientists to upload, analyze, and share their own gene expression data.

3.6.2 Sensitivity Analysis and Independent Validation

A sensitivity analysis is performed in order to prioritize the evaluation of the most promising group profiles. This sensitivity analysis also serves to demonstrate the robustness of the model to off target effects, i.e, changes in the gene expression profile due to factors that are not the focus of study, e.g., vehicle and batch effects, toxicology signatures, etc. Additionally, we can perform a similar sensitivity analysis using alternative methods and compare the results to obtain a better understanding of the robustness of our method across these off target effects compared to other methods that are currently used.

To perform the sensitivity analysis, we randomly divide the group into two equal-sized subgroups: a training group that contains half of the treatment instances from the group and a test group composed of the remainder of the group. A group profile is created for both subgroups, and the top (up-tags) and bottom (down-tags) 100 probes are selected. The number of probes in common between the two subgroups is calculated for both the up- and down-tags respectively. The treatment instances are re-randomized and this process is repeated for a total of 10 iterations. The average number of probes in common across the 10 iterations is calculated for the up- and down-tags. The higher the average number of probes in common (for the up-tags, down-tags, or both up- and down-tags), the more robust we consider the group profile. From this value, i.e., the average number of probes in common, we estimate the probability assuming a binomial distribution.

The most robust ATC level 3 (therapeutic/pharmacological) group profiles are shown in Table 4 for both the up and down signatures together (full results in Supplemental Table 1 and Supplemental Table 2 for the up and down signatures, respectively). The full results for the level 4 ATC (chemical/therapeutic/pharmacological) group profiles for the up and down signatures are shown in Supplemental Table 3 and Supplemental Table 4, respectively. The associated probability for each of these profiles is also listed. The observed probabilities indicate that some of these profiles are not random. Corrections for multiple testing are performed, and the Bonferroni-corrected P values are also included in each of the tables.

At the onset of this chapter, we mention that we are interested in creating a gene expression profile for groups sharing a therapeutic use, and so we focus our analysis on the ATC level 3 groups. There are 36 groups with significant (Bonferroni-corrected $P^* < 0.05$) up-expressed signatures and 28 for the down-expressed signatures. Out of these groups, 25 groups are robust for both up- and down-expressed signatures. While a robust up- or down-expressed signature can independently give novel insight into the underlying shared biological function of a group, we focus on groups that are significant for both because we also want to use these profiles to help predict novel uses of the drugs in our database. The similarity metric that we have adopted requires both the up and down signatures to be used together. We now present a deeper analysis of the most robust profiles. The larger the set of unique drugs that compose a group, the more evidence we have that the therapeutic mechanism is what is being detected in the profile. For this reason, we focus on the significant groups with the largest number of unique drugs. We compare our results to those from an independent dataset using the same method (Table 5).

Table 3.4: The most robust group profiles across the whole database are presented here.

Group	Drugs	Up	P Up*	Down	P Down*	Label
N05A	28	70.6	3.09E-139	49.4	9.59E-86	Antipsychotics
R06A	27	23.7	1.07E-31	12	5.70E-12	Antihistamines for Systemic Use
N06A	25	29.6	5.70E-43	12.1	4.04E-12	Antidepressants
D07A	19	49.8	1.11E-86	19.7	1.64E-24	Corticosteroids, Plain
G01A	18	12.1	4.04E-12	6.5	1.98E-04	Antiinfectives and Antiseptics
D01A	16	10.2	2.42E-09	11.7	1.59E-11	Antifungals for Topical Use
S01B	16	7.9	3.36E-06	8.9	1.56E-07	Antiinflammatory Agents
N03A	11	13.1	1.22E-13	17.8	3.01E-21	Antiepileptics
H02A	11	18.5	1.94E-22	5.2	6.72E-03	Corticosteroids for Systemic Use
R03B	10	15.6	1.35E-17	4.6	3.09E-02	Drugs for Obstructive Airway Diseases, Inhalents
D10A	9	18.7	8.80E-23	6.5	1.98E-04	Anti-Acne Preparations (Topical)
L04A	8	39.9	2.46E-64	31.4	1.47E-46	Immunosuppressants
D07X	8	25.3	1.12E-34	6.7	1.13E-04	Corticosteroids, (Dermatologicals)
G03D	8	11.1	1.22E-10	4.7	2.41E-02	Progestogens
L01X	7	19.9	7.31E-25	11.5	3.16E-11	Other Antineoplastic Agents
L02B	6	19.3	8.12E-24	13.5	2.93E-14	Hormone Antagonists (and related)
R03A	6	9.8	8.89E-09	5.3	5.17E-03	Adrenergics, Inhalents
C08C	6	5.6	2.34E-03	4.5	3.95E-02	Selective Calcium Channel Blockers
G03C	5	30.5	9.35E-45	10.1	3.36E-09	Estrogens
S01C	5	10.3	1.74E-09	5.3	5.17E-03	Anti-inflammatory -infective (Combo)
C08E	4	11.7	1.59E-11	7.3	1.99E-05	Non-selective Calcium Channel Blockers
C01A	3	61.1	2.94E-114	62.2	4.60E-117	Cardiac Glycosides
L01D	3	6.4	2.62E-04	22.9	3.16E-30	Cytotoxic Antibiotics (and related
L01B	3	7.9	3.36E-06	19.8	1.09E-24	Antimetabolites

3.6.3 Antipsychotic Group (N05A)

We start our analysis with the largest group that meets our significance threshold: the antipsychotic group with 28 unique drugs. The ATC level 3 code for this group is N05A. The antipsychotic profile is the most robust result from the ATC level 3 groups when evaluating the up-expressed signature (Bonferroni-corrected P value: P*=3.10E-139). This corresponds to an average of 70.6 probes that are shared between the top 100 probes of two random subgroups. Interestingly, this same group is the second most significant when evaluating the robustness of the down-expressed signature (P*=9.59E-86; Average probes in common = 49.4). In an attempt to discover what the underlying shared biological process is within these antipsychotic agents, we turn to the L2L analysis. The most overrepresented GO Biological Process term is Sterol Biosynthetic Process (GO:0016126; P*=6.45E-20). This is the same term that was found over-expressed within the smaller pilot study and demonstrates that our group profile method can detect the true signature with a small set of samples.

We have the ability to compare this profile with the antipsychotic profile recently published by Polymeropoulos et al. [47]. It is important to note that these two profiles were created by two independent laboratories, with different cell lines and with a different, but overlapping, set of antipsychotics. These two profiles are very similar, and they share 34 probes in common among their top 100 probes (P=6.42E-54). The most significant GO Biological Process term from the Polymeropoulos et al. antipsychotic group profile is Lipid Biosynthesis. Given the significant overlap of the profiles, it is not surprising that this term is actually a grandparent of Sterol Biosynthetic Process (connected through the GO term Steroid Biosynthetic Process). The GO term Lipid Biosynthesis is also highly significant within the CMAP v2.0 antipsychotic group (P*=2.70E-13).

The down-expressed signatures also share several probes in common (Probes=6; P=6.79E-06). The GO Biological Process analysis points to a significant down-regulation of the DNA regulation process (GO:0006260; P*=3.61E-07). Barochovsky et al. have demonstrated in vivo that compounds acting on the central nervous system, specifically those that affect noradrenergic, dopaminergic, and serotoninergic neurotransmitters, reduce brain cell replication [4]. This observation of compounds acting on the CNS was a dose-dependent effect and was seen for both agonists and antagonists. This down-expressed signature, like the up-expressed signature, is well supported by the literature. The antipsychotic profile that we have discovered is robust, both in and across datasets. Furthermore, we have demonstrated the ability of our group profile method to give biological insights into the potentially unknown shared biological process exhibited by a group of drugs.

In an attempt to put these results into perspective we also set out to analyze this same data with one of the more common methods for detecting expressed genes. The LIMMA package (Linear Models for Microarray Data), is an R package that is part of Bioconductor[57]. We followed the standard processing and linear model fitting provided in the examples of the LIMMA documentation. To do this we were forced to only analyze one array type at a time, so we selected the most common array type (HT-HG-U133A) that was used in the CMAP dataset. We performed the

Table 3.5: The most robust profiles were evaluated against an independent dataset (Polymeropoulos et al).

Group	Polymeropoulos et al., PDR Group	Probes In Common	P
N05A	CNS:Antipsychotics	34	6.42E-54
R06A	Resipiratory Agent:Histamine Antagonist	4	1.13E-03
N06A	CNS:Antidepressants	15	1.07E-18
D07A	Dermatological:Corticosteroids	30	7.88E-46

same sensitivity analysis in which we randomly sample the group into two subgroups and determine the number of probes that overlap in the top 100 results. As before, this process is repeated 10 times. The average number of probes in common is 53.6 (compared to 70.6 for our WIMRR method). A t-test shows that this difference is statistically significant (P=5.57E-7). Interestingly, the L2L analysis performed on these top 100 probes points to the same Sterol Biosynthetic Process signature that was demonstrated before as being the most representative, but at much lower confidence (P*=0.0002 compared to P*=6.45E-20 for WIMRR).

3.6.4 Antihistamine Group (R06A)

The second-largest group that meets our significance criteria is the antihistamines (full annotation: Antihistamines for Systemic Use; ATC Code: R06A). This group contains 27 unique drugs. The sensitivity analysis reveals 23.7 probes on average shared within the up-expressed signature and 12 for the down-expressed (P*=1.07E-31 and P*=5.70E-12, respectively). The up-expressed signature exhibits a common underlying theme related to negative regulation of I-kappaB kinase / NF-kappaB cascade (GO:0043124; P=6.08E-05). This GO signature is not as strong as some of the other profiles and is not significant when corrected for multiple testing. However, it is interesting to note that this signature is consistent with the known effect of antihistamines on NF-kappaB. Roumestan et al. have shown that antihistamines inhibit NF-kappaB through both H1 receptor-dependent and independent mechanisms [51]. This profile does not replicate when compared to the equivalent group (Respiratory Agent: Histamine Antagonist) from the dataset presented by Polymeropoulos et al., though a similar trend is seen. The average number of probes in common is four and one respectively, for the up- and down-expressed signatures (P = 1.13E-03 and P = 3.60E-01).

3.6.5 Antidepressant Group (N06A)

Next, we discuss the third-largest group: the antidepressants (ATC Code: N06A). There are 25 unique drugs within this group. The sensitivity analysis results in an average of 29.6 and 12.1 probes in common for the up- and down-expressed signatures (P*=5.70E-43 and P*=4.04E-12, respectively). Evaluating the up-expressed signature, the most overrepresented GO Biological Process term is Sterol Biosynthetic Process (GO:0016126; P*=1.19E-09). This is the same core mechanism seen within the antipsychotic group, but this signature is seen on a smaller scale. Polymeropoulos et al. demonstrated the same relationship between the expression profile of antipsychotic and antidepressant drugs [47]. When we compare our antidepressant profile to the antidepressant profile from the dataset from Polymeropoulos et al., we find 15 probes in common (P=1.07E-18). The down-expressed signature does not reproduce within the Polymeropoulos et. al. dataset, sharing only one probe in common.

3.6.6 Corticosteroid Group (D07A)

The last group that we evaluate in depth is the corticosteroids (N=19; ATC Code: D07A). This profile is also robust according to the sensitivity analysis. The average number of probes in common for the up-expressed signature is 49.8 (P*=1.11E-86). The down-expressed signature has an average of 19.7 probes in common (P*=1.64E-24). Individually, the up- and down-expressed signatures do not exhibit a significant result for any GO Biological Process, but evaluated together they demonstrate an effect on the regulation of the interleukin-6 biosynthetic process (P*=1.38E-02). Corticosteroids are involved in a wide range of physiological systems such as stress response, immune response and regulation of inflammation. Interleukin-6 acts as both a pro-inflammatory and anti-inflammatory cytokine that can be secreted to stimulate response to trauma [18]. There is a significant overlap between this profile and the corresponding profile (Dermatological: Corticosteroids) from Polymeropoulos et al. The up-expressed signatures share 30 probes in common while the down-expressed share nine probes, corresponding to probabilities of P=7.88E-46 and 9.72E-10, respectively.

3.7 Discussion

We have introduced and evaluated our method for creating group profiles from gene expression data. The ability to have reproducible sets of differentially expressed genes from microarray experiments has been a big challenge, and we have demonstrated how our method is able to overcome this obstacle. Furthermore, we have illustrated how to gain biological insight from such group profiles as well as the ability to use them as a signature to query a database. In our example domain of a drug discovery system, this biological insight allows researchers to potentially learn about the etiology of the disease that these compounds are being used to treat and gives them a predictive tool to find novel uses for other drugs.

Though a major focus of this work has been to introduce our method and validate it across independent datasets, we are also releasing all group profiles from the full CMAP 2.0. This includes all corresponding meta-analysis that has been performed: L2L analysis, similarity searching results, etc. We feel that this resource contains a lot of hidden biological insight into many groups of drugs and their target diseases, and we are releasing it for further in-depth research. Another contribution of this work is the independent validation of the common effect of antipsychotics on the biosynthesis and regulation of fatty acids and cholesterol, which supports a key role of lipid homeostasis in schizophrenia.

There are many possible avenues of further improvements and research. Thus far, we have assumed that explicit groups are given a priori. Our sensitivity analysis validates how coherent a group is; however, it does not dictate what to do if the outcome is not positive. For example, a leave-one-out analysis can be done to exclude members that do not fit well within a group. Lastly, it is important to note that our method is focused on determining a reproducible genetic profile for a group of samples; in this case, drugs of a particular class. We provide no guarantee as to the uniqueness of such profiles and instead claim that these profiles can be used to compare groups. We have kept the full ranked list as the profile, and so it is straightforward for extensions to this method to be developed to further refine and learn what genetic components make up a more unique signature if that was the end goal. In keeping the full profile, i.e., the re-ranked list of probesets, we allow further research methods, which are developed for individual expression profiles, e.g., the L2L method, to also be applicable to our group profiles.

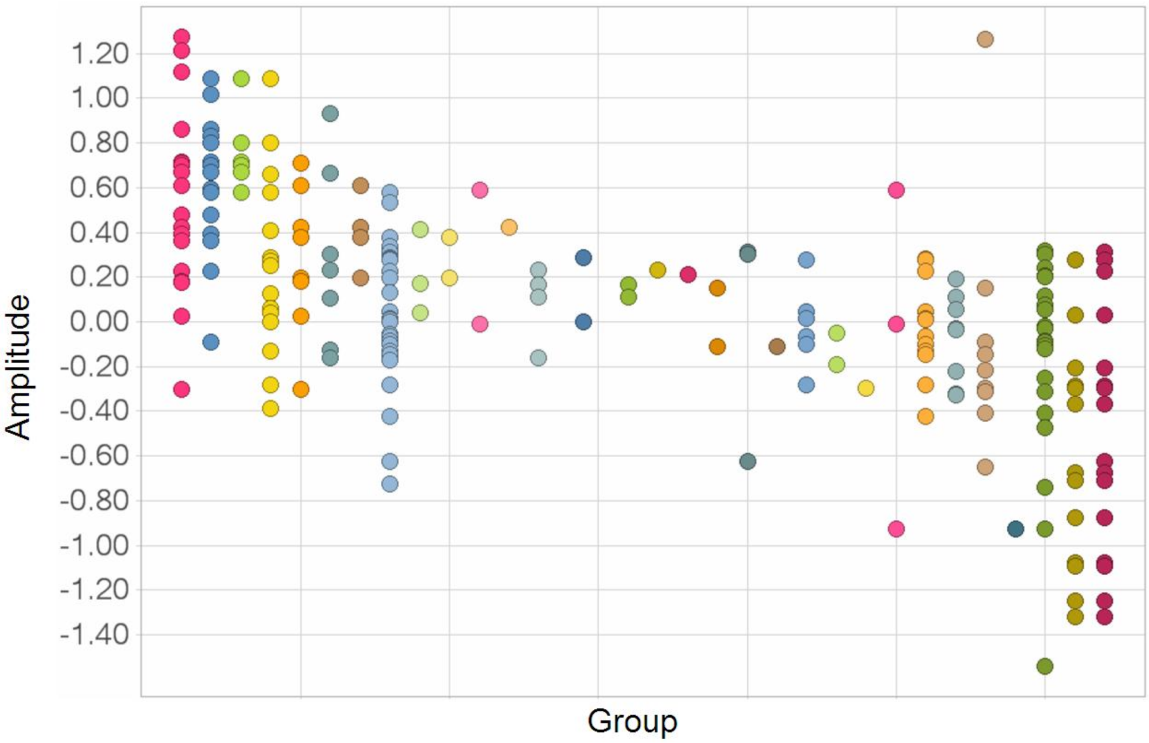


Figure 3.3: Specificity of top probe, BHLHB2, from the group profile method. Each vertical set of points corresponds to a different group in the database. Here we only describe the most similar, and point out that they share some drugs in common: from left to right, the first group is the antipsychotics, the second is the tranquilizers (includes prochlorperazine, fluphenazine, and trifluoperazine), the third group is antiemetics (includes prochlorpromazine and trifluoperazine), and the fourth group is the antineoplastics (includes prochlorpromazine). This probe is specific to the antipsychotic and similar groups, i.e. high amplitude on in antipsychotic and groups sharing properties with antipsychotic agents, and lowe amplitude in other groups.

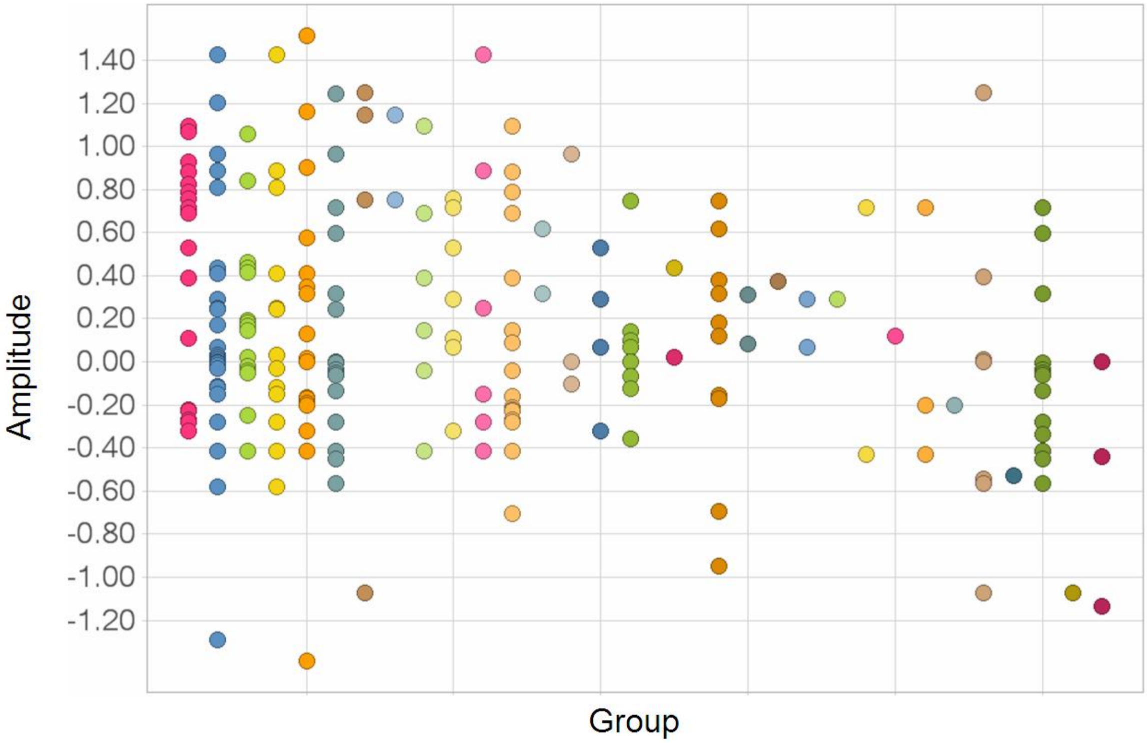


Figure 3.4: The top probe from the fold-change greater than 2 method is not specific to antipsychotics. There is nothing in common among the first couple of groups (sorted by the average score of the group). The first group is the antipsychotics, the second is anti-inflammatory, the third is antineoplastics and the fourth is analgesics. The amplitude values are scattered and show no consistent pattern.

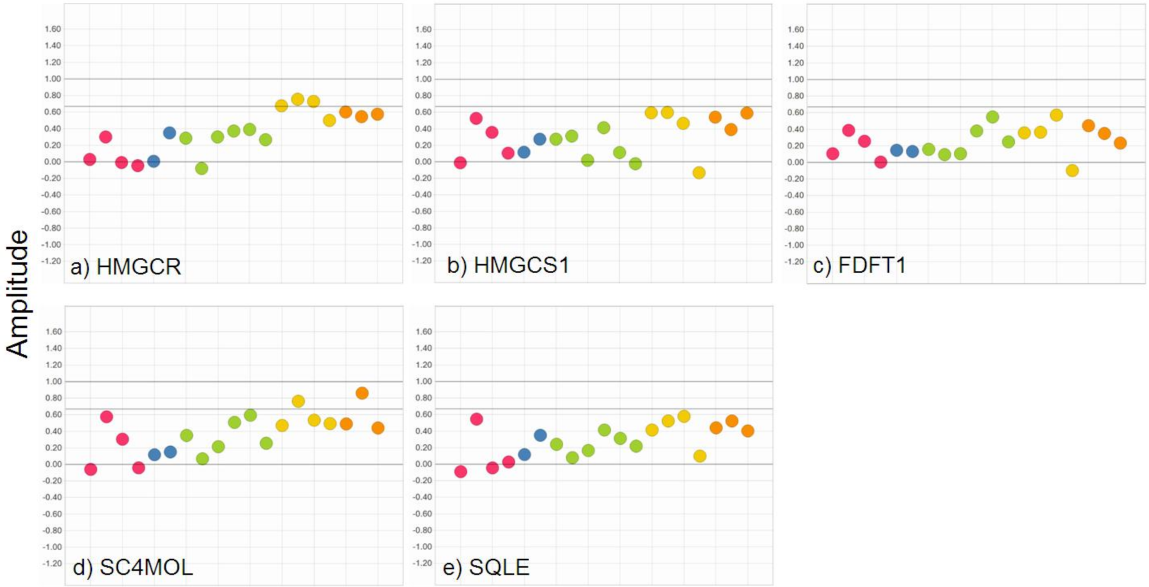


Figure 3.5: The amplitude values for the probes in the most significantly up-expressed GO term for the antipsychotic group: sterol biosynthetic process. The probes correspond to the a) HMGCR, b) HMGCS1, c) FDFT1, d) SC4MOL, and e) SQLE genes. Replicates are designated by the same color.

Chapter 4

Confounder Correction by Profile Subtraction

Since the onset of this work we have been focused on gaining biological insight from gene expression data. In Chapter 2, we focused on the challenge of dealing with off-target effects, e.g., vehicle and batch effects. We concentrated on this challenge applied to calculating reliable and accurate pairwise similarity among gene expression profiles. While pairwise similarity metrics are useful the information gained is somewhat limited. In Chapter 3, the focus was understanding the core gene expression signature shared among a group of profiles. While we demonstrated that our novel method for group profile creation yields informative and reproducible results, we know from past experience that we can improve upon this further by directly tackling the challenge of dealing with confounding effects in this new context of group profiles.

Controlling and correcting batch and other confounding effects is of utmost importance for robust inferences and interpretation of high throughput experiments, including gene expression microarrays [30]. A number of methods have been proposed to do this, and are, for the most part, based on the linear modeling of gene expression as a function of observed biological and technical effects [10, 25, 31, 35]. However, recent work has shown the power of non-parametric methods that estimate gene expression profiles as two-sided ranked lists, including our previously introduced method to estimate profiles that are representative of differential expression in an experimental group of interest (e.g., treatment vs. control, or disease vs. non-diseased). Here we propose an extension to our non-parametric gene expression profile method to correct for observed confounding effects, we refer to our method as profile subtraction. This correction is performed on ranked lists directly and provides a robust alternative to parametric batch profile correction methods.

In this chapter we introduce the concept of confounding groups and we describe our method to remove/subtract out confounding group effects. We compare our approach to linear models. We present an illustrative example of subtraction/removal of confounding group profiles on a small dataset. We evaluate our profile subtraction method on two real world datasets: an Arabidopsis Hormone dataset as well as a dataset consisting of Acute myeloid leukemia (AML) and Acute lympboblastic leukemia (ALL) samples. We compare our proposed method with our uncorrected group profile method and to other current methods including Limma[57], Combat[25] and surrogate variable analysis (SVA)[31]. We show that our profile subtraction method yields more specific and robust group profiles as compared to other these methods. Additionally, we create a group profile generator which has been used to more closely control and evaluate the robustness of our methods.

4.1 Confounding Group Profiles

The goal of gene expression experiments can vary widely but at the core they have a common goal: the detection of a genetic signature in common within a group of interest. This group may be a given treatment compared to controls, a subtype of a disease that has a poor prognosis or not, infected versus healthy individuals, or a profile of a patient who is more likely to respond to a given treatment. A gene expression experiment is designed to evaluate any of these scientific questions, and the labels for the different target groups that we are interested in studying would be known.

However, we recognize that there are often other confounding groupings present in the data. Some of these confounding groups are dictated by the experimental setup and are annotated in the dataset, e.g., vehicle and batch effects. Others possible confounding groups may not be annotated and relate to the samples themselves, e.g., a shared mechanism of action or side effect profile of a drug that is not directly correlated with the therapeutic indication. These confounding groups are non-orthogonal to the target groups of interest and they are not independent.

If the gene expression profiles are not corrected for these confounding groups then the target group profiles that are created are at risk of being incorrect and misleading. Each confounding group contributes to the overall noise in the data resulting in less confidence in the genetic signature discovered. In this chapter, we propose a method to remove the effects of a confounding group from a given gene expression profile.

4.2 Profile Subtraction Method

The goal of our profile subtraction method is to remove the effects of confounding groups. This is a problem that has been the focus of a lot of work and there are several methods that have been developed that have aimed to deal with this task, including using linear models (Limma[57]) and an empirical Bayes framework (Combat[25]). Our goal is to develop a method that can remove confounding effects from nonparametric ranked lists, as compared to these methods that are designed to work on actual expression values. We now introduce our profile subtraction method. In the next section we will provide a comparison of our profile subtraction method to linear models.

Assume we have n gene expression experiments (samples) X_1,X_2,$\ldots$,X_n. For each of the n profiles in the database, there is general information about the experimental conditions of the sample encoded by membership in one or more of groups G_1,G_2,$\ldots$,G_g. We use a binary function, member(X, G), that returns 1 if profile X is a member of group G and returns 0 otherwise to indicate group membership. For each experiment we denote expression data from the microarray itself as follows: each microarray X consists of a collection of probe sets, probes(X); for each probe p in probes(X), there is an absolute expression value EV(p), as well as an amplitude Amp(p) (the difference in expression relative to a control). The control is a reference baseline that is the average expression value calculated from multiple untreated samples run within the same experiment conditions, e.g., the same vehicle and batch.

Rather than measuring the absolute similarity in expression levels, we compare the *ranking* of the probes. We use $rank(p, probes(X))$, or $rank(p, X)$ for short, to denote the rank of probe p in profile X. The size of the group is equal to the number of samples that are members of the group (Equation 4.1). To define a group profile, we rank each probe p in probes(G) by its average rank across all members of the group G (Equation 4.2):

$$size(G) = \sum_{i=1}^{n} member(X, G) \tag{4.1}$$

$$avgRank(p, G) = \frac{\sum_{i=1}^{n} rank(p, X_i) \times member(X_i, G)}{sizeG} \tag{4.2}$$

Given this set of average ranks across the members of a particular group, the probes are now re-ranked according to how consistently they are up- or down-expressed across the group. We define Profile(G) as the probes in probes(G) sorted by their average rank across all members of the group (Equation 4.3).

$$Profile(G) = \{(p_1, avgRank(p_1, G)), (p_2, avgRank(p_2, G)), \ldots, (p_m, avgRank(p_m, G))\} \tag{4.3}$$

In the confounder correction setting we assume there are c confounding groups Z_1, Z_2,...,Z_c. Our goal is to subtract the effect of each of the confounding groups from $rank(p, X)$ for experimental groups of interest. The main idea is to use an important property of the two-sided ranked lists to subtract out these confounding effects by inverting probe ranks in each sample according to the confounding group ranking. We define an inverted rank as described in Equation 4.4. Our profile subtraction method for ranked lists is then defined by Equation 4.5, which describes how to calculate the corrected average ranks.

$$invRank(p, Z) = maxRank(Z) - rank(p, Z) + 1 \tag{4.4}$$

$$\begin{aligned} avgRank(p, X^*) =& [1 - member(X, Z)] * rank(p, X)) \\ & + member(X, Z) \times \frac{(\omega_G \times rank(p, X)) + [(1 - \omega_g) \times invRank(p, Z)]}{2} \end{aligned} \tag{4.5}$$

$$Profile(X^*) = \{(p_1, avgRank(p_1, X^*)), (p_2, avgRank(p_2, X^*)), \ldots, (p_m, avgRank(p_m, X^*))\} \tag{4.6}$$

For each individual expression profile, the effects of each confounding group of which the individual is a member of are removed. This results in corrected profiles for each X, that we denote X^*. The corrected datasets can now be used to create the target group profiles that we are interested in studying. Rank inversion forces probes in the corrected profile to the middle (uninformative portion) of the ranked list that behave the same in a given profile as they are expected to behave based on its membership in a confounding group. Probes that behave differently than expected are forced to move up/down and become more significantly represented in the profile. We have also introduced a weighting term, ω_G, that allows for a weight to be assigned to the original profile when removing the confounding effect (the confounding effects therefore get a weight of $(1 - \omega_G)$. This weighting can be selected from a combination of domain knowledge and by parameter tuning based on labeled data.

4.2.1 Comparison to Linear Models

Linear models are the most commonly used statistical tool to determine differential expression in experimental groups of interest. Here the outcome (y) is modeled as the relationship between a set of fixed effects (x) and a set of learned coefficients (b_1). Additionally an intercept (b_0) and error (e) is also modeled to complete the equation (Equation 4.7). In the case of microarray data, a linear model is fit for each probe on the microarray and a design matrix is used in place of x. The design matrix encodes the classes that each microarray belongs to with each row corresponding to a microarray and each column to the coefficients to be estimated.

$$y = b_0 + b_1 x + e \tag{4.7}$$

A t-statistic can be constructed from the mean effect (b_1) and its standard error estimated by the linear model described above and used as a measure of differential expression. However, standard error estimates in microarray data can be unreliable and empirical Bayes methods are commonly used in order to provide more robust results. It uses a hierarchical model to provide robust variance estimates, pooling measurements across genes/probes to construct a moderated t-statistic from the posterior variance estimated. The hope is that empirical Bayes methods can have better estimates by borrowing information across genes and experimental conditions. (Johnson et al.). Frequently, lists of differentially expressed genes are ranked by this moderated t-statistic.

Observed confounding effects can be included directly in the linear model above. The design matrix can be expanded to reveal the relationship between how a probe behaves in a given target group (b_1) and how it behaves in a given confounding group (b_2) (Equation 4.8)). More specifically, if a probe is consistently upregulated in a particular confounding group(therefore having a large b_2) then the coefficient modeled for the target group will not be increased, as this would otherwise lead to an overestimate of y.

$$y = b_0 + b_1 x_1 + b_2 x_2 + e \tag{4.8}$$

Building on the notion of better estimates of how genes behave by using empirical Bayes methods, Johnson et al., developed Combat to attempt to remove batch/confounding effects in microarray experiments. Unlike the generic linear models/Limma approach, Combat uses empirical Bayes to specifically remove batch effects prior to modeling any target group effects. After the direct removal of these confounding group effects is performed, the signatures of the target groups are discovered by again using a linear model approach.

At a high level our profile subtraction method (Equation 4.5) also has a linear form. Referring back to Equation 4.8, the premise of modeling target groups and confounding groups at once is that the effects among these two groups is additive. In other words, if a confounding group fully explains the final outcome variable of how a probe behaves, then the weight of that probe in the target group profile should be minimal. Similarly, if a probe in a given sample is behaving in the opposite direction than expected based on the confounding groups that it is a member of, then we would like to model an even larger effect based on the target group to counter this. Again, this is analogous to what happens in the linear model but unlike linear models we do not have actual values to just add or subtract, instead we have developed a method to work with the ranked list of probes.

As described earlier, we make use of an important property of the two-sided relative ranked lists that are being used to represent our gene expression profiles. Specifically, that the opposite sides of the ranked list have opposite meanings and that the middle of the list is considered uninformative. Combining the goal that we wish to use the prior information of how a probe/gene may behave in a confounding group with the fact that we are dealing with two-sided relative ranked lists, we are left with a simple solution. We can model the relationship of the ranks with the same general form of the equation from Equation 4.8, but instead using ranks and adding in a weighting factor (Equation 4.9). We can also remove the intercept value, as it is meaningless when dealing with nonparametric ranks.

$$rank(y) = w_1 rank(b_1) x_1 + w_2 rank(b_2) x_2 \tag{4.9}$$

As opposed to estimating the theoretical rankings of our target and confounding groups that best fit our data, we instead directly calculate the most likely ranks based on our data.

4.2.2 Example Profile Subtraction Method

We present a simple example to demonstrate how our profile subtraction method works. Let us assume that there is a gene expression profile of a diabetes drug and that it was run in batch 1. In order to determine what, if any, biases may be introduced by batch 1, we create a group profile of all of the gene expression profiles from batch 1. Any probes that are significantly up-expressed in batch 1 will appear at the top of the list. The stronger and more consistent this signature is across all batch 1 samples, the higher up the list it well be. Analogously, any probes that are significantly down-expressed in batch 1 would be present in the bottom of the ranked list. Probes that remain unchanged are in the middle (non-informative section) of the ranked list.

Referring back to the original diabetes sample that we are interested in studying: in order to determine if a given probe that is highly ranked (up-expressed) may be caused by an anti-diabetic signature we compare the rank in the sample profile to what was expected based on our prior knowledge that this sample is from batch 1. If this probe was also highly ranked in the batch 1 group profile then it is behaving exactly as expected and is not being shifted by the anti-diabetic properties. The probe should then be shifted down towards the middle (uninformative section) of the ranked list. If the probe was unchanged (middle of the list) in the batch 1 group profile then we do not want to make any significant change to the probe in the individual profile. Alternatively, if this probe that is highly up-expressed in the individual profile was actually down-expressed (bottom of the list) in the batch 1 group profile then the probe being unregulated should be treated as being even more substantial than the previous change. The exact mechanism by which the ranks in the individual sample are updated to reflect where they were expected to rank based on the confounding profiles is done as described in the prior section. This profile can now be passed onto further downstream analysis without worrying about the batch effects masking the true signature of the target groups of interest.

4.2.3 Profile Subtraction in Simple Dataset

In order to clearly illustrate the concept of profile subtraction we make use of a simple hypothetical dataset and present a visual summary of the method. A graphical summary of this example gene expression dataset is shown in Figure 4.1. There are two overlapping sets of groups identified by the colors at the bottom of the figure. There are three groups of interest (Y:yellow, B:blue and G:green) and also three confounding groups (1:white, 2:gray, 3:black). The confounding groups are potentially masking the true signature of the groups of interest. Let us assume that we are interested in finding a genetic signature of drugs used to treat diabetes, and that these drugs are represented as the yellow group. Our goal is to remove the effects of the three labelled, confounding groups such that the discovered diabetes group signature is more robust and is more likely to be the true signature.

Referring back to Figure 4.1, in this example dataset, the sample names at the top identify the group membership, e.g., sample Y1 belongs to the yellow group as well as to the white confounding group. Each vertical, gray rectangle represents the full ranked listed created from the gene expression profile of a given sample. The top of the rectangle corresponds to the low ranks (most up-expressed probes), the bottom section corresponds to the highest ranks (most down-expressed probes), and the middle corresponds to the probes without a substantial change in either direction (no or minimal change in expression as compared to control). For simplicity, we focus on a small subset of probes (the colored lines within each rectangle) from the microarray experiment. The location of these different colors represent that probes ranks in the individual gene expression profiles.

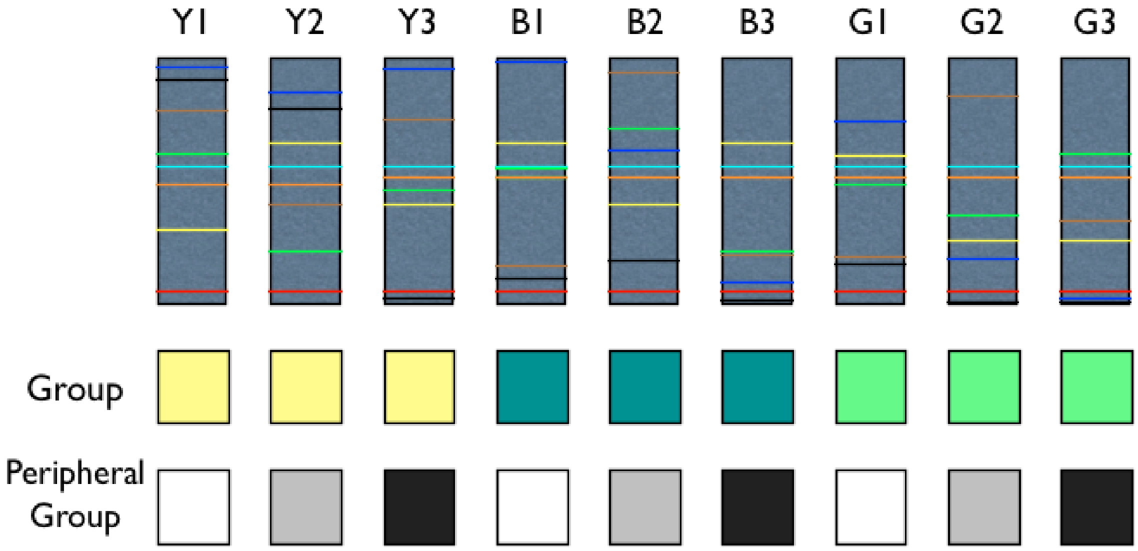


Figure 4.1: Example with target groups of interest (Yellow, Green, Blue) and overlapping confounding groups (White, Gray, Black).

Creating a group profile with WIMRR, introduced in Chapter 3, yields a group profile with the characteristics shown in Figure 4.2. In Figure 4.1 we notice that the

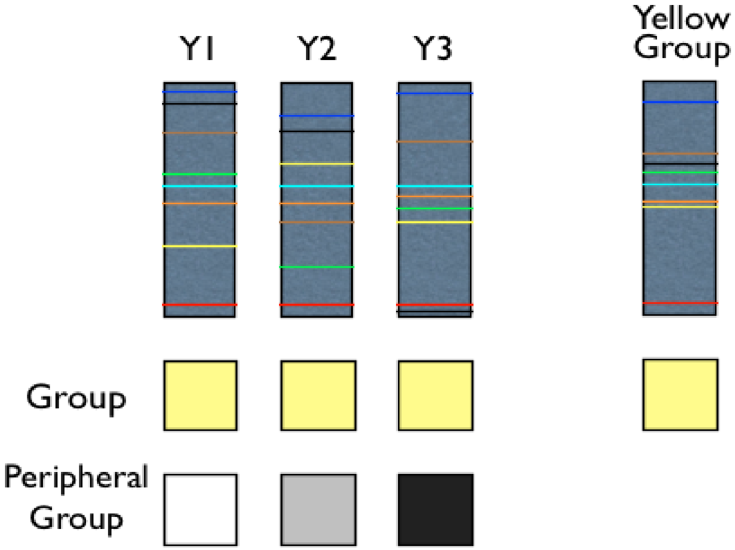


Figure 4.2: Group profile created without correcting for the confounding group effects.

conclusion drawn from some of the probes in this uncorrected group profile method could be biased based on how they behave in other (confounding) groups as well as across the dataset as a whole. For example, a simple correction that our ideal profile subtraction method should make is that since the red probe is consistently at the bottom of all of the individual gene expression profiles so its movement to this same position in the yellow group is not informative. Another interesting case is the black probe. The black probe appears up-expressed in the yellow group, except in Y3 (Y=Yellow:3=Black) in which it appears down-expressed. However, the black probe is consistently down-expressed within all members of the black confounding group. With this prior information about how the probes in the black confounding group behave, we correct the conclusion of how these probes are behaving within Y3. Even though the black probe in Y3 appears down-expressed without any prior information, it is in fact up-expressed when compared to the expected behavior based on the confounding effects.

The creation of the white, gray and black confounding profiles is shown in Figure 4.3. As explained in Section 4.2, the white profile is inverted and we refer to this inverted profile as the white' profile ($rank(p, X^*) = invRank(p, X)$), the gray profile is inverted and becomes the gray' profile, and the black profile is inverted and becomes the black' profile. In order to discover the true signature of the yellow group, we remove the masking effects of the three confounding group profiles. These confounding group profile effects are removed by calculating the average rank of each probe between the original profiles Y1 and white', Y2 with gray' and Y3 with black' which, when the probes are ordered by their average rank scores yields Y1*, Y2*,and Y3* (the profiles of Y1,Y2, and Y3 after removing the confounding group effects) Figure 4.4. From this set of clean profiles Y1*, Y2*,and Y3*, the yellow* group profile is then created (the yellow group profile after removing confounding group effects). The difference between this yellow* profile and the original yellow group profile created without first removing the confounding effects is shown in Figure 4.5.

In the simple example above we assume that groups of interest and the confounding groups have the same expression strength. In the actual implementation we introduced the ability to assign a weighting factor to remove confounding group signatures with varying strengths from the dataset where ω_g is the weight of the group. The optimal weight can be discovered by analyzing the dataset if the group memberships are fully labelled group as in the case of our mock dataset. If the full group memberships are not know, then the weight could be defined by a domain expert or estimated based on the data in the case of either partially or fully missing group labels.

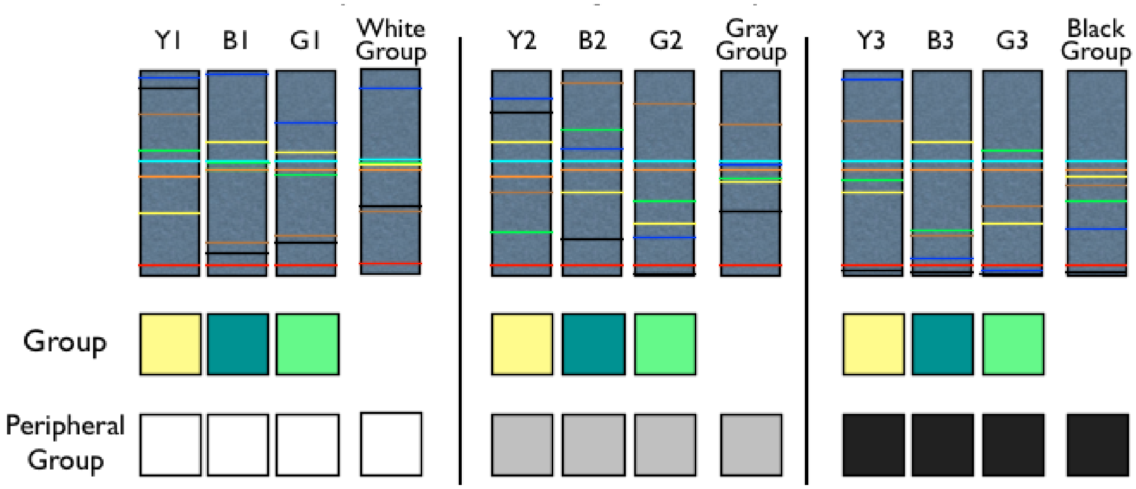


Figure 4.3: Creation of group profiles for the confounding groups (White, Gray, Black).

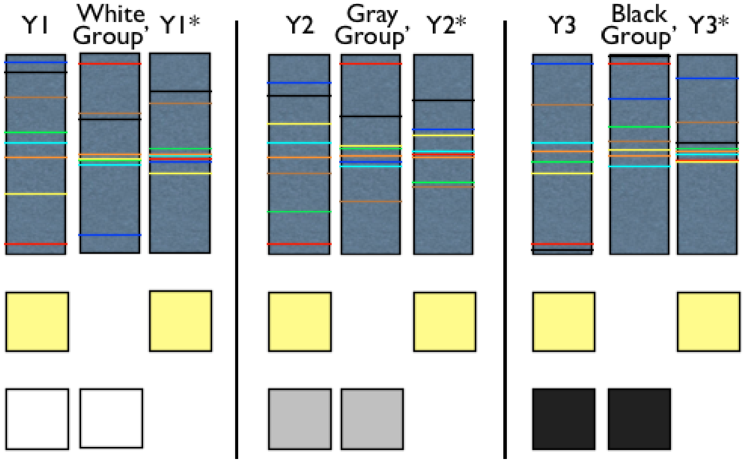


Figure 4.4: Removal of the confounding group effects from each of the members of the yellow group.

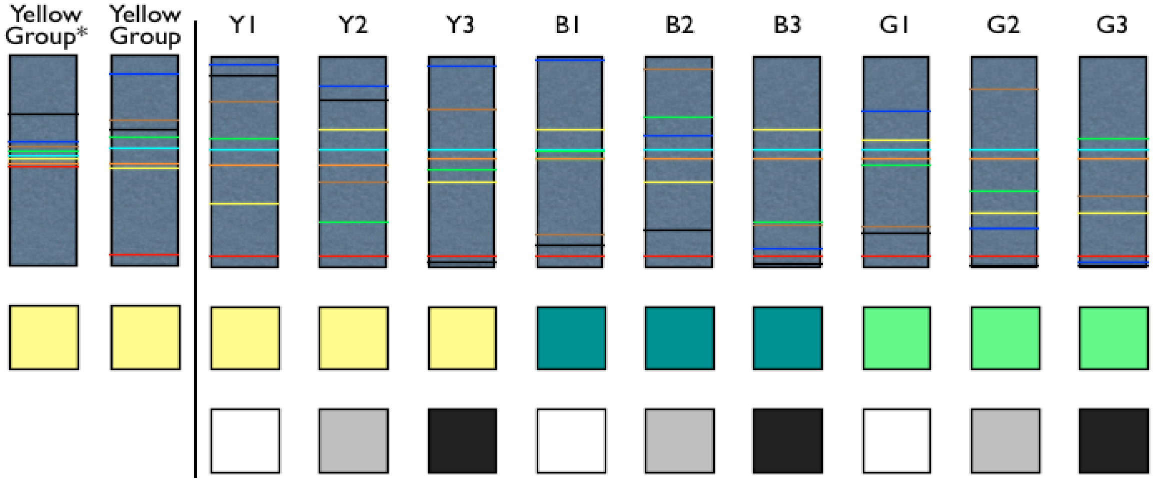


Figure 4.5: Comparison of the yellow group profile (Yellow Group) and the yellow group profile created after removing the confounding group effects (Yellow* Group).

4.3 Evaluation of Profile Subtraction Method

We evaluate our method on two independent datasets. The first dataset consists of gene expression profiles from Arabidopsis treated with different hormones [15]. This dataset is well suited to evaluate our novel profile subtraction method as it contains two clearly annotated sets of overlapping group labels. These correspond to a) the hormone class of a sample and b) the time when the sample was collected. To demonstrate that our profile subtraction method has utility in the absence of a second set of annotated groups we next evaluate a dataset consisting of acute myeloid leukemia (AML) and acute lympboblastic leukemia (ALL) samples[16].

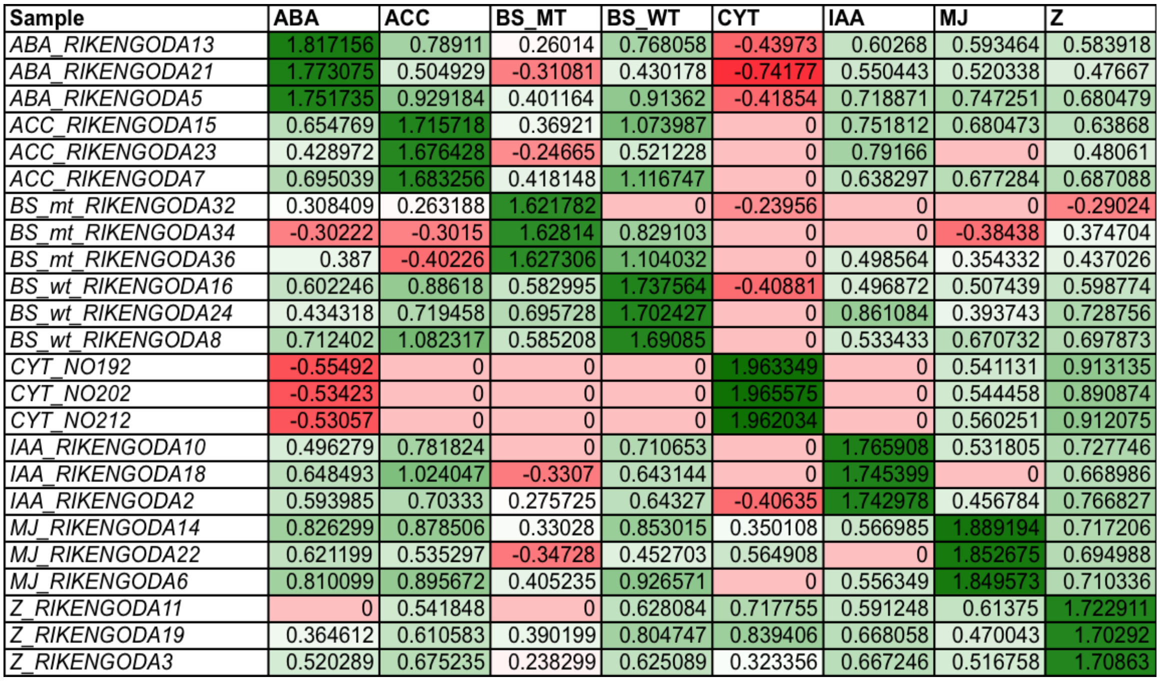
4.3.1 Arabidopsis Hormone Evaluation

The first dataset that we evaluate consists of gene expression profiles from the Arabidopsis Hormone dataset [15]. This dataset contains eight hormone groups.

We are interested in detecting a group signature for each of these hormone groups. The hormone groups are abscisic acid (ABA), aminocyclopropane carboxylic acid (ACC), brassinolide wildtype (BS_wt), brassinolide mutant)(BS_mt), cytokinins (CYT), indoleacetic acid (IAA), methyl jasmonate (MJ), and zeatin (Z). Each group contains three individual samples corresponding to three different collection times (30 min, 1 hour, and 3 hours). One exception to this is that all three samples from the CYT group are from the same time point (3 hours). These samples were left in for the analysis to study how this may or may not change the results.

Our goal in this chapter is to improve upon our group profile method introduced in Chapter 3, while introducing our profile subtraction method that will serve as a nonparametric equivalent to linear models. We compare the group profiles created after using our profile subtraction method to subtract out any time effects to both a) our original group profile method (naive to time) as well as b) group profiles created using linear models. For the linear models analysis we use Limma with both group and time effects together in the model. We also evaluate the group profiles after removing the time effects using Combat as well as after removing the batch effects using SVA.

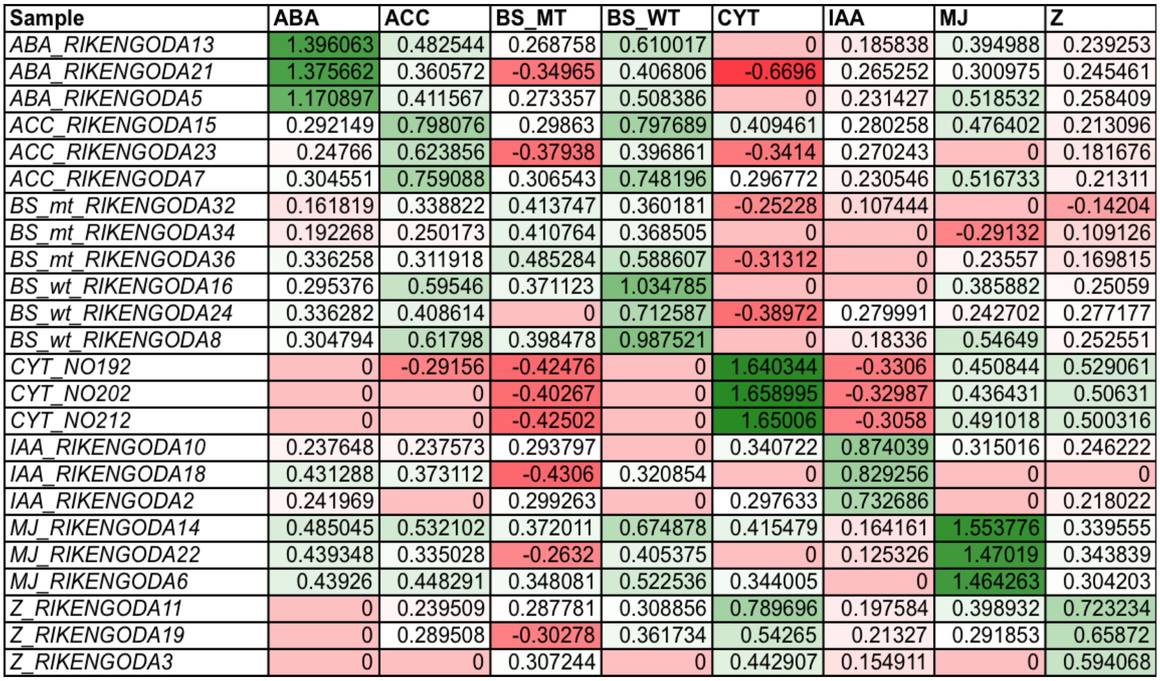
In order to evaluate the group profiles created by each of the methods we compared the measured recall of the individual group members. A group profile was made for each hormone group and each of these group profiles was then used to query the full database of sample profiles resulting in pairwise KS scores (similarity scores for two sided ranked lists first used by [28]). These results represent the ability to use the group profiles created to recall members of the group. The KS scores are



Sample	ABA	ACC	BS_MT	BS_WT	CYT	IAA	MJ	Z
ABA_RIKENGODA13	1.817156	0.78911	0.26014	0.768058	-0.43973	0.60268	0.593464	0.583918
ABA_RIKENGODA21	1.773075	0.504929	-0.31081	0.430178	-0.74177	0.550443	0.520338	0.47667
ABA_RIKENGODA5	1.751735	0.929184	0.401164	0.91362	-0.41854	0.718871	0.747251	0.680479
ACC_RIKENGODA15	0.654769	1.715718	0.36921	1.073987	0	0.751812	0.680473	0.63868
ACC_RIKENGODA23	0.428972	1.676428	-0.24665	0.521228	0	0.79166	0	0.48061
ACC_RIKENGODA7	0.695039	1.683256	0.418148	1.116747	0	0.638297	0.677284	0.687088
BS_mt_RIKENGODA32	0.308409	0.263188	1.621782	0	-0.23956	0	0	-0.29024
BS_mt_RIKENGODA34	-0.30222	-0.3015	1.62814	0.829103	0	0	-0.38438	0.374704
BS_mt_RIKENGODA36	0.387	-0.40226	1.627306	1.104032	0	0.498564	0.354332	0.437026
BS_wt_RIKENGODA16	0.602246	0.88618	0.582995	1.737564	-0.40881	0.496872	0.507439	0.598774
BS_wt_RIKENGODA24	0.434318	0.719458	0.695728	1.702427	0	0.861084	0.393743	0.728756
BS_wt_RIKENGODA8	0.712402	1.082317	0.585208	1.69085	0	0.533433	0.670732	0.697873
CYT_NO192	-0.55492	0	0	0	1.963349	0	0.541131	0.913135
CYT_NO202	-0.53423	0	0	0	1.965575	0	0.544458	0.890874
CYT_NO212	-0.53057	0	0	0	1.962034	0	0.560251	0.912075
IAA_RIKENGODA10	0.496279	0.781824	0	0.710653	0	1.765908	0.531805	0.727746
IAA_RIKENGODA18	0.648493	1.024047	-0.3307	0.643144	0	1.745399	0	0.668986
IAA_RIKENGODA2	0.593985	0.70333	0.275725	0.64327	-0.40635	1.742978	0.456784	0.766827
MJ_RIKENGODA14	0.826299	0.878506	0.33028	0.853015	0.350108	0.566985	1.889194	0.717206
MJ_RIKENGODA22	0.621199	0.535297	-0.34728	0.452703	0.564908	0	1.852675	0.694988
MJ_RIKENGODA6	0.810099	0.895672	0.405235	0.926571	0	0.556349	1.849573	0.710336
Z_RIKENGODA11	0	0.541848	0	0.628084	0.717755	0.591248	0.61375	1.722911
Z_RIKENGODA19	0.364612	0.610583	0.390199	0.804747	0.839406	0.668058	0.470043	1.70292
Z_RIKENGODA3	0.520289	0.675235	0.238299	0.625089	0.323356	0.667246	0.516758	1.70863

Figure 4.6: Uncorrected - Pairwise similarity (KS scores) between uncorrected group profiles for each hormone class and every individual sample in the dataset.

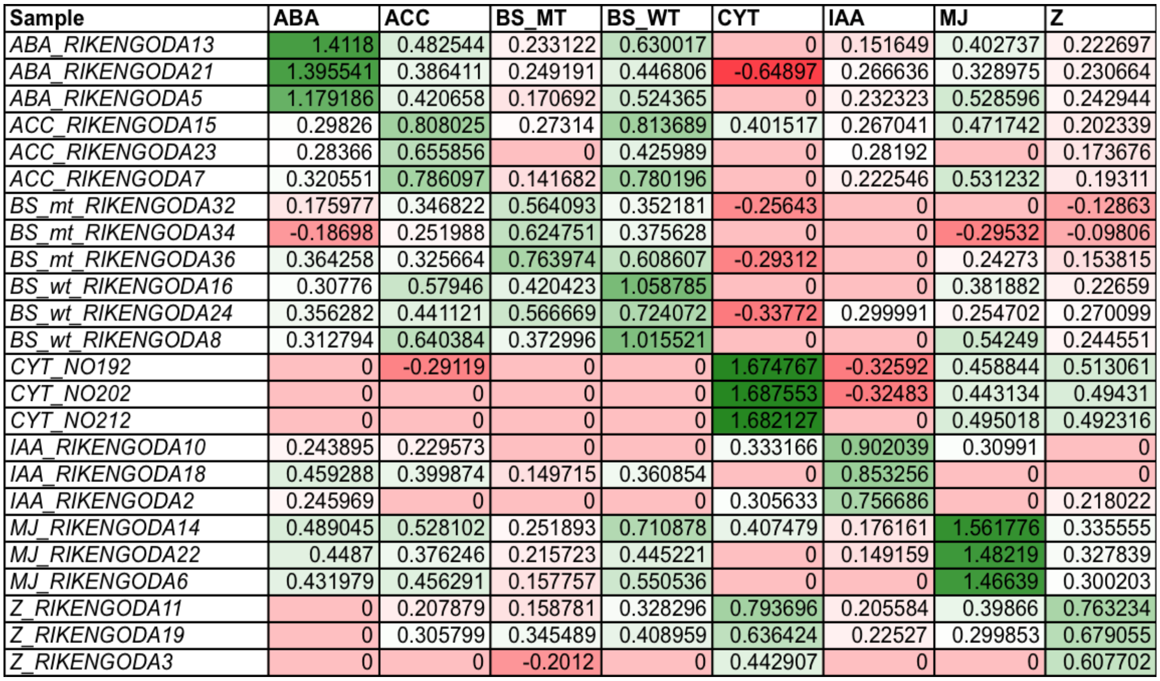
shown in the tables in Figure 4.6, Figure 4.7, Figure 4.8, Figure 4.9, and Figure 4.10, . As can be seen in Figure 4.6, our original group profile method performs adequately. The highest similarity matches for a given group is consistent with true members of that group (shown along the diagonal). It is worth noting that some positive similarities are seen with other samples, so there is room for improvement on the specificity of this method, especially given that we have the knowledge of the group effects. Figure 4.7 contains the results of performing the same analysis with the group profiles created by Limma (with both the hormone group and time in the model). Overall, the similarities are not as strong as our original group profile method but the false positives appear reduced. Normally the best matches are with true member of the group, but with much lower similarity scores. However, the



Sample	ABA	ACC	BS_MT	BS_WT	CYT	IAA	MJ	Z
ABA_RIKENGODA13	1.396063	0.482544	0.268758	0.610017	0	0.185838	0.394988	0.239253
ABA_RIKENGODA21	1.375662	0.360572	-0.34965	0.406806	-0.6696	0.265252	0.300975	0.245461
ABA_RIKENGODA5	1.170897	0.411567	0.273357	0.508386	0	0.231427	0.518532	0.258409
ACC_RIKENGODA15	0.292149	0.798076	0.29863	0.797689	0.409461	0.280258	0.476402	0.213096
ACC_RIKENGODA23	0.24766	0.623856	-0.37938	0.396861	-0.3414	0.270243	0	0.181676
ACC_RIKENGODA7	0.304551	0.759088	0.306543	0.748196	0.296772	0.230546	0.516733	0.21311
BS_mt_RIKENGODA32	0.161819	0.338822	0.413747	0.360181	-0.25228	0.107444	0	-0.14204
BS_mt_RIKENGODA34	0.192268	0.250173	0.410764	0.368505	0	0	-0.29132	0.109126
BS_mt_RIKENGODA36	0.336258	0.311918	0.485284	0.588607	-0.31312	0	0.23557	0.169815
BS_wt_RIKENGODA16	0.295376	0.59546	0.371123	1.034785	0	0	0.385882	0.25059
BS_wt_RIKENGODA24	0.336282	0.408614	0	0.712587	-0.38972	0.279991	0.242702	0.277177
BS_wt_RIKENGODA8	0.304794	0.61798	0.398478	0.987521	0	0.18336	0.54649	0.252551
CYT_NO192	0	-0.29156	-0.42476	0	1.640344	-0.3306	0.450844	0.529061
CYT_NO202	0	0	-0.40267	0	1.658995	-0.32987	0.436431	0.50631
CYT_NO212	0	0	-0.42502	0	1.65006	-0.3058	0.491018	0.500316
IAA_RIKENGODA10	0.237648	0.237573	0.293797	0	0.340722	0.874039	0.315016	0.246222
IAA_RIKENGODA18	0.431288	0.373112	-0.4306	0.320854	0	0.829256	0	0
IAA_RIKENGODA2	0.241969	0	0.299263	0	0.297633	0.732686	0	0.218022
MJ_RIKENGODA14	0.485045	0.532102	0.372011	0.674878	0.415479	0.164161	1.553776	0.339555
MJ_RIKENGODA22	0.439348	0.335028	-0.2632	0.405375	0	0.125326	1.47019	0.343839
MJ_RIKENGODA6	0.43926	0.448291	0.348081	0.522536	0.344005	0	1.464263	0.304203
Z_RIKENGODA11	0	0.239509	0.287781	0.308856	0.789696	0.197584	0.398932	0.723234
Z_RIKENGODA19	0	0.289508	-0.30278	0.361734	0.54265	0.21327	0.291853	0.65872
Z_RIKENGODA3	0	0	0.307244	0	0.442907	0.154911	0	0.594068

Figure 4.7: Limma - Pairwise similarity (KS scores) between Limma group profiles for each hormone class and every individual sample in the dataset. Note that these profiles are time corrected by including time in the linear model.

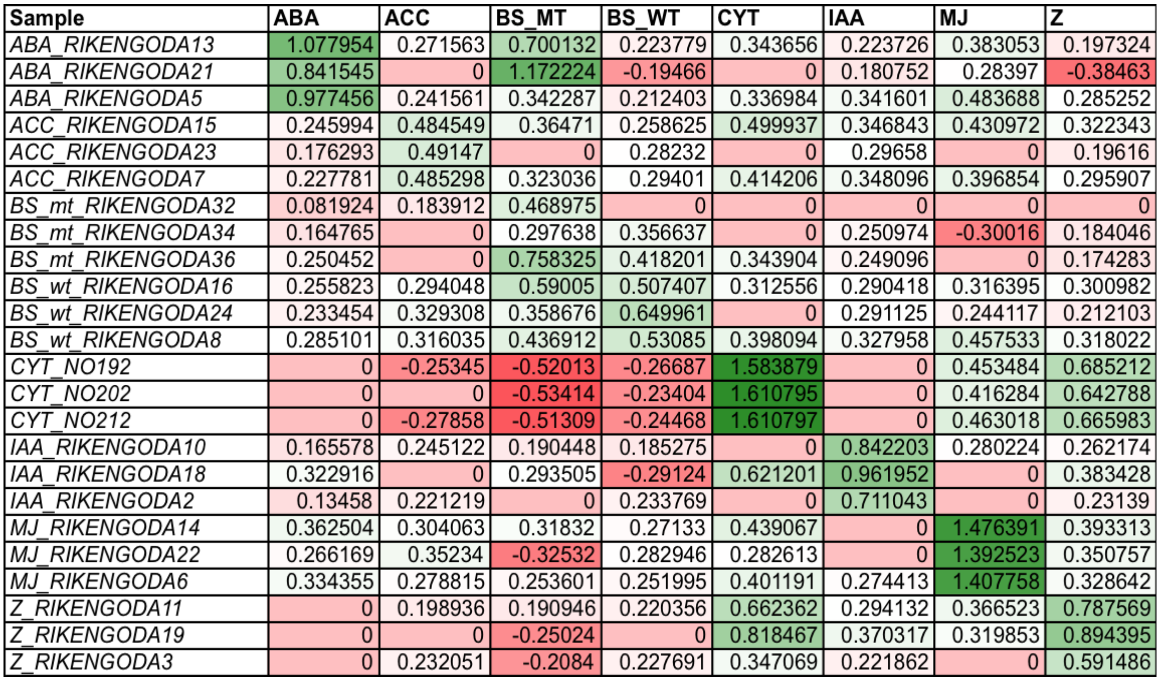
amount of true positives recalled is poor, even though this method appears more sensitive and less likely to produce false positives. An alternative to modeling the hormone group and time in the model concurrently, we can remove the time effects ahead of time with Combat and then create the group profiles(Figure 4.8). These results are very similar to those obtained with Limma. This may be because the confounding effects being removed are equally balanced across the target groups. Both of these methods explicitly model and attempt to remove the confounding effects. An alternative is surrogate variable analysis (SVA[31]). SVA does not require labelled confounders a priori but rather it attempts to discover and remove unlabeled batch effects. The results from first cleaning the dataset with SVA and



Sample	ABA	ACC	BS_MT	BS_WT	CYT	IAA	MJ	Z
ABA_RIKENGODA13	1.4118	0.482544	0.233122	0.630017	0	0.151649	0.402737	0.222697
ABA_RIKENGODA21	1.395541	0.386411	0.249191	0.446806	-0.64897	0.266636	0.328975	0.230664
ABA_RIKENGODA5	1.179186	0.420658	0.170692	0.524365	0	0.232323	0.528596	0.242944
ACC_RIKENGODA15	0.29826	0.808025	0.27314	0.813689	0.401517	0.267041	0.471742	0.202339
ACC_RIKENGODA23	0.28366	0.655856	0	0.425989	0	0.28192	0	0.173676
ACC_RIKENGODA7	0.320551	0.786097	0.141682	0.780196	0	0.222546	0.531232	0.19311
BS_mt_RIKENGODA32	0.175977	0.346822	0.564093	0.352181	-0.25643	0	0	-0.12863
BS_mt_RIKENGODA34	-0.18698	0.251988	0.624751	0.375628	0	0	-0.29532	-0.09806
BS_mt_RIKENGODA36	0.364258	0.325664	0.763974	0.608607	-0.29312	0	0.24273	0.153815
BS_wt_RIKENGODA16	0.30776	0.57946	0.420423	1.058785	0	0	0.381882	0.22659
BS_wt_RIKENGODA24	0.356282	0.441121	0.566669	0.724072	-0.33772	0.299991	0.254702	0.270099
BS_wt_RIKENGODA8	0.312794	0.640384	0.372996	1.015521	0	0	0.54249	0.244551
CYT_NO192	0	-0.29119	0	0	1.674767	-0.32592	0.458844	0.513061
CYT_NO202	0	0	0	0	1.687553	-0.32483	0.443134	0.49431
CYT_NO212	0	0	0	0	1.682127	0	0.495018	0.492316
IAA_RIKENGODA10	0.243895	0.229573	0	0	0.333166	0.902039	0.30991	0
IAA_RIKENGODA18	0.459288	0.399874	0.149715	0.360854	0	0.853256	0	0
IAA_RIKENGODA2	0.245969	0	0	0	0.305633	0.756686	0	0.218022
MJ_RIKENGODA14	0.489045	0.528102	0.251893	0.710878	0.407479	0.176161	1.561776	0.335555
MJ_RIKENGODA22	0.4487	0.376246	0.215723	0.445221	0	0.149159	1.48219	0.327839
MJ_RIKENGODA6	0.431979	0.456291	0.157757	0.550536	0	0	1.46639	0.300203
Z_RIKENGODA11	0	0.207879	0.158781	0.328296	0.793696	0.205584	0.39866	0.763234
Z_RIKENGODA19	0	0.305799	0.345489	0.408959	0.636424	0.22527	0.299853	0.679055
Z_RIKENGODA3	0	0	-0.2012	0	0.442907	0	0	0.607702

Figure 4.8: Combat - Pairwise similarity (KS scores) between Limma group profiles for each hormone class and every individual sample in the dataset after time effects were removed using Combat.

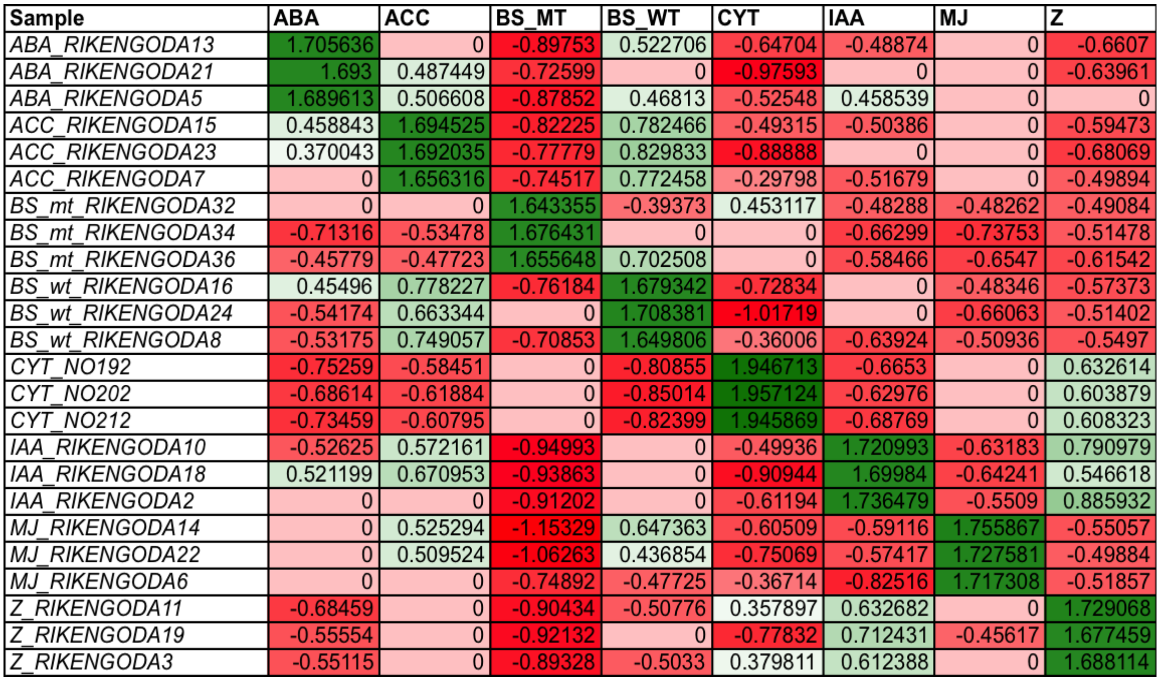
then using linear models to create the group profiles are shown in Figure 4.9. SVA does not perform as well as the other methods, however, it does do well considering it does not use the confounding group labels. It is therefore a viable alternative to consider when the confounding groups are unknown – similar to how our original group profile method works. The last table (Figure 4.10) contains the results from hormone group profiles created after using our profile subtraction method to remove the confounding time effects. The results are an improvement on the original group profile method. The biggest difference is the reduction in possible false positives. These results also demonstrate an improvement over the results obtained with other methods, i.e., Limma, Combat and SVA.



Sample	ABA	ACC	BS_MT	BS_WT	CYT	IAA	MJ	Z
ABA_RIKENGODA13	1.077954	0.271563	0.700132	0.223779	0.343656	0.223726	0.383053	0.197324
ABA_RIKENGODA21	0.841545	0	1.172224	-0.19466	0	0.180752	0.28397	-0.38463
ABA_RIKENGODA5	0.977456	0.241561	0.342287	0.212403	0.336984	0.341601	0.483688	0.285252
ACC_RIKENGODA15	0.245994	0.484549	0.36471	0.258625	0.499937	0.346843	0.430972	0.322343
ACC_RIKENGODA23	0.176293	0.49147	0	0.28232	0	0.29658	0	0.19616
ACC_RIKENGODA7	0.227781	0.485298	0.323036	0.29401	0.414206	0.348096	0.396854	0.295907
BS_mt_RIKENGODA32	0.081924	0.183912	0.468975	0	0	0	0	0
BS_mt_RIKENGODA34	0.164765	0	0.297638	0.356637	0	0.250974	-0.30016	0.184046
BS_mt_RIKENGODA36	0.250452	0	0.758325	0.418201	0.343904	0.249096	0	0.174283
BS_wt_RIKENGODA16	0.255823	0.294048	0.59005	0.507407	0.312556	0.290418	0.316395	0.300982
BS_wt_RIKENGODA24	0.233454	0.329308	0.358676	0.649961	0	0.291125	0.244117	0.212103
BS_wt_RIKENGODA8	0.285101	0.316035	0.436912	0.53085	0.398094	0.327958	0.457533	0.318022
CYT_NO192	0	-0.25345	-0.52013	-0.26687	1.583879	0	0.453484	0.685212
CYT_NO202	0	0	-0.53414	-0.23404	1.610795	0	0.416284	0.642788
CYT_NO212	0	-0.27858	-0.51309	-0.24468	1.610797	0	0.463018	0.665983
IAA_RIKENGODA10	0.165578	0.245122	0.190448	0.185275	0	0.842203	0.280224	0.262174
IAA_RIKENGODA18	0.322916	0	0.293505	-0.29124	0.621201	0.961952	0	0.383428
IAA_RIKENGODA2	0.13458	0.221219	0	0.233769	0	0.711043	0	0.23139
MJ_RIKENGODA14	0.362504	0.304063	0.31832	0.27133	0.439067	0	1.476391	0.393313
MJ_RIKENGODA22	0.266169	0.35234	-0.32532	0.282946	0.282613	0	1.392523	0.350757
MJ_RIKENGODA6	0.334355	0.278815	0.253601	0.251995	0.401191	0.274413	1.407758	0.328642
Z_RIKENGODA11	0	0.198936	0.190946	0.220356	0.662362	0.294132	0.366523	0.787569
Z_RIKENGODA19	0	0	-0.25024	0	0.818467	0.370317	0.319853	0.894395
Z_RIKENGODA3	0	0.232051	-0.2084	0.227691	0.347069	0.221862	0	0.591486

Figure 4.9: SVA - Pairwise similarity (KS scores) between Limma group profiles for each hormone class and every individual sample in the dataset after removing unlabeled batch effects using SVA

This analysis demonstrates how in a dataset with confounding effects, our profile subtraction method can lead to improvements over other methods. However, the nature of this analysis is somewhat biased. Any of these methods will in fact be able to learn across the off-target signatures, e.g., the time effects in this case, when they are specifically trained on samples across such barriers as we have done. Specifically, by creating a hormone group profile consisting of one sample from each time point, the method is forced to learn what is common across the time points. This may be how such a method will be used sometimes in real life, but we recognize that at other times the dataset will not be so balanced. The true grouping may in fact not even be known a priori. To further compare and test these methods we



Sample	ABA	ACC	BS_MT	BS_WT	CYT	IAA	MJ	Z
ABA_RIKENGODA13	1.705636	0	-0.89753	0.522706	-0.64704	-0.48874	0	-0.6607
ABA_RIKENGODA21	1.693	0.487449	-0.72599	0	-0.97593	0	0	-0.63961
ABA_RIKENGODA5	1.689613	0.506608	-0.87852	0.46813	-0.52548	0.458539	0	0
ACC_RIKENGODA15	0.458843	1.694525	-0.82225	0.782466	-0.49315	-0.50386	0	-0.59473
ACC_RIKENGODA23	0.370043	1.692035	-0.77779	0.829833	-0.88888	0	0	-0.68069
ACC_RIKENGODA7	0	1.656316	-0.74517	0.772458	-0.29798	-0.51679	0	-0.49894
BS_mt_RIKENGODA32	0	0	1.643355	-0.39373	0.453117	-0.48288	-0.48262	-0.49084
BS_mt_RIKENGODA34	-0.71316	-0.53478	1.676431	0	0	-0.66299	-0.73753	-0.51478
BS_mt_RIKENGODA36	-0.45779	-0.47723	1.655648	0.702508	0	-0.58466	-0.6547	-0.61542
BS_wt_RIKENGODA16	0.45496	0.778227	-0.76184	1.679342	-0.72834	0	-0.48346	-0.57373
BS_wt_RIKENGODA24	-0.54174	0.663344	0	1.708381	-1.01719	0	-0.66063	-0.51402
BS_wt_RIKENGODA8	-0.53175	0.749057	-0.70853	1.649806	-0.36006	-0.63924	-0.50936	-0.5497
CYT_NO192	-0.75259	-0.58451	0	-0.80855	1.946713	-0.6653	0	0.632614
CYT_NO202	-0.68614	-0.61884	0	-0.85014	1.957124	-0.62976	0	0.603879
CYT_NO212	-0.73459	-0.60795	0	-0.82399	1.945869	-0.68769	0	0.608323
IAA_RIKENGODA10	-0.52625	0.572161	-0.94993	0	-0.49936	1.720993	-0.63183	0.790979
IAA_RIKENGODA18	0.521199	0.670953	-0.93863	0	-0.90944	1.69984	-0.64241	0.546618
IAA_RIKENGODA2	0	0	-0.91202	0	-0.61194	1.736479	-0.5509	0.885932
MJ_RIKENGODA14	0	0.525294	-1.15329	0.647363	-0.60509	-0.59116	1.755867	-0.55057
MJ_RIKENGODA22	0	0.509524	-1.06263	0.436854	-0.75069	-0.57417	1.727581	-0.49884
MJ_RIKENGODA6	0	0	-0.74892	-0.47725	-0.36714	-0.82516	1.717308	-0.51857
Z_RIKENGODA11	-0.68459	0	-0.90434	-0.50776	0.357897	0.632682	0	1.729068
Z_RIKENGODA19	-0.55554	0	-0.92132	0	-0.77832	0.712431	-0.45617	1.677459
Z_RIKENGODA3	-0.55115	0	-0.89328	-0.5033	0.379811	0.612388	0	1.688114

Figure 4.10: Profile Subtraction - Pairwise similarity (KS scores) between time corrected (time group profiles subtracted out) group profiles for each hormone class and every individual sample in the dataset.

explore the task of rediscovering the true hormone groups from a dataset with only the times labelled.

For this comparison, we implement a custom hierarchical clustering based on our group profile method. The similarity score at each level of the clustering is the pairwise KS statistic. The two closest profiles are merged by creating a new group profile from the two profiles. As in any agglomerative clustering, this process is repeated until one large cluster is left. In order to evaluate the accuracy of each of the three methods (group profile, linear models using Limma, group profiles after removal of time effects), we count clustering mistakes after performing the hierarchical group clustering. A clustering mistake is defined as the clustering that merges two or more different groups before each of the individual groups are fully clustered separately. In other words, if we had three profiles from group A and three from group B, the only way for there to be no clustering mistakes is for one cluster to be created that includes all three A's and a second cluster to be created that includes all three B's before these two clusters could be joined together. If a cluster of three A's was iteratively joined with each of the individual B profiles then this would result in 3 clustering mistakes.

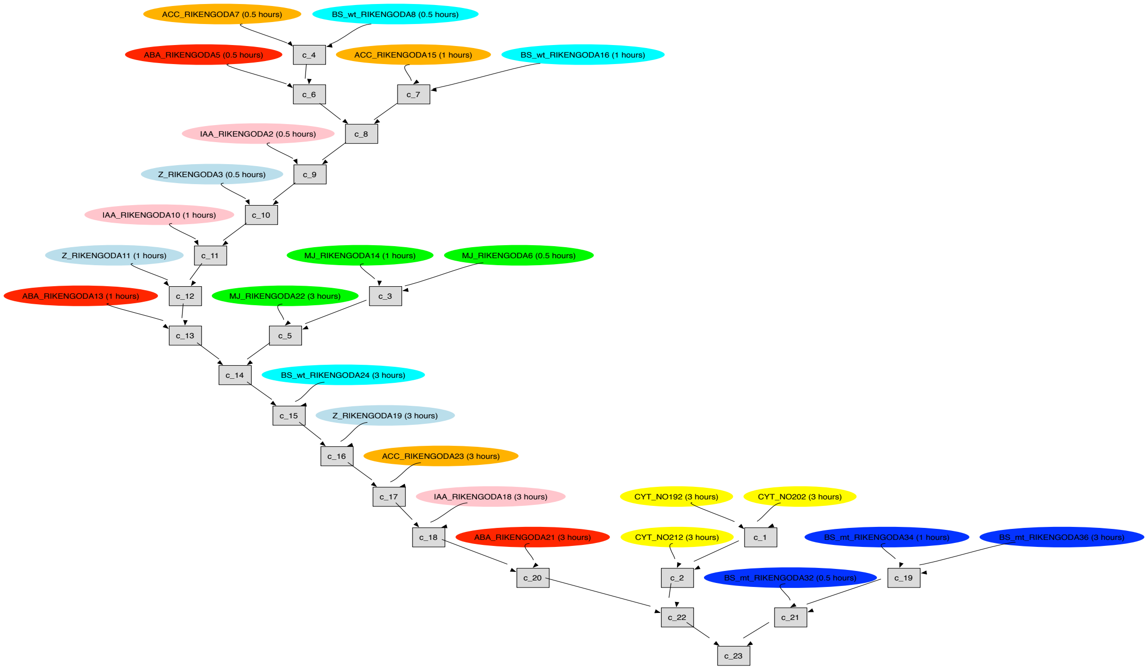

ACC_RIKENGODA7 (0.5 hours)
BS_wt_RIKENGODA8 (0.5 hours)
ABA_RIKENGODA5 (0.5 hours)
c_4
ACC_RIKENGODA15 (1 hours)
BS_wt_RIKENGODA16 (1 hours)
c_6
c_7
IAA_RIKENGODA2 (0.5 hours)
c_8
Z_RIKENGODA3 (0.5 hours)
c_9
IAA_RIKENGODA10 (1 hours)
c_10
Z_RIKENGODA11 (1 hours)
c_11
MJ_RIKENGODA14 (1 hours)
MJ_RIKENGODA6 (0.5 hours)
ABA_RIKENGODA13 (1 hours)
c_12
MJ_RIKENGODA22 (3 hours)
c_3
c_13
c_5
c_14
BS_wt_RIKENGODA24 (3 hours)
c_15
Z_RIKENGODA19 (3 hours)
c_16
ACC_RIKENGODA23 (3 hours)
c_17
IAA_RIKENGODA18 (3 hours)
CYT_NO192 (3 hours)
CYT_NO202 (3 hours)
c_18
ABA_RIKENGODA21 (3 hours)
CYT_NO212 (3 hours)
c_1
BS_mt_RIKENGODA34 (1 hours)
BS_mt_RIKENGODA36 (3 hours)
c_20
c_2
BS_mt_RIKENGODA32 (0.5 hours)
c_19
c_22
c_21
c_23

Figure 4.11: Hierarchical group profile based clustering of uncorrected individual Arabidopsis Hormone profiles.

The group profile based hierarchical clustering of the uncorrected profiles is shown in Figure 4.11. Certain hormone groups cluster correctly, e.g., the mutant strain that we expect to be very different from others (BS_mt), the CYT group that has no time effects modeled, as well as the MJ group. It is interesting to note that certain other clusters are based on the time effects. For example, three of the 30 minute samples cluster perfectly together first (ABA 0.5 hours, ACC 0.5 hours, and BS_wt 0.5 hours). Two of the 1 hour samples also cluster together (again ACC and BS-wt). After all of these are merged the IAA and Z 0.5 hour profiles are added into the hierarchical grouping, which in turn bring in their 1 hour counterparts. The MJ cluster is clustered next, which consisting of all 3 hour samples it next brings in all of the rest of the 3 hour samples.

In Figure 4.12 we present the group profile based hierarchical clustering of the Arabidopsis Hormone dataset using the time corrected profiles from Limma. Interestingly this method appears to over compensate for the time effects and the clusters align more with the time the samples were collected. The only hormone group that clusters correctly is the CYT group which is all collected at the same time point (3 hours).

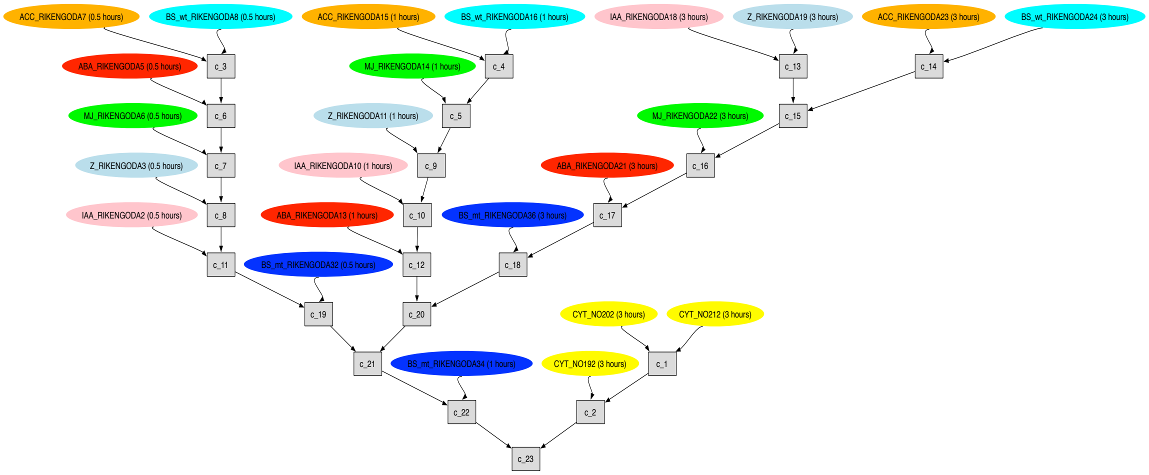

ACC_RIKENGODA7 (0.5 hours)
BS_wt_RIKENGODA8 (0.5 hours)
ACC_RIKENGODA15 (1 hours)
BS_wt_RIKENGODA16 (1 hours)
IAA_RIKENGODA18 (3 hours)
Z_RIKENGODA19 (3 hours)
ACC_RIKENGODA23 (3 hours)
BS_wt_RIKENGODA24 (3 hours)
ABA_RIKENGODA5 (0.5 hours)
c_3
MJ_RIKENGODA14 (1 hours)
c_4
c_13
c_14
MJ_RIKENGODA6 (0.5 hours)
c_6
Z_RIKENGODA11 (1 hours)
c_5
MJ_RIKENGODA22 (3 hours)
c_15
Z_RIKENGODA3 (0.5 hours)
c_7
IAA_RIKENGODA10 (1 hours)
c_9
ABA_RIKENGODA21 (3 hours)
c_16
IAA_RIKENGODA2 (0.5 hours)
c_8
ABA_RIKENGODA13 (1 hours)
c_10
BS_mt_RIKENGODA36 (3 hours)
c_17
c_11
BS_mt_RIKENGODA32 (0.5 hours)
c_12
c_18
c_19
c_20
CYT_NO202 (3 hours)
CYT_NO212 (3 hours)
c_21
BS_mt_RIKENGODA34 (1 hours)
CYT_NO192 (3 hours)
c_1
c_22
c_2
c_23

Figure 4.12: Hierarchical group profile based clustering of individual Arabidopsis Hormone profiles after correcting for time effects as modeled by Limma.

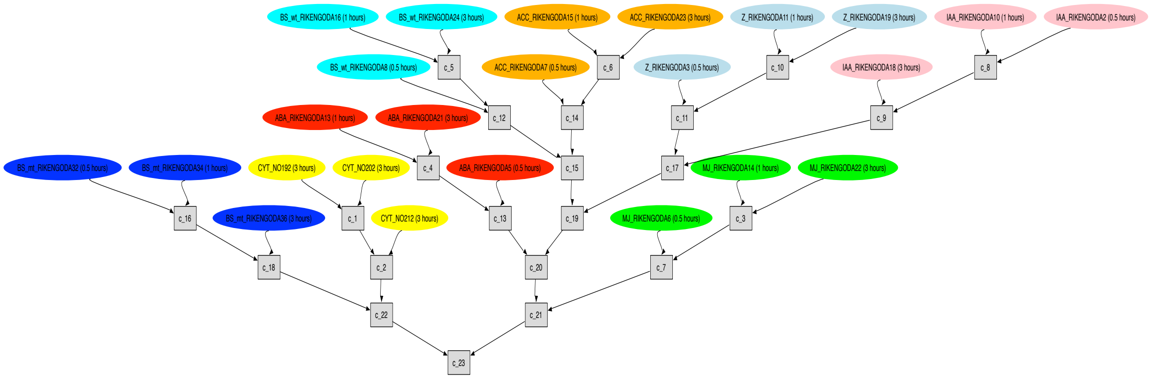

BS_wt_RIKENGODA16 (1 hours)
BS_wt_RIKENGODA24 (3 hours)
ACC_RIKENGODA15 (1 hours)
ACC_RIKENGODA23 (3 hours)
Z_RIKENGODA11 (1 hours)
Z_RIKENGODA19 (3 hours)
IAA_RIKENGODA10 (1 hours)
IAA_RIKENGODA2 (0.5 hours)
BS_wt_RIKENGODA8 (0.5 hours)
c_5
ACC_RIKENGODA7 (0.5 hours)
c_6
Z_RIKENGODA3 (0.5 hours)
c_10
IAA_RIKENGODA18 (3 hours)
c_8
ABA_RIKENGODA13 (1 hours)
ABA_RIKENGODA21 (3 hours)
c_12
c_14
c_11
c_9
BS_mt_RIKENGODA32 (0.5 hours)
BS_mt_RIKENGODA34 (1 hours)
CYT_NO192 (3 hours)
CYT_NO202 (3 hours)
c_4
ABA_RIKENGODA5 (0.5 hours)
c_15
c_17
MJ_RIKENGODA14 (1 hours)
MJ_RIKENGODA22 (3 hours)
c_16
BS_mt_RIKENGODA36 (3 hours)
c_1
CYT_NO212 (3 hours)
c_13
c_19
MJ_RIKENGODA6 (0.5 hours)
c_3
c_18
c_2
c_20
c_7
c_22
c_21
c_23

Figure 4.13: Hierarchical group profile based clustering of individual Arabidopsis Hormone profiles after subtracting out time group profiles.

Lastly, the hierarchical clustering of the profiles for which the profile subtraction method was used to remove the time effects is shown in Figure 4.13. No clustering mistakes are made. All of the hormone groups are clustered together before being joined with any other hormone group. The underlying patterns and observations seen previously still hold, namely that the mutant strain and the CYT group are distant from the other groups. However, the results are clear and the hormone groups have been rediscovered without mistake. If this dataset contained unknown groups we have demonstrated that in order to make important discoveries you must first remove all the know, labelled confounding group effects, e.g., time, vehicle, and batch effects.

One could argue that this evaluation is still slightly biased in that we used our custom group profile based agglomerative clustering method. We believe that this approach is meaningful and in itself is a useful contribution in dealing with novel group detection in gene expression experiments. However, in order to provide an unbiased demonstration of how our method works we include a full off-the-shelf hierarchical clustering of the uncorrected (Figure 4.14) and time effect profile subtracted (Figure 4.15) profiles respectively (done using Spotfire[1]).

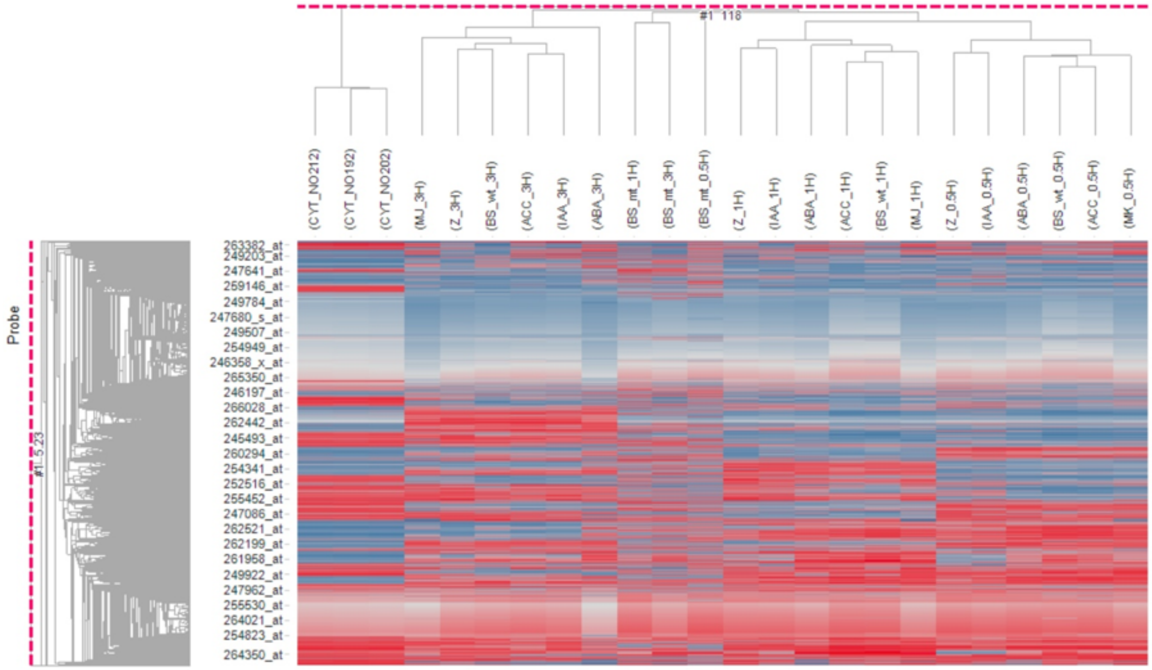


Figure 4.14: Hierarchical clustering of uncorrected individual Arabidopsis Hormone profiles. The full ranked list is used for this clustering.

4.3.2 Acute Leukemia Prediction

In the previous section we have demonstrated how our profile subtraction method improves on both our group profile method and a linear models approach (as implemented in Limma). The dataset for this was one that was optimally suited for this comparison in that it contained two sets of overlapping groups: hormone groups and time groups. Our method can be useful even when this second set of groups is not as obvious. We evaluate a gene expression dataset consisting of

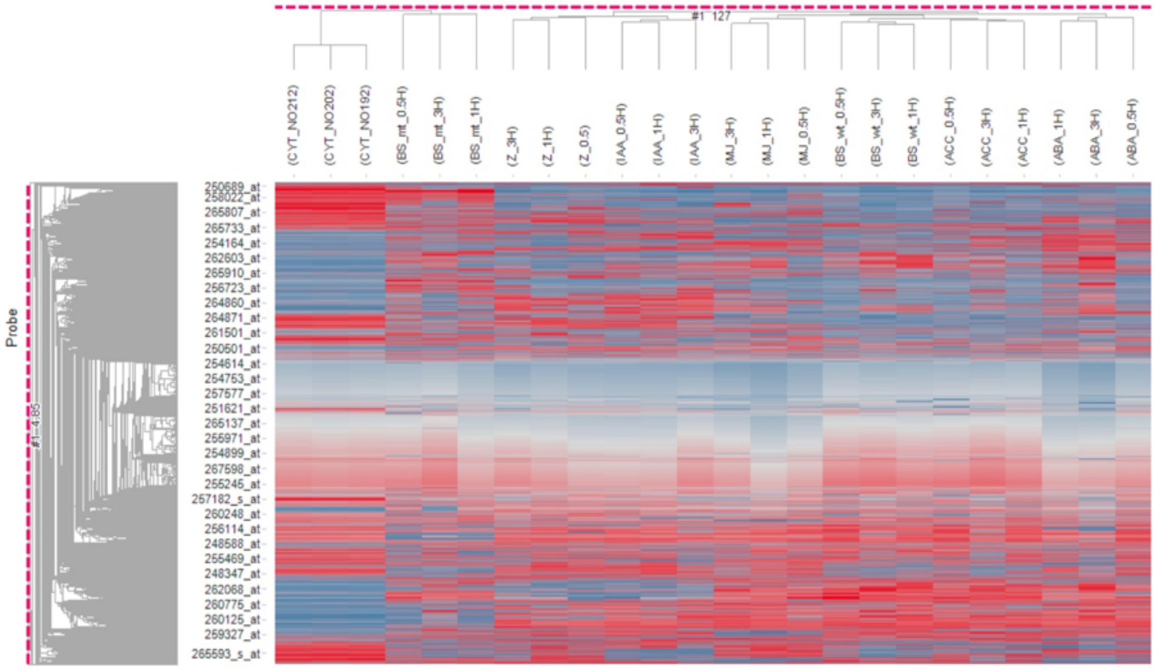


Figure 4.15: Hierarchical clustering of individual Arabidopsis Hormone profiles after subtracting out time group profiles. The full ranked list is used for this clustering.

two subtypes of acute leukemia, namely acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL)[16].

The task that we are interested in is to see if we can rediscover the two subtypes of leukemia and correctly segment the dataset correctly. To do this we again use our group profile based hierarchical clustering method. The results of the clustering based on the uncorrected gene expression profiles is shown in Figure 4.16. To improve on this we again use our profile subtraction method, but unlike before there are no specific, annotated confounding groups. All of the samples are very similar and have large pairwise KS scores across the dataset because they all are leukemia samples and this large generic leukemia signature is masking out the subtype signatures. We therefore create a super leukemia group containing all of the samples and remove this generic leukemia profile from all of the profiles. After removing the generic leukemia signature from the individual profiles we then rediscover the two subtypes of leukemia Figure 4.17. These subtypes of leukemia are perfectly clustered without any further knowledge needed.

4.4 Gene Expression Profile Generator

We have demonstrated how our group profile subtraction method can work on real world datasets. We have shown how this method is an improvement over both our original group profile method and linear models. For these methods there are two main parameters to deal with, the weight to assign to the confounding group profile to remove and the tagsize. The tagsize is only used for the hierarchical clustering and not for the profile subtraction itself. Assuming that we had fully labelled group and confounding group memberships there is not much issue in exploring varying these weights and evaluating the impact on the results. Even if the dataset is partially labelled we can take this approach and evaluate correct assignment of groups to minimize the mistakes. If the labels are missing a domain expert could help estimate how these parameters should be set. We acknowledge that sometimes there may be no labels or domain expert to help tune these parameters so in order to better understand the tradeoff of tuning these two variables we have created a gene expression profile generator. The goal of doing this is to be able to understand the relationship between groups of interest and confounding groups in a completely controlled fashion. More specifically to we are interested in three main topics: understanding the impact of different weights of signatures, the selection of the optimal weight factor in the profile subtraction implementation, and to evaluate the impact on differing probe sizes used in all of the calculations (up/down tagsizes).

4.4.1 Creation of groups and confounding groups

There are several parameters that can be set in the gene expression profile generator including: Number of group profiles selected (NG), Number of confounding (side effect) profiles selected(NSE), and Number of individual profiles selected.

For each group (and confounding group) a random set of GO terms is selected to be up-expressed. For each of these GO terms a weight is randomly selected using the following Gaussian distribution $f(x) = ae^{-\frac{(x-\omega)^2}{2c^2}}$. Analogously a random set of GO terms is selected to be down-expressed and a weight is assigned from $f(x) = ae^{-\frac{(x+\omega)^2}{2c^2}}$.

4.4.2 Creation of individual profiles

For each individual profile s to be created, an initial random profile is generated. To do this each probe is randomly selected from Gaussian distribution (variance $= 0.5^2$) $f(x) = ae^{-\frac{x^2}{2\times 0.5^2}}$. Then for each group/confounding group that this individual is a member of and for each probe that is linked to a gene in the GO id for this group, we average together the original random expression value with the GO term weighted expression. This gives us the effect that probes belonging to genes in our GO terms selected as being part of the up (or down) expressed signatures are more likely to be up (or down) expressed, but they do not necessarily have to be. This is repeated for all the groups that this individual is a member with and this causes the effects to be cumulative (if a individual belong to two groups with the same GO term as either both up- or down-expressed) or diminishing (if the individual belongs to two groups with the same GO term in opposite signatures – up- and down-expressed).

For our evaluation we keep the group weight constant at 1.0. Confounding weights are evaluated over the range between 0.5 and 1.5 in increments of 0.1. We fix the number of groups (NG) to 3, and the number of confounding groups (NSE) to 3. In order to guarantee a group worth evaluating we additionally fix the minimum number of GO terms to 10. The number of individual profiles created for each group/confounding group is set to 3 (total n=27). The Up/Down tagsizes are evaluated across several values of t in the set (25,50,100,250,500). For modeling purposed c is held constant at 0.2. Additionally, we evaluate profile subtraction weights, ω_g , from 50 to 100 in increments of 5.

In order to evaluate the accuracy of each of these simulations, we again perform hierarchical group clustering and count clustering mistakes. For each of these simulations we also create confounding group profiles (from our designated side effect profiles) and calculate pairwise KS scores with each individual profile in the dataset. From this we can evaluate the relationship between these scores and optimal weights

Figure 4.18 demonstrates that for a normal tagsize of 100, there are confounding group weightings that result in perfect clustering regardless of the modeled confounding profile weight. The results for varying tagsizes is shown in Figure 4.19

by showing the results broken down by the confounding group size modeled. These results clearly demonstrated that when the confounding group effects are less than the target group effects (first panel), the tagsize chosen has no impact on the results. The tagsize becomes more important as the modeled confounding group effects increases (relative to the group of interest), but it never becomes a driving force in the accuracy of the method. We therefore should expect a tagsize in the 100 range to perform reasonable well. As previously stated, a domain expert could help decide this or this weighting could be more accurately selected in the presence of any labelled group membership annotations.

4.5 Discussion

In this chapter we have introduced our gene expression profile subtraction method. The goal of this work was to extend our original group profile method in a way such that we would be able to solve the problem of removing confounding group effects as these confounding effects may mask the true signatures of the groups being studied. We created a method tangential to what one can do with linear models, but for ranked lists, and more specifically two-sided relative ranked lists. We introduced an extension to our non-parametric gene expression profile method to correct for observed confounding effects. This correction is performed on ranked lists directly and provides a robust alternative to parametric batch profile correction methods.

We have evaluated our profile subtraction method and demonstrated how it improves on our original group profile method in two different datasets when dealing with confounding effects. We further have provided a comparison to alternative methods including Limma, Combat, and SVA. The results from our method exhibits a high level of true positive recall similar to our group profile method in addition to a reduction in false positives similar to a linear model type of approach.

For this work we have focused on the ability to rediscover known groups of interest as our comparison metric, however, the outcome of all of this work is the actual group profiles being created. The accuracy obtained in the clustering gives us confidence that the signatures obtained are in fact meaningful. We can now use these signatures to both find new members of a given group as well as to gain biological insight into the shared genetic signature of our groups as was demonstrated in Chapter 3.

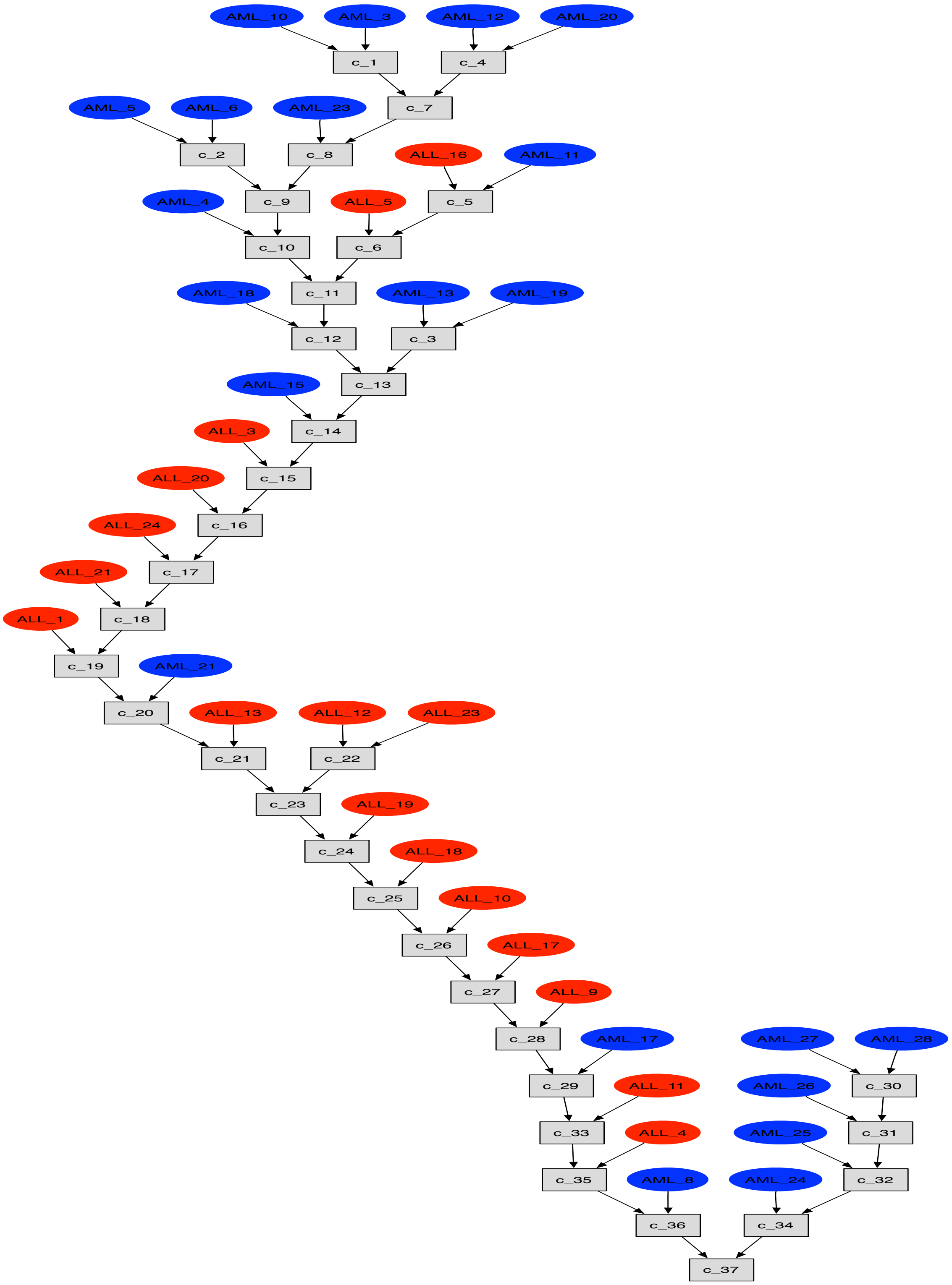


Figure 4.16: Hierarchical group profile based clustering of uncorrected individual AML (blue) and ALL (red) profiles.

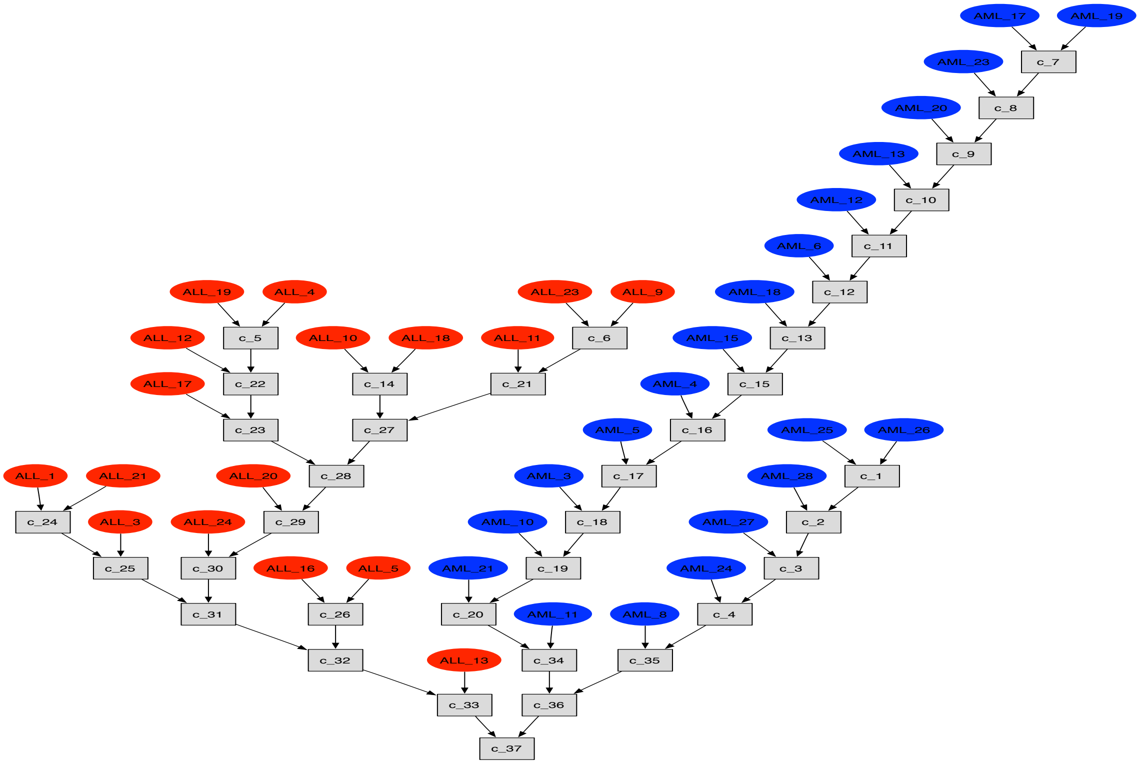


Figure 4.17: Hierarchical group profile based clustering of individual AML (blue) and ALL (red) profiles after subtracting out an overarching cancer group profile.

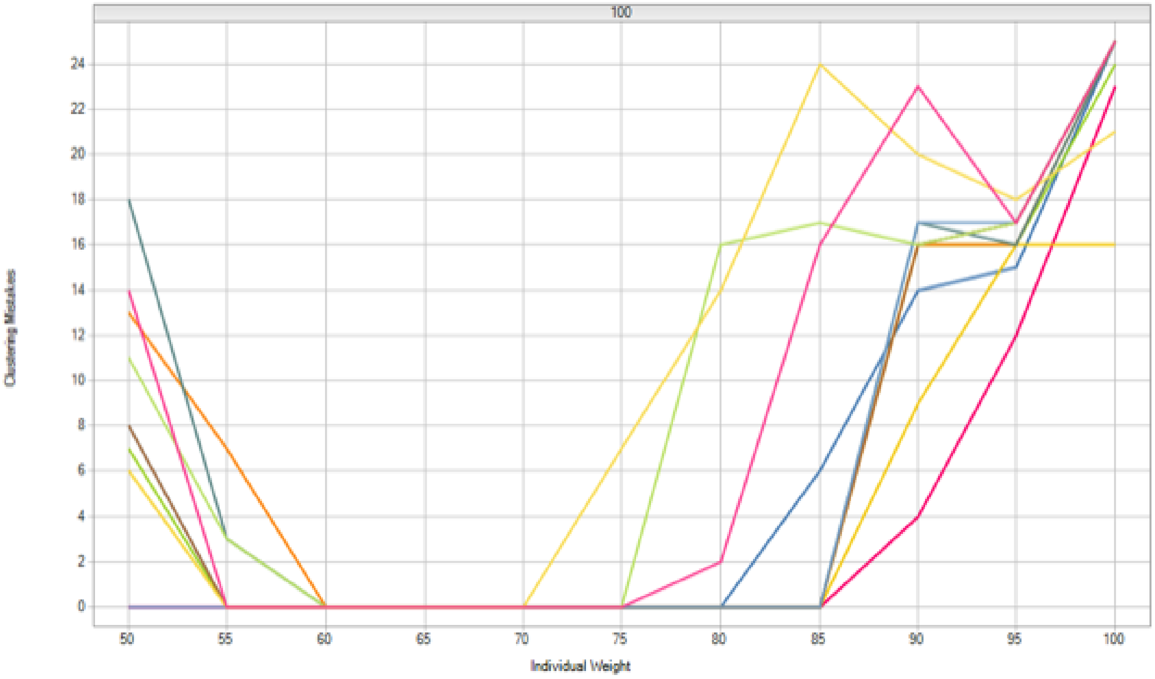


Figure 4.18: Evaluation of the impact of the profile subtraction weighting factor on clustering accuracy across varying modeled true confounding group profile strengths (corresponding to different colors). Here we show the results using a tagsize of 100.

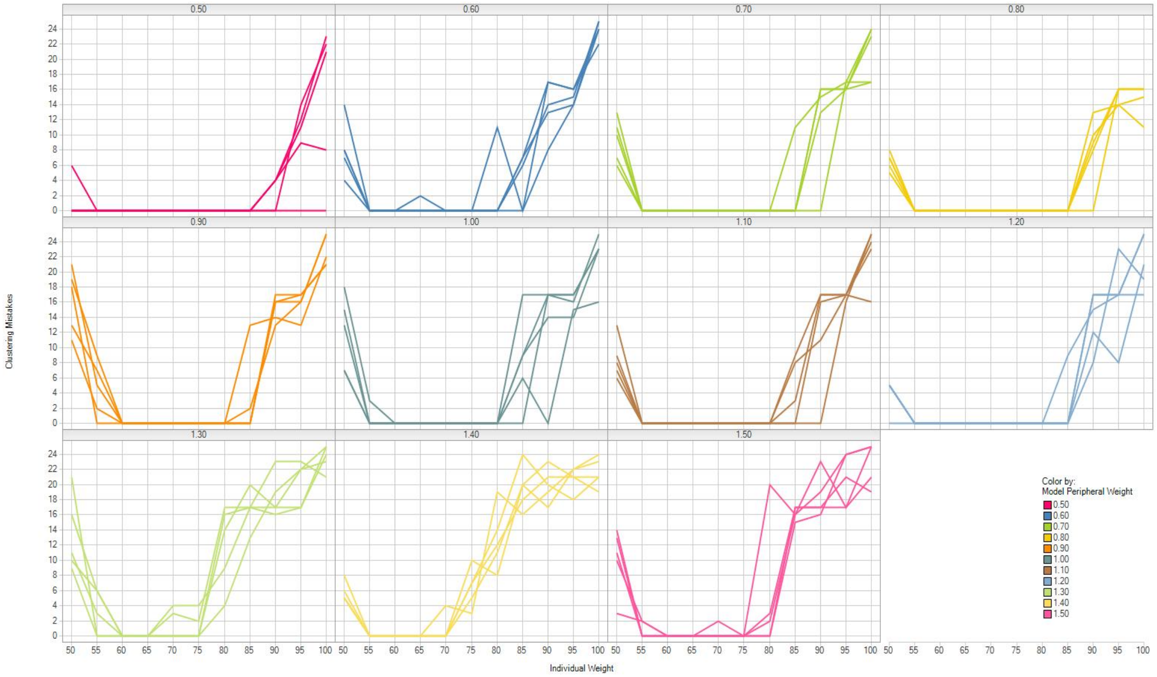


Figure 4.19: Negligible impact of varying the tagsize on the clustering accuracy across varying modeled true confounding group profile strengths. Each subgraph contains the results from a given modeled true confounding weight and each line corresponds to a given tagsize (25,50,100,250,500). Each line represents a different tagsize.

Chapter 5

Conclusions and Future Work

Gene expression microarrays are used to answer a variety of scientific questions. These questions include gaining a better understanding of a drug, segmenting a disease, and predicting an optimal therapeutic response. The amount of gene expression data publicly available is extremely large and continues to grow at an increasing rate. However, this rapid growth of gene expression data from laboratories across the world has not had the full impact on the scientific community that it is capable of achieving. This shortcoming is because a lot of the data cannot be combined and used all at once. Even within a closely controlled gene expression experiment there are confounding factors that may mask the true signatures when analyzed with current methods. We are interested in three core tasks that we believe are important, namely similarity search, signature detection, and confounder correction. We have developed novel methods that address each of these tasks.

In Chapter 2, we focus on similarity search within gene expression data. More specifically, we are interested in methods for similarity search that overcome confounding effects, e.g., vehicle and batch effects. We introduce this topic through the recent work from Lamb et al. [28] in which they present their tool, the Connectivity-Map (CMAP) which tackles the problem of comparing gene expression profiles generated under diverse experimental conditions. They use a distribution statistic to compare the rankings of the probes and they ignore the raw expression values which is a deviation from previous methods. Overall, the CMAP approach is robust and provides good similarity measures. However, as it only evaluates a subset of probes, it can fail when faced with severe vehicle and batch effects. In order to improve upon this method we introduce the notion of an indirect similarity measure. The indirect similarity measure uses the combined knowledge from throughout the database to improve on the pairwise similarity calculation similar to how GPS works. This is done to minimize the error bounds of a single estimate. We presented an evaluation of this new indirect similarity method on a large proprietary dataset. Through this evaluation we showed that this new method is able to overcome experimental noise by obtaining a 97.03% improvement in recall of similar drugs over the direct CMAP approach. In addition, we validated the robustness of this method on the publicly available CMAP dataset, with an improvement in recall of 49.44%.

In Chapter 3, we dealt with the challenge of determining a robust genetic signature to represent a group of gene expression profiles. We motivated and described our weighted influence model rank of ranks (WIMRR) method for group profile creation. We demonstrated how to gain biological insight into the underlying function of a group of compounds by evaluating overrepresentation of GO terms within the top/bottom of the group signature. As we showed, this in turn can potentially lead to a new hypothesis into the etiology of the disease that they treat. We explained how to perform similarity searching with these group profiles and showed how such a profile can be used for classification, e.g., classifying subtypes of a disease. A case study of the antipsychotic group was presented to demonstrate the power of this group profile method. A sensitivity analysis and independent validation of the group profile method was performed demonstrating scientifically meaningful results. We concluded this chapter with an analysis of a large dataset consisting of over 200 therapeutic classes. We have created a website (GEPedia.org) that hosts all of these group profiles. This website also includes all of the downstream analysis for each group profile, including both an evaluation of the biological signatures as well as the similarity score of each compound in the database.

In Chapter 4, we addressed the issue of confounding effects in gene expression experiments. At a high level, our goal was to develop a method that will behave similar to the parametric methods that are currently employed in this field, e.g., Limma, Combat, SVA, etc. We have proposed an extension to our non-parametric gene expression profile method to correct for observed confounding effects. This correction is performed on ranked lists directly and provides a robust alternative to parametric batch profile correction methods. The premise of modeling target groups and confounding groups at once is that the effects among these two groups is additive. In order to remove the effects of a confounding group profile we first identify all of the members of the group and create the group profile of our confounding group. Given the unique properties of our two-sided ranked lists, in order to subtract the confounding profile from each of the samples we first invert the ranked list. Then to complete the subtraction of this confounding groups effects the original profile (rank(y) values) and the ranks of this inverted ranked lists representing our confounders can just be averaged together. We are left with the individual gene expression profiles after having the confounding effects removed. We evaluated our novel profile subtraction method on an Arabidopsis dataset that contained a set of hormone samples treated for different times. We successfully were able to remove the confounding time effects and improve on the results observed with other methods when creating the hormone group profiles. We also were able to rediscover the hormone groups perfectly. We provided an additional evaluation on an independent dataset consisting of AML and ALL samples, for which we also were able to cluster the two subtypes of leukemia correctly. We concluded this work with the creation of a gene expression profile generator and a discussion on the robustness of the method to the tuning parameters.

5.1 Future Work

We have introduced methods to deal with the three tasks that are important when analyzing gene expression experiments. However, it is important to note that unlike many methods for dealing with similarity search within gene expression data, our methods works regardless of the data representation. For the first method dealing with indirect similarity, any pairwise distance metric can be swapped in in place of the KS statistic that we evaluated. The only requirement of the group profile and profile subtraction methods is that the data can be represented as two-sided relatively ranked lists. We believe that these methods would prove useful if evaluated on other datasets, especially datasets with a large number of observations that can be ranked. This includes any system whose state can be represented by a large set of sensors/probes. For example, a sensor network containing tens of thousands of sensors could be represented as a two-sided ranked lists by first normalizing everything as changes from one state to another, and then ranking these in the same way that was done for the gene expression profiles.

We have briefly demonstrated that these methods are robust to the parameter tuning but further work could be done to automate the optimal selection of some of these parameters, e.g., the tag size (number of probes) used for the search, weight of a confounding effect to remove, etc. We have shown how our methods work well when a dataset contains two sets of overlapping groups (target groups and confounding groups) that are known and correctly annotated. We have also demonstrated that when only the confounding groups are known, the groups of interest can be discovered. Another interesting task would be the collective discovery of both sets of overlapping groups, assuming both the target groups and confounding groups are unknown. A simple step in this direction would be to deal with a partially labelled set of group memberships or a set of group labels that contains errors. An updated version of these methods could attempt to automatically correct these issues. Alternatively, an active learning approach could be employed in which the algorithm could present the user with classifications for which there are questions.

We discussed at the onset of this work that one of our motivations is to be able to use more of the information available in the public domain to drive new scientific discovery. We believe that the methods introduced in this work allow for the combination of gene expression data from multiple sources as they are robust to vehicle, batch, and other confounding effects. However, we have assumed that the data is all from similar microarrays. There is some work on mapping one microarray technology to another, as well as mapping the results of gene expression data from one organism to another (through GO terms and other mechanisms). It would be worthwhile to explore the ability of our methods to work across these barriers and to evaluate what changes would be required to have them perform optimally across such large barriers. Combining all of these possible avenues of future research would lead to discovery of novel scientific information from data that has already been generated and that sits undiscovered in the public domain. The possibilities to gain a better understanding of diseases and of new potential therapies to treat them serves as a motivation to continue developing and refining these methods.

Bibliography

[1] C Ahlberg. Spotfire: an information exploration environment. *SIGMOD Rec.*, 25:25–29, December 1996.

[2] DE Arking, A Pfeufer, W Post, WHL Kao, C Newton-Cheh, M Ikeda, Kristen West, Carl Kashuk, Mahmut Akyol, Siegfried Perz, Shapour Jalilzadeh, T Illig, C Gieger, Chao-Yu Guo, MG Larson, HE Wichmann, EM án, CJ O'Donnell, JN Hirschhorn, S Kääb, PM Spooner, T Meitinger, and A Chakravarti. A common genetic variant in the nos1 regulator nos1ap modulates cardiac repolarization. *Nat Genet*, 38(6):644–51, Jun 2006.

[3] M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, JT Eppig, MA Harris, DP Hill, L Issel-Tarver, A Kasarskis, S Lewis, JC Matese, JE Richardson, M Ringwald, GM Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–9, May 2000.

[4] O Barochovsky and A J Patel. Effect of central nervous system acting drugs on brain cell replication in vitro. *Neurochem Res*, 7(9):1059–74, Sep 1982.

[5] M Benito, J Parker, Q Du, J Wu, D Xiang, C M Perou, and JS Marron. Adjustment of systematic microarray data biases. *Bioinformatics*, 20(1):105–14, Jan 2004.

[6] I Bhattacharya and L Getoor. Iterative record linkage for cleaning and integration. *DMKD '04: Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, Jun 2004.

[7] I Bhattacharya and L Getoor. Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Jan 2007.

[8] M Bilgic, L Licamele, L Getoor, and B Shneiderman. D-dupe: An interactive tool for entity resolution in social networks. *Visual Analytics Science And Technology*, Jan 2006.

[9] E Chávez, G Navarro, R Baeza-Yates, and J Marroquín. Searching in metric spaces. *ACM Computing Surveys (CSUR*, 33(3), Sep 2001.

[10] C Chen, K Grennan, J Badner, D Zhang, E Gershon, Li Jin, and C Liu. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS ONE*, 6(2):e17238, Jan 2011.

[11] ER DeLong, DM DeLong, and DL Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–45, Sep 1988.

[12] JL DeRisi, VR Iyer, and PO Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–6, Oct 1997.

[13] DR Donohoe, EJ Aamodt, E Osborn, and FS Dwyer. Antipsychotic drugs disrupt normal development in caenorhabditis elegans via additional mechanisms besides dopamine and serotonin receptors. *Pharmacol Res*, 54(5):361–72, Nov 2006.

[14] S Gerard. Reviewing medications for bipolar disorder: understanding the mechanisms of action. *The Journal of clinical psychiatry*, 70(1):e02, Jan 2009.

[15] H Goda, E Sasaki, K Akiyama, A Maruyama-Nakashita, K Nakabayashi, W Li, M Ogawa, Y Yamauchi, J Preston, K Aoki, T Kiba, S Takatsuto, S Fujioka, T Asami, T Nakano, H Kato, T Mizuno, H Sakakibara, S Yamaguchi, E Nambara, Y Kamiya, H Takahashi, M Yokota Hirai, T Sakurai, K Shinozaki, K Saito, S Yoshida, and Y Shimada. The atgenexpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access. *Plant J*, 55(3):526–42, Aug 2008.

[16] TR Golub, DK Slonim, PTamayo, C Huard, M Gaasenbeek, JP Mesirov, H Coller, ML Loh, JR Downing, MA Caligiuri, CD Bloomfield, and ES Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–7, Oct 1999.

[17] M Gupta, C Chauhan, P Bhatnagar, S Gupta, S Grover, PK Singh, M Purushottam, O Mukherjee, S Jain, SK Brahmachari, and R Kukreti. Genetic susceptibility to schizophrenia: role of dopaminergic pathway gene polymorphisms. *Pharmacogenomics*, 10(2):277–91, Feb 2009.

[18] PC Heinrich, I Behrmann, S Haan, HM Hermanns, G Müller-Newen, and F Schaper. Principles of interleukin (il)-6-type cytokine signalling and its regulation. *Biochem J*, 374(Pt 1):1–20, Aug 2003.

[19] T Hongo, S Yajima, M Sakurai, Y Horikoshi, and R Hanada. In vitro drug sensitivity testing can predict induction failure and early relapse of childhood acute lymphoblastic leukemia. *Blood*, 89(8):2959–65, Apr 1997.

[20] E Hubbell, W Liu, and R Mei. Robust estimators for expression analysis. *Bioinformatics*, 18(12):1585–92, Dec 2002.

[21] T R Hughes, M J Marton, A R Jones, C J Roberts, R Stoughton, C D Armour, H A Bennett, E Coffey, H Dai, Y D He, M J Kidd, A M King, M R Meyer, D Slade, P Y Lum, S B Stepaniants, D D Shoemaker, D Gachotte, K Chakraburtty, J Simon, M Bard, and S H Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–26, Jul 2000.

[22] NB Ivanova, JT Dimos, C Schaniel, JA Hackney, KA Moore, and IR Lemischka. A stem cell molecular signature. *Science*, 298(5593):601–4, Oct 2002.

[23] X Jiang, F Tian, Y Du, NG Copeland, NA Jenkins, L Tessarollo, X Wu, H Pan, XZ Hu, K Xu, H Kenney, SE Egan, H Turley, AL Harris, AM Marini, and RH Lipsky. Bhlhb2 controls bdnf promoter 4 activity and neuronal excitability. *J Neurosci*, 28(5):1118–30, Jan 2008.

[24] B John, AJ Enright, A Aravin, T Tuschl, C Sander, and DS Marks. Human microrna targets. *PLoS Biol*, 2(11):e363, Nov 2004.

[25] WE Johnson, C Li, and A Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–27, Jan 2007.

[26] GJ Kaspers, R Pieters, CH Van Zantwijk, ER Van Wering, A Van Der Does-Van Den Berg, and AJ Veerman. Prednisolone resistance in childhood acute lymphoblastic leukemia: vitro-vivo correlations and cross-resistance to other drugs. *Blood*, 92(1):259–66, Jul 1998.

[27] GJ Kaspers, AJ Veerman, R Pieters, CH Van Zantwijk, LA Smets, ER Van Wering, and A Van Der Does-Van Den Berg. In vitro cellular drug resistance and prognosis in newly diagnosed childhood acute lymphoblastic leukemia. *Blood*, 90(7):2723–9, Oct 1997.

[28] J Lamb, ED Crawford, D Peck, JW Modell, IC Blat, MJ Wrobel, J Lerner, J Brunet, A Subramanian, KN Ross, M Reich, H Hieronymus, G Wei, SA Armstrong, SJ Haggarty, PA Clemons, R Wei, S A Carr, ES Lander, and TR Golub. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–35, Sep 2006.

[29] J Lamb, S Ramaswamy, H Ford, and B Contreras. A mechanism of cyclin d1 action encoded in the patterns of gene expression in human cancer. *Cell*, Jan 2003.

[30] JT Leek, RB Scharpf, HC Bravo, D Simcha, B Langmead, WE Johnson, D Geman, K Baggerly, and RA Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*, 11(10):733–9, Oct 2010.

[31] JT Leek and JD Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9):1724–35, Sep 2007.

[32] L Licamele and L Getoor. Indirect two-sided relative ranking: a robust similarity measure for gene expression data. *BMC Bioinformatics*, 11:137, Jan 2010.

[33] L Licamele and L Getoor. A method for the detection of meaningful and reproducible group signatures from gene expression profiles. *J Bioinform Comput Biol*, 9(3):431–51, Jun 2011.

[34] G Linden, B Smith, and J York. Amazon. com recommendations: item-to-item collaborative filtering. *Internet Computing*, Jan 2003.

[35] J Listgarten, C Kadie, EE Schadt, and D Heckerman. Correction for hidden confounders in the genetic analysis of gene expression. *Proc Natl Acad Sci USA*, 107(38):16465–70, Sep 2010.

[36] J Lu, S Guo, BL Ebert, H Zhang, X Peng, J Bosco, J Pretz, R Schlanger, JY Wang, RH Mak, DM Dombkowski, FI Preffer, DT Scadden, and TR Golub. Microrna-mediated control of cell fate in megakaryocyte-erythrocyte progenitors. *Dev Cell*, 14(6):843–53, Jun 2008.

[37] PY Lum, CD Armour, SB Stepaniants, G Cavet, MK Wolf, JS Butler, JC Hinshaw, P Garnier, GD Prestwich, A Leonardson, P Garrett-Engele, CM Rush, M Bard, G Schimmack, JW Phillips, CJ Roberts, and DD Shoemaker. Discovering modes of action for therapeutic compounds using a genome-wide screen of yeast heterozygotes. *Cell*, 116(1):121–37, Jan 2004.

[38] J Luo, M Schumacher, A Scherer, D Sanoudou, D Megherbi, T Davison, T Shi, W Tong, L Shi, H Hong, C Zhao, F Elloumi, W Shi, R Thomas, S Lin, G Tillinghast, G Liu, Y Zhou, D Herman, Y Li, Y Deng, H Fang, P Bushel, M Woods, and J Zhang. A comparison of batch effect removal methods for enhancement of prediction performance using maqc-ii microarray gene expression data. *Pharmacogenomics J*, 10(4):278–91, Aug 2010.

[39] W m Liu, R Mei, X Di, T B Ryder, E Hubbell, S Dee, TA Webster, CA Harrington, M h Ho, J Baid, and SP Smeekens. Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics*, 18(12):1593–9, Dec 2002.

[40] DE Martin, P Demougin, MN Hall, and M Bellis. Rank difference analysis of microarrays (rdam), a novel approach to statistical analysis of microarray expression profiling data. *BMC Bioinformatics*, 5:148, Oct 2004.

[41] RM Miller, LM Callahan, C Casaceli, L Chen, GL Kiser, B Chui, TM Kaysser-Kranich, TJ Sendera, C Palaniappan, and HJ Federoff. Dysregulation of gene expression in the 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine-lesioned mouse substantia nigra. *J Neurosci*, 24(34):7445–54, Aug 2004.

[42] JW Newcomer and MJ Sernyak. Identifying metabolic risks with antipsychotics and monitoring and management strategies. *The Journal of clinical psychiatry*, 68(7):e17, Jul 2007.

[43] JC Newman and AM Weiner. L2l: a simple tool for discovering the hidden significance in microarray expression data. *Genome Biol*, 6(9):R81, Jan 2005.

[44] CM Perou, T Sørlie, MB Eisen, M van de Rijn, SS Jeffrey, CA Rees, JR Pollack, DT Ross, H Johnsen, LA Akslen, O Fluge, A Pergamenschikov, C Williams, SX Zhu, PE Lønning, AL Børresen-Dale, PO Brown, and D Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–52, Aug 2000.

[45] R Pieters, DR Huismans, AH Loonen, KHählen, A Van Der Does-Van Den Berg, ER Van Wering, and AJ Veerman. Relation of cellular drug resistance to long-term clinical outcome in childhood acute lymphoblastic leukaemia. *Lancet*, 338(8764):399–403, Aug 1991.

[46] M Pirooznia, JY Yang, M Qu Yang, and Y Deng. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics*, 9 Suppl 1:S13, Jan 2008.

[47] MH Polymeropoulos, L Licamele, S Volpi, K Mack, SN Mitkus, ED Carstea, L Getoor, A Thompson, and C Lavedan. Common effect of antipsychotics on the biosynthesis and regulation of fatty acids and cholesterol supports a key role of lipid homeostasis in schizophrenia. *Schizophr Res*, 108(1-3):134–42, Mar 2009.

[48] SL Pomeroy, P Tamayo, M Gaasenbeek, LM Sturla, M Angelo, ME McLaughlin, JYH Kim, LC Goumnerova, PM Black, C Lau, JC Allen, D Zagzag, JM Olson, T Curran, C Wetmore, JA Biegel, T Poggio, S Mukherjee, R Rifkin, A Califano, G Stolovitzky, DN Louis, JP Mesirov, ES Lander, and TR Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–42, Jan 2002.

[49] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[50] M Ramalho-Santos, S Yoon, Y Matsuzaki, RC Mulligan, and DA Melton. "stemness": transcriptional profiling of embryonic and adult stem cells. *Science*, 298(5593):597–600, Oct 2002.

[51] C Roumestan, C Henriquet, C Gougat, A Michel, F Bichon, K Portet, D Jaffuel, and M Mathieu. Histamine h1-receptor antagonists inhibit nuclear factor-kappab and activator protein-1 activities via h1-receptor-dependent and -independent mechanisms. *Clin Exp Allergy*, 38(6):947–56, Jun 2008.

[52] B Sarwar, G Karypis, J Konstan, and J Reidl. Item-based collaborative filtering recommendation algorithms. *WWW '01: Proceedings of the 10th international conference on World Wide Web*, Apr 2001.

[53] E Segal, N Friedman, D Koller, and A Regev. A module map showing conditional activity of expression modules in cancer. *Nat Genet*, 36(10):1090–8, Oct 2004.

[54] KP Seiler, GA George, MP Happ, NE Bodycombe, HA Carrinski, S Norton, S Brudz, JP Sullivan, J Muhlich, M Serrano, P Ferraiolo, NJ Tolliday, SL Schreiber, and PA Clemons. Chembank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res*, 36(Database issue):D351–9, Jan 2008.

[55] J Shi, JK Wittke-Thompson, JA Badner, E Hattori, JB Potash, VL Willour, FJ McMahon, RS Gershon, and C Liu. Clock genes may influence bipolar disorder susceptibility and dysfunctional circadian rhythm. *Am J Med Genet B Neuropsychiatr Genet*, 147B(7):1047–55, Oct 2008.

[56] AH Sims, GJ Smethurst, Y Hey, MJ Okoniewski, SD Pepper, A Howell, CJ Miller, and RB Clarke. The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets - improving meta-analysis and prediction of prognosis. *BMC Med Genomics*, 1:42, Jan 2008.

[57] GK Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3, Jan 2004.

[58] PK Tan, TJ Downey, EL Spitznagel, P Xu, D Fu, DS Dimitrov, RA Lempicki, BM Raaka, and MC Cam. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res*, 31(19):5676–84, Oct 2003.

[59] OG Troyanskaya, ME Garber, PO Brown, D Botstein, and Russ B Altman. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 18(11):1454–61, Nov 2002.

[60] LJ van 't Veer, H Dai, MJ van de Vijver, YD He, AAM Hart, M Mao, Hans L Peterse, Karin van der Kooy, Matthew J Marton, Anke T Witteveen, George J Schreiber, RM Kerkhoven, C Roberts, PS Linsley, René Bernards, and Stephen H Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–6, Jan 2002.

[61] I Vastrik, P D'Eustachio, E Schmidt, G Joshi-Tope, G Gopinath, D Croft, B de Bono, M Gillespie, B Jassal, S Lewis, L Matthews, G Wu, E Birney, and L Stein. Reactome: a knowledge base of biologic pathways and processes. *Genome Biol*, 8(3):R39, Jan 2007.

[62] J F Waring, R A Jolly, R Ciurlionis, P Y Lum, J T Praestgaard, D C Morfitt, B Buratto, C Roberts, E Schadt, and R G Ulrich. Clustering of hepatotoxins based on mechanism of toxicity using gene expression profiles. *Toxicol Appl Pharmacol*, 175(1):28–42, Aug 2001.

[63] G Wei, D Twomey, J Lamb, K Schlis, J Agarwal, RW Stam, JT Opferman, SE Sallan, ML den Boer, R Pieters, TR Golub, and SA Armstrong. Gene expression-based chemical genomics identifies rapamycin as a modulator of mcl1 and glucocorticoid resistance. *Cancer Cell*, 10(4):331–42, Oct 2006.

[64] M Xiu, L Hui, Y Dang, T De Hou, C Zhang, Y Zheng, D Chen, T Kosten, and X Zhang. Decreased serum bdnf levels in chronic institutionalized schizophrenia on long-term treatment with typical and atypical antipsychotics. *Prog Neuropsychopharmacol Biol Psychiatry*, Aug 2009.