Bicluster-Based Identification of Gene Sets Through Multivariate Meta-Analysis (MVMA)

Tim Wu

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2018

Reading Committee:
Peter Tarczy-Hornoch, Chair
Roger E. Bumgarner
Shuai Huang

Program Authorized to Offer Degree:
Biomedical and Health Informatics

University of Washington

Abstract

Bicluster-Based Identification of Gene Sets
Through Multivariate Meta-Analysis (MVMA)

Tim Wu

Chair of the Supervisory Committee:
Peter Tarczy-Hornoch, MD, FACMI
Department of Biomedical Informatics and Medical Education

Omics technologies are among the most exciting developments in biology and medicine in recent decades. They offer a whole new way of investigating a sample or a patient by taking comprehensive molecular-level snapshots. These snapshots, in the form of massive amount of data, provide important hints about the pathophysiological state of the target.

Despite the promises of the omics technologies, their usefulness hinges upon proper translation of the data into knowledge. This dissertation is focused on mining of public gene expression data to discover gene sets that may be parts of biological pathways. It tries to answer these two **overall questions**: (1) what is the data mining method best suited for finding gene sets? (2) how best to utilize multiple datasets in order to increase statistical strength?

Biclustering has been proven to be highly effective for identifying gene sets. Compared to traditional clustering methods, biclustering recognizes a list of genes that are up- or down-regulated under a subset of the conditions, as opposed to the whole spectrum of the conditions. A

large number of biclustering algorithms have been applied to analysis of gene expression data. Condition-dependent Correlation Subgroups (CCS), as one of these algorithms, is chosen for the current study. Identifying individual biclusters using CCS is the task of **Aim 1**.

Most public expression datasets have relatively small sample sizes. Making inference on these datasets may be error prone, which motivates the use of multiple datasets to increase the statistical power. This study makes use of multiple related datasets by adapting the approach of meta-analysis. More specifically, a group of biclusters, each coming from a separate dataset, are identified. Meta-analysis is then applied to these biclusters. Hence, the biclusters are analogous to the individual studies in a traditional meta-analysis. The goal is to identify a gene set, through combining the evidence in the individual biclusters.

Since each gene in this group of bicluster is modeled as an endpoint (equivalent to outcome in traditional meta-analysis context), and the correlations among the endpoints are taken into consideration, the approach of multivariate meta-analysis (MVMA) is taken. Using MVMA to combine biclusters from separate datasets is the focus of **Aim 2**.

Despite the fact that biclustering has significantly reduced the dimension, analyzing the stack still faces the difficulty of high dimensionality (p) and small number of available datasets (n), which is the well-known p >> n problem. The traditional MVMA methods, either within the Bayesian or the Frequentist framework, are not effective when p is over 50. Since a typical bicluster stack has a dimension in the range of 70 - 150, it renders the traditional methods impractical in the current context. A previous study by Jackson and Riley [1] proposed an interesting two-step procedure for MVMA to tackle the issue of data scarcity. It involves estimation of the between-study covariance matrix as the step 1, following by making inference

about the overall effect sizes as the step 2. In step 2, multivariate t rather than normal distribution is used in order to take the uncertainty of the between-study variance estimate into account.

Jackson's method is implemented and tested in the current study. Unfortunately, it is found to be still slow for moderate or high dimensions, mainly because of method of moments (MM) used in step 1. To overcome this constraint, an alternative step 1 method is proposed, which involves using weighted sample covariance matrix, subject to matrix regularization, to approximate the between-study variance/covariance. A series of simulation studies have shown that the improved two-step procedure performs favorably compared to the traditional MVMA methods as well as Jackson's original routine. Given these results, the new two-step procedure is applied to analysis of real bicluster stacks, which leads to a series of candidate gene sets.

The candidate gene sets are then analyzed in **Aim 3** by enrichment-based analyses using public pathway knowledge bases. The specific methods used include Over Representation Analysis (ORA), Gene Set Enrichment Analysis (GSEA), and Network Topology-based Analysis (NTA). **A key finding** is that high-certainty effect size estimates derived from MVMA are often associated with significant enrichment results from the pathway analysis, especially when the size of bicluster stack is big enough. In other words, effect size estimates are predictive of the biological relevance of the gene sets, which is perhaps the most significant result of the current study.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would first like to express my sincere gratitude to my dissertation committee for their continuous support. This work would not have been possible without their advice, discussions, and feedback. In particular, I would like to thank Dr. Peter Tarczy-Hornoch, who is my advisor and dissertation committee chair, for his ongoing guidance and assistance.

I decided to make a change in my career from software engineering to data science, with the understanding that it was not going to be an easy task. During the period of my graduate research, I went through some major obstacles and challenges. Peter has been providing me with unwavering support for which I am forever grateful. His guidance was instrumental in helping me to define my research topic and keeping me focused on the specific aims. In addition, I benefitted tremendously from my discussions with Peter, which allowed me to re-think and refine some of the key strategies and research methodologies.

I would also like to thank the Department of Anesthesiology for my RA position of working as a software engineer, which has been the source of financial support for my research and tuition coverage.

This study makes use of some algorithms that have long computation time. Thanks to the Shared Scalable Compute Cluster for Research (provided by UW IT) and the Microsoft Azure Cloud Services, I was able to complete the runs that were crucial for this dissertation.

Finally, I'd like to thank my family, particularly my wife, for their understanding, love, and support.

Chapter 1 Executive Summary

1.1 Overview

High-throughput technologies are among the most important developments in biology and medicine in recent decades. With the technical advancements, comprehensive measurements can be made to quantify various types of molecules, resulting in massive amount of high-resolution data for biological samples or patients.

This dissertation focuses on the gene expression data. In particular, microarray data are chosen as an example to illustrate the methodological framework. The microarray technique is a very well-developed and widely used high-throughput technique that allows expression levels of thousands of genes to be measured simultaneously. Routine use of the technique for investigating biological problems or diseases has produced enormous amount gene expression data. A large portion these data has been made available through public databases, allowing researchers and data mining practitioners to explore for knowledge discovery.

When multiple datasets are analyzed separately, it can be error prone due to a number of reasons: (1) high dimensionality of the data. For example, Affymetrix Human Genome U133 Set Plus, which is one of the most commonly used microarray platforms, contains probes for 47,000 different transcripts [2]. Experiments using such platform will produce very high-dimensional data; (2) small sample sizes. Human samples are rare and expensive to obtain, contributing to the small sample sizes in the studies; (3) low signal-to-noise ratio. Samples used in gene expression experiments often contain non-homogenous cell types. The results derived from such samples tend to have mixed expression profiles and low signal-to-noise ratios. All these factors make it very difficult to retrieve the true patterns from the individual datasets. A possible solution to

overcome the issues is to pool together multiple independent, but related datasets to increase the statistical strength.

A primary goal of mining of gene expression data is to discover gene sets that are functionally relevant. It is widely recognized that in biological systems, genes tend to work together by forming pathways or networks to initiate or sustain biological processes [3][4][5][6][7].

In light of the discussion above, this dissertation tries to answer the following two question: (1) what is the data mining method best suited for finding gene sets? (2) how to utilize multiple datasets jointly in order to increase statistical strength?

Some of most commonly used clustering methods have been proven to be effective in finding gene sets from gene expression data [8]. They include *k*-means clustering, hierarchal clustering, and self-organizing feature map (SOFM). However, they all suffer from two major drawbacks: (1) they form gene clusters by including all the samples to calculate the inter-gene distances. But in reality, genes are often related to each other in some subset, not all of the samples. Including all the samples may skew the distance estimation; (2) clusters are not allowed to overlap. It is not uncommon that the same genes participate in multiple biological processes, which will lead to overlaps of the gene clusters.

Biclustering is a newer clustering method, designed to overcome the above limitations. A bicluster is a rectangular submatrix with distinct statistical property, obtained by simultaneously clustering on both dimensions of a data matrix (hence the term “biclustering”). In the case of gene expression data, a bicluster is a subset of the genes exhibiting some coherent pattern in some subset of the samples. The term “coherent” can be defined differently, depending on the

biclustering algorithms that make different assumptions about the target patterns. It has been shown that different algorithms can lead to discovery of different gene sets on the same expression dataset [9][10]. The diversity of bicluster types is a significant advantage compared to the more traditional clustering methods discussed above.

To utilize multiple independent datasets, the current study adopts the approach of meta-analysis. Within the framework of meta-analysis, the random-effects model allows estimation of the overall effect size, while accounting for the heterogeneity between the studies. Furthermore, multivariate meta-analysis (MVMA) [11] permits use of multiple outcomes to assess an effect size, which opens the possibility of combining high-dimensional data patterns such as biclusters.

The current study is to apply meta-analysis to a group of biclusters. Each bicluster is analogous to an individual study in a clinical trial meta-analysis. The goal is to identify a gene set, through combining the evidence in the individual biclusters.

1.2 Motivation for this dissertation

To summarize from the above discussion, public data have not been fully explored or utilized. The target knowledge to be uncovered from public data in this case is gene sets because they may correspond to pathways which are the functional units in biological systems. Public research data have a number of features that make them difficult to tap into, including data heterogeneity and smaller sample sizes. Thus, designing a statistically sound framework to find knowledge in heterogeneous data sets is critically important for utilizing the data. This leads to the overall two questions that are already mentioned in the section above: (1) what is the data

mining method best suited for finding gene sets? (2) how to utilize multiple datasets jointly in order to increase statistical strength?

1.3 Research aims and questions

The goal of this dissertation is to devise a statistically sound procedure to extract functionally relevant gene sets from heterogeneous, publicly available gene expression data. The overall strategy is to first identify a large number of biclusters from selected independent datasets, identify overlapped biclusters to form bicluster stacks, then conduct meta-analysis on the bicluster stacks to combine the evidence of the embedded gene sets. The specific aims are as follows:

Aim 1: Determine optimal method for constructing bicluster stacks.

Research questions: Which biclustering algorithm has good potential of revealing real gene sets? What is the run time of the algorithm? What is the most appropriate effect size definition for the biclusters? How to stack biclusters and what is rationale behind the stacking scheme? How to estimate the individual effect sizes?

Specific sub-aims:

1. Evaluate a pool of biclustering algorithms and pick one of them that has good potential of finding real gene sets, can output consistent results on successive runs and identify overlapped biclusters.

2. Select a number of datasets related to a particular cancer type, generated using the same microarray platform but from heterogeneous sample types such as cell lines, patient tissues, and xenografts.
3. Identify individual biclusters and implement a brute-force search algorithm to construct bicluster stacks.
4. Propose and validate a method for measuring effect sizes for biclusters

Aim 2: Determine suitability of meta-analysis techniques to pool biclusters and assess performance.

Research questions: How the traditional MVMA methods perform in meta-analysis of biclusters? Can they perform satisfactorily when the sample sizes are small and the data heterogeneity is high? Do they scale well when dimension increases? If they have a performance issue in terms long computation time, are there alternative methods?

Specific sub-aims:

1. Evaluate traditional MVMA methods for bicluster meta-analysis by varying sample sizes, levels of heterogeneity, dimensions.
2. Assuming that the traditional methods are computationally too demanding when the dimension is relatively high, propose a more efficient alternative method. Then compare the new with the older methods.
3. Apply the new method to analysis of real bicluster stacks.

Aim 3: Assess potential utility of gene sets identified in Aim 2 using pathway analysis.

Research questions: Why is pathway analysis important? What are the major types of pathway analysis for gene sets, and what are their differences? What are the rationales behind the analyses, and how to interpret the results?

Specific sub-aims:

1. Perform Over-Representation Analysis (ORA) on the identified gene sets
2. Perform Gene Set Enrichment Analysis (GSEA) on the identified gene sets.
3. Perform Network Topology-based Analysis (NTA) on the identified gene sets.

The specific aims above are devised based on a two of hypotheses:

1. The CCS biclustering algorithm is effective in uncovering gene sets that may correspond to biological pathways.
2. Meta-analysis applied to multiple biclusters can increase the chances of finding true gene sets that are otherwise hidden in the individual dataset, while at the same time minimize the probability of getting false discoveries, which tend to occur with higher frequency when the individual biclusters are analyzed separately.

1.4 Outline of this dissertation

This dissertation is organized as below: Chapter 2. Background and significance. In this chapter, an overview is provided about the main challenges in analysis of high throughput gene expression, namely high dimensionality, data heterogeneity, small sample sizes, and in many cases data scarcity. Then, it will discuss identification of gene sets and why biclustering is chosen as the pertinent technique. Finally, it will introduce the idea of meta-analysis as a means for combining biclusters to overcome the above mentioned challenges.

Chapter 3. The focus of this Chapter is Aim 1: Determine the optimal method for constructing bicluster stacks. Specifically, it will describe Condition-dependent Correlation Subgroups (CSS), which is the chosen biclustering algorithm, and an implementation of a brute-force search scheme for constructing bicluster stacks. Then it will introduce and discuss effect sizes in meta-analysis and how to adapt the concept to biclusters. Finally, results will be presented about the individual biclusters identified, the stacks of overlapped biclusters, and characterization of the bicluster-level effect sizes.

Chapter 4. This chapter focuses on Aim 2: Determine suitability of meta-analysis techniques to pool biclusters and assess performance. It will first consider multivariate meta-analysis and how it can be adapted to biclusters. Traditional multivariate meta-analysis (MVMA) methods will be evaluated for combining biclusters. Their limitations will be highlighted, particularly their abilities in dealing high dimensionalities. Then, an alternative two-step MVMA method will be proposed, implemented and evaluated. Specifically, the two-step method will be assessed through simulation studies, and applied to analysis of real bicluster stacks.

Chapter 5. This chapter focuses on Aim 3: Assess potential utility of gene sets identified in Aim 2 using pathway analysis. A challenge is the absence of a gold standard (e.g. biological or external validation). Proxy approaches are used. The chapter consists of two parts: first, the biological relevance of the gene sets will be assessed using three popular types of pathway analyses. Second, the relationship between the effect size estimates of the gene sets and their pathway enrichment outcomes will be examined. The goals of the second part is to inspect whether the effect sizes can be used to predict the biological pertinence for the gene sets.

Chapter 6. Conclusion. This final chapter will summarize the results from the previous chapters, highlight the significance and potential contributions. In addition, it will discuss the

limitations of the proposed method or framework, and provide some suggestions on future research directions.

1.5 Contributions

The current study may bring about a number of contributions to the field of bioinformatics. First, to the best of my knowledge, it is the first attempt to apply multivariate meta-analysis to biclusters, which thus a) allows utilization of multiple datasets to increase the statistical strength and leverage of biclustering to identify significant gene sets, and b) allows estimates of the effect sizes and their confidence intervals for the member genes in a gene set. Our results suggest that the effect size estimates can be used to predict the biological relevance of the gene sets, which is the most significant finding of the current study.

Second, methodologically, multivariate meta-analysis (MVMA) has been gaining popularity in many domains, but it has been mainly applied to situations of low dimensionality due to its computational demand. A previously published two-step procedure allows MVMA to be applied to high dimensional data, but it is based on a biased estimator. The current study proposes an alternative estimator that leads to a notable improvement on meta-analytic performance, which is significant given the usual small number of datasets available.

Third, Gene Set Enrichment Analysis (GSEA) has become a major pathway-based analysis for assessing gene sets. It requires the candidate gene sets to be ranked typically by degree of differential expressions. To the best of my knowledge, the current study is the first attempt to rank the gene sets by effect size estimates, which opens the possibility of enriching the

gene sets based on evidence that comes from different studies and weighing it by the strength of this evidence.

Chapter 2 Background and significance

This chapter provides an overview of high-throughput gene expression data, describes the goals of the data analysis, highlights the difficulties and challenges, and proposes to use biclustering and meta-analysis to identify significant gene sets from the vast amount of available research data.

2.1 Overview of high-throughput omics technologies

Thanks to decades of advancement in "omics" technologies, high-throughput experiments are now routinely performed in laboratories and increasingly in hospitals. What makes the technologies unique is that they provide a comprehensive view of the molecules that constitute a cell, tissue, or organism. They aim to universally quantify genes, mRNA, proteins and metabolites in a given sample in a non-targeted and non-biased manner. As a result, they are also referred to as high-dimensional biology [12][13].

Omics technologies can be categorized into a number of specialty areas, including genomics, transcriptomics, proteomics, and metabolomics, etc. They aim to profile a sample by focusing on various molecular aspects. More specifically, next-generation sequencing (NGS) can sequence a human genome within a day [14][15]. It has transformed genomic research and promises to facilitate precision medicine. The genome-scale expression analysis allows the expressions of thousands of genes to be quantified on a "gene chip", [16][17]. Similarly, proteomics [18][19] and metabolomics [20] focus on measurements of whole sets of proteins and metabolites, respectively, present in an organism, cell, or tissue.

Omics techniques have transformed how we conduct biological research. Traditionally, biological research is largely hypothesis driven or reductionist, meaning that the researchers start with a specific hypothesis based on some prior knowledge, then conduct experiments to verify, refute, or expand the hypothesis. With the omics technologies, researchers can now pursue holistic investigations that are hypothesis generating. In such studies, there are no prescribed hypotheses. Instead, large amount of data are acquired to define hypotheses that are subject to further testing [21].

Omics technologies can be applied to medical processes by playing a role in disease screening, diagnosis and prognosis, as well as helping to enhance our understanding of diseases aetiology [22][23]. In addition, they have found themselves increasingly being leveraged in drug discovery and assessment of drug toxicity and efficacy [24][25]. Furthermore, pharmacogenomics, which is the intersection of pharmacology and genomics, is becoming increasingly important in biology and medicine[26].

Omics technologies promise to transform biological research and medicine. However, its power hinges upon proper translation from data to knowledge. The following sections will discuss data analysis and the associated challenges. This dissertation is concerned with mining of gene expression data. As an example of gene expression data, microarray data are chosen to illustrate the methodology. Therefore, the discussion below focuses on analysis of microarray data.

2.2 Gene expression analysis and its applications

As aforementioned, high throughput expression analysis allows expressions of the genes in a whole genome to be measured using a "gene chip". The power of the technique lies in comprehensive survey of the expression levels of all the genes in the cells, which provides molecular snapshots at the transcriptional level. For example, by comparing the transcriptional profiles between a normal and cancer samples, the snapshots can lead to insight about the transcriptional disruption during cancer formation, which has important impacts on cancer research and diagnosis [27][28]. If the experiments are performed using time series samples, the revealed evolution of the transcriptional profiles can shed light on infection, disease development, drug mechanism, etc.[29][30].

In a high throughput expression experiment, mRNA is extracted from a sample of interest and reverse-transcribed into cDNA. The cDNA is then fluorescently labeled and hybridized to an array of known sequence embedded on a chip. The expression levels of the genes are then determined based on the hybridization intensities. A typical gene expression experiment usually includes multiple conditions, for example: normal, cancer, drug-treated cancer, etc. Each condition is often repeated a number of times using separately prepared samples, resulting in biological replicas.

The intensity numbers from all the conditions and replicas, after pre-processing and normalization, are then pooled together to form an expression matrix. By convention, the gene probe sets are listed vertically while the samples are organized horizontally. Thus, an entry E_{ij} in the matrix represents the expression value for gene probe set i in sample j. A row in the matrix presents the expression profile of the gene probe set across the conditions. Similarly, a column

represents the expressions of all the probe sets under that condition. Table 2.1 below illustrates such an expression matrix.

	Condition 1			Condition 2			Condition 3			Condition 4			Condition 5			Condition 6		
Gene Probe	**Sample 1**	**Sample 2**	**Sample 3**	**Sample 1**	**Sample 2**	**Sample 3**	**Sample 1**	**Sample 2**	**Sample 3**	**Sample 1**	**Sample 2**	**Sample 3**	**Sample 1**	**Sample 2**	**Sample 3**	**Sample 1**	**Sample 2**	**Sample 3**
1053_at	8.9488	9.0048	9.0021	8.9801	9.2193	9.1141	9.1266	8.5259	8.9765	8.8499	9.1652	9.0451	9.2299	9.2642	9.4003	9.2552	9.1299	9.1441
1431_at	6.9138	7.0488	6.7472	7.0089	6.7667	6.6923	6.5765	6.4109	6.5485	6.7888	6.7803	6.8752	6.8329	6.9005	6.8998	7.0309	6.8169	6.9999
1552291_at	10.4692	10.5245	10.6276	9.8142	10.2057	10.0157	10.8184	10.882	10.3815	10.2032	10.6869	10.2512	9.6663	9.9992	10.063	9.9403	10.0641	10.0547
1552296_at	6.6378	6.2793	6.2489	5.8467	6.0876	5.509	5.7667	5.8232	6.0484	6.2183	6.1182	6.1994	6.0531	6.1569	6.0348	6.1772	6.0745	6.1772
1552329_at	9.4387	9.3747	9.3399	9.3668	9.598	9.1955	9.1568	9.2356	8.8514	8.6526	9.016	8.8932	8.8999	8.8503	9.0664	9.115	8.9545	8.8158
1552359_at	6.0142	6.0478	6.0655	5.8444	5.9659	5.8362	5.7977	5.5141	5.9309	5.8448	5.9049	6.1571	6.0149	6.0409	6.2784	6.3431	6.1955	6.0775
1552370_at	8.7844	8.9388	8.6811	8.3316	8.4638	8.3779	8.637	8.6425	8.5252	8.4942	8.7559	8.5891	8.3177	8.3759	8.5906	8.678	8.4499	8.4959
1552422_at	6.6923	6.7726	6.7511	6.4414	6.7162	6.842	6.6441	7.0107	6.3568	6.6102	6.8219	6.5799	6.3845	6.2136	6.332	6.1408	6.3779	6.5513
1552427_at	6.5368	6.5521	6.7076	7.0989	7.5457	7.2845	6.89	7.0742	6.7731	6.6464	6.9148	6.7025	7.1443	7.4309	7.6628	7.1375	7.2246	7.1847
1552582_at	4.5671	4.6955	4.6397	4.8554	4.8828	5.0397	4.9133	5.2877	4.8245	4.939	5.0205	4.9393	4.8698	4.9656	5.1732	5.1772	4.9377	5.0985
1552587_at	6.8462	6.9465	6.8345	5.6157	5.1816	6.0386	7.3091	7.0357	7.1287	7.1087	7.2533	7.0887	6.2315	5.9932	6.1958	6.4135	6.3245	6.5052
1552612_at	10.0786	10.2831	10.3164	9.7643	9.2788	9.9622	10.1582	9.9162	10.3282	10.017	10.2683	10.1916	9.9799	9.919	10.0282	10.2062	10.1033	10.2165
1552673_at	4.7755	4.732	4.7597	5.2354	5.087	4.9537	4.718	4.5768	4.5177	4.8092	4.5911	4.7655	5.0558	4.8289	4.6003	4.656	4.6857	4.7026
1552729_at	7.6207	7.7148	7.5514	7.7336	7.7269	7.7363	7.6171	7.3883	7.6413	7.8076	7.7427	7.8159	8.0587	7.9251	7.8327	7.9758	8.0252	7.8098
1552742_at	9.6522	9.6102	9.751	9.8188	9.5892	9.7222	9.4555	9.3234	9.3881	9.0504	9.3111	9.402	9.8506	9.8209	10.0776	9.9379	9.9719	9.8307
1552766_at	5.0494	5.1188	5.1644	4.4692	4.5067	4.5021	4.8904	4.586	5.3821	5.3912	5.1817	5.2951	4.8502	4.7475	4.9454	5.2315	4.9083	5.0032
1552835_at	7.6928	7.4803	7.4719	6.9666	7.2514	6.8865	7.7427	7.7996	7.743	7.7884	8.0882	7.6245	7.2868	7.2104	7.0333	7.3167	7.2362	7.2704
1552878_at	4.647	4.4119	4.6034	4.2473	4.1922	4.3822	4.5852	4.5335	4.6483	4.8975	4.6795	4.6704	4.6501	4.593	4.5882	4.7069	4.597	4.766
1552904_at	9.6403	9.7707	9.8334	8.0505	8.4342	8.0794	9.9662	10.0123	9.692	9.3298	9.7872	9.6714	8.4884	8.825	9.1842	9.1121	8.9981	8.9739
1552919_at	5.3075	5.6037	5.4079	4.6046	4.4985	5.1119	5.2139	4.6113	5.1045	5.4056	5.2901	5.4581	5.3304	5.0741	5.7532	5.6755	5.377	5.2851
1552927_at	5.8682	5.718	5.9294	6.1783	5.7354	6.1482	5.2983	5.5793	5.8419	5.7093	5.7451	5.927	6.2773	6.346	6.4786	6.4677	6.1673	6.1726
1552930_at	9.4949	9.4656	9.5141	10.0983	10.1928	10.0062	9.712	9.6784	9.458	9.5086	9.7378	9.5199	9.9768	9.9568	9.9093	9.7495	9.756	9.5584
1552979_at	5.3253	5.1277	5.3102	5.5995	6.072	5.786	5.4541	5.6285	5.5327	5.3992	5.4541	5.5688	5.561	5.6789	5.5857	5.5059	5.4527	5.5893
1553011_at	6.6792	6.791	6.5951	6.4692	6.1694	6.2593	6.3903	6.4595	6.236	6.3018	6.3079	6.4265	6.3632	6.2471	6.4935	6.132	6.494	6.5161

Table 2.1: Illustration of microarray data from dataset GSE44905

Traditionally, the goal of a gene expression experiment is to detect individual genes that exhibit differential expression between states or samples. Now the focus has largely switched to detection of sets of genes, based on the understanding that pathways are the actual functional units, as discussed in greater details in 2.3 below.

2.3 Pathways and gene sets as functional units

This section starts to address the first question: what is the data mining method best suited for finding gene sets? Aim 1 will delve more into this. Specifically, Aim 1 is determining the optimal method for constructing bicluster stacks, which is the focus of Chapter 3.

It is widely recognized that genes do not act alone; instead they tend to work together to form pathways to underlie biological processes such as DNA repair, cell division and cell cycle,

cancer metastasis, etc. As a result, there has been a great deal of interest in discovering gene sets using high-throughput expression analyses [3][4][5][6][7].

In biological systems, how the genes interact with each other can be dauntingly complex. Functionally related genes may form pathways, co-locate in a common cellular component, or share chromosomal locations. There are many possible ways by which the genes may interact with each other. A biological process can have a large number of underlying genes, forming a complex regulatory network. Furthermore, the topology of the network may dynamically change, depending on the pathophysiological state of the system.

It has been shown that many diseases are associated with modest changes in expression of groups of genes, rather than a dramatic increase in individual genes [31][32]. Hence, it is important to be able to detect subtle changes in activities of groups of genes in order to identify the gene sets.

So far, a large body of pathway knowledge has been accumulated, which has led to development of pathway databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG)[33], Reactome Pathway Database [34][35], Gene Ontology [36][37], The Biogrid [38], etc. These repositories are now widely used as the basis for knowledge-based analyses of gene sets extracted from data. It is important to point out that the analyses should not be seen as a gold standard, because the knowledge may be incomplete and in some cases inaccurate. Matching a gene set with existing pathways does not necessarily validate the gene set.

Gene set mining may uncover previously unknown pathways or new variants of existing pathways that have no match with current knowledge bases. In such cases, the only way to validate the gene sets is through additional lab research.

To summarize, it is desirable to identify gene sets for knowledge discovery and for detection of diseases-related changes. However, tapping into public data is not a readily achievable task, due to a number of difficulties and challenges as discussed below in 2.4.

2.4 Challenges in analysis of public expression data

Technical improvements and commercialization has enabled gene experiment analyses to be performed effectively and inexpensively. Despite the power of the technology, analyzing the data to identify the real and meaningful patterns remains a challenge. The primary constraints include: (1) high dimensionality (*p*) arising from the need to consider a large number of genes, (2) relatively small sample size (*n*) due to scarcity of the samples and the high cost of obtaining and processing the samples. In most cases, *p* is much larger than *n*, resulting in high rate of false positives in pattern recognition. This $p >> n$ problem, commonly known as "curse of dimensionality", is prevalent in omics studies, and has become a major area of research in biostatistics and bioinformatics [39][40].

Data generated from omics research are often deposited to some of the public databases by the authors. In the case of gene expression data, major databases include Gene Expression Omnibus (GEO) [40], ArrayExpress [41], Stanford Microarray database[42], etc. These large data sources may give a false impression of unlimited abundance of data. But in reality, most datasets are small and highly heterogeneous, thus creating data scarcity.

Previous studies have shown that sample size affects expression analysis results; small sample sizes result in unstable gene lists and poor prediction accuracy [43][44].

Besides the small sample sizes, another issue with public data is heterogeneity. The datasets are derived by different laboratories for investigation of different problems. In addition, the types of samples used, the analysis platforms, and the statistical procedure used to process the data, etc., all contribute to the data heterogeneity. Thus, simply merging the datasets to enlarge the sample size is not generally applicable.

The Table 2.2 below lists the GEO datasets used in the current research, which are all related to prostate cancers. Despite the fact that they are related in terms of general area of research, the small sample sizes and heterogeneity are obvious. Small datasets like these are common in GEO. They are selected because they share a common general topic of research.

GEO dataset	Title of the primary research	Sample size
GSE7868	Expression data from LNCaP cell line	9
GSE17044	Expression data from androgen treated LNCaP cells	6
GSE22483	Hormone-independence of prostate cancer cells is supported by the androgen receptor without binding to classical response elements	6
GSE22606	Identification of an SRF- and androgen-dependent gene signature in prostate cancer	12
GSE29232	Identification of androgen-regulated genes in RWPE-1-AR cells	12
GSE34589	Elk1 directs a critical component of growth signaling by the androgen receptor in prostate cancer	8
GSE44905	Expression data from LNCaP cells treated with DHT and enzalutamide	18
GSE56908	LNCaP and C42B LSD1 knockdown microarray gene expression data and C42B androgen (DHT) stimulation microarray gene expression data	4
GSE3325	Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression	19
GSE7708	Suppression of androgen receptor mediated gene expression by a sequence-specific DNA binding polyamide	14
GSE51524	LNCaP prostate cancer cell lines overexpressing wild-type or GARRPR-mutant Bag-1L	18
GSE55945	Gene expression profiling of prostate benign and malignant tissue	21
GSE94580	Discovery and mechanistic characterization of A-485, a potent p300/CBP catalytic inhibitor	22

Table 2.2: Prostate cancer related datasets from GEO chosen for the current study

Given the small sample sizes and heterogeneity, the task is to uncover the true knowledge hidden in some of these datasets, which will serve as a demonstration for the methodology framework developed in the current study. The following section 2.5 intends to review some of the most commonly used mining methods for gene expression data, which will form the basis for the work done in Aim 1 (determining the optimal method for constructing bicluster stacks). Section 2.6 introduces biclustering which addresses the limitations of the commonly used mining methods for gene expression data.

2.5 Commonly used data mining methods and their limitations

In most cases, little is known about the data prior to the data analysis. The researcher usually aims to discover "interesting" rather than real patterns because of lack of knowledge about the ground truth in the data. This type of data analysis falls into the category of unsupervised learning or data mining. A variety of unsupervised learning methods have been applied to gene expression data. In this section, I will review three prominent clustering algorithms: *k*-means clustering, hierarchical clustering, and self-organizing map (SOM). They can all potentially lead to discovery of gene sets, but unfortunately they suffer from some severe limitations, as summarized at the end of section 2.5.

2. 5.1 K-means clustering

The *k*-means algorithm aims to partition *n* data points into *k* (usually user-defined) clusters based on some distance measure. It is among the simplest, fastest, the most widely used clustering algorithms [45][46][47][48].

Figure 2.1 below illustrates the *k*-means clustering using simulated data with three clusters of Gaussian distribution (http://pypr.sourceforge.net/kmeans.html#k-means-example).

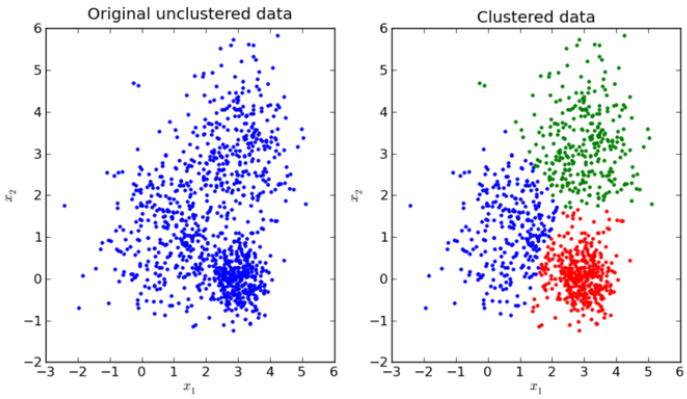


Figure 2.1: Illustration of k-means clustering

The algorithm takes the number of clusters, *k*, as a user input and starts by randomly choosing *k* data points as the initial centers of clusters. The process continues by alternating between two steps until convergence. The first step is to take all the data points and calculate their distances to the cluster centers. The distances are then used for allocating each data point to a cluster. The second step is to adjust the centroid for each cluster based on the current data point allocations. The coordinates of new centroid is usually the mean of the coordinates of the members. Since the centroids have been shifted now, it is necessary to repeat step one by recalculating the distances between the data points and new centroids, and re-assigning each data point to the clusters. The alternation of the two steps continue until convergence, which is when the cluster centroids are settled such that no data point moves from one cluster to another.

There are a number of quality measures for *k*-means clustering. They are: (1) the sizes of the clusters versus inter-cluster distances; (2) distances between the members of a cluster and cluster center; (3) the diameter of the smallest sphere that includes all member of a given cluster.

The complexity of the *k*-means clustering algorithm is O(c*N), where N is the number of genes, and c is a constant that depends on *k*. This means that the computation time is linear in the number of genes, which explains the high efficiency of the algorithm.

K-means clustering has both pros and cons. Its main advantages are simplicity and computational efficiency. However, a problem with *k*-means clustering is that the result may depend on the choice of the starting cluster centers, and thus can change between successive runs of the algorithm. To minimize this pitfall, a common practice is to choose the starting centers in areas that are more densely populated. Another problem with *k*-means clustering is that that number of *k* has to be supplied by the user before the run. The choice of *k* is often ambiguous, and the run will always dutifully produce *k* clusters. Thus, improperly chosen *k* will lead to incorrect clustering results. Many studies have been devoted to estimating a number of clusters in the data [49][50].

2.5.2 Hierarchical clustering

Hierarchical clustering [51][52][53][54] has been applied to expression data since the inception of the microarray technique [48]. When applied to either the vertical or the horizontal dimension of the data matrix, it produces a tree-like hierarchical structure in which the leaves are the genes or the samples, and the branches represent the groupings of the objects at the lower levels. How the objects are grouped depends on the distance metric used. Figure 2.2 below shows an example of a hierarchical clustering result from the combined data of three studies [55]. As expected, it shows three distinct clusters horizontally as a result of the three studies. Vertically, the genes are clustered into two distinct groups in each study.

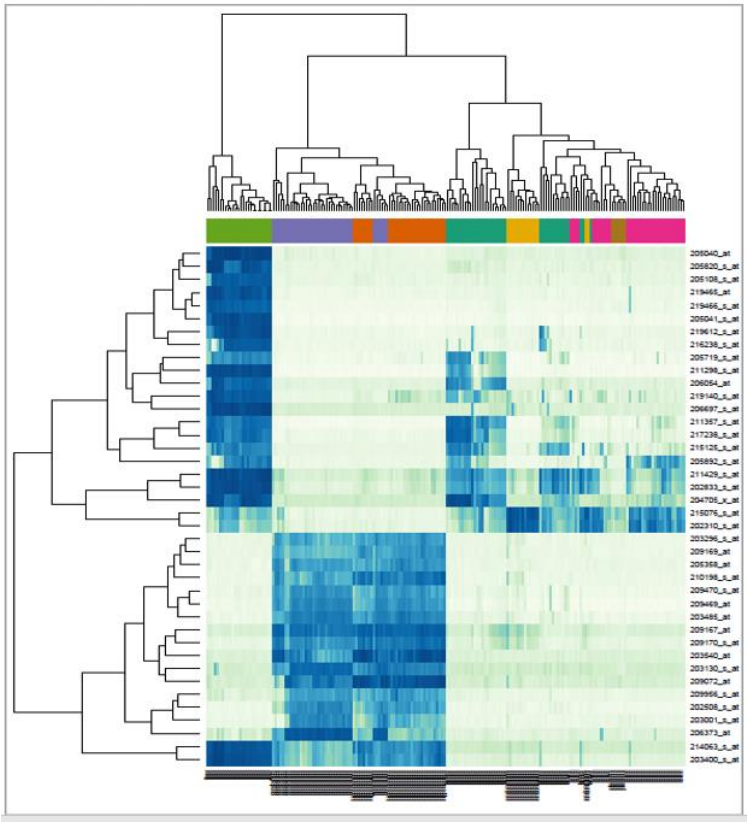


Figure 2.2: A hierarchical dendrogram that combines the microarray results

The tree of hierarchical clustering can be constructed either in a top-down or a bottom-up fashion. The former tends to be faster. The complexity of the top-down approach requires between $n\log(n)$ and n^2 computations, while bottom-up requires between n^2 and n^3 steps.

Hierarchical clustering has both pros and cons. As discussed earlier, the result of *k*-means clustering is *k* groups of genes. All these groups, as well as all elements within each group, are on the same level. No information about the relationship between any two groups is available. In contrast, hierarchical clustering produces hierarchically organized clusters. In addition, it avoids the problem of estimating the value of *k*, because no assumption is made about the number of clusters.

Another difference is that hierarchical clustering algorithm is completely deterministic, meaning that when applied to a given dataset using the same distance metric, hierarchical clustering will always produce the same tree. On the contrary, *k*-means clustering is stochastic: successive runs may produce different results. Nevertheless, hierarchical clustering shares a major disadvantage with k-means clustering as summarized below in 2.5.4.

2.5.3 Self-organizing feature map (SOFM)

The self-organizing feature map (SOFM), also called Kohonen map, was proposed by Finnish professor Teuvo Kohonen in the 1980's [56][57][58][59] . The SOFM works by mapping the input space into a feature space in which the neighborhood relationship reflects the degree of similarity between the original data points. After the SOFM is constructed and the mapping is done, the distances and relationships measured on the feature map are proportional to the distances and the relationships between the original data points according to the similarity metric chosen.

The SOFM is actually a neural network. It differs from other artificial neural networks in that it utilizes competitive learning as opposed to error-correction learning because of lack of labeled data (unsupervised learning). The map space is defined before the training starts, usually as a finite two-dimensional plane in which the nodes are arranged as a grid. The grid evolves as weights for the coordinates of the nodes are updated when sample vectors are iteratively fed to the network. The node corresponding to the best matching unit (BMU), determined by competitive learning, gets its weight updated the most. The neighboring nodes are updated as well, but to a lesser extent. As a result, the topology of the network is maintained during the training.

Similar to the other clustering methods, SOFM has both advantages and disadvantages. First, SOFM preserves the information about the relationships and reciprocal positions of the data points in the original input space, which is not the case with k-means or hierarchical clustering, although hierarchical clustering is more informative than *k*-means.

The SOFM has drawbacks too. First, it is not deterministic. Different initiations can result in different clustering results. The learning can be sped up by initializing the weights of the nodes by sampling the subspace spanned by the two largest principal component eigenvectors. Second, the size of the feature map (number of nodes and their layout) has to be chosen beforehand. Therefore, the training has a heuristic component, which is often tackled by trial and error.

2.5.4 Limitations of the traditional clustering methods

Despite the usefulness of the above discussed methods, they share a number of drawback. First, they consider all the samples when clustering the genes and vice versa. It has been shown that genes may not be active or triggered by all the conditions, and thus, they are not necessarily correlated to every sample [60]. Using all the samples to evaluate a set of genes may mask the significance of the target gene set and thus reduce to chance of uncovering it. To avoid this pitfall, it is necessary to cluster the genes and the samples simultaneously.

Second, another limitation with the traditional clustering methods is that they assign each gene exclusively to the non-overlapped groups. But in reality, a gene may belong to several clusters because it can participate in multiple pathways, depending on the cellular condition or the experimental context [61].

The approach of biclustering is meant to address these limitations, with the goal to uncover gene sets from the data, as discussed in the next section. The key benefit of biclustering is that it clusters the dataset on both the horizontal and vertical dimensions, allowing identification of a group of genes that are active in a subset of samples.

2.6 Biclustering

In this section, I will first give an overview of biclustering, followed by a discussion on diversity of biclustering algorithms. At the end, I will make the case that biclustering is a powerful technique for identifying gene sets.

2.6.1 Overview of biclustering

As described above, in the context of gene expression data, a bicluster can be defined as a subset of genes that show coherent behavior in a subset of samples. It exhibits as a rectangular submatrix whose rows and columns are not necessarily adjacent to each other. It is identified by simultaneously clustering the rows and columns of a 2-dimensional data matrix. As a data mining technique, biclustering is also called block clustering, co-clustering, or two-way clustering.

The idea of biclustering was originally put forward by J. A. Hartigan in 1972 [62]. Y. Cheng and G.M. Church first applied the technique to gene expression data. They define a bicluster as a low variance submatrix, and use a deterministic greedy algorithm to search for biclusters with low variance as defined by a threshold mean squared residue (MSR) [63]. In the past two decades, biclustering has been attracting high research interest within data mining

community. As a result, a large number of biclustering algorithms have been proposed and implemented. Many of them have been applied to expression data to search for gene sets. These algorithms define the bicluster patterns differently, and often employ different search schemes to find the target patterns.

Besides gene expression data, biclustering has also been applied to text mining for classification [64]. In that case, the rows of the data matrix represents documents, while the columns denote the words in the dictionary. The data entry D_{ij} denotes occurrence of word j in document i. The goal of the biclustering procedure is to find blocks in D that correspond to a set of documents characterized by a set of words.

2.6.2 Diversity of biclustering algorithms

Biclustering as a data mining method can identify statistically diverse bicluster patterns, due to a large number of available algorithms that make different assumptions about the structures of the target patterns. This is another key advantage of the technique compared to the more traditional clustering methods. In this section, I will provide examples to illustrate the diversity.

2.6.2.1 Cheng and Church Algorithm

The algorithm proposed by Cheng and Church is among the most cited ones because of its original use in gene expression data. The pattern is defined by mean squared residue (MSR) [63]:

$$MSR(\mathcal{B}) = \frac{1}{|I| \cdot |J|} \sum_{i=1}^{|I|} \sum_{j=1}^{|J|} (b_{ij} - b_{iJ} - b_{Ij} + b_{IJ})^2$$

where b_{ij} is the expression value for gene i at sample j; b_{iJ}, b_{Ij} and b_{IJ} are the means for row i, column j, and the whole bicluster respectively. MSR has been shown to be useful for identifying constant biclusters, constant row and column biclusters, and shift biclusters. However, this metric tends to miss scale and shift-scale biclusters [65][66].

2.6.2.2 Order-preserving submatrix problem (OPSM)

The OPSM algorithm [67][9][10] assumes a bicluster as an order-preserving submatrix, in which there exists a linear ordering of the columns shared by all the rows, as illustrated in Table 2.3 below:

Gene \ tissue	t_1	t_2	t_3	t_4	t_5
g_1	7	13	19	2	50
g_2	19	23	39	6	42
g_3	4	6	8	2	10
Induced permutation	2	3	4	1	5

Table 2.3: Illustration of the OPSM algorithm

In the above table, for each gene, the expression values follow the same permutation of (2, 3, 4, 1, 5). OPSM aims to find such pattern using a deterministic greedy search scheme. Since constant columns, shifting, scaling and shift-scale bicluster models are all order-preserving, OPSM is advantageous over the Cheng and Church algorithm. However, OPSM has an obvious drawback: it ignores the original expression values, which may represent a significant information loss.

2.6.2.3 The plaid model

The plaid model [68][69][9] is among the most flexible biclustering models. It simulates the biclusters as layers that can be overlapped, as illustrated in the following diagram:

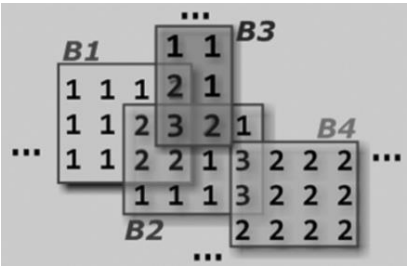


Figure 2.3: Illustration of the plaid model

More specifically, it assumes that each expression value is the sum of $k \geq 0$ layers representing *k* biclusters, plus a background layer and some Gaussian noise. The formulation is expressed by the following equation:

$$X_{ij} = \theta + \sum_{k=1}^{K} (\mu_k + \alpha_{ik} + \beta_{jk}) \rho_{ik} \kappa_{jk} + e_{ij},$$

where θ is the background effect, μ_k is the effect from bicluster, α_{ik}, β_{ik} represent the row and column effects for bicluster *k*, respectively. The indicator $\rho_{ik} = 1$ if and only if gene i belongs to bicluster k. Similarly, $\kappa_{jk} = 1$ if and only if sample j belongs to bicluster k. Finally, e_{ij} is the noise. The algorithm fits this model by iteratively updating each parameter to minimize the mean squared errors (MSE) between the expected values and the true values.

A key advantage of the plaid model is that it allows different biclusters to overlap because each bicluster is treated as a layer. In addition, positive and negative column effects are

allowed to reflect up- or down-regulation, respectively. The row and column effects combined can accommodate the shifting and scaling patters, thus providing maximum flexibility. However, this algorithm requires substantial efforts in parameter tuning. Moreover, slightly different parameter settings would likely result in different clustering results.

2.6.2.4 Correlated Pattern Biclusters Algorithm (CPB)

The bicluster pattern sought by CPB is characterized by high row-wise correlation according to the Pearson Correlation Coefficient (PCC) [70][9][66]:

$$PCC = \frac{\sum_{j \in J}(a_{ij} - a_{iJ})(a_{\ell j} - a_{\ell J})}{\sqrt{\sum_{j \in J}(a_{ij} - a_{iJ})^2 \sum_{j \in J}(a_{\ell j} - a_{\ell J})^2}} \quad (1)$$

where α_{iJ} and α_{lJ} respectively denote the means of rows *i* and *l* over the columns in the bicluster. The assumption is that the pair-wise correlations in expressions of the genes can be modeled by PCC.

It has been shown that this model can capture shifting, scaling, and shift-scale patterns, but it is ineffective in discovering constant biclusters or constant row patterns [9]. Furthermore, the algorithm assumes genes belonging to the same pathway are linearly correlated in expression, but in reality they may be correlated in a non-linear or even anti-correlated fashion.

2.6.2.5 Biclustering by Gibbs sampling

Gibbs sampling is one of the best known Markov chain Monte Carlo methods. An algorithm was proposed to identify bicluster patterns using Gibbs sampling [71][72][9]. It starts

with discretization of the expression data into multiple (l) bins. Two vectors of Bernoulli random variables are used to indicate whether a gene or a sample belongs to a bicluster:

$$g = \begin{bmatrix} g_1 & g_2 & \cdots & g_n \end{bmatrix}^T$$
$$\text{and } c = \begin{bmatrix} c_1 & c_2 & \cdots & c_m \end{bmatrix}^T,$$

A bicluster is modeled using a mixture of multinomial distributions:

$$\Theta = \begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,j} & \cdots & \theta_{1,w} \\ \theta_{2,1} & \cdots & \theta_{2,j} & \cdots & \theta_{2,w} \\ \vdots & & \vdots & & \vdots \\ \theta_{l,1} & \cdots & \theta_{l,j} & \cdots & \theta_{l,w} \end{bmatrix}$$
$$0 \le \theta_{i,j} \le 1, \sum_i \theta_{i,j} = 1$$
$$\text{for } i = 1, \ldots, l;\ j = 1, \ldots, w$$

where w denotes the total number of samples in the bicluster. A Gibbs sampling procedure is then carried out to infer the parameters within the Bayesian framework with conjugate priors.

An advantage with this algorithm is the ability to allow prior knowledge to be incorporated into the prior distribution, but it comes with a high computational cost.

2.6.3 Summary

To summarize this section, bicluster as a data pattern includes a variety of sub-patterns with diverse statistical properties. Different biclustering algorithms seek different sub-patterns based on the assumed statistical structures of the target sub-patterns. Biclustering is advantageous over the traditional mining methods because it can simultaneously cluster the data on both dimensions. The biclustering algorithm chosen for this study is called Condition-dependent Correlation Subgroups (CCS), which will be thoroughly discussed in Chapter 3.

Pontes et al. survey a large number of biclustering algorithms and the quality measures used [73][74]. They classify these methods into two broad categories: those based on evaluation measures, and those that are non metric-based. A subset of these methods are listed in Appendix I.

Tanay et al. [75] proved that biclustering is an NP-hard problem, and thus much more complex than the traditional clustering methods discussed above. Therefore, most of the proposed algorithms are based on optimization procedures as the search heuristics.

Despite the fact that biclustering has been successfully applied to analysis of gene expression data, significant gap remains. When the datasets are small, the biclusters found have a high tendency of being false positives. The issue of how to minimize the false positive rate has not been sufficiently addressed.

The goal of this dissertation is to discover gene sets from gene expression data that are biologically meaningful. Biclustering is a naturally suited for this task for the following reasons. First, as discussed above, a gene set may be active only in a subset of the samples or subjects in an expression study. Biclusters exactly model such behavior.

Second, in biological systems, the inter-gene relationships can be dauntingly complex. There are many possible ways by which the genes interact with each other. These diverse relationships may manifest themselves as biclusters with dissimilar statistical properties on an expression dataset. The large number of available biclustering algorithms can serve as a powerful tool to recognize the patterns and to extract the gene sets. In fact, it has been shown that different biclustering algorithms tend to pick up different gene sets from the same expression datasets [9][10].

For the above two reasons, biclustering is chosen to be the approach in the current research for finding functionally related genes in gene expression data. It addresses the first of the two overall questions mentioned above: what is the data mining method best suited for finding gene sets?

One might be concerned that biclustering adds nothing new compared with the traditional clustering methods such as k-means clustering. First, as discussed above, biclustering aims to uncover "local" patterns as opposed to more "global" patterns identified by k-means, hierarchical clustering, or self-organizing feature map, as shown in Figure 2.4 below. Since a pathway may respond to a subset of the conditions, it is likely to exhibit a local pattern on a gene expression dataset. Second, an experiment can be readily done to show the utility of biclustering by using an artificial dataset with a known bicluster implanted. A biclustering algorithm should be able uncover the bicluster, while k-means may fail, especially when the noise level is high. Therefore, biclustering can indeed add something new compared to the traditional clustering method.

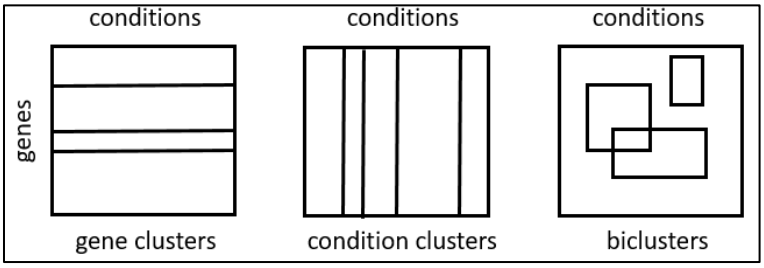


Figure 2.4: Biclusters vs. gene or condition clusters

The second overall question, which is how to utilize multiple datasets jointly in order to increase statistical strength, is addressed in section 2.7 below. As a reminder, the second overall question is what Aim 2 (determining suitability of meta-analysis techniques to pool biclusters and assess performance) tries to solve.

2.7 Approaches to combining biclusters

This section discusses utilization of multiple datasets by combining the biclusters found in the datasets. In section 2.7.1 below, a review is first provided to describe some of the major methods for utilizing multiple relevant datasets.

2.7.1 Current approaches for utilizing multiple datasets

Making use of multiple datasets in order to enhance the evidence for the common truth shared by the datasets is a challenging problem in statistics. In broad sense, there are two approaches: integrative data analysis (IDA) [76][77] and synthesis of summary statistics (SOSS) drawn from multiple studies [78][79][80].

In [77], IDA is defined as "the statistical analysis of a single data set that consists of two or more separate samples that have been pooled into one". The motivating example used was the Cross Study funded by NIH. The project was to investigate how parental alcoholism impacts the development of children by pooling three separate studies together. The authors argue that IDA offers a number of advantages, including:

1. IDA provides a direct mechanism to test whether the same findings replicate across independent studies;
2. IDA can potentially increase the statistical power for testing research hypotheses by pooling multiple datasets together.
3. IDA often provides more heterogeneous pooled samples, which may overcome the issue of under-representation associated with individual studies.

4. IDA can encourage more efforts in data sharing.

However, IDA suffers some major drawbacks. First, the original data may not be available from the studies, in which case IDA is not possible to conduct. Second, heterogeneity across the studies may prevent the process of combining the original data. In the case the public gene expression data, the heterogeneity can result from different design characteristics and platforms used, different sample types, different signal-to-noise ratios, etc. Aggregating the highly heterogeneous datasets into a larger one may lead to information loss.

An alternative to IDA is to synthesize some summary statistics of interest, which are drawn from the individual studies. Biclusters represents a type of summary statistic. Synthesizing summary statistics is typically done by meta-analysis, which has been widely used in many domains for combining statistical evidence. In the clinical domain, its success in combining clinical trials is well documented [81]. Formally, meta-analysis is a statistical method to combine the results of multiple studies that are conceptually similar, based on the assumption that there is a common truth behind the studies. This common truth can be enriched via effect size that serves as a standardized measure for the strength of evidence across the individual studies.

Multi-task learning (MTL) [82][83], a recent development in machine learning, represents a third option besides the two main approaches above. The idea of MTL is to solve multiple learning tasks synchronously, while allowing the commonalities and the differences to be explored at the same time. The promise of the technique is to allow the tasks to inform each other in a knowledge transfer fashion. When applied to multiple datasets, mining each dataset represents a task. Multi-task learning in such case can be seen as multi-dataset mining.

In [83], the authors proposed a Robust MultiTask Feature Learning algorithm (rMTFL). It aims to capture a common set of features among relevant tasks and to identify outlier tasks simultaneously. The overall strategy is to decompose a weight matrix W consisting of the prediction models of all tasks into two parts: P and Q. P captures the shared features among the tasks, while Q identifies the outlier tasks. The rMTFL model is formulated as:

$$\min_{W,P,Q} \sum_{i=1}^{m} \frac{1}{mn_i} \left\| X_i^T \mathbf{w}_i - \mathbf{y}_i \right\|^2 + \lambda_1 \|P\|_{1,2} + \lambda_2 \left\| Q^T \right\|_{1,2}$$
$$s.t.\ W = P + Q,$$

where λ_1 and λ_2 denote two group lasso regularization parameters. Unfortunately, although rMTFL makes sense in theory, the required step of tuning λ_1 and λ_2 is very difficult. Improper settings of the parameters can easily lead to inaccurate estimates of the shared features and the outliers.

To summarize, both IDA and MTL have severe limitations that make them impractical for combining public expression data. Meta-analysis on summary statistics represents a proven and reliable option, although it is not without its own issues. Furthermore, there appears to be no literature of using meta-analytic approach on biclusters which are a form of summary statistics. Section 2.7.2 below provides an overview for meta-analysis and how the concept could be adapted to combining summary statistics from gene expression data.

In this section, I will talk about meta-analysis and the reasons why it is useful for achieving the goal in this dissertation. I will first give an overview of meta-analysis, following by the comparison between fixed-effect and random-effect meta-analysis. Then, I will consider multivariate meta-analysis (MVMA) and how it can be applied to biclusters.

2.7.2 Overview of meta-analysis

Statistical meta-analysis [78][79][80] provides a framework for combining evidence from multiple studies, while accounting for the heterogeneity among the studies. Formally, meta-analysis is a quantitative statistical analysis on multiple independent but similar experiments or studies in order to derive a more accurate and precise estimate of the effect size. Effect size is simply a quantification of the difference between two groups of samples. For example, if the two groups are non-treated and treated patients, the effect size quantifies the effect of the treatment. This subject will be discussed in more detail later. The assumption behind meta-analyses is that there is a common truth about the effect size in the individual studies, but the individual estimates may be imprecise or biased. Multiple factors can contribute to the biases. One of them is sample sizes. Smaller sample sizes usually lead to more varying estimates than bigger sample sizes. Thus, the estimates from smaller studies are usually given a smaller weights. Likewise, results from larger studies carry larger weights. The aim of meta-analysis is then is to use statistical approaches to derive a pooled estimate closest to the unknown common truth while accounting for the variations among the individual estimates.

Meta-analysis has been widely used in the clinical area [84][85], often to summarize the effect (or side effect) of a treatment from multiple studies. For example, Karthikesalingam et al. [86] conducted a meta-analysis by pooling 17 studies together to quantify the effect of endovascular repair (EVR) on deterioration in renal function. The result is shown as a forest plot in Figure 2.5 below:

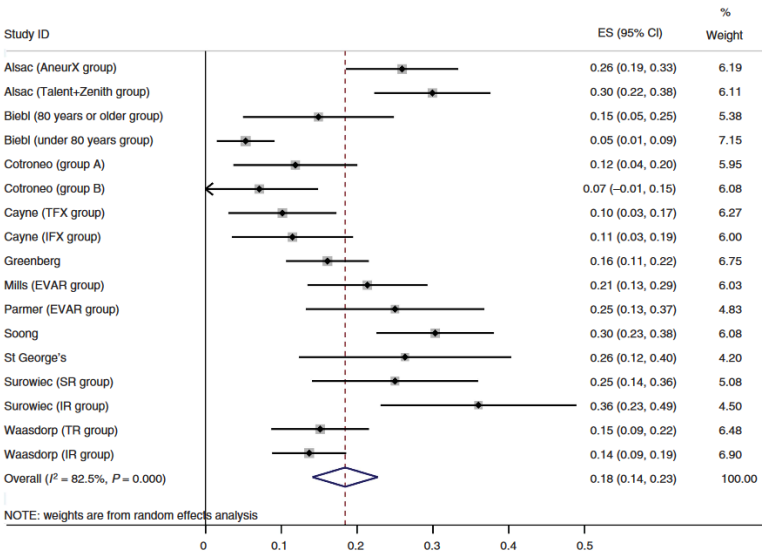


Figure 2.5: Forest plot from a meta-analysis (ES: effect size, CI: confidence interval)(source: [86])

Figure 2.5 shows estimate of the effect size and its confidence interval (CI) for each individual study. The associated weights are also calculated. At the bottom, the overall effect size is given, based on the meta-analysis of the individual studies.

The current study is to apply meta-analysis to a group of biclusters. Each bicluster is analogous to an individual study in a clinical trial meta-analysis. The goal is to identify a gene set, through combining the evidence in the individual biclusters.

To summarize, the framework of meta-analysis has been well established. Meta-analysis to combine clinical trials has been routinely performed with proven efficacy. However, meta-analysis has limitations too, just as any other statistical method.

There are two models for meta-analysis: fixed-effects and random-effects models, as described in section 2.7.3 below. Limitations of these models are pointed at the end of the section.

2.7.3 Fixed-effect versus random-effects models

Under the fixed-effect model [87], it is assumed that the true effect size is identical across all studies. The separately estimated effect sizes vary only because of sampling errors or the within-study variances.

In contrast, under the random-effects model [87][88], the true effect size is different for each study. There are two sources for the variation of effect size: within-study and between-study variances, as represented in the following equation:

$$y_i = \mu + \delta_i + \epsilon_i \quad (2)$$

where μ is the average effect size, δ_i is the true effect size for study i as a result of between-study variance, and ε_i is the sampling error (or within-study variance) for study i. The goal of random-effect meta-analysis is twofold: to estimate the average effect size, and to account for the distribution of the effect sizes δ_i.

In general, unless there is a reason to believe the effect size is the same across the studies, random-effect meta-analysis should be the method of choice. In addition, the fixed-effect model can be considered as a special case of random-effect model where $\delta_i = 0$. In the current study, the random-effects model is chosen for the meta-analysis.

The fixed-effect model assumes zero heterogeneity across the participating studies, which is not realistic in most cases. In contrast, the random-effects model is more general because it takes between-study heterogeneity into account as a source of variance. For the public gene expression data, the heterogeneity among the datasets should be considered and taken into account. For this reason, the random-effects model is chosen in the current study to meta-analyze the biclusters.

2.7.4 Multivariate meta-analysis (MVMA)

Many interventions or risk factors have multiple outcomes. When a meta-analysis is conducted to assess the effect of such an intervention or rick factor, it is necessary to include all the outcomes as the measuring criteria. A typical example is the use of systolic and diastolic blood pressures for evaluating an anti-hypertension treatment. Another example is evaluation of a new teaching method, which can be assessed by multiple outcomes corresponding to the subjects, such as mathematics, English, and biology, etc. In such cases, meta-analysis generally are conducted by including multiple outcomes [89].

One approach for multi-outcome meta-analysis is to assess each outcome separately and ignore the possible correlations among them, then combine the individual estimates to form a summary estimate. This approach is referred to as univariate meta-analysis (UVMA).

There are a couple of pitfalls with UVMA. First, outcomes are often statistically correlated, especially in the case of gene expression data. The correlation structure is useful information. It may convey how the genes are dependent on each other, and thus may allow inference to be made about an inter-gene dependency graph. Disregarding the correlations would result in loss of significant information. Second, previous studies have shown that ignoring the correlation structures can lead to overestimate of the variance of the summary effect sizes, and increase the chances of finding spuriously significant treatment effects [90][91][92].

An alternative approach is multivariate meta-analysis (MVMA), which considers all the outcomes joint instead of separately. Because MVMA takes the correlations into account, it allows borrowing of strength across outcomes to derive the pooled estimates of the effect sizes.

For example, synthesis of outcome 1 can utilize the available data for outcome 2 via correlation, and vice-versa. Borrowing strength can lead to increased precision in MVMA, and can be particularly useful if not all the studies include the full set of the outcomes [93]. In addition, one may be interested in making inferences about a linear combination of the estimated effects. The ability of the multivariate approach in allowing this is one of its advantages [94].

The need for multivariate meta-analysis has been largely driven by a variety of medical applications:

(1) Diagnostic test meta-analysis. One of the most common medical application of multivariate meta-analysis is the bivariate meta- analysis in diagnostic test accuracy [95][96]. In these tests, the purpose is to derive the effect sizes of sensitivity and specificity, which are the two outcomes. For example, Kertai et al. [97] conducted a bivariate meta-analysis to quantify the sensitivity and specificity of exercise electrocardiography for predicting cardiac events in patients undergoing major vascular surgery by pooling 7 studies together.

(2) Multiple effects in randomized controlled trials or observational studies. In this area, clinical trials or observational studies report more than a single outcome of interest, thus multivariate meta-analysis may be used [98][1].

(3) Multiple parameter models for exposure in observational studies. The goal of the applications is to extend multivariate meta-analysis to multivariate regression. For example, a large collaborative study [84] pooled together 39 studies to evaluate the magnitude of association of diabetes mellitus and fasting glucose concentrations with risk

of coronary heart disease and major stroke subtypes. In this study, multiple covariates were included into a multivariate regression model, including sex and study group, and adjusted for age, smoking status, BMI and systolic blood pressure.

(4) Network meta-analysis. This type of analyses involves comparisons of three or more treatments using direct comparisons of interventions within randomized controlled trials, or indirect comparisons across trials based on a common comparator.

The outcomes or transformations of them are usually modeled as random variables. MVMA not only induces the overall effect size of the intervention, but also outputs an estimate of the correlation structure for the outcomes. In the case of gene expression data, if a gene is modeled as an outcome, then MVMA can output the inter-gene correlations, which may lead to a gene network.

To summarize, MVMA is more general and useful than UVMA. The latter can be seen as a special case of the former. MVMA has been gaining popularity in many areas because of its flexibility. In addition, the design of the current study models each gene as an endpoint, and correlations among the endpoints are to be taken into account due to the presumed interactions between the genes. For the above reasons, the current study adopts the framework of MVMA.

The results from meta-analysis convey the effect size estimates, including the point estimates the associated confidence intervals. While these numbers reveal the statistical significance, they do not necessarily imply biological relevance of the gene sets. Therefore, validation of the gene sets is necessary. Pathway analysis based on prior knowledge provides a promising approach for the validation.

2. 7.5 Literature review on the use of meta-analysis for gene expression data

The idea of utilizing multiple datasets through meta-analysis is not new. Rhodes et al. [99] performed a meta-analysis on four microarray datasets related to prostate cancer. All four studies were about gene expression profiling that compared clinically localized prostate cancer samples versus benign prostate samples. The results found some pathway dysregulation in prostate cancer. This meta-analysis was done on gene expression data, but did not employ effect size.

Another study by Choi et al. [100] represents an important improvement over the above one. In their study, the concept of effect size was first used to monitor the progress and measure the result of the meta-analysis. The reasons for using effect size, as suggested by the authors, include: (1) it provides a standardized measurement for statistical strength across different studies, including studies based on different microarray platforms; (2) the use of effect size is based on a well-established statistical frame-work for combining evidences; (3) it allows modeling of between-study variability. In addition, this study employed fixed-effects, random-effects models, and Bayesian inference for the meta-analysis.

Both of the above studies were limited to data with binary outcomes, namely cancer vs. normal. As a result, both meta-analyses focused on measuring the differential expression between the conditions by using standardized difference of means as the effect size. Furthermore, in both studies, the inter-gene correlations were ignored and univariate rather than multivariate meta-analysis was used.

Multivariate meta-analysis (MVMA) on gene expression data has been largely unexplored despite its potentials. The lack of attempt is not surprising. High dimensionality in gene expression data as well as small sample sizes make it very difficult to accurately infer the effect sizes. In this dissertation, I first apply biclustering as a means to identify interesting patterns and to reduce dimensionality at the same time. Then the individual biclusters are meta-analyzed to combine the evidence, which may lead to identification of gene sets that are part of biological pathways.

2.8 Pathway analysis of gene sets

The results from meta-analysis should be interpreted with prudence for the following reasons:

(1) The participating studies selected for a meta-analysis may not be relevant or related. Selection of suitable studies requires careful examination of how each study was conducted, the type of samples or patients used, and how the data were collected, etc. This issue becomes more acute when the meta-analysis is applied to data mining. Due to lack of ground truth knowledge about the data, there is a high degree of uncertainty about every pattern identified on the individual datasets. Thus, the patterns may be unrelated. Combining them through meta-analysis would lead to false discoveries.

(2) The most likely reason for failures in meta-analyses is bias. The methods for calculating effect size are often biased when applied to finite samples. Combining the effect size estimates may reduce the bias, but unlikely to make the bias disappear, especially when

the number of studies is small. Similarly, in the case of data mining, if the number of datasets is small, the estimate of the overall effect size may likely to be biased too.

Given the above the reasons, it is important to exercise caution when interpreting the results of meta-analysis, especially when the technique is applied to data mining as in the current research, because of lack of knowledge about the ground truth in the data.

In addition, even if a gene set have significant effect sizes, it does not necessarily translate into biological relevance. Pathway analyses may provide hints about the biological functions, but they do not necessarily validate the gene sets. Ideally, one would have a collection of datasets where the ground truth is known so that a method can be validated by mining these datasets. However, such datasets don't exist. Even if they did exist, if the new approach found additional gene sets of interest that were not found by prior approaches this would not necessarily mean these additional gene sets were not valid. In fact they might represent new real findings due to the superiority of the new approach. Absent a gold standard (biological or clinical experimental validation) we propose alternate proxy methods.

As mentioned in section 2.3, many public knowledge bases that store curated pathway information in searchable formats have been developed, making it possible to conduct pathway analyses through enrichment tests.

Wang et al. [101] categorize the enrichment analyses into three classes, which are Over-Representation Analysis (ORA) [102], Gene Set Enrichment Analysis (GSEA) [32], and Network Topology-based Analysis (NTA) [103][104].

ORA statistically evaluates the fraction of genes from a pre-known pathway found among the set of genes to be assessed. The test is also referred to as "2x2 table method" [105]. The most

commonly used statistics used by ORA are the hypergeometric distribution, binomial distribution, chi-squared distribution, etc. There are a number of limitations with ORA: (1) It treats all the genes in the gene set equally, despite some of the genes may be more differentially expressed than the others; (2) The gene set to be assessed excludes other genes that may belong to the same pathway because of a threshold used to determine the gene set membership; (3) The genes within the gene set are assumed to be independent. Disregard of the inter-gene dependence represents a significant information loss.

GSEA is an improvement over ORA. It takes a gene list, as well as the ranks of the genes in the list, as inputs and produces an enrichment score with a pre-known pathway. The ranks of the genes are based on some gene-level statistics, such as correlation of expression measurements with phenotype [106], ANOVA [107], Q-statistic [108], signal-to-noise ratio [32], t-test [107][109], and Z- score [110]. In other words, GSEA treats the genes differentially according to their ranks when calculating the enrichment score, unlike ORA that assigns all the genes with an equal weight. However, similar to ORA, GSEA still ignores the interdependencies among the genes in the gene set.

Network Topology-based Analysis (NTA) takes advantage of the pathway topology in protein-protein interaction databases. An implementation from Wang et al. [111][101] is particularly interesting. First, because most pathways have some hierarchical structures, they identified nearly 1000 hierarchical modules from the human protein–protein interaction networks. Then they implemented a random-walk algorithm [112] to identify the best partition of the network that maximizes a modularity score [113].

To summarize this section, applying meta-analysis to data mining is challenging, mainly due to lack of ground truth knowledge about the data. Hence, the mining results should be subject to

validation. Thanks to the growing number of public knowledge bases, it is now increasingly possible to apply a variety of statistical tests to assessing the biological relevance of gene sets extracted from data. Unfortunately a biological validation of these gene sets is not possible within the scope of available data and this dissertation.

It is possible that some gene sets are real, but have no match yet in the knowledge bases. In such cases, the findings may serve as a hint to guide future biological research.

2.10 Overall strategy

Given the above discussion, we propose the overall strategy for the current research as illustrated in Figure 2.6 below. The steps match the specific aims outlined above in the Executive Summary. As a reminder, the three aims are:

- Aim 1: Determine optimal method for constructing bicluster stacks.
- Aim 2: Determine suitability of meta-analysis techniques to pool biclusters and assess performance.
- Aim 3: Assess potential utility of gene sets identified in Aim 2 using pathway analysis.

They try to answer these overall questions: (1) what is the data mining method best suited for finding gene sets? (2) how to utilize multiple datasets jointly in order to increase statistical strength?

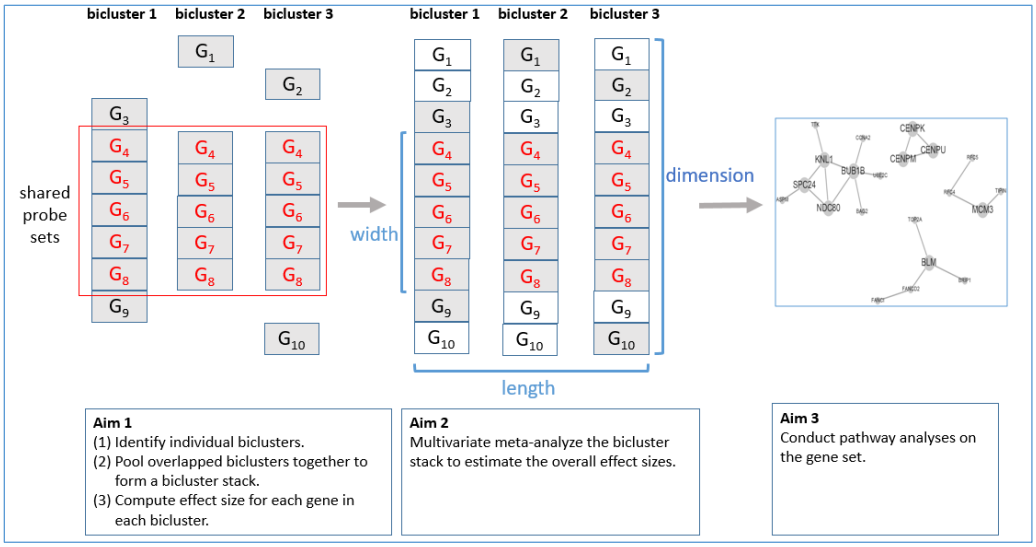


Figure 2.6: The overall strategy

In this strategy, overlapped biclusters are pooled together to form a stack based on the number of genes that they share (Step 1 in Figure 2.6). This is a process similar to selection of participating studies in a traditional meta-analysis. Thus, each bicluster is treated as a participating study. The goal is to combine the evidence about the candidate gene set from the biclusters, assuming that the biclusters are related.

The size of a bicluster stack is defined by two factors: the length (the number of biclusters in the stack) and width (the number of genes shared by ALL the biclusters). In the illustration of Figure 2.5, the length and width are 3 and 5, respectively. In addition, some genes are present in some biclusters but not in the others. When a stack is formed, they are added to the biclusters where the genes are missing, so that the member biclusters will have same set of genes. For example, in Figure 2.6, G_3 is present in bicluster 1 but not in biclusters 2 and 3, so G_3 is added to biclusters 2 and 3 during the stacking. The total number of genes becomes the dimension of the stack. In Figure 2.6, the dimension is 10.

In step 2 of Figure 2.6, the meta-analysis is then performed on the biclusters (rather than on the source datasets). During the meta-analysis, each gene is treated as an endpoint because they collectively reflect the strength of the gene set being turned on or off, just as systolic and diastolic blood pressures together reflecting the effectiveness of a medical treatment. In addition, it is well established that genes in a pathway tend to work together in a coordinated fashion. Thus, the statistical correlations among the genes should not be ignored. Given these considerations, multivariate meta-analysis (MVMA) is adopted here to combine the biclusters.

When multiple biclusters are pooled together to form a stack, an implicit assumption made is that within each bicluster, the genes' activities change because of the same effect, which is analogous to the intervention is a traditional meta-analysis. Obviously, this assumption may not hold true, due to the heterogeneity in public studies in terms of design and data collection process. Hence, the biclusters in a stack may not be related. MVMA aims to statistically assess the stacks, but it does not validate the embedded gene sets. Aim 3 (step 3 in Figure 2.6) deals with pathway analyses of the gene sets. The analyses should not be considered as a gold standard and thus the results do not necessarily validate the gene sets. Nevertheless, they provide insight into possible functions of the gene sets, and increase the degree of confidence that the gene sets may be biologically relevant, as explained in Chapter 5.

Chapter 3 Biclustering and stacking of biclusters

This chapter is focused on the methods and results exploring Aim 1, which is to determine optimal method for constructing bicluster stacks. This is part of addressing the first question: what is the data mining method best suited for finding gene sets? As discussed in

Chapter 2, a bicluster in gene expression data is a pattern displayed by a group of genes in a subset of samples. Hence, gene sets can be identified through biclustering. A single bicluster may not have sufficient statistical evidence on whether the gene set is real (part of a real biological pathway), or whether all the member genes actually belong to the gene set provided it is real. Combining multiple biclusters would increase the statistical power and thus give us more confidence about authenticity of the gene set.

3.1 Background, motivations, and strategy

As discussed in Section 2.7.1, there are two main approaches for utilizing multiple datasets: (1) integrative data analysis (IDA) [76][77], which is simply to pool the datasets together, and (2) synthesis of summary statistics drawn from multiple studies [78][79][80]. This section is to expand the discussion in the context of gene expression data.

The IDA approach is straightforward and sounds more intuitive, but it suffers a number of limitations. First, the datasets may not be combinable. In the case of gene expression data, the datasets may be generated by different platforms, such as cDNA [16], long oligonucleotide [114](Operon 70-mer), and short 25-mer [2](Affymetrix) array platforms. Merging the data derived from different platforms may suffer significant information loss.

Second, even if the data are generated using the same platform, it may not be desirable to combine them into one large data matrix. The reason is related to data heterogeneity. For example, when the biclustering algorithm Condition-dependent Correlation Subgroups (CCS) is performed on the separate datasets, the correlation coefficient threshold can be individually defined based on the signal-to-noise ratio in the dataset, allowing the biclusters to be identified on a per-dataset basis. In contrast, if CCS is applied to an aggregated large dataset, the same

correlation coefficient has to be used on all the combined samples, which may lead to some of the biclusters undetected and thus information loss. Third, in some cases, the researchers may choose to publish summary statistics as opposed to making the raw datasets available. Biclusters can be seen as a form of summary statistics. Apparently, combining datasets is not possible in such cases, and combining the summary statistics is the only option.

To summarize, although combining multiple datasets to increase the sample sizes sounds intuitive, it is not generally feasible or desirable. Instead, combining summary statistics is a more general approach. In the current study, the summary statistic is bicluster.

The agenda of this chapter is as follows: Section 3.2 discusses identification of individual biclusters from different datasets; Section 3.3 focuses on stacking of biclusters; Section 3.4 is concerned with estimation of bicluster effect sizes; Section 3.5 discusses combining data versus combining biclusters; Section 3.6 provides a summary and a discussion.

3.2 Identification of individual biclusters

This section considers several specific topics including selection of a biclustering algorithm for this study, performance of the algorithm, and the results of individual biclusters.

3.2.1 Pre-processing and normalization of microarray data

Thirteen datasets from GEO (listed in Table 2.2) are selected for the current study based on three criteria: (1) they are all related to prostate cancer; (2) they all use the same platform Affymetrix Human Genome U133 Set Plus; (3) the number of samples is at least 4. The reason for choosing GEO as the data source is that GEO is the largest microarray data repository.

The raw microarray data are downloaded from the GEO website. After pre-processing that includes background correction, the data are then quantile normalized using the Robust Multi-array Average (RMA) procedure, which is primarily designed for analyzing of data from Affymetrix arrays [115].

Figure 3.1 below shows the distribution of the array data for dataset GSE29232 before and after normalization.

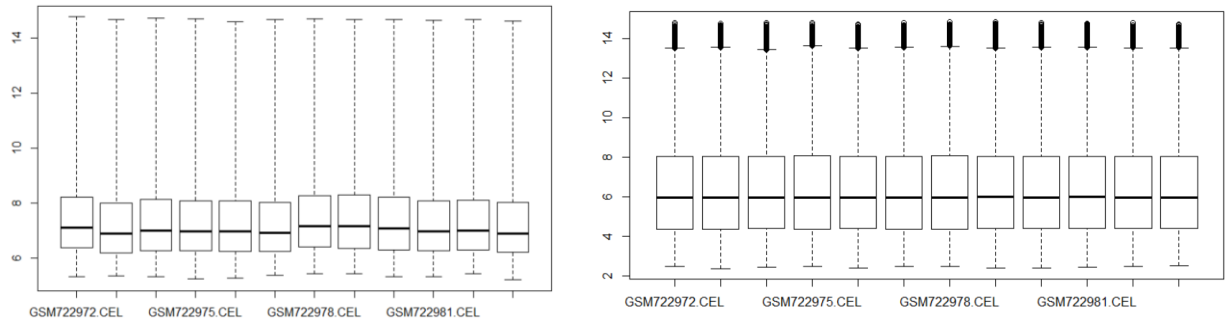


Figure 3.1: Distributions of the expression values for GSE29232 before and after normalization

The boxplots in Figure 3.1 show that the data from GSE29232 have been properly pre-processed and normalized. All other datasets exhibit similar pattern of distribution change. For reference of GSE29232 and other datasets, please refer to Table 2.2 and the GEO website.

3.2.2 Selection of biclustering algorithm

There exists a variety of biclustering algorithms that recognize distinct statistical features in the target patterns, as discussed in Chapter 2 and illustrated in Appendix I. Among these algorithms, those that are based on correlations between genes are chosen in the current study for the following reasons:

(1) Pearson Correlation Coefficient (PCC), shown in equation (1) above, can capture shifting, scaling, and shift-scale patterns, making it an ideal measure for quantifying differential expressions. As explained in [65], shifting patterns mean that the rows have different overall values but they are correlated. Scaling patterns mean that the rows scale differently cross the columns, resulting in less correlated rows. Shift-scale patterns have rows that scale differently and have different overall values at the same time.

(2) PCC is widely used in many practical situations including engineering and medicine, and can be easily interpreted [116].

(3) Correlation based biclusters can naturally fit into the framework of multivariate meta-analysis (MVMA), as explained in details in the next chapter.

(4) More importantly, it has been shown that the expression of the genes in a pathway are often statistically correlated [117][118]. The interactions among the genes in a pathway are biologically complex, but their expression profiles can be statistically characterized by correlation coefficient.

Several correlation based algorithms have been proposed and applied to gene expression data. They are Correlated Pattern Biclusters Algorithm (CPB) [70], Biclustering by Correlated and Large number of Individual Clustered seeds (BICLIC) [119], and Condition-dependent Correlation Subgroups (CCS) [120]. They all have advantages and disadvantages. CPB initiates the search on random seeds, resulting in poor reproducibility. Successive runs can lead to different results. In addition, CPB does not allow overlapped biclusters and can only identify positively correlated modules. The biggest strength of CPB is its computational efficiency.

BICLIC starts the search on comprehensive set of seeds, and can handle overlapped biclusters. But similar to CPB, it only recognizes positively correlated modules.

In contrast, CCS comprehensively searches all the matching modules, leading to completely consistent and reproducible results [120]. It can identify overlapped biclusters and allow the degree of overlap to be specified by the user. Furthermore, it can recognize both positively and negatively correlated gene sets, making it more versatile than the above two algorithms.

Reproducibility and flexibility of CCS is particularly important to the current study. Given the limited data available, it is important to uncover as many matching patterns as possible in each dataset. The reproducibility of CSS and its ability in allowing the user to specify matching criteria may lead to high recall of the target patterns. For the above reasons, CSS is chosen in this study to illustrate biclustering for identification of gene sets.

However, the strengths of CCS do come with a major cost, which is the computation time. It is much more demanding in computation power and run time, particularly when compared with CPB. Overall, CCS is an ideal biclustering algorithm that is part of the solution for the first overall question: what is the data mining method best suited for finding gene sets?

3.2.3 Parameter tuning and performance of CCS

CCS is based on Pearson Correlation Coefficient (PCC) [120]. It recognizes biclusters in which every gene pair shows expression correlation above a threshold PCC. Before the run of the algorithm, the user needs to supply the initial values for these three parameters: minimum number of genes, minimum number of samples, and the threshold PCC.

In the current study, the minimum number of genes is chosen to be 30 for all datasets. The rationale behind the choice is that the gene sets being sought are of at least 60 genes (dimension >= 60). This is based on a preliminary observation: if the number of genes falls below 60, the gene sets rarely have a match from the pathway analyses. Assuming 50% of the genes in a target gene set are recognized by CCS, then the bicluster should have a gene count of at least 30.

The minimum of samples is set to be 6. This is because in a typical dataset, it contains multiple conditions, with each condition often including 3 or more samples or replicas. If differential expression is to be detected across two conditions, then 6 samples need to be included in a bicluster. For simplicity, the CCS algorithm treats the within-condition replicas and between-condition replicas the same.

With regard to the threshold PCC, it is set to be the 90^{th} percentile of all the pair-wise PCCs in the datasets. It is estimated by sampling a large number of gene pairs and computing their PCCs across 6 samples that are randomly selected. The 90^{th} percentile is chosen because it seems to strike a good balance between a sufficient number of biclusters to be found and the computation time needed. If the threshold is higher than 90^{th} percentile, too few biclusters may be identified. If the threshold is below the 90^{th} percentile, it may result in forbiddingly long computation time and increase the chances of getting false positive results.

With regard to the performance of CCS, based on some small scale preliminary tests that are not presented here, the computation time appears to grow exponentially as the numbers of genes and samples increase, which is consistent with the previous finding that biclustering is a NP-hard problem as mentioned above. For the selected datasets, time periods ranging from 5 to 10 days are typically needed to complete one run of CCS on a Red Hat Linux (v7.4) server

equipped with 2 virtual CPUs and 4 GB memory, provided by the Microsoft Azure cloud services.

3.2.4 Results on individual biclusters

Table 3.1 below lists the numbers of individual biclusters identified by runs of CCS on the datasets. The term "probe sets" that appear in the table refers to the probe sets used in the "Affymetrix Human Genome U133 Set Plus" platform that relies on 25-mer oligonucleotide probes. In general, multiple probes are mapped a single transcript. As a result, they form a probe set. The expression numbers in a dataset are summaries of probes mapped to the same transcript [121][122]. See Table 2.2 for descriptions of what the focus of each dataset is.

Dataset	# of probe sets	# of samples	Threshold PCC used	# of biclusters found
GSE7868	9346	9	0.78	3451
GSE17044	9341	6	0.90	1293
GSE22483	9346	6	0.97	973
GSE29232	9339	12	0.84	1878
GSE34589	9352	8	0.96	1612
GSE44905	9315	18	0.85	2809
GSE56908	9349	4	0.93	319
GSE3325	9342	19	0.89	4632
GSE7708	9338	14	0.90	5604
GSE55945	9342	19	0.79	7302
GSE94580	9354	14	0.98	1257
GSE51524	9324	18	0.92	3274
GSE22606	9323	12	0.94	1400

Table 3.1: Summary of CCS biclusters found in the 13 datasets

3.3 Stacking of biclusters

The general problem explored next is the issue of how to combine data across experiments/datasets with a particular focus on meta-analysis approaches (as discussed in Chapter 2).

Prior to combining data across studies in a meta-analysis, the participating studies need to be carefully selected to avoid bias or misleading results. The selection process is usually done by review of relevant literature. Though this is not identical to combining data sets from different experiments there are parallels as described below.

The focused goal of this section is to present a procedure for pooling biclusters together to form bicluster stacks. The member biclusters in a stack, each coming from a separate microarray dataset, are functionally equivalent to the participating studies in a meta-analysis. The stacking process is analogous (and functionally equivalent) to the process of selecting participating studies prior to a meta-analysis.

In Section 3.3.1 below, I will first describe a method for finding stackable (or overlapped) biclusters based on a brute-force search procedure. In Section 3.3.2, I will present results of stacking the individual biclusters identified in Section 3.2.

3.3.1 A procedure for stacking biclusters

Stacking of biclusters is illustrated as the step 1 of the overall strategy in Figure 2.6 in Chapter 2. It ensures that at least a pre-defined number of genes are shared by all member biclusters.

As shown in Figure 3.2 below, a procedure (implemented in Java) exhaustively searches for every possible new bicluster to add to a given stack. An advantage of such brute-force search algorithm is that it outputs all the possible stacks. In addition, it produces consistent, reproducible result. The downside is the long computation time. It took about 21 days to complete the search of the all individual biclusters summarized in Table 3.1 using a Windows 10 server machine.

```
Input:   a list of datasets, with individual biclusters existing in each dataset;
         a pre-specified stack size k (smaller than number of datasets);
         a pre-specified minimum number of genes shared by all member biclusters in the stack;

Output:  a list of bicluster stacks;

1. Construct all possible combinations of datasets, with # datasets in each combo = k;

2. For each combo i
        for each bicluster j in the first dataset in combo i
            for each subsequent dataset m in combo i
                for each bicluster n in dataset m
                    if n shares at least k genes with all previous members in the stack
                        add n to the stack;
                        skip dataset m, continue to dataset (m+1)  in combo i;
                    else
                        select bicluster (n+1) in dataset m;
                end for
            end for
        end for
    end for

3.  Output all candidate bicluster stacks;
```

Figure 3.2: A brute-force search procedure for finding overlapped biclusters

3.3.2 Results on bicluster stacking

The size of a stack is determined by two user-defined parameters: the number of biclusters (length) and the number of genes shared by ALL the member biclusters (width). Please see Figure 2.6 for illustration of bicluster stack size. In Figure 2.6, three biclusters are shown to

form a stack, and five genes are shared by all the biclusters. Therefore, the stack has a length of 3 and a width of 5. After an initial stack is found, for each member bicluster, those genes that are missing but are detected by the other member biclusters are added to the original bicluster, as shown in the middle portion of Figure 2.6. This ensures all the member biclusters have the same dimension (number of genes) at the end.

Before running the above described searching procedure for stacking biclusters, the target length and width need to be pre-defined and provided as inputs. The values are determined somewhat arbitrarily, and a compromise is needed between the two. If a bigger length is desired, then the width will need to be reduced in order to find a sufficient number of stacks. The initial settings are length=7 and width=10, which lead to only about 6 unique stacks to be found.

Later, in order to find more stacks, the length is reduced to 5 and the width is kept to be 10 for consistency. This lead to about 200 stacks have been identified from the prostate cancer data sets. Since many of them are overlapped or redundant, roughly 50 of them are possibly unique based on a rough estimation. Currently I do not have a systematic way to filter out the overlapped stacks. The selection of candidate stacks is largely done manually on a case-by-case basis.

Six of the stacks from the ~50 unique stacks are selected for demonstration purpose. Table 3.2 below gives the size summary of the stacks. The complete information about the probe sets in the two stacks is given in Appendices II and III.

Stack Label	Length	Width	Dimension of the gene set	Source datasets
ProsBicSta01	7	11	83	GSE56908, GSE22483, GSE17044, GSE22606, GSE44905, GSE7868, GSE7708
ProsBicSta02	7	10	117	GSE17044, GSE22606, GSE34589, GSE44905, GSE7868, GSE3325, GSE7708

ProsBicSta03	7	10	86	GSE56908, GSE17044, GSE22606, GSE44905, GSE7868, GSE3325, GSE7708
ProsBicSta06	7	10	74	GSE56908, GSE22483, GSE94580, GSE22606, GSE29232, GSE7868, GSE3325
ProsBicSta12	5	10	175	GSE56908, GSE22483, GSE17044, GSE44905, GSE3325
ProsBicSta19	3	30	107	GSE17044, GSE44905, GSE7868

Table 3.2: Summary of selected real bicluster stacks

As aforementioned, not all the probe sets appear in all the biclusters (for a description about Affymetrix probe sets, please refer to Section 3.2.4). Some occur more frequently than the others. For example, in ProsBicSta01, 11 of the total 83 probe sets occur in all 7 biclusters, while 39 probe sets occur in only 4 biclusters. The bar plots in Figure 3.3 below depict the size breakdown of the six stacks.

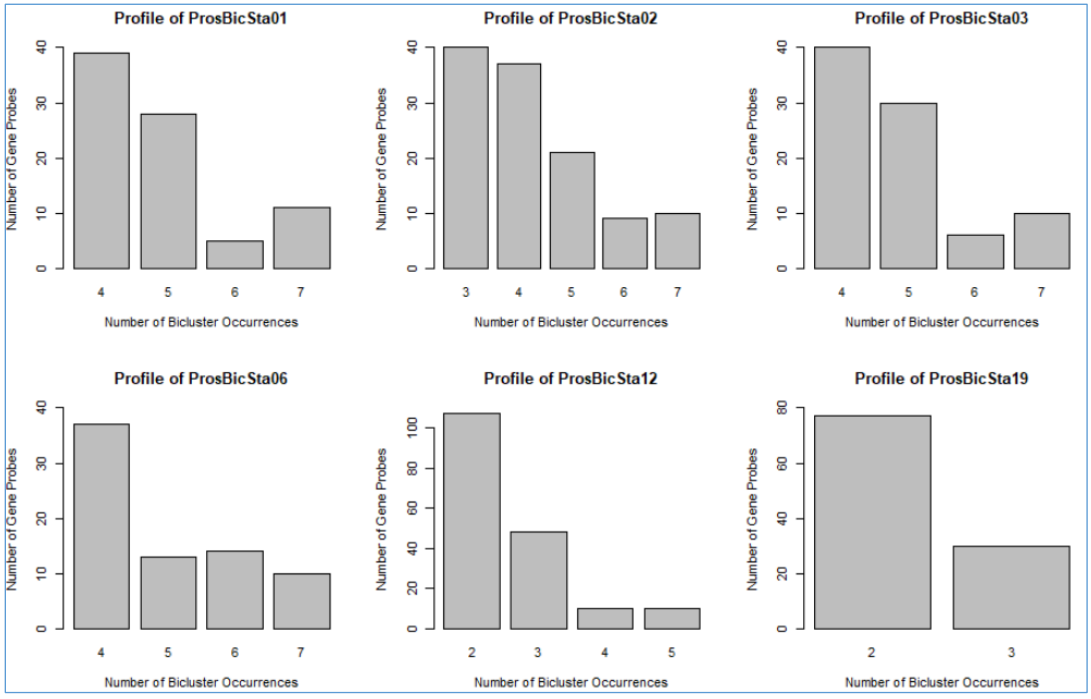


Figure 3.3: Size breakdown of the six bicluster stacks

As shown in Figure 3.3, ProsBicSta12 and ProsBicSta19 includes probe sets that appear in only two biclusters, while ProsBicSta01, ProsBicSta02, ProsBicSta03, ProsBicSta06 only include probe sets that appear at least in three biclusters. This is because I tried to limit the dimensions of the stacks to be not more than 200 for consistency as well as computational concerns. For example, if ProsBicSta01 includes probe sets that appear in 3 biclusters, the dimension of the stack would exceed 200.

These six bicluster stacks will continue to be used to demonstrate the MVMA and the pathway analysis in the next two chapters.

3.4 Effect sizes for biclusters

This section is aimed to describing how to compute the effect sizes for a bicluster. In Section 3.4.1, a general introduction to effect sizes will be given. In Section 3.4.2, the uniqueness of bicluster effect size is discussed. In Section 3.4.3, a new method for computing bicluster effect size is proposed and validated through a simulation study. Finally in Section 3.4.4, the estimated effect sizes for all six bicluster stacks will be presented.

3.4.1 Introduction to effect sizes

In simple terms, effect size is a measure of the difference between two groups of samples. For example, if the groups are before- and after-treatment patients, the calculated effect size can be used to quantify the effect of the treatment or intervention. Effect sizes are widely used, especially in clinical trials and educational studies, to measure the effect of new treatments or teaching methods.

Compared to a p-value that informs whether there is a statistically significant difference between the groups, effect size conveys the magnitude of the difference. In addition, if the sample sizes are large enough, p-value will always be significant even if the difference is negligible. In such cases, p-values are simply meaningless or misleading. For the above reasons, effect size is increasingly used in various areas including educational, social, and medical studies, and has become the foundation of meta-analysis.

Meta-analysis aims to systematically review multiple studies and produce a consensus measure for the common effect, provided the common effect exists among the studies. As illustrated in Figure 2.6 in Chapter 2, the effect sizes in the individual studies tend to be more uncertain, characterized by the wider confidence intervals. A properly conducted meta-analysis may lead to an overall effect size that has smaller confidence interval than those from the contributing studies. Evolution of the effect size as more data are added to the meta-analysis represents our changing confidence for the effect. In other words, weaker evidences combined can lead to a stronger evidence. Effect size plays a key role in this process by providing a measure for the strength of the evidence, and serves as a yardstick for monitoring the progress or the success of the meta-analysis [123][124].

Effect size can be defined differently. One of the most classic and widely used effect sizes is Cohen's d, suggested by Jacob Cohen [125]. It is defined as the difference between two means divided by a standard deviation for the data (standardized difference in means):

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s} = \frac{\mu_1 - \mu_2}{s},$$

where *s* is the pooled standard deviation:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}},$$

where s_1^2 and s_2^2 are the variances for groups 1 and 2, n_1 and n_2 are the sample sizes for groups 1 and 2, respectively.

A problem with Cohen's d is that when sample sizes are small (usually below 20), the effect size estimate is somewhat overstated. To correct for this bias, Hedges' g [126] which a slight adjustment to Cohen's d, is recommended. The estimated effect size in Hedges' g is also based on standardized difference in means using pooled standard deviation:

$$g = \frac{\bar{x}_1 - \bar{x}_2}{s},$$

The bias can be approximately corrected through multiplication by a factor

$$g^* = J(n_1 + n_2 - 2)g$$

The exact form for the correction factor $J()$ involves the gamma function:

$$J(a) = \frac{\Gamma(a/2)}{\sqrt{a/2}\,\Gamma((a-1)/2)}$$

In the current study, because of the smaller sizes, Hedges' g is used to estimate the effect sizes.

3.4.2 Effect size for biclusters

Effect size [123][127][124] has gained increasing popularity as a measure for the effect of an interventions. However, the concept is relatively new in data mining, and to the best of my knowledge, has not been used to characterize biclusters. In the current study, effect size is used to measure the relative "strength" of biclusters.

Control samples are usually needed to compute an effect size. In the case of bicluster, ideally, the control samples come from a submatrix in the same dataset that shares the same genes with the bicluster, but under different conditions. Unfortunately, such samples are often

not available due to the small number of samples of the datasets. Alternative source for the control needs to be sought.

For a given bicluster, each gene has a corresponding effect size. The bicluster level effect sizes will be combined to produce the overall effect sizes through meta-analysis, which is the main goal of this dissertation.

Since the biclustering algorithm CCS searches for modules that have high inter-gene correlations, the effect size for a gene is defined by how strongly the gene is correlated in expression with the other genes in the same bicluster, compared to the correlations in the control samples. Each effect size estimate has a corresponding standard error and thus a confidence interval. In addition, the covariance between each pair of genes is estimated as well to provide the within-study covariance matrix for the bicluster, which will be used in the later MVMA.

To summarize, a bicluster is measured by multiple effect sizes, one for each gene. Due to the small sample sizes in most datasets, the control samples required for computing the effect sizes may not be available. Alternative source for the control needs to be sought.

The next section will cover the specific method for computing bicluster effect sizes and the within-study covariance matrix.

3.4.3 A proposed procedure for computing bicluster effect sizes

In this section, I will first propose a simple and effective method for estimating bicluster effect size. It is suited for data with small sample sizes. Then, I will present a simulation study to validate the method.

3.4.3.1 Description of the proposed method

A typical gene expression experiment often includes contrasting samples such as cancer vs. normal samples. Ideally, if a bicluster covers the cancer samples only, then the normal samples can be used as the corresponding control. This horizontal arrangement is illustrated in the left panel in Figure 3.4 blow.

Unfortunately, this arrangement is often not possible for biclusters identified by CCS, especially when the sample sizes are small. As described earlier, CCS recognizes correlated rows based on Pearson Correlation Coefficient. If a group of genes exhibits differential expression across the normal and cancer conditions, then both of these conditions will become part of the CCS bicluster, leaving no more conditions or samples left to serve as a control. Since most of the public datasets have small numbers of conditions/samples, it is often not possible to find a control horizontally given a bicluster..

To overcome the issue of lack of control data, I propose a strategy for approximating bicluster effect sizes. It involves obtaining the control samples from different genes outside of the bicluster (vertical arrangement, as shown in the right panel of Figure 3.4 below). I would argue that this approach is justifiable, because bicluster is a result of the interaction between a group of genes and a group of samples. Thus, genes and samples are equivalent. Comparing genes, just as comparing samples, can lead to valid approximation of the bicluster effect sizes.

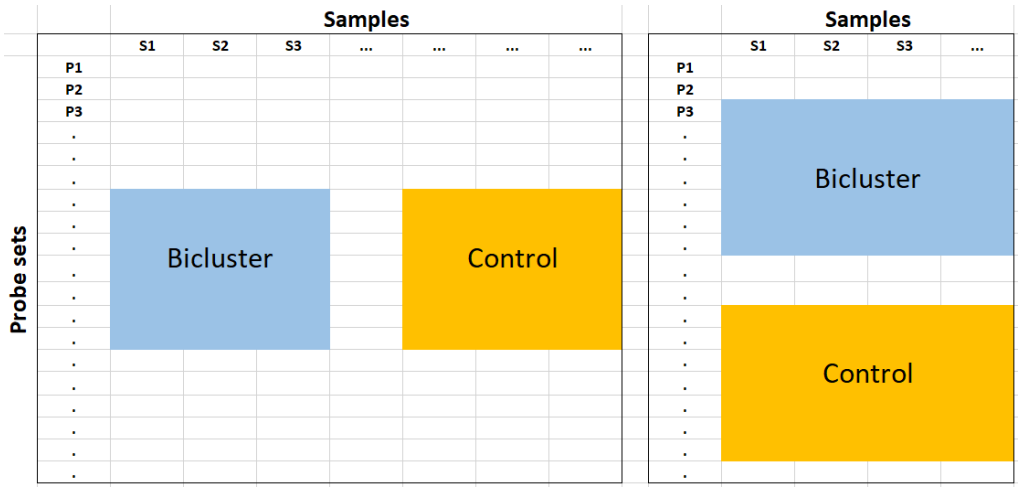


Figure 3.4: Two arrangements of treatment vs. control in a microarray dataset

As discussed above, an effect size quantifies the magnitude of difference between the treatment samples and the control samples. In this case, the treatment corresponds to a bicluster, while the matching control is sought based on strategy described above. Since the CCS algorithm recognizes highly correlated rows in the datasets [120], the effect sizes should reflect the strength of a correlations within the bicluster compared to those in the control. Since each gene is treated as an endpoint, it has its own effect size. The following bootstrapping procedure is designed and used to compute the gene-level effect sizes for a given bicluster:

1. For each gene (A) within a bicluster, randomly sample with replacement another gene (B ≠ A) within the same bicluster, and compute the Pearson Correlation Coefficient between A and B.
2. Step 1 is repeated 1000 times to obtain a pool of correlations for gene A.
3. Randomly sample with replacement two genes outside of the bicluster and compute their correlation.
4. Step 3 is repeated 1000 times to obtain a pool of correlations for the control.

5. The two pools of correlations figures are then used to compute the effect size for gene A by Cohen's d as described above.
6. Repeat steps 1-5 for all other genes within the same bicluster.

The above procedure results in a collection of gene specific effect sizes for a given bicluster. This collection is referred to as individual or bicluster-level effect sizes in the later discussion.

To obtain the within-study covariance for a bicluster, the sample covariance matrix is computed and used with the bootstrap samples generated above.

To summarize, public expression data typically have small sample sizes. As a result, a novel procedure based on bootstrapping is designed and implemented to estimate the individual (or bicluster level) effect sizes. The next section focuses on validation of the procedure.

3.4.3.2 Validation of the proposed method

The goal of this section is to establish the validity of the above procedure. The concern is that this is the first time, to the best of my knowledge, that the concept of effect size is applied to biclusters. It is unclear whether the resulting estimates would satisfy some basic requirements of effect sizes.

A valid effect size estimate should satisfy two requirements: (1) it reflects the magnitude of the difference between the treatment samples and the control samples; (2) it responds positively to the increase in sample size, meaning that the magnitude of the estimate should increase as the sample size increases. Intuitively, the second requirement means that the more samples we have, the stronger the effect size is.

The first requirement is automatically satisfied based on the procedure described above. The second one is verified by a simulation study and results from real bicluster stacks as described below.

First, in an experiment, a series of synthetic biclusters are generated as follows. First, two submatrices are generated to represent a bicluster and a control. Second, each row in both the bicluster and the control contains two conditions, and each condition contains varying number of samples. Third, for each condition within the same row, the samples are normally distributed with an incremented mean and a fixed standard deviation. Finally, the mean expression value in the bicluster is set to be 2 higher than that in the control.

Using these artificial data, 10,000 bootstrap samples are generated to compute the effect sizes based on Cohen's d. The forest plot in Figure 3.5 below shows the relationship between effect size and sample size (number of samples in each condition).

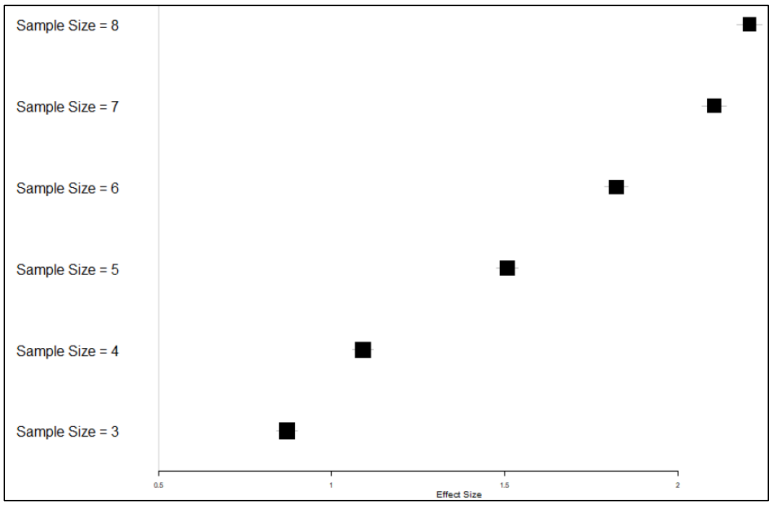


Figure 3.5: A forest plot that shows the relationship between sample size and effect size

The result in Figure 3.5 clearly shows a positive relationship between sample size and effect size. It indicates that the bicluster effect sizes computed by the above described procedure meet a key requirement for effect sizes suitable for meta-analysis.

Besides using artificial data, the individual effect sizes estimated for the six bicluster stacks also exhibit the similar positive relationship with sample size, as shown in the scatter plots in Figure 3.6 below. Recall that individual effect sizes (a.k.a. bicluster-level effect sizes) need to be estimated prior to the meta-analysis that aims to infer the overall effect sizes. Thus, for a given stack, each member bicluster has a collection of effect sizes, one for each gene. Since each bicluster has its sample size, it is possible to inspect the relationship between the effect sizes and the sample sizes, which is what Figure 3.6 illustrates.

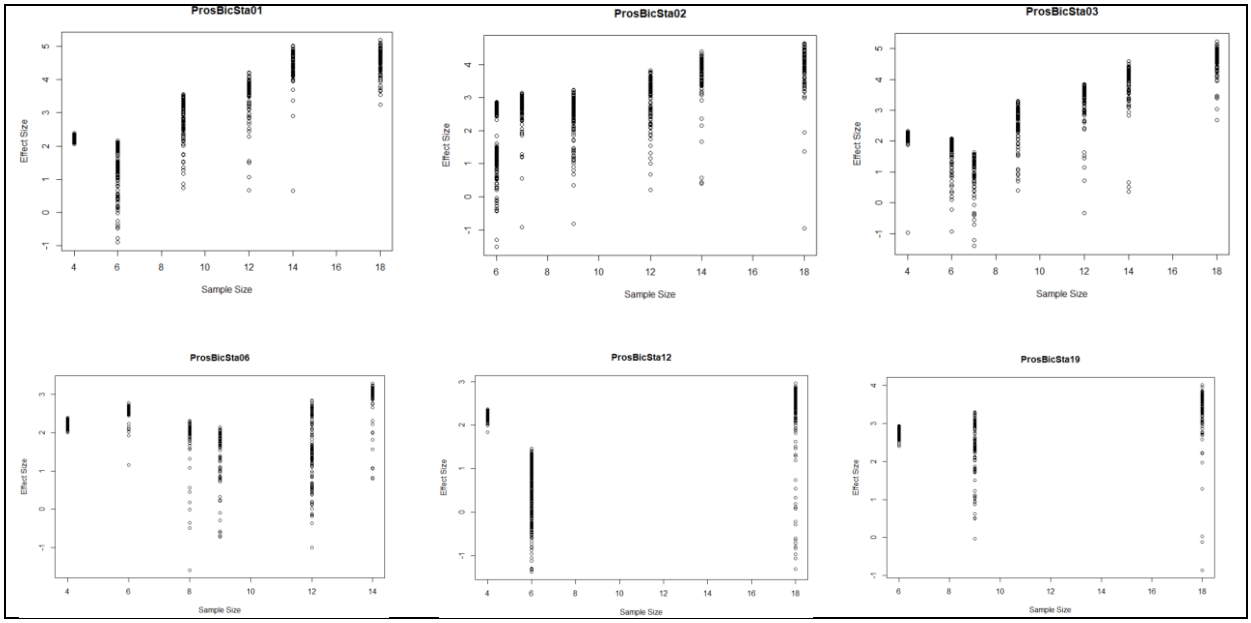


Figure 3.6: Scatter plots that illustrate the relationship between the estimated effect sizes and the sample sizes in six bicluster stacks

The results in Figure 3.6 show that the individual effect sizes tend to increase as the sample sizes increases, which is consistent the result from the simulation study described in the first part of this section. Thus, two sets of results (one on simulated data and one on real data) demonstrate that the effect size estimates respond positively to the increase in sample sizes.

To summarize, the small sample sizes in public expression data necessitate a novel method for estimating the effect sizes for biclusters. A bootstrap-based procedure is proposed in the previous section. Results from a simulation study and from six bicluster stacks show that effect size estimates satisfy two basic requirements for an effect size measure, thus validating the method proposed in 3.4.3.1.

3.4.4 Distribution of effect sizes within bicluster stacks

Given that the proposed method for approximating bicluster effect size has been validated, it can now be applied to estimation of the individual effect sizes for the six bicluster stacks.

The effect sizes within a bicluster vary from gene to gene. Moreover, the distribution of the effect sizes changes from bicluster to bicluster within the same stack. The histograms in Figures 3.7 and 3.8 below show the effect size distributions within two stacks: ProsBicSta01 and ProsBicSta06, respectively. The same results for the other four stacks are given in Appendix III.

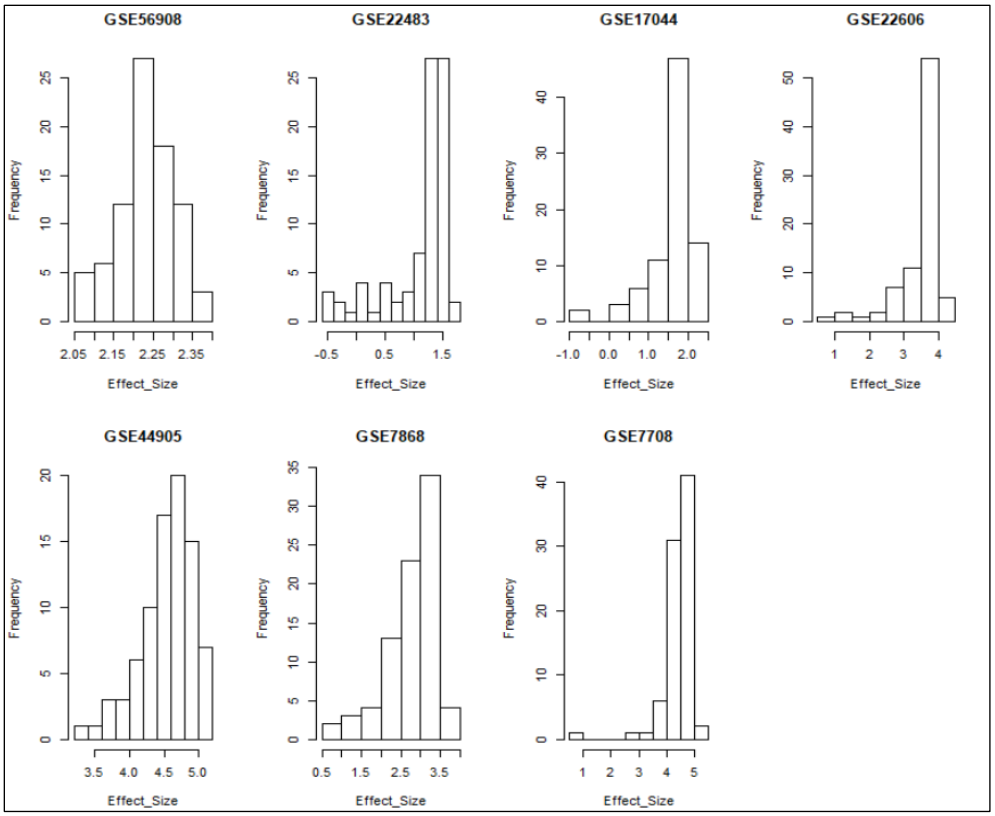


Figure 3.7: Histograms of the effect sizes for the seven biclusters in stack ProsBicSta01

(Note: Since each bicluster comes from a separate dataset, it adopts the label of its host dataset. For example, in stack ProsBicSta01, one of the source datasets is GSE56908. Hence, the bicluster from GSE56908 is labeled as “GSE56908”). Thus this shows the effect size of each of the 7 bi-clusters that makes up the stack ProsBicSta01

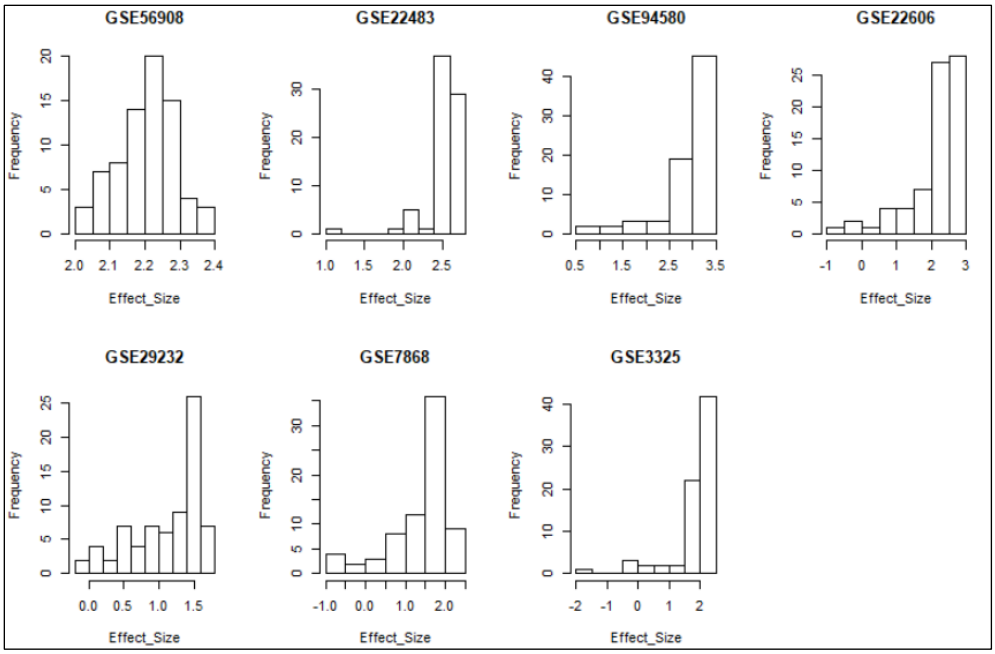


Figure 3.8: Histograms of the effect sizes for the seven biclusters in stack ProsBicSta06

From the histograms above, it appears that ProsBicSta01 is more heterogeneous than ProsBicSta06 in terms of both the means and the variances of the bicluster-level effect sizes. The within-stack heterogeneity can also be observed in the other stacks, as shown in Appendix III.

3.5 Combining data vs. combining biclusters

As discussed in Section 2.7.1, there are two broad approaches for utilizing multiple datasets: integrative data analysis (IDA) [76][77] and synthesis of summary statistics (SOSS)[78][79][80]. The current study chooses the second approach for the reasons mentioned in 2.7.1. The earlier sections in this chapter adopt the approach of SOSS by identifying biclusters from the individual datasets and constructing stacks of biclusters.

To recap the discussion in 2.7.1, the idea of IDA is relatively straightforward, which is to combine the source datasets to form a larger one, followed by mining on the aggregated dataset. The alternative approach, SOSS, involves extracting summary statistics of interest individually from the datasets, with subsequent synthesis of the statistics. SOSS offers a number of benefits: (1) It is a more flexible approach because it does not require access to the original data; (2) It can take data heterogeneity into account by assigning weights to the summary statistics; (3) It allows different parameters to be used on different datasets when the summary statistics are extracted, which is not generally feasible in IDA. Based on these benefits, SOSS is chosen as the method for harnessing multiple datasets in this study.

The goal of this section is to experimentally demonstrate that SOSS is advantageous over IDA because of the third benefit mentioned above. The overall strategy is to apply IDA and SOSS to a group of synthetic datasets with known biclusters implanted. In the case SOSS, the CCS biclustering algorithm is separately run on the individual datasets. In contrast, when IDA is applied, CCS is run on the combined dataset. In either case, a threshold Pearson Correlation Coefficient (PCC) needs to be pre-determined before the CCS run. The two approaches are compared based on how well they retrieve the known biclusters. The measurement for the retrieval is based on Jaccard Index, as explained in the following paragraphs.

Jaccard index ([128], also known as Jaccard similarity coefficient) will be used. It aims to compare the similarity and diversity of two sample sets, and is formally defined as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Thus, Jaccard Index is simply the ratio of the intersection over the union between two sample sets, as illustrated in the left of Figure 3.9 below. When used to evaluate the performance of biclustering, Jaccard index measures the overlap between the known bicluster and the bicluster identified by a specific algorithm (CCS in this case). Higher values of the index correspond to larger overlaps between the two patterns.

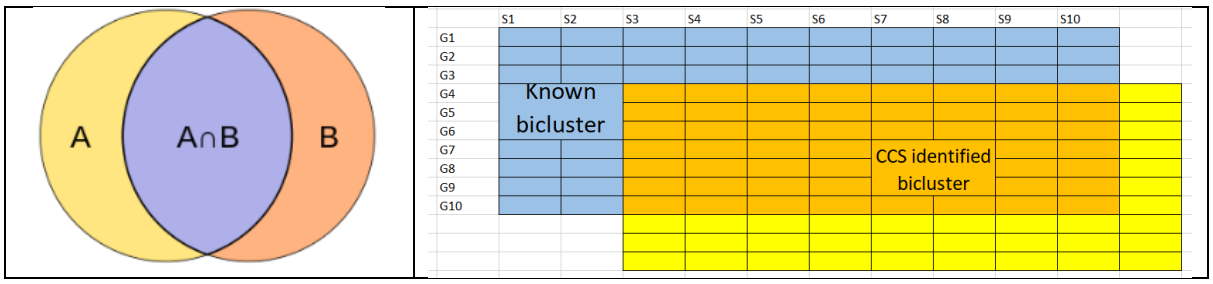


Figure 3.9: Jaccard Index used for measuring overlap (Left: general illustration of Jaccard Index. Right: Jaccard Index used to evaluate biclustering.)

In the example shown on the right panel of Figure 3.9, the known bicluster contains 100 data entries (10 genes x 10 samples), while the CCS identified bicluster has 90 entries (10 genes x 9 samples). There are 56 entries shared between them. Thus, the Jaccard Index is 56/ (100 + 90 – 56) = 0.42.

The design of the simulation study is as follows: six artificial datasets with dimension of 100 x 100 are generated. In each dataset, a single known bicluster of dimension 50 x 50 is implanted in the center of the data matrix. The data entries outside of the bicluster in each dataset are normally distributed with mean = 0 and standard deviation = 0.5 to represent the background noise (in other words, the background noise is the same across the datasets). Within each bicluster, the shift-scale pattern (explained in Section 3.2.2) is simulated. More specifically, in a given bicluster, the entries in the same column follow a normal distribution with mean μ_c that is

column specific, and standard deviation σ_b that is bicluster specific. The values of μ_c are uniformly distributed between 4 and 5, and σ_b is used to simulate the signal strength within the bicluster. Higher values of σ_b (see Table 3.3 below) correspond to weaker signals and thus lower signal-to-noise ratios in the biclusters.

The six biclusters in the six datasets, with decreasing signal-to-noise ratios, are retrieved by either IDA or SOSS, and the results are compared. When SOSS is applied, the threshold correlation coefficients, one for each bicluster, are separately determined by sampling the known biclusters using bootstrapping. The 30^{th} percentiles of the bootstrap samples are chosen as the thresholds for CCS to be run on the individual datasets. In contrast, when IDA is performed, the datasets are combined, and the average of the six threshold correlation coefficients is used for the CCS run on the combined dataset. Table 3.3 below summarizes the results.

	SSOS						IDA
	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6	Combined dataset
σ_b	0.05	0.1	0.15	0.2	0.25	0.3	
Threshold PCC	0.96	0.86	0.68	0.61	0.51	0.37	0.67
Jaccard Index	0.998	0.998	0.998	0.996	0.996	0.721	0.740

Table 3.3: Comparison of SSOS and IDA for performance of bicluster retrieval measured by Jaccard Index

As shown in Table 3.3, increasing values of σ_b are used to generate the artificial biclusters. Consequently, the known biclusters have growing levels of noise. And the threshold correlation coefficients used in the CCS runs, which are determined by bootstrap sampling as described above, are moving downward because the signals in the biclusters become weaker.

As shown in the table, the Jaccard Index figures are close to 100% in 5 of the 6 datasets when SSOS is performed, indicating that the biclusters are almost perfectly retrieved despite the varying noise levels in the biclusters. This result shows that SSOS is highly effective in

uncovering the biclusters because it allows different thresholds of PCC to be applied to the datasets. In contrast, when IDA is performed, the same PCC value (i.e. 0.67) is applied to the entire combined datasets, resulting in a poor retrieval of the biclusters as shown by the lower Jaccard Index (i.e. 0.740).

An alternative experiment for comparing IDA with SOSS is to start with a large dataset with a known large bicluster, randomly break down the large dataset into smaller ones, then compare the CCS biclustering results between before and after the breaking down. Since the smaller datasets are random partitions of the large dataset, the original bicluster is now randomly split and distributed into the smaller datasets, resulting in the same signal-to-noise ratios in the smaller datasets as in the original large dataset. Thus, the same correlation threshold for CCS can be used for both the large and the small datasets. Based on the results from the previous experiment above, one can expect the same Jaccard Index will be achieved between the large dataset and the smaller sub-parts of it. In other words, there is no advantage of performing IDA or SOSS in this case.

Although the above results show that SOSS tends to outperform IDA, this may not be always the case. As pointed out in Section 2.7.1, IDA does offer some advantages. The main reason for SOSS being favored here is related to the biclustering operation. The implementation of the CCS algorithm always perform biclustering on the entire dataset using the same correlation threshold, despite the possible presence of the internal heterogeneity. This is the reason that IDA has a low Jaccard Index in the first experiment above.

To summarize, data heterogeneity is often exhibited as varying signal-to-noise levels across different datasets. Compared to the approach of IDA, SSOS allows biclustering parameters to be separately defined based on the signal-to-noise ratios in the datasets. Hence,

SOSS maximizes the chance of uncovering the true patterns in the data. The results presented in the section further justify the use of SSOS in the current study.

3.6 Summary and discussion

This chapter is mainly concerned with identification and stacking of individual biclusters, and characterization of the bicluster stacks.

First, we identified the optimal biclustering algorithm in Section 3.1. Second, in Section 3.3.2, the selected CCS biclustering algorithm is run on thirteen microarray datasets related to prostate cancer, and thousands of biclusters are identified, with the results presented in 3.2.4. Finally, the stacks of biclusters are generated using a brute-force search strategy.

Biclustering aims to cluster the rows and the columns at the same time. It is thus naturally suited for identification of gene sets in gene expression data. CCS recognizes biclusters in which the rows show high-level of correlations as defined by Pearson Correlation Coefficient. It is chosen mainly because genes within a pathway are often co-regulated, resulting in statistically correlated expressions [117][118]. The ability in clustering two dimensions at the same time is remarkable, but it does come with a computational cost. Since CCS is not relying on random seeding and heuristic search to find the target biclusters, it takes a long time to complete a run on a regular microarray dataset.

Second, a brute-force search algorithm is designed and implemented, as described in Section 3.3.1. The search procedure is then used to find overlapped biclusters in a comprehensive manner to form stacks. The results of bicluster stacking are presented in Section 3.3.2.

Stacking of overlapped biclusters is an attempt to maximize the chance of finding related biclusters. It is a process analogous and functionally equivalent to selection of participating studies in a traditional meta-analysis. The size of a stack is determined by two factors: the number of member biclusters in the stack, and the number of genes shared by ALL the member biclusters (width). Intuitively, the bigger size of the stack, the more likely the embedded gene set is real, meaning that the gene set may be part of a biological pathway.

Finally, the stacks are characterized, with a focus on the analysis of the effect size distribution within each stack. Specifically, a bootstrapping-based method is designed to estimate the bicluster level effect sizes (Section 3.4.3.1). The method is validated using both synthetic and real data (3.4.3.2). The resulting effect size estimates and their distributions are analyzed (Section 3.4.4).

Unlike in a traditional meta-analysis, estimating the effect sizes for a bicluster has some unique challenges. First, the control samples may not be available for the bicluster. Second, a new and unique method needs to be developed to estimate the bicluster effect sizes. The current chapter has addressed these challenges. The solutions represent potential contributions to mining of gene expression data.

This chapter also includes a section that focuses on comparing two approaches for utilizing multiple datasets. A simulation study shows that the approach of synthesizing summary statistics performs better than integrative data analysis in the context of biclustering.

Chapter 4 Meta-analysis on bicluster stacks

The goal of this dissertation is to address two **overall questions**: (1) what is the data mining method best suited for finding gene sets? (2) how to utilize multiple datasets jointly in order to increase statistical strength? While the first question has been addressed in the previous chapter, the second one is the focus of the current chapter. Specifically, in this chapter, I will present the details of applying the multivariate meta-analysis (MVMA) framework to bicluster stacks to assess the significance of the gene sets.

In Chapter 3, individual biclusters are identified in order to form stacks. A stack is similar to a collection of participating studies in a traditional meta-analysis. The statistical significance of the stack is to be evaluated by a MVMA framework proposed in this chapter.

Although MVMA has been widely used in clinical trials to make inference about the overall treatment effects, applying the technique to high dimensional genome scale data is uniquely challenging.

To tackle the main challenge of high dimensionalities, a two-step method previously proposed [1] is evaluated. Despite its advantage in non-iterative estimation, it is found that that the original formulation of the method is still slow when the dimensions are relatively high. The cause of the long computation time is in step 1 of the two-step process. To speed up the estimation, an alternative method is proposed to improve step 1, while step 2 continues to adopt the original formulation. The resulting two-step procedure is then evaluated and applied to analysis of real bicluster stacks.

This chapter is organized as follows. In Section 4.1, the general idea of MVMA is reviewed, the unique challenges in the context of the current study are emphasized, and the

overall strategy is outlined. In Section 4.2, various MVMA methods are evaluated using a series of simulation experiments. In Section 4.3, an improved two-step MVMA method is applied to analysis of real bicluster stacks. Finally in Section 4.4, conclusion and discussion is given.

4.1 Background, motivations, and strategy

In this section, I will first expand the discussion on MVMA and the popular hierarchical model for random effect meta-analysis. Then I will discuss the challenges of applying MVMA to expression data, which include small sample sizes, high dimensionality, and data heterogeneity. Finally, I will outline a proposed strategy for applying MVMA to bicluster stacks.

4.1.1 Multivariate meta-analysis formulation

As introduced in section 2.7.3, an intervention may have multiple outcomes. To evaluate the intervention through meta-analysis, the outcomes can be considered either separately or jointly. If the outcomes are considered separately, the meta-analysis is univariate. On the other hand, if the outcomes are considered jointly by taking their correlations into account, the process is called multivariate meta-analysis (MVMA). The limitations of UVMA and the potentials of MVMA have been discussed in Section 2.7.3. To statistically assess a bicluster stack, each gene is modeled as an endpoint, each bicluster effectively is a study, and each stack of biclusters is similar to a collection of related studies with related endpoints. Since the genes may be correlated, and ignoring the correlation can lead to bias in effect size estimation as discussed in Chapter 2, the approach of MVMA is adopted here.

The multivariate expansion of equation (2) in Section 2.7.3 is [11]:

$$\begin{pmatrix} y_{i1} \\ \vdots \\ y_{ip} \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}, \begin{pmatrix} \sigma_{i1}^2 + \tau_1^2 & \cdots & \rho_i\sigma_{i1}\sigma_{ip} + \rho_{\tau(1,p)}\tau_1\tau_p \\ \vdots & \ddots & \vdots \\ \rho_i\sigma_{i1}\sigma_{ip} + \rho_{\tau(1,p)}\tau_1\tau_p & \cdots & \sigma_{ip}^2 + \tau_p^2 \end{pmatrix}\right) \quad (3)$$

where MVN stands for multivariate normal distribution, p is the number of endpoints (dimension), y_i is the vector of observed effect sizes for study i, μ_j is the true effect size for effect j, σ_{ij} is the within-study standard deviation for study i and effect j, τ_j is the between-study standard deviation for effect j, ρ_i is the within-study correlation, and ρ_τ is the between-study correlation. This formulation can be seen as a hierarchical model because of the two-level specification for the effect sizes.

4.1.2 Model fitting with small samples and high dimensionalities

There are numerous methods for fitting random-effects meta-analysis models, which can be broadly classified into frequentist and Bayesian methods. In brief, the frequentist school assumes some hypothesis is correct and the parameters specifying the hypothesis are fixed. The observed data are drawn from the assumed distribution. It does not depend on a prior distribution which can be somewhat subjective. In contrast, the Bayesian framework models the uncertainty about a hypothesis by treating the model parameters as random variables. It depends on prior distributions of the parameters.

For the frequentist framework, a natural choice is the Maximum Likelihood (ML) estimator, which is a procedure of finding the values of a set of parameters that maximize a known likelihood function [129].

It has been shown that in the random effects model, the ML estimates (MLE) for the variances are typically biased; they systematically under-estimate the variance parameters. Restricted Maximum Likelihood (REML) [130] corrects this bias by adopting the mixed effect models:

$$observation = (fixed\ effects\ part) + (random\ effects\ part) + error$$

REML works by first getting regression residuals for the observations modeled by the fixed effects portion. To achieve this, all variance components are being ignored. Then, the fixed effect part is taken out, and the residuals are used to estimate the variance of the random effects through maximum likelihood estimation. These steps result in unbiased estimates of the variance components. Because of this feature, REML has become the default frequentist method for MVMA [131][132].

However, REML is problematic when the number of studies is very small, because the likelihood function tends to be flat and numeric problems would arise as a result when attempting to maximize the function [1]. The number of studies is equivalent to the length of a bicluster stack, which is 7 maximum in the current study. Given the smaller numbers of biclusters, REML may be problematic.

With regard to the Bayesian approach, a key advantage is its ability to allow external knowledge to be incorporated in the model via informative priors. In addition, the posterior distribution provides full information about the estimate, making it relatively easy to derive the credible interval. In a full Bayesian framework, all parameters are treated as random variables, which can be estimated by Markov chain Monte Carlo (MCMC) using suitable prior distributions [11][133].

In [11], an excellent review is given about the Bayesian method applied to MVMA. A key advantage described in that article is that it allows easy implementation of the hierarchical model widely used for random-effects meta-analysis. When the sample sizes or the number of studies is small, the Bayesian approach provides uncertainty information about the estimates by posterior distributions. In addition, in the case when prior knowledge about the parameters (either the overall effect sizes or the between-study covariance) is available, the Bayesian method can easily allow incorporation of the knowledge. However, the downside is that the MCMC-based procedures widely used in Bayesian inference are very time-consuming.

4.1.3 Challenges of applying MVMA to high dimensional data

As mentioned earlier, the strategy of adopting MVMA to analysis of biclusters is to treat each bicluster in a stack as an individual study, and each gene in the stack as an endpoint. The intervention to be evaluated is implicitly implied as the effect that causes the genes to change their activities in the samples covered by the biclusters.

Despite the successful and extensive use of MVMA in clinical trials, adopting the same technique to data mining on high-dimensional public data poses a number of challenges. First, public data come from different laboratories and are generated with diverse research contexts. Thus, heterogeneity needs to be considered and properly modeled.

Second, the current study aims to uncover gene sets from genome scale gene expression data. A typical gene set can contain over a hundred genes. Analyzing the corresponding stack by traditional MVMA can be computationally too demanding. For example, suppose a bicluster stack has a dimension (i.e. number of genes, which corresponds to the number of endpoints in a

traditional MVMA) of 100, then there will be 100 effect sizes plus 100 * (100 + 1)/2 = 5050 free elements from the between-study covariance structure to be estimated. The rationale behind the count of the free elements is that a covariance matrix is a symmetric matrix. The total number (100 + 5050) of parameters is prohibitively large to be inferred by a typical MCMC procedure on a regular desktop computer. This challenge applies not only to Bayesian inference, but also to REML.

Third, as mentioned above, data scarcity is a common problem in data mining. In the current research, the number of available datasets is limited. As a result, the estimated covariance matrix may not be positive definite.

In the next section, a strategy is proposed to tackle the challenges of applying MVMA to bicluster stacks, including data heterogeneity, high dimensionality, and data scarcity as discussed above.

4.1.4 Overview of the proposed strategy

Given the aforementioned challenges, a possible solution is to adopt a two-step strategy as opposed to the traditional procedures that iteratively estimate all the parameters, which is referred to as the one-step approach. The two-step process involves first estimating the between-study covariance non-iteratively, followed by calculation of the overall effect sizes.

In step 1, the between-study covariance matrix is estimated by an efficient non-iterative procedure. DerSimonian and Laird first proposed to use method of moments to estimate the variance between studies [85]. Later, Jackson et al. extended the method to the multivariate settings [134]. Jackson's method will be explained and evaluated in Section 4.3.1. Given the

small number of biclusters and the high dimensionality in a typical stack, it may be necessary to apply regularization to the covariance matrix to impose structural sparsity and ensure positive definiteness.

In step 2, the estimated covariance matrix will then be used to make inference about overall effect sizes. In this step, it may be necessary to take the uncertainty of the between-study covariance into account.

In Section 4.2 below, the traditional one-step methods is first evaluated for combining biclusters. Then in 4.3, a two-step method proposed by Jackson and Riley [1] is introduced and assessed. An improvement over Jackson's original method (improving on the first step) is suggested and compared with the other methods. Finally in 4.4, the improved two-step method is applied to analysis of six real bicluster stacks.

4.2 Evaluation of traditional MVMA methods

Before getting into the details of the two-step method, it is necessary to first evaluate the traditional one-step approach and highlight its limitations in the context of bicluster meta-analysis.

This section is organized as follows. First, I will describe the preparation of synthetic data, then apply the traditional one-step methods to analyze low-dimensional bicluster stacks using the synthetic data, followed by demonstration of their limitations. Next, I will describe a two-step strategy and compare it with the one-step counterpart. Finally, I will expand the two-step procedure to analysis of high-dimensional bicluster stacks. All the experiments are conducted through simulations using synthetic data.

4.2.1 Preparation of simulated data

The design of the experiments in this and the later sections is schematically illustrated in Figure 4.1 below.

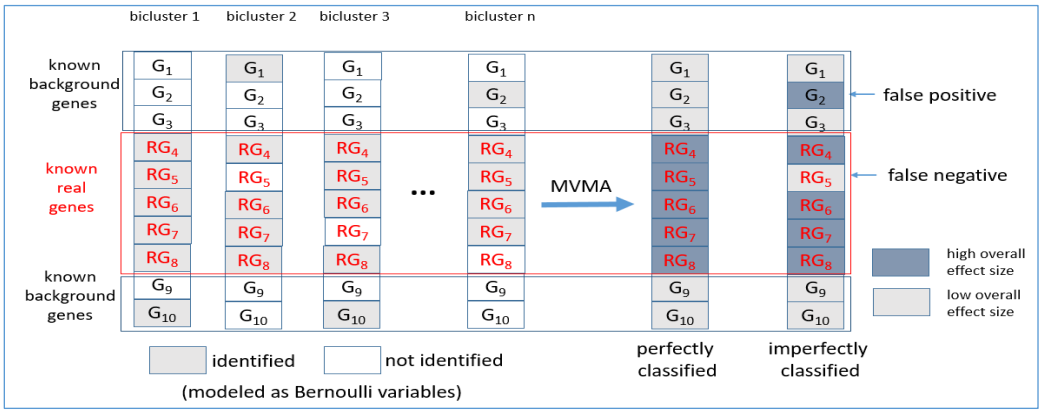


Figure 4.1: Schematic representation of the simulation study of MVMA on a bicluster stack (dimension = 10)

The synthetic data are prepared to simulate a bicluster stack. In each bicluster in the stack, a fixed subset of genes are designed as real genes (RGs, red-highlighted in the figure), while the others are background genes. In a perfect situation, all the real genes will be identified by the biclustering algorithm CCS. However, due to the small sample sizes and the noise in realistic data, some of the real genes may be missed, and some of the background genes may be falsely identified. To simulate this situation, each gene is modeled as Bernoulli random variable with two endpoints: identified and not identified. The real genes, with probability p = 0.75, is more likely to be identified than the background genes (p = 0.25). The identified genes, marked by a gray background in the figure, have a known effect size of 2, while the non-identified genes have a zero effect size. Note that there are both real genes (RG, red) and background genes (G)

in each bicluster that are identified genes. Each bicluster has a known within-study covariance matrix. A bicluster-specific diagonal design matrix is used to indicate which genes are identified.

The synthetic data is generated by following the aforementioned two-level hierarchical model that incorporates the design matrices. Specifically,

$$y_i \sim MVN\ (\theta_i, \Sigma_i) \quad (4)$$

$$\theta_i \sim MVN\ (X_i\mu, X_i\Omega X_i^T) \quad (5)$$

where X_i is the design matrix for bicluster i, and *MVN* stands for multivariate normal distribution.

This hierarchical model can be fitted by either the Bayesian or the frequentist paradigm. If the Bayesian framework is chosen, the inverse Wishart distribution, which is the conjugate and most frequently used prior for covariance matrix Ω [135], is used:

$$\Omega \sim W^{-1}(\Psi, \nu)$$

where Ψ is a scale matrix, and ν is the degree of freedom, which is chosen as the dimension + 1. The choice for the scale matrix is an identity matrix.

4.2.2 Performance of the traditional MVMA methods

The main goal of this section is to examine various traditional MVMA methods (either Bayesian or frequentist) for their abilities in combining low-dimensional biclusters. Then, attempt will be made to expand their application to high-dimensional biclusters. Finally, their limitations will be highlighted and discussed.

First, a standard Markov Chain Monte Carlo (MCMC) procedure is carried out to estimate the parameters using the software tool JAGS (Just Another Gibbs Sampler [136]).

Figure 4.2 below shows convergence of the Markov chains. The visual inspection of the traces suggest that the MCMC settings are appropriate.

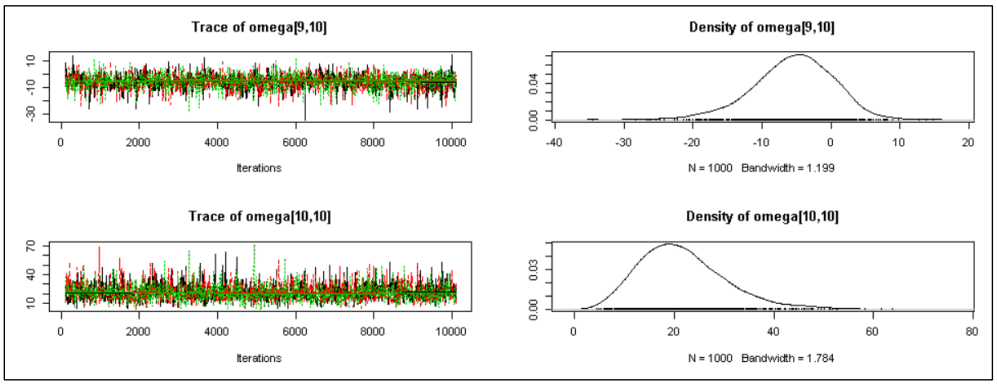


Figure 4.2: Convergence of the Markov chains in Bayesian estimation of the MVMA parameters

(Note: the vertical axis displays the sampled values of the parameters being estimated, while the horizontal axis displays the iteration indices)

Now that MCMC procedure has been properly set up, a second experiment is carried out to demonstrate the effectiveness of the Bayesian procedure in combining biclusters,

As discussed above, meta-analysis is meant to combine the effect sizes in the participating studies. If properly designed and conducted, a meta-analysis usually results in reducing width of the confidence intervals for the effect sizes as more data are included [137]. The narrowing of the confidence interval indicates that we are gaining more confidence about the effect size. The boxplot in Figure 4.3 below shows such observation. The boxplot is drawn using the highest density intervals of the effect size estimates from independent MCMC outputs. This result demonstrates that the Bayesian procedure is capable of combining biclusters.

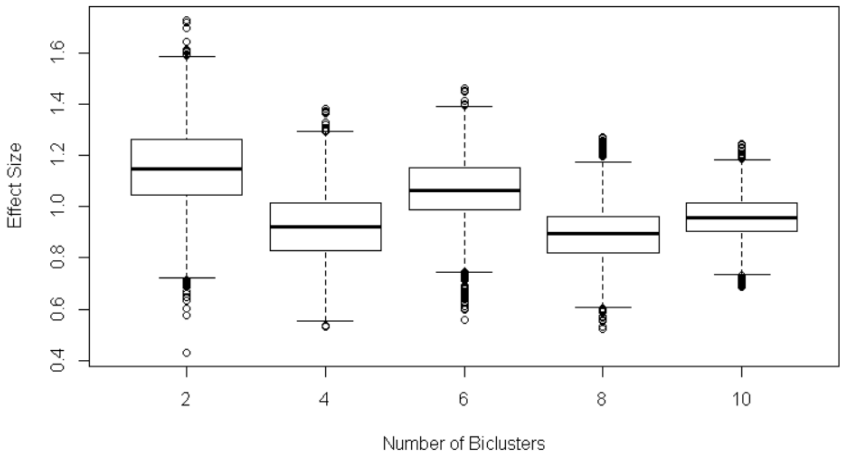


Figure 4.3: Change of effect size estimated by the one-step Bayesian method as the number of biclusters increases

Given the capability of the Bayesian MVMA method in combining biclusters, the next step is to quantify the performance so that later comparison with the two-step counterpart can be made. Since real genes are implanted in each synthetic biclusters as described above, the MVMA performance is measured by how well the procedure uncovers the known real genes while excluding the background genes. Thus, MVMA can be seen as a classifier that aims to classify real genes vs. background genes. A key factor in doing this study with simulated data was to have an absolute gold standard of what the correct classification is of each gene. Frequently used performance measures for classification, including recall, precision (or truth positive rate), and specificity (or truth negative rate), are used here. Formally speaking, recall (also known as sensitivity) is the fraction of successfully identified positive instances among all the positive instances existing in the data. Precision (also called positive predictive value) is the fraction of true positives among all the instances identified as positives, and specificity measures the proportion of identified negatives among all the negative instance existing in the data. (For a review on recall, precision, and specificity, please refer to [137])

In the context of the current study, recall is defined as the fraction of uncovered real genes among all the real genes in the data, precision is the fraction of the real genes among all the genes identified by the procedure, and specificity is the proportion of the actual background genes among those recognized as background genes by the procedure.

In addition to the Bayesian procedure, a number of frequentist methods are also considered for their performance in meta-analyzing biclusters, including Maximum Likelihood (ML) [138], Method of Moments (MM) [139], and Restricted Maximum Likelihood (REML) [140]. In the case of Bayesian inference, the highest density interval (HDI) is derived from the posterior distribution of effect size of each gene. If the 95% HDI contains the known effect size, then the gene is considered as being recognized as real gene by the MVMA procedure. In the case of the frequentist approach, the confidence interval of the fixed-effects coefficient is used instead [141].

Experiment 1: Effect of increasing number of biclusters in a stack on recall, precision and specificity

In this experiment, various one-step methods are examined for how they respond increasing number of biclusters in a stack. Figure 4.4 shows the change of recall.

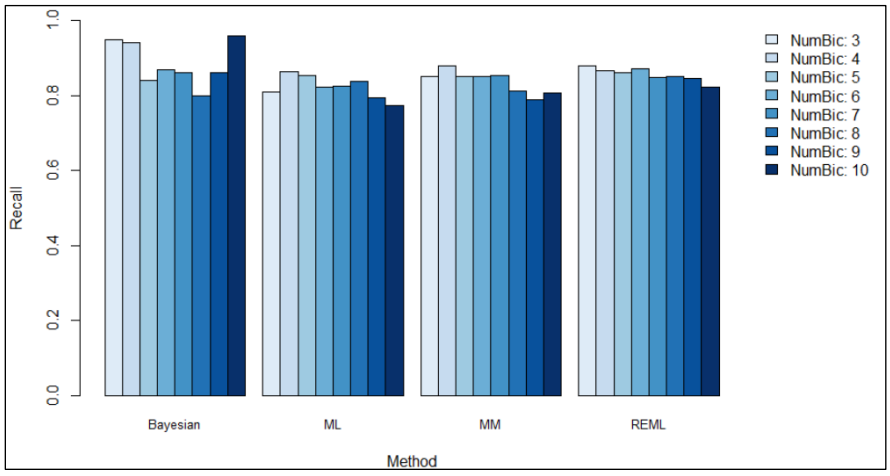


Figure 4.4: Change of recall as the number of biclusters increases

From Figure 4.4 above, all the methods perform reasonably well in terms of recall. The frequentist methods appear to decrease slightly as the number of biclusters increases, while the Bayesian approach seems to have initial drop, followed by a rebound. It is unclear what causes the rebound of recall as the number of biclusters increases when the Bayesian approach is taken. Furthermore, this non-monotonic change in recall is not seen in the two-step MVMA approach. Section 4.3.2 below provides some speculation and discusses what can be done to investigate the discrepancy.

In addition to the recall, precision and specificity are also looked at in the same experiment. Figures 4.5 and 4.6 below show the results:

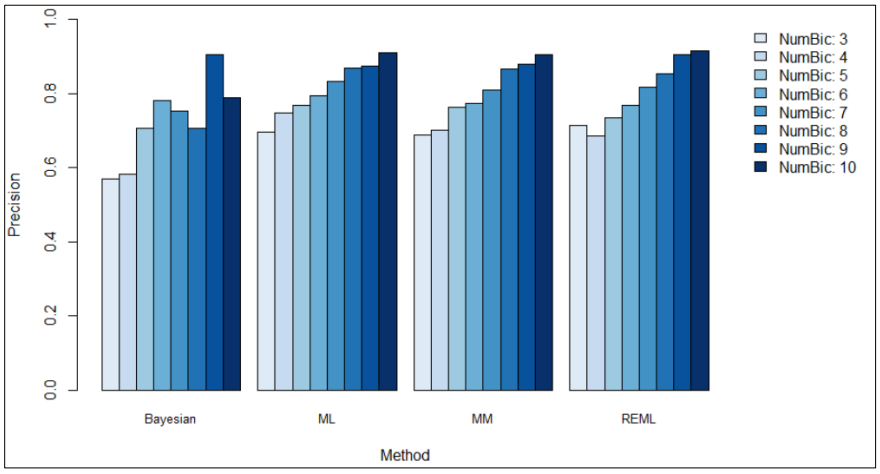


Figure 4.5: Change of precision as the number of biclusters increases

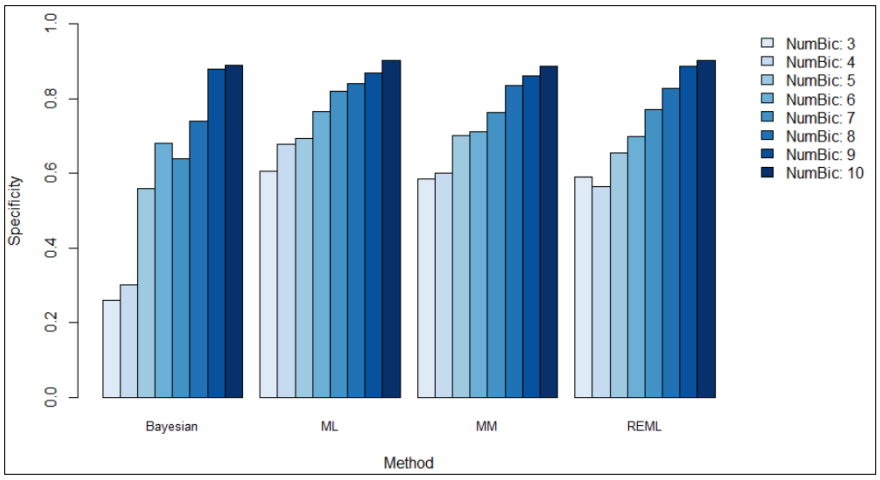


Figure 4.6: Change of specificity in identifying real genes as the number of biclusters increases

From Figures 4.5 and 4.6 above, it is clear that both precision and specificity improve significantly as the number of biclusters increases. So far, the traditional one-step MVMA methods have been shown to be effective in combining biclusters. However, there was no further investigation to compare the various one-step methods (Bayesian, MEML, ML, and MM) in terms of performance, due to the poor scaling of the one-step methods as the dimension increases, as described in Experiment 2 below.

Experiment 2: Effect of increasing dimension of stack on performance of one-step methods

Although the one-step methods are capable of combining biclusters, their performance deteriorate quickly as the dimension increases, which is shown in the Figure 4.7 below. In this experiment, the run times are recorded for completing estimation runs under different dimensions.

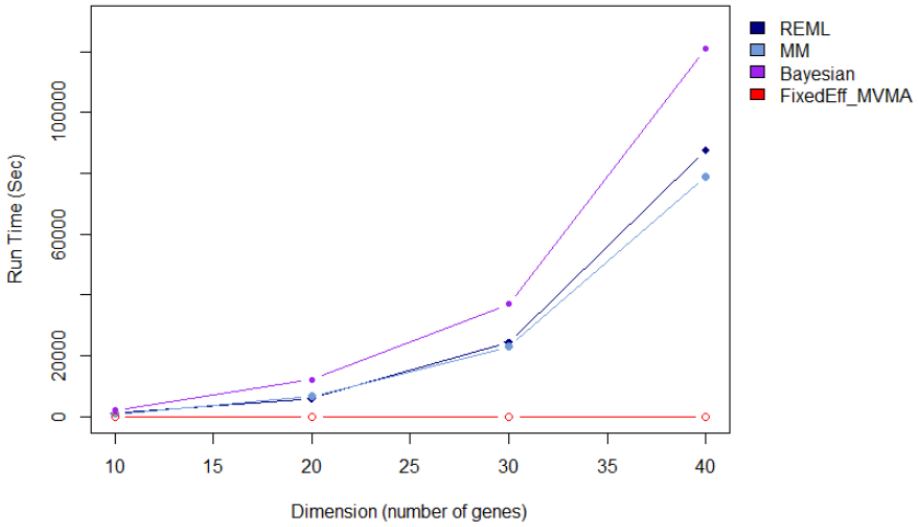


Figure 4.7: Change of run time as the dimension increases

As the result shows, the run time appears to grow exponentially as the dimension (or number of genes) increases. For example, with regard to the Bayesian method, when the dimension = 40, it takes more than 33 hours on a Windows desktop to complete a run of the MCMC-based estimation procedure. Since the dimension of a typical bicluster stack can reach over 100, the run time, estimated to be 7-10 days by projection, would be prohibitively too long.

4.2.3 Summary of the traditional MVMA methods

To summarize, based on the simulations presented above, the traditional MVMA methods are effective under low-dimensional settings. However, their computation times become unmanageable as the dimension increases. This observation makes the traditional MVMA methods unpractical even for biclusters that are of moderate dimensions. As shown in Table 3.2, the dimensions of the six selected bicluster stacks are in the range of 83-175. This obstacle leads to the adoption of a two-step approach, as discussed in the next section.

4.3 A two-step MVMA method

A two-step approach involves estimating the between-study covariance and the overall effect sizes sequentially rather than concurrently. The goal of this section is to evaluate a previously proposed two-step strategy and suggest an improvement.

The two-step method needs to solve a couple of problems. First, how to efficiently estimate the between-study covariance matrix that is as accurate as possible, because over- or under-estimate of the variances can lead to biased inference for the overall effect sizes. Second, the one-step method naturally incorporates the uncertainty of the between-study covariance estimate when inferring the overall effect sizes. How can the two-step method take the uncertainty of between-study covariance into account?

The agenda of this section is as follow. I will first describe a two-step procedure proposed by Jackson and Riley [1] (in 4.3.1), evaluate its performance, and point out its limitation based on the results from simulation experiments (in 4.3.2). Then, I will suggest an improvement over step 1 of the original two-step procedure. The suggested improvement is then evaluated by plugging into the original two-step framework (in 4.3.3). Finally, the altered procedure is applied to analysis of six bicluster stacks.

4.3.1 The two-step method proposed by Jackson and Riley

As mentioned earlier, DerSimonian and Laird proposed to use method of moments to estimate the variance between studies in 1986 [85], which since then has become a popular and

default method for quantifying heterogeneity among the participating studies in a meta-analysis. Jackson et al. later extended the method to the multivariate settings [134]. In another article by Jackson and Riley [1], they proposed a two-step MVMA procedure that aimed to tackle situations with small numbers of studies.

Specifically, in their procedure, the between-study covariance is first estimated by method of moments [134] as the step 1, which is then used to infer the overall effect sizes as the step 2, based on multivariate t-distribution rather than normal distribution as assumed by the hierarchical model commonly used in random-effects meta-analysis. The rationale behind the step 2 is that the effect size estimates are better approximated by t distribution when the number of available studies is small. Thus, Jackson's method takes the uncertainty of the between-study covariance into account when making inference about the overall effect sizes. It makes sense statistically and is appealing in settings where the numbers of studies are small.

Before looking into the details of Jackson's method, it is helpful to review mapping of MVMA clinical trial design terms to the gene expression terms in the current study. A bicluster is equivalent to a participating study in a traditional meta-analysis. A stack of biclusters, each coming from a separate microarray dataset, is analogous to the collection of the participating studies. Furthermore, each gene is modeled as an endpoint, the total number of genes is the dimension of the stack and of the gene set embedded in the stack. The primary goal of the MVMA here is to make inference about the effect sizes of the genes in the gene set.

Let's take a close look at the details of the Jackson's method as described in their paper[1]. First, the model in their method considers the weights of the study outcomes are: $w_i = 1/(\sigma_i^2 + \hat{\tau}^2)$ instead of the usual σ_i^{-2}. If both the within-study variance (σ_i^{-2}) and the between-study variance estimate ($\hat{\tau}^{-2}$, estimated by MM) are available, then

$$(n-1)H^2 = \sum w_i(Y_i - \hat{\mu})^2 \sim \chi_i^2 \quad (6)$$

Since $\hat{\mu}$ and H^2 are independent, then

$$\frac{(\hat{\mu} - \mu)\sqrt{\sum w_i}}{H} \sim t_{n-1} \quad (7)$$

Several previous studies have suggested using the above t-distribution, instead of the conventional normal distribution, to make reference about μ [142][143][144]. Since a meta-analysis can be considered a special case of meta-regression, Jackson and Riley adopts the meta-regression formulation as below:

$$Y_i = N(X_i\beta, S_i + \Sigma)$$

for i = 1, 2, …, n, where n is the number of studies (biclusters in this case), and Y_i denotes d x 1 vector of outcomes associate with study i (d = dimension). Furthermore, S_i is the d x d corresponding within-study covariance matrix, Σ is the d x d between-study covariance matrix. For a MVMA, if study *i* provides all outcomes, then X_i is the d x d identity matrix and β is the d x 1 average outcome or effects vector.

Let Y denote the stacked vector of the observed entries of Y_i and let X denote its design matrix. Further, let $\text{Var}(Y) = \Delta^{-1}$, where Δ incorporates both the within and the estimated between-study variances, then

$$\widehat{\beta} = (X^t\Delta X)^{-1}X^t\Delta Y$$

which is approximately normally distributed with covariance matrix:

$$Var(\hat{\beta}) = C = (X^t\Delta X)^{-1}$$

The conventional procedure for making inferences about β is:

$$C^{-\frac{1}{2}}(\hat{\beta} - \beta) \sim Z_p$$

where Z_p denotes a standard multivariate normal distribution of dimension p. However, this formulation does not take into account the uncertainty in estimation of Σ, which leads to the multivariate extension of (6) proposed by Jackson and Riley:

$$(N-p)H^2 = \left(Y-\hat{Y}\right)^t \Delta \left(Y-\hat{Y}\right) \sim \chi^2_{(N-p)} \quad (8)$$

where N is the total number of estimates, which equals to n x d, if there is not missing outcome, and $\hat{Y} = X(X^t \Delta X)^{-1} X^t \Delta Y$, which is the fitted vector of Y. Multivariate generalization of (7) leads to:

$$\frac{C^{-1/2}\,(\hat{\beta} - \beta)}{H} \sim t(I_p)_{(N-p)} \quad (9)$$

where H^2 is now given by (8), and $t(R)_{(N-p)}$ denotes a central multivariate t distribution, with correlation matrix R and degrees of freedom (N-p) and I_p denotes the p x p identity matrix. (9) can now be used to make inference about the overall effect sizes using the R package mvtnorm that implements the multivariate t distribution as described in [145][146][147].

Given above formulation by Jackson and Riley [1], I implement it in the R language and evaluate its performance for bicluster meta-analysis, as detailed in the next section.

4.3.2 Performance and limitations of the implementation Jackson's method

First, an experiment is carried out to confirm whether Jackson's method is able to combine biclusters. The method of moments as originally proposed by the authors is used to estimate the between-study covariance matrix, which is then plugged into the hierarchical model to estimate the overall effect sizes based on multivariate t distribution.

Similar to the one-step method discussed above in section 4.2.1.1, synthetic biclusters are used with the same setup. For each of the bicluster stack sizes, 1000 runs are repeated. Figure 4.8 below shows how the effect size estimates change as the number of biclusters in a stack increases.

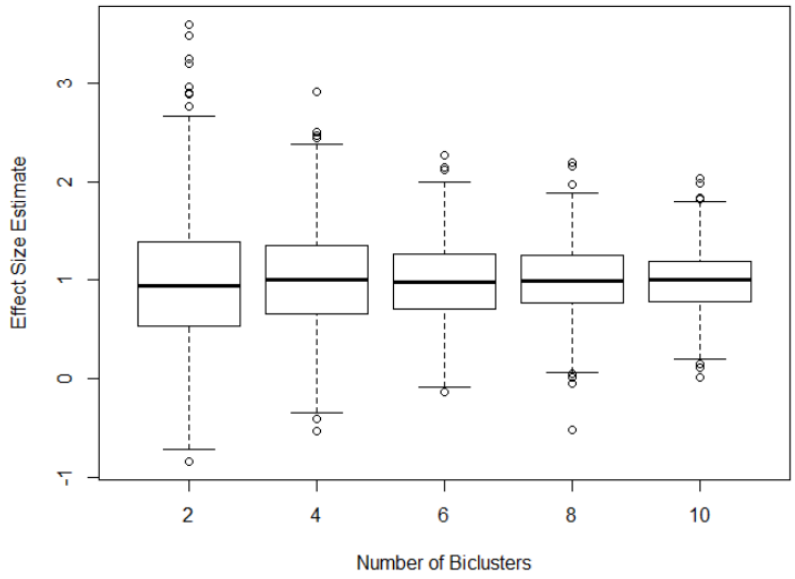


Figure 4.8: Change of effect size estimated by Jackson's method as the number of biclusters increases

The result is similar to that of the one-step method: the width of the confidence interval for the effect size estimates decreases as the number of biclusters increases, indicating the capability in combining biclusters by Jackson's method.

Next, the same criteria, including recall, precision, and specificity, are used to evaluate Jackson's method. As shown in Figure 4.9 below, recall decreases as the number of biclusters increases. This is because of the shrinking of the confidence interval as more biclusters are added to the stack. When the confidence intervals are narrowing, some of real genes in the synthetic

stack are more likely to be missed, resulting in the decreasing recall. On the other hand, the specificity improves as a result of the narrowing confidence intervals.

As mentioned earlier, in the one-step Bayesian method, the recall seems to bounce back as the number of biclusters continue to increase. This result is not observed with Jackson's method. The cause for the discrepancy is not entirely clear at this point. Perhaps, the re-bounce of recall observed in the Bayesian method may have something to do with the random-walk nature of the Markov Chain Monte Carlo (MCMC) procedure. In Jackson's method, it is non-iterative and is solely based on the assumption that the study-specific parameters are multivariate t distributed. As a result, the recall monotonically decreases without re-bounce as the number of biclusters increases. Further investigation using different MCMC settings and t distribution parameters may shed more light on what exactly causes the discrepancy. This will be part of the future work as we expand the current research.

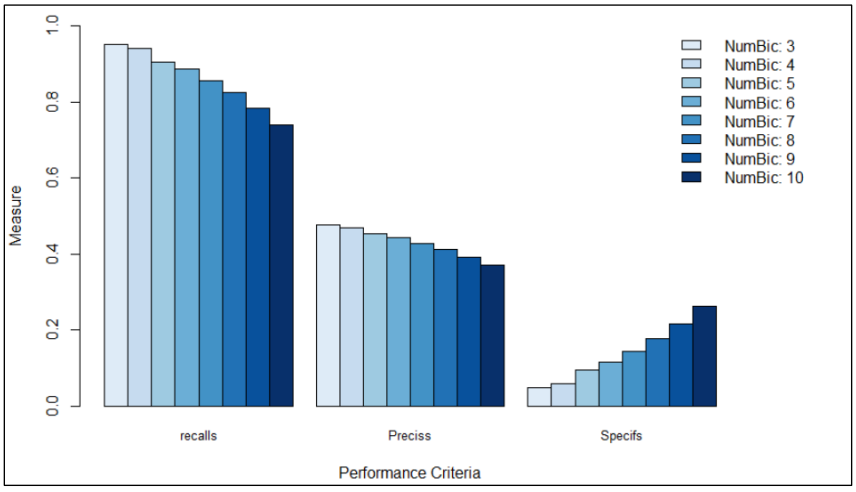


Figure 4.9: Performance of Jackson's method in combining biclusters

A significant benefit of Jackson's method is that it is computationally efficient because both steps are non-iterative. However, it is a well-known fact that method of moments used in step 1 often leads to a biased estimator [148]. Since a biased estimate for between-study covariance can lead to an inaccurate estimate for the overall effect sizes in step 2, we consider an alternative non-biased step 1 method that is based on sample covariance matrix, weighted by the within-study variances. The resulting weighted sample covariance matrix is then subject to lasso regularization to impose structural sparsity due to the p >> n situation.

To summarize, although the method proposed by Jackson and Ridley is statistically sound and has shown promise in tackling the issues of small numbers of studies and high dimensionalities, its step 1 is based on method of moments, which can lead to a biased estimate for the between-study covariance matrix. The following section is devoted to discussion and evaluation of an alternative step 1 method.

4.3.3 An improved method for estimating between-study covariance

Given the downside of Jackson's method as discussed above, I propose a new step 1 method for approximating the between-study covariance structure. The new method is based on sparse estimates of the between-study covariance by regularizing the weighted sample covariance matrix. The obtained estimate is then used to make reference about the overall effect sizes following Jackson's original framework.

4.3.3.1 Description of the new step 1 method

Estimation of covariance matrix is of critical importance in many machine learning applications [149][150]. For example, it is needed for the estimation of principal components and eigenvalues in order to obtain an interpretable lower dimensional data representation. Linear Discriminant Analysis (LDA) [151] also requires estimation of covariance matrix for classification of Gaussian data. In addition, a precision matrix (inverse of covariance matrix) can provide useful insight into how variables are conditionally independent or dependent.

Estimation of covariance matrix is not trivial and has sparked major research interest. The primary challenges include (1) the positive-definiteness requirement, and (2) high-dimensionality that causes the number of parameters to be estimated to grow quadratically.

Although the sample covariance matrix $S = \frac{1}{n}\sum_{i=1}^{n} Y_i Y_i'$ is an unbiased estimate of the population covariance matrix Σ, it is a poor estimator when the dimension far exceeds the sample size (p >> n). It tends to be singular and may misrepresent the eigenstructure of Σ by introducing more spread-out eigenvalues [152].

As described earlier in section 4.3.1, Jackson and Riley propose to use method of moments (MM) to estimate the between-study covariance matrix in their two-step procedure [1]. Unfortunately, MM is too memory and CPU exhaustive when the dimension is relatively high, making the method not directly useable in bicluster multivariate meta-analysis. An improvement has to be made over Jackson's step 1 method. Here, I propose to use weighted sample covariance matrix to approximate the between-study covariance in the first step of the two-step procedure. Then regularization is imposed on the estimated between-study covariance matrix due to the p >> n situation. The following paragraphs are intended to provide detailed description of the proposed method.

As mentioned above in Chapter 2, a meta-analysis achieves its goal by assembling a collection of related studies and deriving a consensus among the studies. Typically, the participating studies usually have varying weights in terms of strength of evidence due to different sample sizes and sample qualities. Thus, the overall effect size is usually the weighted mean of the individual effect sizes from the participating studies, as described by Hedges and Olkin [153]:

$$\bar{T} = \frac{\sum W_i T_i}{\sum W_i}$$

where T_i is the effect size computed from the ith study, and W_i is the weight assigned to the effect size in the *i*th study.

Similar to weighted mean, I argue that the between-study variance/covariance can be approximated by weighting the participating studies. In other words, the participating studies do not contribute to the variance equally, just as they do not contribute to the overall effect size equally. This leads to the proposal of using weighted sample covariance matrix, defined as below, to estimate the between-study covariance matrix.

$$q_{jk} = \frac{1}{1 - \sum_{i=1}^{N} w_i^2} \sum_{i=1}^{N} w_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

The weighted estimates for the mean and covariance have potential pitfalls. Specifically, the sample mean and sample covariance are not robust statistics. Hence, they are sensitive to outliers. In addition, as mentioned above, in high dimensional settings, sample covariance matrix is a poor estimate for the population covariance matrix. It tends to be singular and may misrepresents the eigenstructure of Σ by introducing more spread-out eigenvalues [152]. A solution is to impose regularization to the sample covariance matrix. The most popular technique is perhaps the graphical lasso algorithm proposed by Friedman et al [154].

The graphical lasso aims to estimate sparse undirected graphical models through L1 (lasso) regularization. The basic model assumes the observations follow a multivariate Gaussian distribution with mean μ and covariance matrix Σ. If the *ij*th component of Σ^{-1} is zero, then variables *i* and *j* are conditionally independent, given the other variables. This is the rationale for imposing an L1 penalty for the estimate of Σ^{-1} to raise the sparsity.

Suppose we have N data points that are multivariate normally distributed with dimension p, mean μ and covariance Σ. Let $\Theta = \Sigma^{-1}$, and let S be the empirical covariance matrix, the problem is to maximize the penalized log-likelihood

$$\log\det\Theta - \mathrm{tr}\,(S\Theta) - \rho\left|\left|\theta\right|\right|_1$$

over non-negative definite matrices Θ [155]. The tuning parameter ρ controls the level of sparsity. The R package glasso is a popular and fast implementation of the algorithm. It allows one to efficiently build a series of models with different values of the tuning parameter.

In the next few sections, numerous simulations will be conducted to evaluate the new step 1 method. Specifically, the performance of the new two-step procedure will be assessed on how it responds to varying sample sizes, increasing levels of data heterogeneity and dimensions.

4.3.3.2 Effect of numbers of biclusters on performance

In a typical meta-analysis, sample size primarily means the number of participating studies. In the current research, the number of participating studies is equivalent to the number of biclusters in a bicluster stack. To inspect whether increased numbers of biclusters in a stack would boost the classification performance, four stacks are generated, with the number of

biclusters as 2, 4, 7, and 10, respectively. The method for generating the synthetic data is described earlier in Section 4.2.1.

For each stack, 13 data points are collected using various Bernoulli probabilities for the real and background genes. For example, for data point #1, the average Bernoulli probabilities for the real and background genes are 0.34 and 0.11, respectively. These data points can be classified into three groups: (1) low signal, low noise; (2) high signal, low noise; (3) high signal, high nose. (Table 4.1). For each data point, 1000 runs are repeated to obtain the average recall and false positive rates.

	Bernoulli probabilities used in the synthetic data												
	Low signal, low noise				High signal, low noise					High signal, high noise			
	1	2	3	4	5	6	7	8	9	10	11	12	13
Real Genes	0.01	0.18	0.21	0.26	0.34	0.41	0.46	0.51	0.56	0.61	0.64	0.71	0.91
Background Genes	0.01	0.02	0.03	0.06	0.11	0.16	0.21	0.26	0.31	0.36	0.41	0.51	0.71

Table 4.1: The Bernoulli probabilities used for generating the synthetic bicluster stacks

The classification results are summarized as receiver operating characteristic (ROC) curves as shown in Figure 4.10 below.

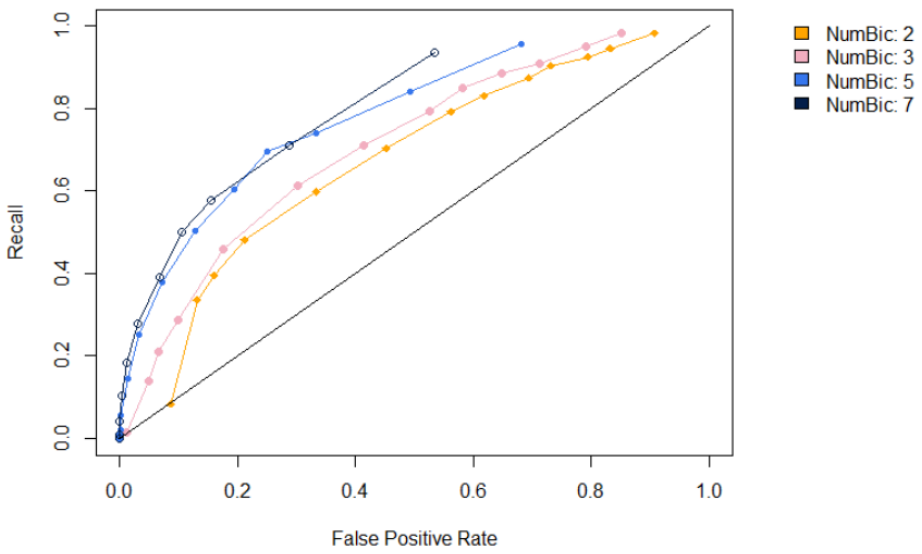


Figure 4.10: Effect of numbers of biclusters (lengths of the stacks) on classification performance

The result in Figure 4.10 is consistent with that in Figure 4.8. The effect size estimates, assumed to be t distributed, have relatively wide confidence intervals when the number of bicluster is small. Wider CI's leads to higher recall and false positive rate (FPR). As the number biclusters increases, the CI's become narrower, resulting in lower recall, but lower FPR at the same time. In other words, as the number of biclusters increases, the classifier becomes more conservative.

Since the curves do not cover the same spectrum of FPRs, it is difficult to compare the performance of the stacks with different lengths by AUC (area under the curve). A sensible interpretation of the result is perhaps that as the stacks become longer, the classifier turns more conservative, as stated above.

4.3.3.3 Effect of data heterogeneity on performance

In the context of the current research, data heterogeneity have two sources: (1) different biclusters may have different effect sizes (bicluster-level effect sizes), which is caused by between-study variances; (2) different biclusters may have different within-study variances, and thus different weights. To examine the impact of data heterogeneity, two experiments are conducted as described below.

Experiment 1: heterogeneity from between-study variances

In this experiment, a simulation is conducted to examine the impact of varying bicluster-level effect sizes on the classification. Three synthetic stacks, each containing two biclusters, are generated with known between-study variances as 0.2, 2, and 20. The biclusters within a stack are more divergent in terms of effect size as the between-study variance increases.

As in the previous experiment, 13 data points with the same Bernoulli probabilities are collected. For each parameter and each data point, the procedure is repeated 1000 times to obtain the average recalls and false positive rates. The ROC curve result is shown in Figure 4.11 below:

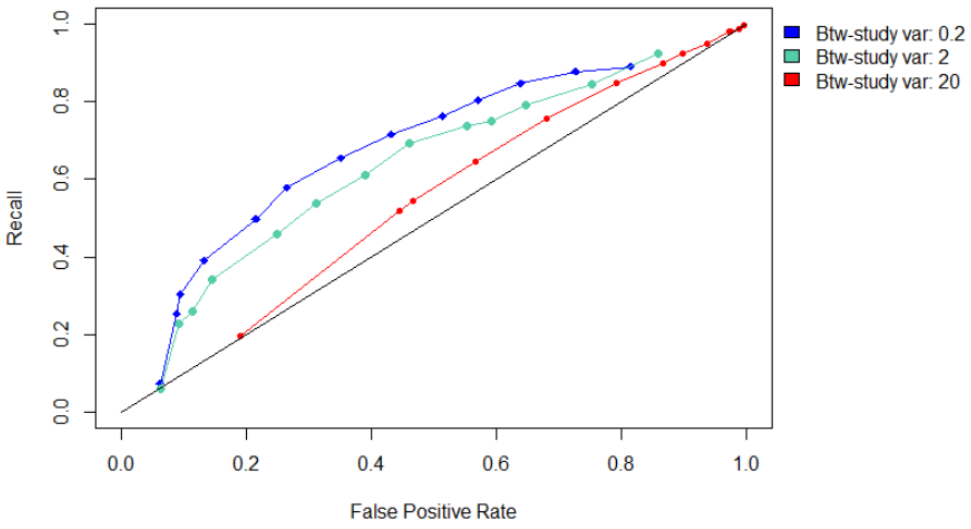


Figure 4.11: Impact of data heterogeneity on classification performance

(Note: Btw-study var: between-study variance where variance is set to 0.2, 2, and 20)

In this experiment, the data heterogeneity comes from the between-study variances. Higher variance means higher heterogeneity. The result shows that as the data becomes more heterogeneous, the classification performance deteriorates. This is because the effect sizes within the same stack deviate more across the biclusters when the between-study variance increases. The divergent effect sizes lead to wider confidence intervals for the overall effect size estimates, which then increase the recall but also the false positive rate at the same time.

Experiment 2: heterogeneity from within-study variances

The goal of this experiment is to compare standard vs. weighted sample covariance matrices as an estimate for between-study variance/covariance.

As described earlier in section 4.3.3.1, the proposed two-step procedure estimates the between-study covariance by weighted sample covariance matrix, as opposed by method of moments as in Jackson and Ridley [134]. The weight of each bicluster is the inverse of the within-study variance. The idea of the experiment is to generate artificial stacks in which the weights of the member biclusters are as diverse as possible. In other words, some biclusters carry significant more weights than the others in the same stack. To maximize the weight divergence, the within-study variances are uniformly distributed between 0 and 1. The between-study covariance matrix is then estimated by either standard or weighted sample covariance matrix, and the classification performance is then compared.

The comparison is done under various settings, including three known between-study variances: 7.5, 15, and 30. The results are illustrated as three pairs of ROC curves in Figure 4.12 below.

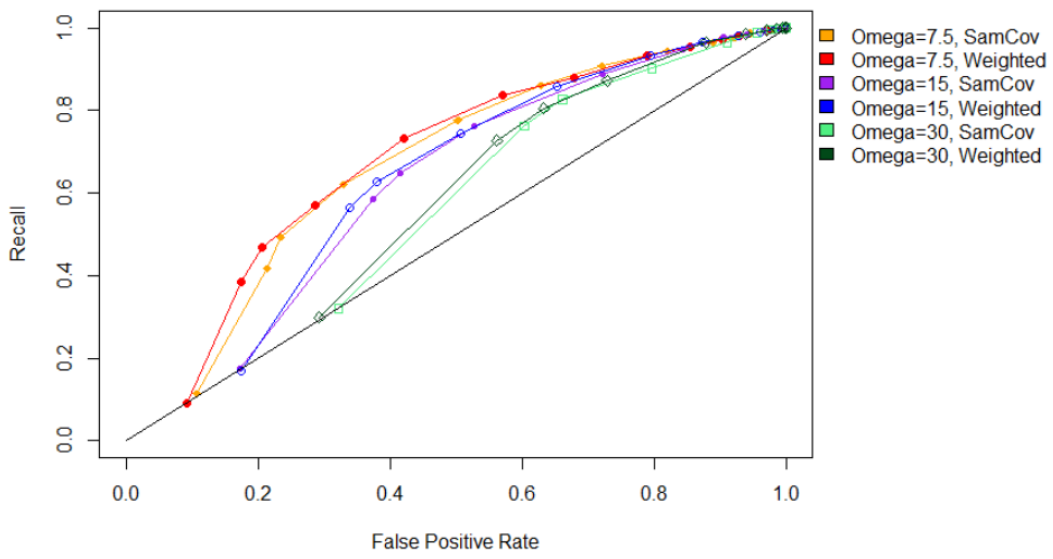


Figure 4.12: Comparison of standard vs. weighted sample covariance matrices as the estimate for between-study variance/covariance

As shown in Figure 4.12, for all the between-study variances Ω chosen, the performance of the weighted estimate is slightly better than the non-weighted one. This result demonstrates that weighted sample covariance matrix can be used to approximate the between-study variance/covariance.

4.3.3.3 Effect of regularization on the between-study covariance matrix

As discussed above, the graphical lasso (glasso) algorithm is being used to estimate a sparse between-study covariance matrix due to the p >> n situation in this research. In glasso, the regularization parameter rho controls the level of sparsity in the matrix. It is thus necessary to examine whether the different levels of sparsity have an impact on the classification performance, and to find the optimal sparsity given some known facts about the data.

In this experiment, five bicluster stacks are generated, each containing 3 biclusters with a dimension of 10. The same 13 data points are collected under different values of rho or no regularization at all. In the case of no regularization, the nearest positive definite form of the weighted sample covariance matrix is used. The result is shown in Figure 4.13 below.

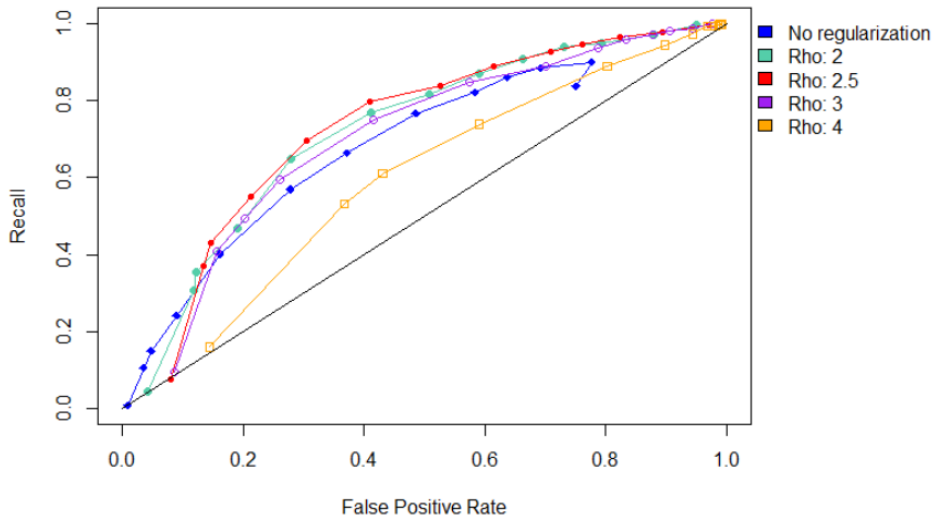


Figure 4.13: Effect of the graphical lasso regularization parameter on classification performance

From the result displayed in Figure 4.13, it is clear that sparsity of the estimated between-study covariance matrix has an impact on how well the MVMA classifier works. The classifier underperforms when there is no or too much (rho = 4) regularization imposed on the weighted sample covariance matrix. For the setting in this experiment, the optimal rho value appears to be 2.5.

4.3.3.4 Effect of dimension on computation

Experiment 1: Impact of dimensions on the classification performance

In this experiment, different dimensions are inspected for how they affect the classifier performs. Five synthetic bicluster stacks are generated, each with seven biclusters. The dimensions are 10, 20, 40, 80, and 100 respectively. The same 13 data points are collected as in the previous experiments. For each dimension, the estimated between-study covariance matrix is regularized by graphical lasso with $\rho = 0.1$.

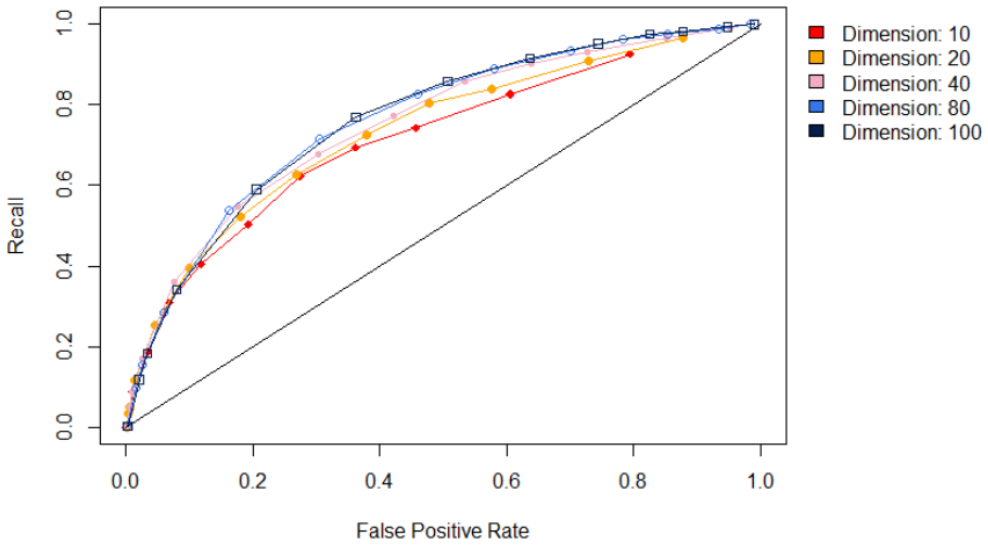


Figure 4.14: Effect of dimensions on the classifier

From the ROC curves shown in Figure 4.14 above, it appears that dimension does not represent a significant factor to the MVMA-based classifier. In fact, increasing the dimension seems to slightly enhance the classification performance. This observation is somewhat surprising, because as the dimension increases, the issue of dimension curse (i.e. $p >> n$) should become more severe, which makes the estimated between-study covariance more likely to be biased. However, the result here does not point to increasing bias of the estimates.

Experiment 2: Effect of dimension on run time

In this experiment, several dimensions, including 10, 20, 40, 80, and 100, are chosen to demonstrate their effects on run time using a Windows 7 desktop computer. For each dimension, 1000 runs are repeated to obtain the total time for that dimensions. The total times are then plotted against the dimensions as shown in Figure 4.15 below.

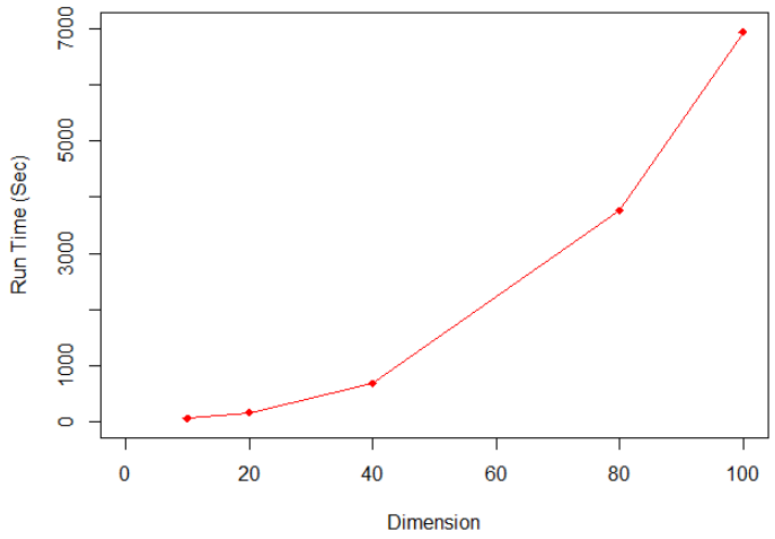


Figure 4.15: Effect of dimensions on run time

(Note: each run time number is the sum of 1000 repetitions)

From Figure 4.15, it is clear that the run time increases at a superlinear rate as the dimension increases. Nevertheless, the two-step MVMA classifier is much more efficient than the one-step counterpart described section 4.2.2. For example, the run times are 120953.85 seconds for one-step Bayesian inference, and 0.8 seconds for the two-step procedure proposed above when the dimension is 40. This means that the new two-step MVMA method is about 1.5 x 10^5 more efficient than the traditional Bayesian framework for MVMA.

4.3.4 Comparing the two-step method with the traditional one-step counterpart

So far numerous simulations have been performed to characterize the new two-step MVMA method, but it remains unclear how it matches the traditional one-step method in terms of classification performance. The goal of this section is to compare the two procedures under a low-dimensional setting thorough simulation. The reason that low-dimensional setting is chosen is because all one-step methods cannot be completed within acceptable timeframe when the dimension is moderate or high.

In this experiment, the full Bayesian inference described earlier is chosen as the one-step method. It uses inverse Wishart distribution as the prior for the between-study covariance matrix. For each data point, 150 runs are repeated to obtain the average recall and false positive rate. For the altered two-step method, the between-study covariance is approximated by the weighted sample covariance matrix subject to regularization by graphical lasso with $\rho=1$. The regularized matrix is then used to estimate the overall effect sizes based on Jackson's procedure. In both cases, the number of biclusters is set to be 7 and dimension is set to be 10.

In addition, Jackson's original two-step routine that uses method of moments to estimate between-study covariance is also included in the comparison. The ROC curves in Figure 4.16 below show the result of the comparison.

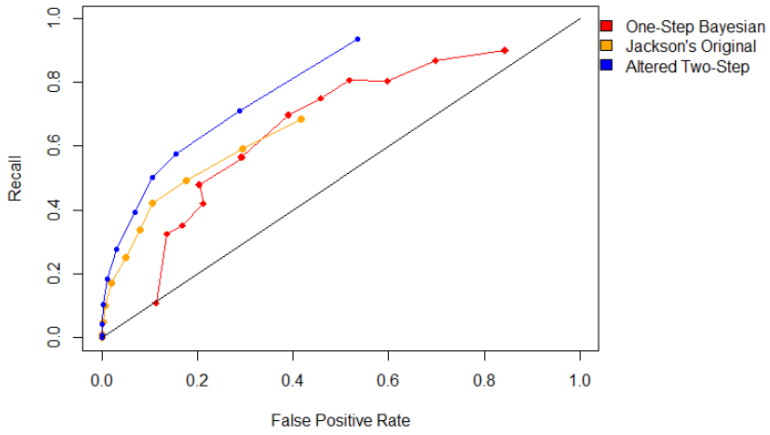


Figure 4.16: Comparison of three MVMA methods: traditional Bayesian, Jackson's method, and the new two-step method

From Figure 4.16, a couple of observations can be made. First, the altered two-step procedure clearly outperforms Jackson's original method. Second, when comparing the new two-step method with the traditional one-step Bayesian framework, the former appears to perform better when the stacks are of high signal and high noise (please see Table 4.1 above for description of the data points). In general, both two-step methods tend to be more conservative than the one-step Bayesian counterpart. They have lower recalls, but lower FPR's as well. This is because the multivariate t-distribution, assumed by the model in Jackson's method, has a narrower confidence interval with the number of biclusters used in the experiment.

4.4 MVMA on real data bicluster stacks

The above simulation studies provide evidence that a two-step strategy for MVMA, which is an altered version of Jackson's original method, is effective for meta-analyzing biclusters. The goal of this section is to apply what has been learned from the simulation studies to the analysis of real bicluster stacks.

It is now ready to pool together the member biclusters within a stack and to estimate the between-study covariance followed by the overall effect sizes by taking the two-step process discussed above. The resulting overall effect size estimates for the six stacks are illustrated as a forest plots in Figure 4.17 below.

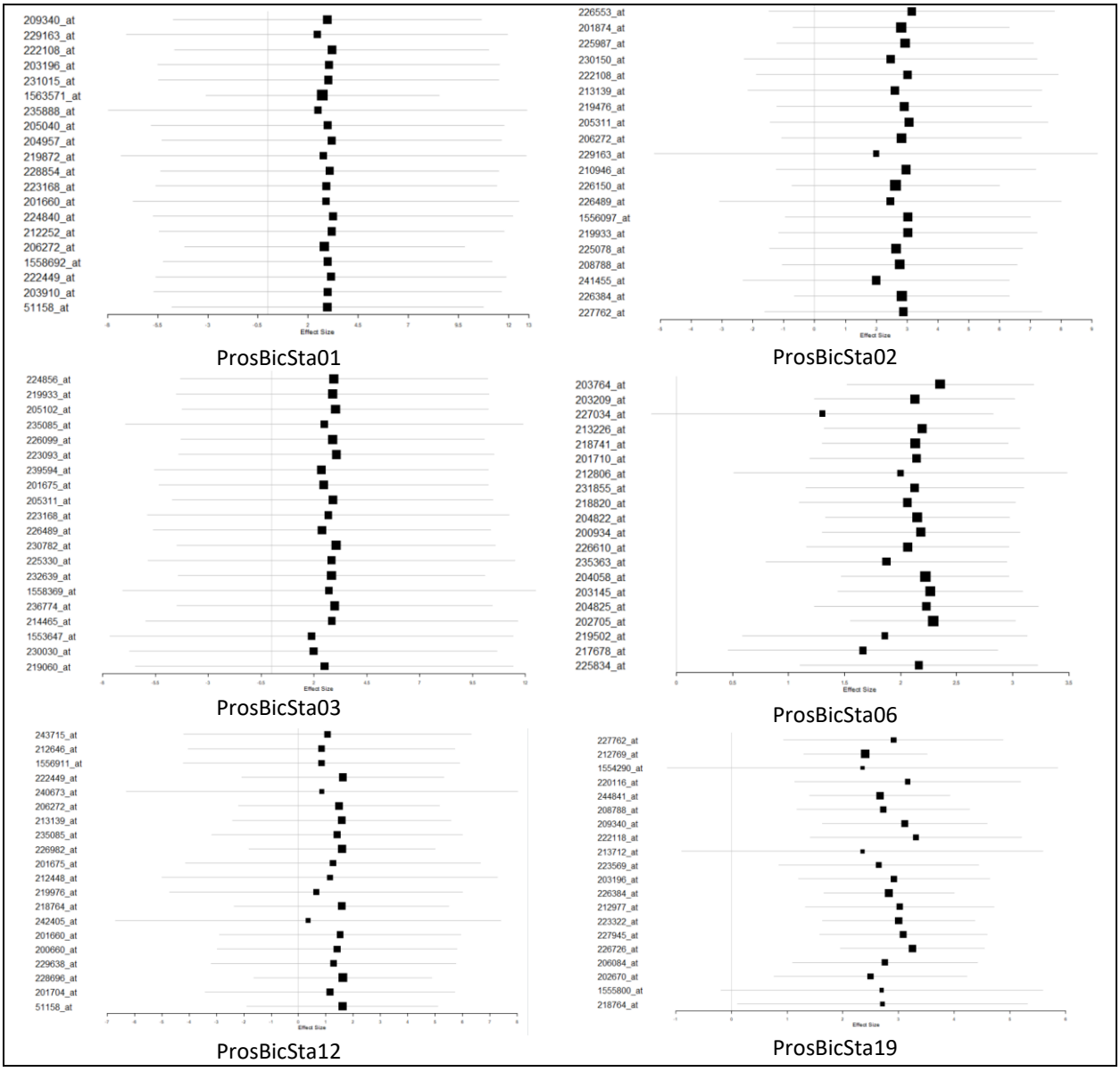


Figure 4.17: Forest plots that show the estimated overall effect sizes of 20 randomly selected probe sets from the six bicluster stacks

As shown in Figures 4.17 above, the estimated overall effect sizes vary considerably from one stack to another, as do the associated confidence intervals. For example, stack ProsBicSta01 have much wider confidence intervals than those of ProsBicSta06. More precisely, the widths are 17.55 and 2.11, respectively, which lead to the conclusion about the higher certainty on the estimates in stack ProsBicSta06 than that in ProsBicSta01.

The effect size estimates reflect the statistical "strengths" of the genes inside the biclusters relative to the corresponding "strengths" in the controls. The information does not however convey whether the gene sets are biologically relevant. It is necessary to assess the gene sets in terms of biological meaning, and it would be interesting to inspect whether the effect sizes of the genes are somehow connected to their biological interpretation, which is the focus of the next chapter.

4.5 Summary and conclusion

The focus of this chapter has been to apply multivariate meta-analysis (MVMA) to the biclusters in a stack. The goal is to address the second overall question of the dissertation: how to utilize multiple datasets jointly in order to increase statistical strength?

A treatment or intervention often has multiple outcomes. For example, systolic and diastolic blood pressures are the two outcomes to be measured when a hypertension drug is evaluated. Similarly, to evaluate a biological pathway, it is natural to measure the activities of all the member genes in the pathway. The more member genes shown to be up- or down-regulated, the more confidence we have about the activity change of the pathway. Therefore, the genes are

treated as endpoints when evaluating the "effect" that causes the pathway to change its activity. In this chapter, we try to evaluate a gene set, which may be part of a pathway, based on the evidence in the stack that harbors the gene set. Thus, the strategy here is to model each gene in the gene set as an endpoint.

If these outcomes are statistically correlated but their correlations are ignored during the analysis, then the process is called univariate meta-analysis (UVMA). Otherwise, it is referred to as MVMA. Previous studies have shown that ignoring the correlation structures can lead to overestimate of the variance of the summary effect sizes, and increase the chances of finding spuriously significant treatment effects [90][91][92]. In addition, UVMA can be considered as a special case of MVMA, where the inter-outcome correlations equal to zero.

Since genes in a pathway do not work alone, their functional relationship may result in statistical correlation. Thus, to evaluate the activity of a pathway, the genes should be considered jointly rather than separately, which is the reason that MVMA is adopted in the current study rather than UVMA.

MVMA has been well-established and well applied, especially in the clinical domain to combine multiple clinical trials in order to derive the overall effect of a treatment. Despite this fact, applying MVMA to gene expression data is not straightforward. The complications arises from multiple sources, including high-dimensionalities, small samples, and data heterogeneity.

The simulation studies presented in this chapter show that the traditional MVMA methods, either within the Bayesian or the frequentist framework, can perform satisfactorily in meta-analyzing biclusters (Section 4.2.4). However, they do not scale well when the dimension increases. Specifically, they quickly become computationally too demanding and thus not

practical. Thus, an alternative method had to be sought in order for biclusters to be combined and analyzed.

Jackson and Riley proposed a two-step MVMA method to address the issue of data scarcity [1]. Their approach involves estimating the between-study covariance using method of moments as step 1, and making inferences about the overall effect size as step 2. Furthermore, they proposed to use multivariate t-distribution rather than normal distribution to estimate the effect size in order to take into account the uncertainty of the between-study covariance estimate, which is a result of small sample sizes. Their method is statistically sound, but it also is computationally slow in the case of moderate or high dimensionalities, due to the use of method of moments in step 1.

To overcome the problem with method of moments, I proposed to use weighted sample covariance matrix to approximate the between-study covariance matrix. Each bicluster in a stack has a weight, which is the average within-study variance for that bicluster. Because of the high dimension and low sample size situation ($p >> n$), Regularization based on the graphical lasso algorithm is applied to the weighted sample covariance matrix to impose structure sparsity and to ensure positive definiteness. The second step continues to follow Jackson's method that is based on multivariate t-distribution to make reference about the overall effect sizes.

This change in step 1 of Jackson's original method was tested by a series of simulation studies. The performance was evaluated by treating the altered two-step MVMA procedure as a classifier that aims to classify the real genes from the background genes in artificial bicluster stacks. The results from the experiments show that the new two-step method is effective in meta-analyzing biclusters. Furthermore, it runs extremely fast even in high dimensions, especially when compared with the Bayesian version of the traditional MVMA method that relies on

Markov Chain Monte Carlo to make inference about the parameters. More importantly, the performance comparison shows favorable result for the altered two-step procedure. It outperforms Jackson's original procedure. When compared with the traditional Bayesian MVMA, it tends to be more conservative (both the recalls and the false positives rates are smaller). The smaller recall means reduced capability in detecting the genes when they are real. It can be amended by changing the significance level of the t-test used in the second step, which is a topic of future research. Overall, I consider the two-step method by Jackson and Riley a favorable solution. The new step 1 method proposed in this chapter further extends it utility to situations of $p >> n$.

The altered two-step procedure is applied to the analysis of six real bicluster stacks that are derived from prostate cancer expression data (Section 4.4). The results of the overall effect sizes are presented as forest plots, from which it is clear the effect size estimates are not equal across the stacks. Some stacks have estimates with high certainty than the other stacks. It would be interesting to investigate whether the effect sizes would allow us to predict the biological significance of the gene sets, which will be attempted in the next chapter.

To summarize, an improvement is suggested over Jackson's original two-step method to meet the needs of meta-analyzing higher-dimensional data. It performs comparatively well, especially in terms computation time. And it has been applied to analysis of six real bicluster stacks.

In order to begin to assess the value and utility of the most statistically significant bicluster stacks vs. a set of negative controls, we propose to use pathway analysis in Aim 3, as described in the next chapter.

Chapter 5 Pathway analysis of statistically significant gene sets

The overall goal of this dissertation is to develop a methodological framework for discovering gene sets from multiple gene expression datasets. Chapter 3 addressed the aim of identifying a biclustering algorithm suited for gene set identification. Building on this, Chapter 4 focused on utilization of multiple datasets to increase the statistical power. The key findings include: (1) the CCS algorithm is effective in recognizing correlated genes (Chapter 3); (2) the traditional MVMA methods can be used to combine multiple CCS biclusters, but their time efficiency deteriorates quickly as dimension increases (Chapter 4); (3) a two-step MVMA procedure previously proposed, with an improvement suggested in the Chapter 4, shows good promise in analyzing higher dimensional data. The remaining question addressed in Chapter 5 is whether there is evidence that the methods presented thus far are finding gene sets of interest.

This dissertation is in the realm of knowledge discovery from the data. So far the study has been focused on identifying and evaluating stacks of biclusters. Despite the fact that some stacks may be statistically significant as characterized by the effect size distributions, we have no knowledge whether the embedded gene sets are biologically relevant. Our ignorance has multiple facets. First, we do not know whether the individual biclusters contain genes that are parts of a real pathway. Second, when we pool together multiple biclusters based on how many genes that they share, we assume these biclusters are "related", but in reality they may not be. Third, even if the pooled biclusters are related, the resulting gene sets from the meta-analysis may mistakenly include genes that are false positives and miss others are true positives. Therefore, it is important to assess the gene sets in terms of a proxy for functional relevance, given that biological functional validation would be beyond the scope of an informatics dissertation.

Thanks to extensive efforts that have been made to curate the pathway knowledge into public databases, we now have the pathway information in searchable formats, which provides a foundation for validating the gene sets extracted from data.

The chapter is organized as follows: first, I will demonstrate three popular pathway analyses using six selected bicluster stacks. Then I will apply the analyses to additional stacks of different sizes that have been meta-analyzed, in order to investigate whether the MVMA results will allow us to predict the biological relevance of the gene sets.

5.1 Demonstration of three pathway analyses

Earlier in Section 2.8, I gave an introduction to three major types of pathway analysis, namely Over-Representation Analysis (ORA) [102], Gene Set Enrichment Analysis (GSEA) [32], and Network Topology-based Analysis (NTA) [103][104]. Ideally, the pathway analyses can be used as a gold standard for validating gene sets. However, the results from pathway analyses only provide an "enrichment score". A larger enrichment score means a higher level of confidence that the input gene set may match a known pathway. Thus, pathway analyses should not be considered as a gold standard.

In this section, I will give an in-depth discussion of the analyses, followed by demonstrations of their usage with the same six bicluster stacks listed in Table 3.2. The results reveal varying degrees of enrichment, as presented in Sections 5.1.1 through 5.1.3.

5.1.1 Over-Representation Analysis (ORA)

The Over-Representation Analysis (ORA) aims to statistically evaluate the fraction of genes in a pre-known pathway among the set of genes to be assessed. It is also referred to as "2x2 table method" [105]. The most commonly used statistics used by ORA are the hypergeometric distribution, binomial distribution, chi-squared distribution.

The ORA tool used in the current research is called WebGestalt [111][101]. It adopts the hypergeometric test to evaluate the significance of enrichment for a category C in gene set A. Suppose the gene set (A) to be evaluated contains n genes, and the reference gene set (B) contains m genes. Further, if A and B have k (out of n) and j (out of m) genes, respectively, in a given category (C) (e.g. a GO category [36][37], a KEGG pathway [33], a BioCarta [156] pathway etc.). Based on the reference gene set, the expected value of k is (n/m) x j. If k is bigger than the expected value, then category C is considered to be enriched, and the enrichment ratio r is k/k_e. If B is the population from which A is drawn, then the hypergeometric probability mass function (pmf) is given by:

$$p = \sum_{i=k}^{n} \frac{\binom{m-j}{n-i}\binom{j}{i}}{\binom{m}{n}}$$

WebGestalt allows the significance level and the minimum number of genes in a significant category to be specified by the user. For example, the user can decide that at least 3 genes from a category are statistically enriched.

Table 5.1 below summarizes the ORA results for the six bicluster stacks. For each stack, the top 5 enriched Gene Ontology pathways are listed. With regard to the individual ORA outcomes, the Shared Gene Count (k) denotes the number of genes in the input gene set that are shared with the pathway. Enrichment Ratio denotes k/k_e, as described above. The P-Value is the

hypergeometric test p-value, and FDR denotes the false discovery rate from the Benjamini–Hochberg procedure. Higher ORA significance is associated with higher shared gene count, higher enrichment ratio, lower p-value, and lower FDR.

Stack Label	Avg. CI Width of ES Estimates	Pathway Name	Shared Gene Count (k)	Enrichment Ratio	P-Value	FDR
ProsBicSta01	17.55	insulin-like growth factor receptor signaling pathway	3	26.57	1.93E-04	8.57E-01
		positive regulation of B cell receptor signaling pathway	2	91.08	2.01E-04	8.57E-01
		bicellular tight junction assembly	3	21.25	3.76E-04	1.00E+00
		apical junction assembly	3	17.71	6.44E-04	1.00E+00
		regulation of B cell receptor signaling pathway	2	45.54	8.57E-04	1.00E+00
		Average	**2.6**	**40.43**	**4.54E-04**	**0.9428**
ProsBicSta02	9.23	negative regulation of phosphorylation	9	6.44	9.10E-06	7.78E-02
		negative regulation of protein phosphorylation	8	6.27	3.62E-05	1.28E-01
		negative regulation of phosphate metabolic process	9	5.03	6.36E-05	1.28E-01
		negative regulation of phosphorus metabolic process	9	5.02	6.45E-05	1.28E-01
		response to hormone	11	3.98	7.47E-05	1.28E-01
		Average	**9.2**	**5.348**	**4.96E-05**	**0.11796**
ProsBicSta03	15.83	positive regulation of B cell receptor signaling pathway	2	82.95	2.42E-04	8.48E-01
		insulin-like growth factor receptor signaling pathway	3	24.19	2.55E-04	8.48E-01
		regulation of antigen receptor-mediated signaling pathway	3	20.74	4.04E-04	8.48E-01
		regulation of chondrocyte differentiation	3	19.35	4.96E-04	8.48E-01
		bicellular tight junction assembly	3	19.35	4.96E-04	8.48E-01
		Average	**2.8**	**33.316**	**3.79E-04**	**0.848**
ProsBicSta06	2.12	mitotic cell cycle	32	8.77	0.00E+00	0.00E+00
		nuclear division	26	12.09	0.00E+00	0.00E+00
		sister chromatid segregation	17	21.07	0.00E+00	0.00E+00
		cell cycle	40	6.33	0.00E+00	0.00E+00
		chromosome segregation	22	17.82	0.00E+00	0.00E+00
		Average	**27.4**	**13.216**	**0**	**0**
ProsBicSta12	9.17	regulation of cell proliferation	14	3.16	7.32E-05	3.79E-01
		intracellular receptor signaling pathway	6	7.82	1.10E-04	3.79E-01
		regulation of ossification	5	9.51	1.74E-04	3.79E-01
		positive regulation of B cell receptor signaling pathway	2	96.77	1.78E-04	3.79E-01
		cell proliferation	15	2.65	2.76E-04	3.84E-01
		Average	**8.4**	**23.982**	**1.62E-04**	**0.38**
ProsBicSta19	3.94	lipid biosynthetic process	15	4.73	4.84E-07	4.13E-03
		lipid metabolic process	21	3.21	1.19E-06	4.39E-03
		cellular lipid metabolic process	18	3.49	2.53E-06	4.39E-03
		single-organism biosynthetic process	21	3	3.49E-06	4.39E-03
		fatty acid elongation, saturated fatty acid	3	86.02	4.11E-06	4.39E-03
		Average	**15.6**	**20.09**	**2.36E-06**	**0.004338**

Table 5.1: Summary of ORA results for six bicluster stacks

From Table 5.1, it is clear that distinct groups of pathways are found to be enriched by the six stacks with varying degrees of significance according to the enrichment outcomes. In addition, it appears that the confidence intervals of the effect size estimates are associated with the ORA outcomes. Specifically, the narrower the intervals, the more significant the ORA results are.

WebGestalt allows the enriched GO pathways to be mapped to the GO tree and provides the hierarchical visualization. By inspecting a mapping, one can learn about the relationships among the GO pathways in terms of how closely they are related with each other. Figures 5.1 through 5.6 below depict the mappings of the six bicluster stacks. The high-resolution source images are available from https://students.washington.edu/thwu/pathway/.

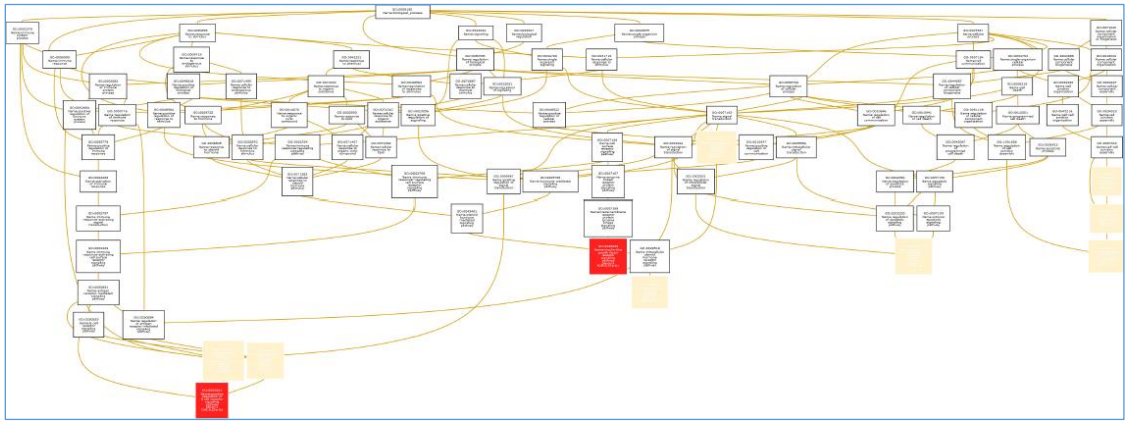


Figure 5.1: Mapping of enriched pathways to GO tree for stack ProsBicSta01

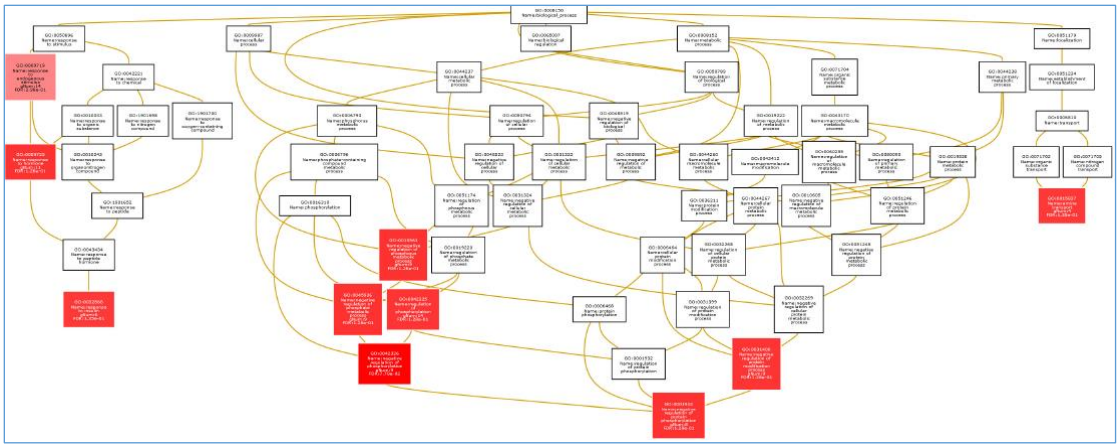


Figure 5.2: Mapping of enriched pathways to GO tree for stack ProsBicSta02

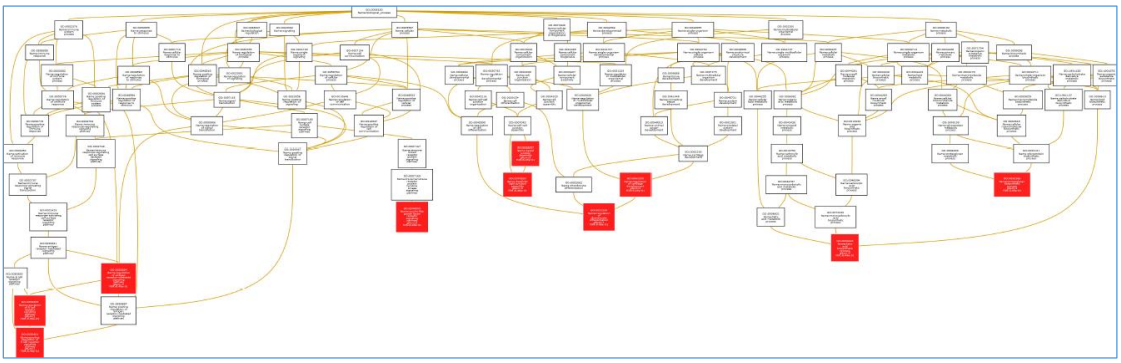


Figure 5.3: Mapping of enriched pathways to GO tree for stack ProsBicSta03

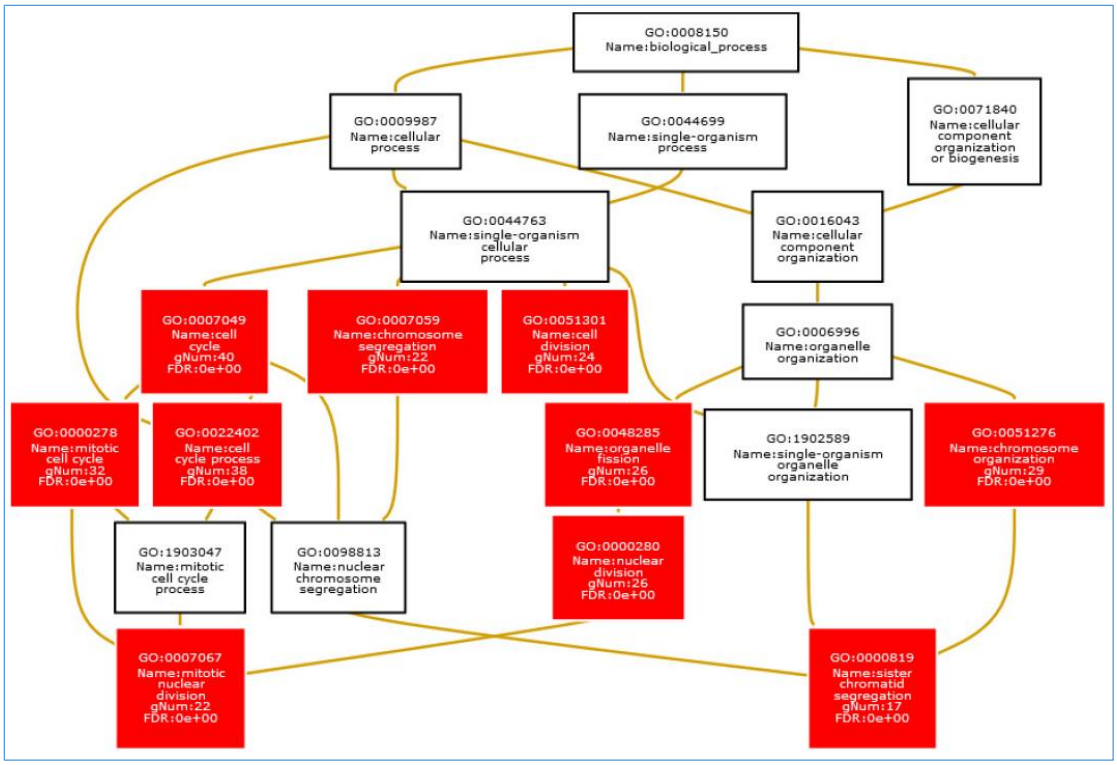


Figure 5.4: Mapping of enriched pathways to GO tree for stack ProsBicSta06

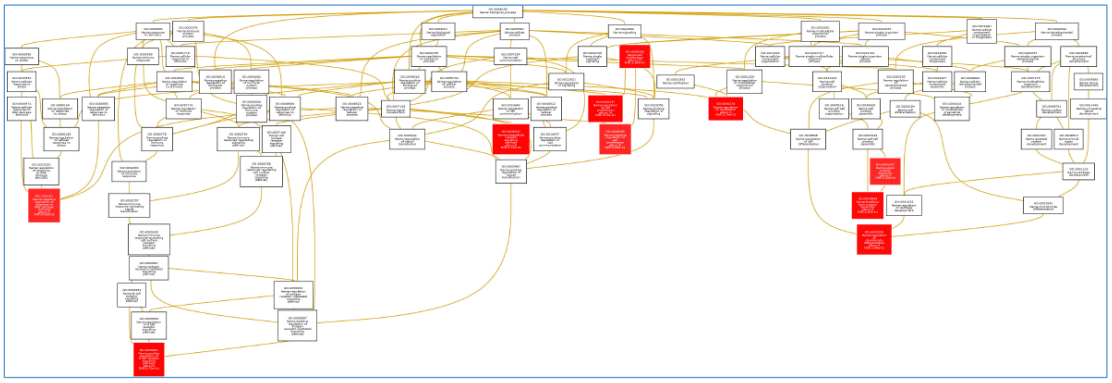


Figure 5.5: Mapping of enriched pathways to GO tree for stack ProsBicSta012

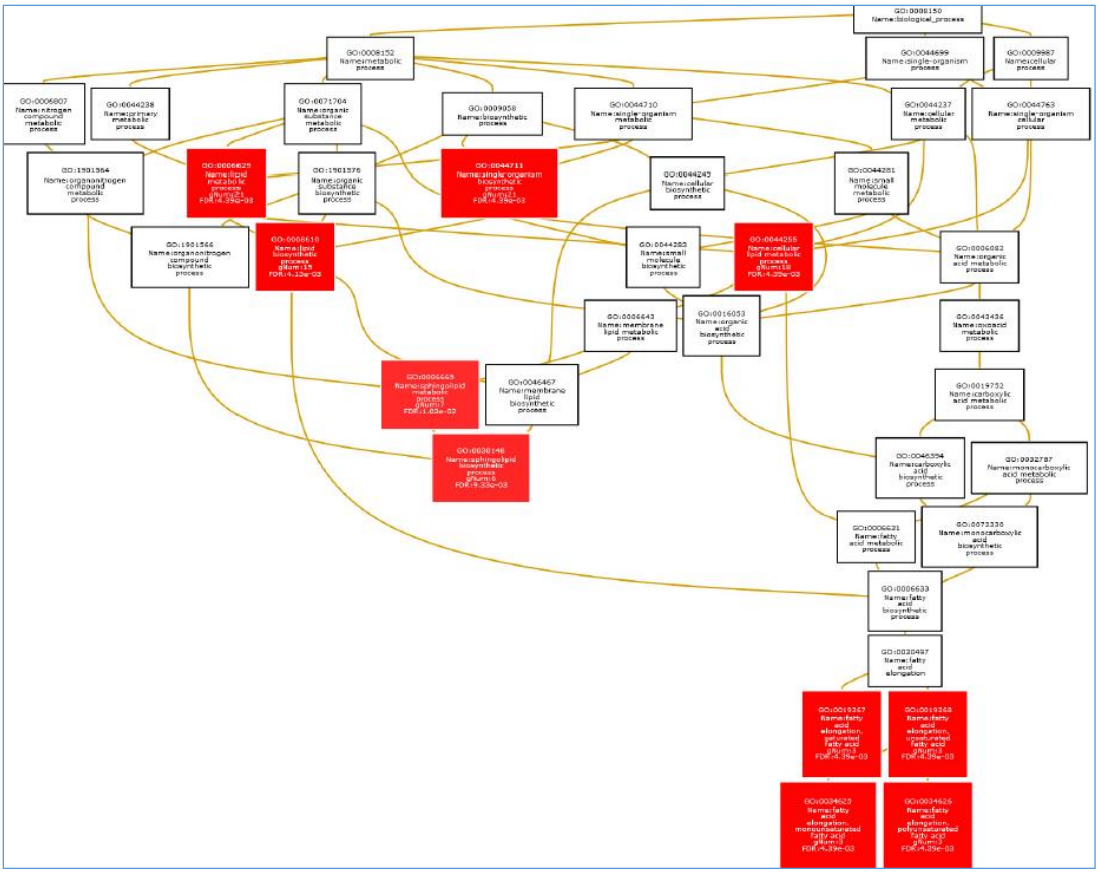


Figure 5.6: Mapping of enriched pathways to GO tree for stack ProsBicSta19

Two observations can be made based on Figures from 5.1 to 5.6 above. First, different stacks have different mappings on the GO tree. For example, ProsBicSta06 and ProsBicSta19 have more concentrated pathways mapped to the GO tree compared to ProsBicSta01, suggesting that the gene sets in ProsBicSta06 and ProsBicSta19 may have more closely related functions than that in ProsBicSta01.

Second, the GO tree mapping results appear to be correlated with the confidence intervals of the effect size estimates shown in Table 5.1. Specifically, more concentrated mappings are associated with narrower intervals, and vice versa.

To summarize the Over-Representation Analysis, the six bicluster stacks exhibit varying degrees of significance measured by enrichment-related measures and GO tree mappings. While

ProsBicSta06 has the most significant ORA results, ProsBicSta01 is on the other end of the significance spectrum. More importantly, the effect size estimates of the stacks seem to be correlated with the ORA results in terms of both the enrichment measures and GO tree mappings. Higher confidence estimates of the effect sizes correspond to more significant ORA enrichments and more closely related potential functions of the gene sets.

5.1.2 Gene Set Enrichment Analysis (GSEA)

Gene Set Enrichment Analysis (GSEA) is a powerful analytical method for interpreting gene expression data [32]. Suppose we have list of genes that can be sorted into a ranked list L, according to their differential expression between the classes. In addition, Given an *a priori* defined set of genes S, which can be genes encoding products in a pathway, or located in the same cytogenetic band, or sharing the same GO category, the goal of GSEA is to determine whether the members of S are randomly spread across L or primarily located at the top or bottom. In a typical GSEA scenario, L is sorted based on differential expression of the genes: the up-regulated genes are placed to the top, and the down-regulated genes to the bottom. In such cases, the gene expression data and the associated phenotype labels need to be supplied as inputs to the analysis.

According to the original paper that proposed GSEA [32], there are three steps involved in the analysis. In the first step, an enrichment score (ES) that measures the degree to which S is overrepresented at the extremes (top or bottom) of the entire ranked list L. The score is calculated by walking down the sorted list L. When a gene in S is encountered, a running-sum statistic is increased. Likewise, if a gene not present in S is met, the running-sum statistic is

decreased. The degree of increment or decrement depends how much the expression of the gene is correlated with the conditions or phenotypes in the expression data. In the second step, the significance level of the enrichment score (ES) is estimated. This is done by permuting the phenotype labels and recalculating the ES on the permuted data to generate the null hypothesis. In the final step, the estimated significance levels are adjusted to account for multiple hypothesis testing.

GSEA allows L to be sorted by measures not directly related to differential expression. In such cases, L is pre-ranked by any user-specified criterion. This version of GSEA is used in the current research to evaluate the gene sets in the bicluster stacks. The ranking is based on the overall effect sizes derived from MVMA as described in the previous chapter.

The tool used here is implemented by the Broad Institute (http://www.broad.mit.edu/gsea), which is perhaps the most popular GSEA platform within the bioinformatics community. The backend database of the system is called the Molecular Signature Database (MSigDB) that includes a large collection of pathway information [32][157].

To use the tool, the Reactome pathway knowledgebase (v6.1) [35] is used for consistence to validate all the bicluster stack gene sets. In addition, the minimum number of genes shared between the gene set and a pre-known pathway is set to be 15. All pathways that share less than 15 genes with the gene set are excluded from the analysis.

As in previous section, the same six bicluster stacks are analyzed by GSEA. The results are summarized in the Table 5.2 below.

Stack Label	Pathway found	Size	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX

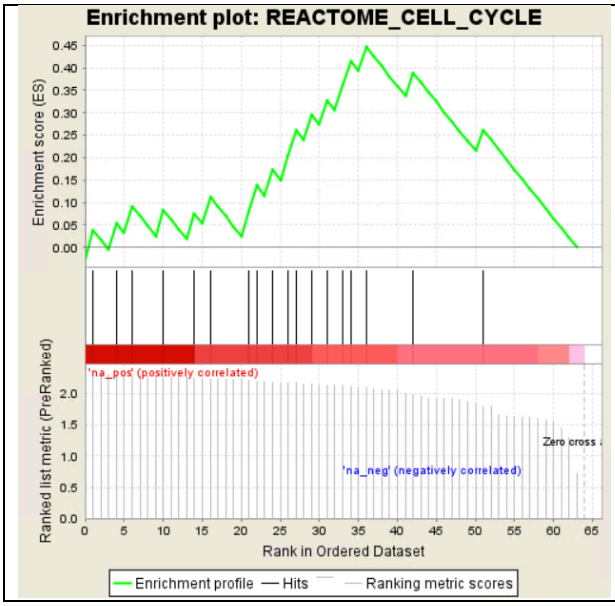
ProsBicSta01	No significant pathway found							
ProsBicSta02	No significant pathway found							
ProsBicSta03	No significant pathway found							
ProsBicSta06	REACTOME_CELL_CYCLE	18	0.45	1.83	0.014	0.030	0.018	36
	REACTOME_CELL_CYCLE_MITOTIC	16	0.41	1.63	0.033	0.034	0.045	36
ProsBicSta12	No significant pathway found							
ProsBicSta19	REACTOME_METABOLISM_OF_LIPIDS_AND_LIPOPROTEINS	10	-0.28	-1.06	0.374	0.374	0.138	37

Table 5.2: Summary of GSEA result for six bicluster stacks

(Size: number of genes shared between the input gene set and the corresponding pathway; ES: enrichment score; NES: normalized enrichment score; NOM p-val: nominal p-value; FDR q-val: false discovery rate q-value; FWER p-val: family-wise error rate p-value; RANK AT MAX: the position in the ranked list at which the maximum enrichment score occurred)

The results in Table 5.2 show that only ProsBicSta06 has significant enrichment with two pathways. For ProsBicSta19, a pathway is found but the statistical significance is relatively low, according to both the nominal and FWER p-values which are above the commonly used threshold 0.05. The other stacks are missed in the table because their sizes (number of genes shared with any pathway) are less than 10.

Figures 5.7 and 5.8 below provide details of the GSEA results for ProsBicSta06.

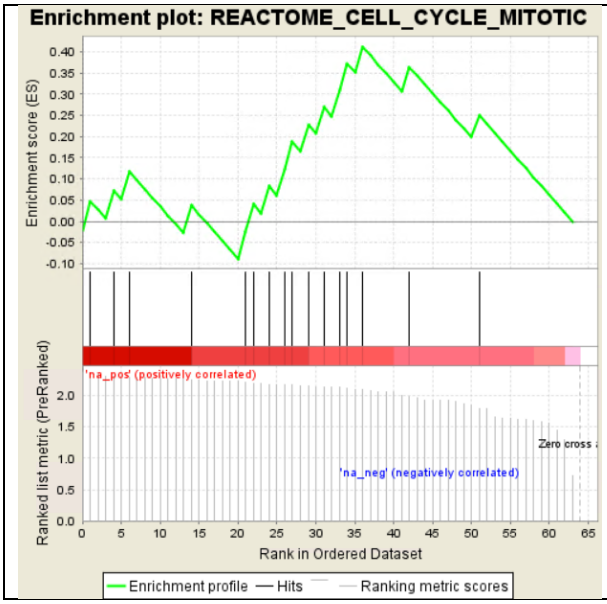


Standard name: REACTOME_CELL_CYCLE
Systematic name:M543
Brief description: Genes involved in Cell Cycle
Collection C2: curated gene sets
CP:REACTOME: Reactome gene sets

Organism: Homo sapiens
Gene members: 430 members mapped to 421 genes

External link: http://www.reactome.org/cgi-bin/eventbrowser_st_id?ST_ID=REACT_115566

Figure 5.7: Enrichment score distribution along the gene set of ProsBicSta06 and additional details of the first enriched pathway



Standard name: REACTOME_CELL_CYCLE_MITOTIC
Systematic name:M5336
Brief description: Genes involved in Cell Cycle, Mitotic
Collection C2: curated gene sets
CP:REACTOME: Reactome gene sets

Organism: Homo sapiens
Gene members: 333 members mapped to 325 genes

External link: http://www.reactome.org/cgi-bin/eventbrowser_st_id?ST_ID=REACT_152

Figure 5.8: Enrichment score distribution along the gene set of ProsBicSta06 and additional details of the second enriched pathway

To conclude, most of the bicluster stacks do not show significant enrichment with existing pathways according to the GSEA observations. The overall results are consistent with the ORA data. The stacks found to be significant in GSEA are also significant according to ORA.

5.1.3 Network Topology-based Analysis (NTA)

Some more recent pathway knowledge bases provide information beyond simple lists of genes for each pathway. The new information includes how the genes interact (e.g., activation, inhibition, etc.) and where they interact (e.g., cytoplasm, nucleus, etc.). Network Topology-based Analysis (NTA) aims to utilize this additional information to assess the significance of candidate gene lists. Some well-known network knowledge bases that enable NTA include KEGG [33],

MetaCyc [158], Reactome [34][35], RegulonDB [159], STKE (http://stke.sciencemag.org), BioCarta (http://www. biocarta.com), PantherDB [160].

NTA was implemented and added to WebGestalt in its 2013 release. Prior to this addition, a few tools were already existing that supported network-based enrichment analysis using protein–protein interaction networks, but they typically relied on single-level gene lists derived from network decomposition without considering the hierarchical structure of the network. Since it is been shown that biological functions at the molecular level are often executed in hierarchical manner, it is thus necessary to approach network analysis by taking the hierarchical structure into account. The NTA implementation in WebGestalt aims to achieve exactly that through computational network analysis.

The NTA implementation of WebGestalt relies on a random-walk algorithm to find the most relevant network modules [111]. The default reference network is the BioGrid protein-protein interaction database [38]. The enrichment process is outlined as below: first, it identifies the best partition of the network by maximizing the modularity score [113] using a random walk-based algorithm [161]. Second, it uses the edge switching algorithm to create 1000 random networks that match the protein-protein interaction network, then identify the best partition and the corresponding modularity score for each random network. Finally, if the modularity score for the candidate interaction network is significantly higher than those for the 1000 random networks ($P < 0.05$), then the candidate interaction network is considered to have a modular organization as the identified best partition. To construct the hierarchical structure, the above three steps are repeated iteratively for each sub-network until none of them shows a modular reorganization.

WebGestalt provides a graphical visualization for the resulting network. As in the previous two sections, the same six bicluster stacks are analyzed by NTA. Except for stacks ProsBicSta01 and ProsBicSta12, all other four stacks show significantly enriched networks. The results are illustrated in Figures 5.9 through 5.12. The Reference knowledge base is the BioGrid protein-protein interaction database [38]). The high-resolution source images are available from https://students.washington.edu/thwu/pathway/.

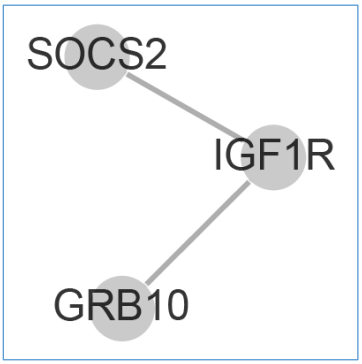


Figure 5.9: Result of Network Topology-based Analysis (NTA) for stack ProsBicSta02

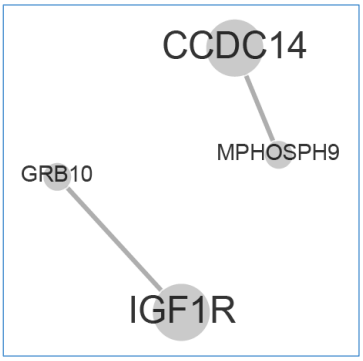


Figure 5.10: Result of Network Topology-based Analysis (NTA) for stack ProsBicSta03

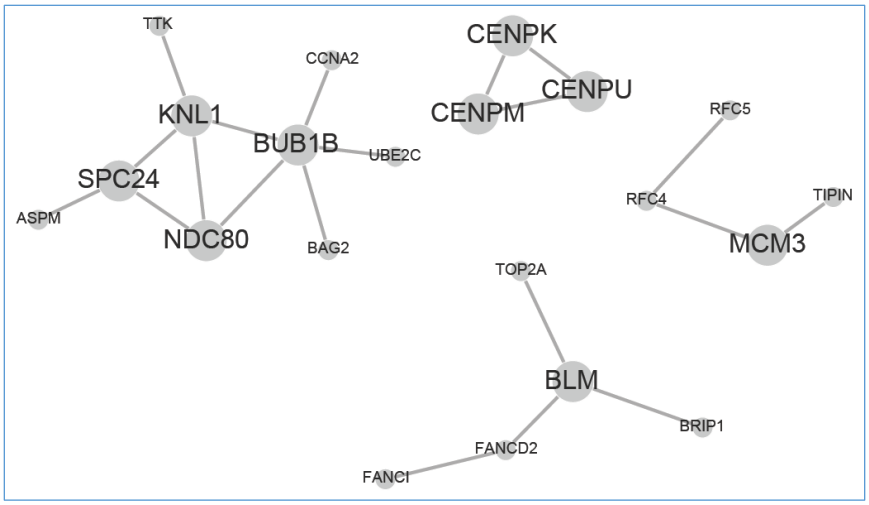


Figure 5.11: Result of Network Topology-based Analysis (NTA) for stack ProsBicSta06

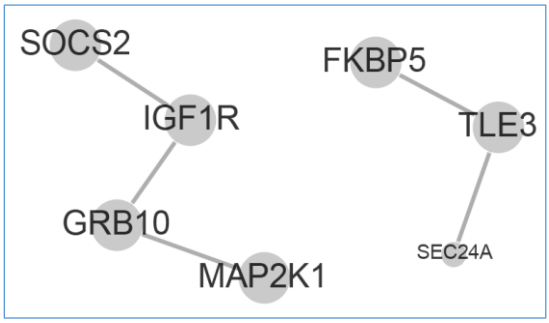


Figure 5.12: Result of Network Topology-based Analysis (NTA) for stack ProsBicSta19

From the figures above, the first observation is that different stacks show different levels of enrichment with some target protein-protein interaction network. Higher levels of enrichment are associated with higher numbers of genes and more connectivity among the genes.

The second observation is the levels of network enrichment are consistent with the ORA and GSER enrichment results. Those stacks that are found to be significant in ORA and GSEA tend to have higher levels of network enrichment. In other words, the effect size estimates are again correlated with the enrichment results of NTA.

5.1.4 Summary of pathway analyses demonstration

To summarize Section 5.1, three types of pathway analyses have been demonstrated using six selected bicluster stacks. The analyses are Over-Representation Analysis (ORA) [102], Gene Set Enrichment Analysis (GSEA) [32], and Network Topology-based Analysis (NTA) [103][104].

The six stacks show different degrees of pathway enrichment according to the analyses. One of them, ProsBicSta06 consistently has high level of enrichment across all three analyses. In the case of ORA, ProsBicSta06 matches a few cell-cycle related pathways with very high statistical significance (virtually zero p-value and FDR). Furthermore, the significant pathways are found to be closely related according to their mappings to the GO tree. In GSEA, similar cell-cycle pathways are discovered with high enrichment scores. In NTA, ProsBicSta06 exhibits high level of networking in terms of number of genes and connectivity among the genes in the network.

In contrast, ProsBicSta01 shows poor enrichment results by all three analyses. Thus, ProsBicSta01 can be seen as a negative control. The other four stacks fall between ProsBicSta01 and ProsBicSta06. They show enriched pathways according ORA, but the degrees of significance are lower than that of ProsBicSta06. In addition, they fail to enrich any pathway according GSEA and have smaller matched networks according NTA.

A more important observation is that there appears to be a correlation between the effect size estimates and the pathway analysis results for the six stacks. The narrower the confidence intervals of the effect size estimates, the higher level of enrichment significance according to all three pathway analyses. Since this correlation is observed in only six bicluster stacks, the next section aims to expanding the investigation by including more bicluster stacks.

5.2 Connecting the results from MVMA and pathway analyses

The previous sections demonstrate the three pathway analyses using six selected stacks. They show varying degrees of pathway enrichment: from ProsBicStc01 that shows near zero significant enrichment to ProsBicStc06 that has consistent enrichment with some cell cycle pathways across all three analyses. Furthermore, an initial observation appears that the confidence levels of the effect size estimates are correlated with the significance levels of the pathway analysis results. This section expands the discussion by including more biclusters with varying lengths as well as varying confidence levels of the effect size estimates.

As a reminder, the size of a bicluster stack is determined by two factors: the number of biclusters (length) and the number of genes shared by all the biclusters (width).

The goals of this section are twofold: first, it will examine whether stacking of biclusters can increase the chance of finding real biological pathways (in Section 5.2.1). This will be done by using bicluster stacks with varying lengths as well as the gene lists with randomly selected genes (to serve as a negative control). If longer stacks show better enrichment results than shorter ones and randomly selected genes, then we may begin to infer that the stacking process itself, without estimating the effect sizes of the genes, can lead to finding of gene sets of biological significance.

The second goal is to investigate whether the effect size estimates of the gene set would allow us to predict its biological relevance as determined by the pathway analyses (in Section 5.2.2). In other words, we wish to answer the question: what is the value of the MVMA procedure. If the effect sizes are a predictor of biological relevance, then the MVMA procedure

is valuable because it can serve as a screening mechanism for candidate gene sets by providing estimated effect sizes.

The second goal will be approached by examining the relationship between the confidence interval (CI) of the effect size estimates and some enrichment measures. If the CI width is statistically associated with the enrichment outcomes, based on fitting of a linear regression model, then we may conclude that the effect size estimates are a predictor of the biological relevance of the gene set.

5.2.1 Pathway analysis for bicluster stacks with varying lengths

This section aims to determine whether increasing the length of a bicluster will boost the chance of finding real gene sets. Stacks with lengths of 3, 5, 6, and 7 are included in the comparison. As a negative control, randomly selected probe sets are used. For consistence, all the stacks have a width of at least 10, and the random gene lists have a gene count of 100. For each stack or gene list, five replicas are taken. Due to the ease of performing Over-Representation Analysis (ORA) and the several available outcome variables associated with the test, ORA is chosen as the main enrichment method for the comparison. The ORA outcome variables include: the number of genes matched, the enrichment score, the ORA p-value, and the false discovery rate (FDR). Based on the results from the previous section, most stacks do not show GSEA enrichment. Even for stacks with length of 7, there is only one stack (i.e. ProsBicSta06) that can successfully enrich pathways. Similarly, the results for NTA have been sporadic: only longer stacks tend to show NTA-based enrichments. Thus, GSEA and NTA are

not suitable for the quantitative tests in this section, leaving ORA as the only option for the tests described below.

Tables 5.3 and 5.4 below list the details of the ORA results for the random gene lists and the stacks with length of 7. For each stack or gene list, the top five Gene Ontology pathways with the lowest p-values are presented. For stacks of length = 3, 5, and 6, please refer to Appendices V, VI, and VII for their ORA results.

Gene list label	Gene Ontology ID	Pathway Name	Number of genes matched	Enrichment score	ORA p-value	FDR
R_100_01	0050922	negative regulation of chemotaxis	2	0.12	6.31E-03	1.00E+00
	0044283	small molecule biosynthetic process	5	1.2	6.46E-03	1.00E+00
	1902041	regulation of extrinsic apoptotic signaling pathway via death domain receptors	2	0.13	7.31E-03	1.00E+00
	0001915	negative regulation of T cell mediated cytotoxicity	1	0.01	1.16E-02	1.00E+00
	0006549	isoleucine metabolic process	1	0.01	1.16E-02	1.00E+00
R_100_02	0006497	protein lipidation	3	0.26	2.19E-03	1.00E+00
	0042158	lipoprotein biosynthetic process	3	0.28	2.84E-03	1.00E+00
	2000008	regulation of protein localization to cell surface	2	0.09	3.82E-03	1.00E+00
	0006506	GPI anchor biosynthetic process	2	0.1	4.61E-03	1.00E+00
	0006505	GPI anchor metabolic process	2	0.1	4.88E-03	1.00E+00
R_100_03	0031053	primary miRNA processing	2	0.03	2.78E-04	1.00E+00
	0038166	angiotensin-activated signaling pathway	2	0.03	4.24E-04	1.00E+00
	2000765	regulation of cytoplasmic translation	2	0.04	5.99E-04	1.00E+00
	0017148	negative regulation of translation	4	0.42	7.79E-04	1.00E+00
	1904385	cellular response to angiotensin	2	0.05	9.16E-04	1.00E+00
R_100_04	0006378	mRNA polyadenylation	3	0.11	1.71E-04	7.87E-01
	0043631	RNA polyadenylation	3	0.11	1.84E-04	7.87E-01
	0031124	mRNA 3'-end processing	3	0.23	1.56E-03	1.00E+00
	0018146	keratan sulfate biosynthetic process	2	0.07	2.30E-03	1.00E+00
	0031123	RNA 3'-end processing	3	0.3	3.38E-03	1.00E+00
R_100_05	0051569	regulation of histone H3-K4 methylation	2	0.08	2.86E-03	1.00E+00
	0006101	citrate metabolic process	2	0.1	4.25E-03	1.00E+00
	0015872	dopamine transport	2	0.1	4.77E-03	1.00E+00
	0072350	tricarboxylic acid metabolic process	2	0.11	5.32E-03	1.00E+00
	2000736	regulation of stem cell differentiation	2	0.12	6.50E-03	1.00E+00

Table 5.3: ORA results for randomly selected genes

Stack Label	CI width for effect size estimate	Gene Ontology ID	Pathway Name	Number of genes matched	Enrich ment score		ORA p-value	FDR
ProsBicSta01	17.55	0048009	insulin-like growth factor receptor signaling pathway	3	0.11		1.93E-04	8.57E-01
	17.55	0050861	positive regulation of B cell receptor signaling pathway	2	0.02		2.01E-04	8.57E-01
	17.55	0070830	bicellular tight junction assembly	3	0.14		3.76E-04	1.00E+00
	17.55	0043297	apical junction assembly	3	0.17		6.44E-04	1.00E+00
	17.55	0050855	regulation of B cell receptor signaling pathway	2	0.04		8.57E-04	1.00E+00
ProsBicSta02	9.23	0042326	negative regulation of phosphorylation	9	1.4		9.10E-06	7.78E-02
	9.23	0001933	negative regulation of protein phosphorylation	8	1.28		3.62E-05	1.28E-01
	9.23	0045936	negative regulation of phosphate metabolic process	9	1.79		6.36E-05	1.28E-01
	9.23	0010563	negative regulation of phosphorus metabolic process	9	1.79		6.45E-05	1.28E-01
	9.23	0009725	response to hormone	11	2.76		7.47E-05	1.28E-01
ProsBicSta03	15.84	0050861	positive regulation of B cell receptor signaling pathway	2	0.02		2.42E-04	8.48E-01
	15.84	0048009	insulin-like growth factor receptor signaling pathway	3	0.12		2.55E-04	8.48E-01
	15.84	0050854	regulation of antigen receptor-mediated signaling pathway	3	0.14		4.04E-04	8.48E-01
	15.84	0032330	regulation of chondrocyte differentiation	3	0.16		4.96E-04	8.48E-01
	15.84	0070830	bicellular tight junction assembly	3	0.16		4.96E-04	8.48E-01
ProsBicSta04	9.15	0061035	regulation of cartilage development	4	0.19		4.12E-05	3.53E-01
	9.15	0061036	positive regulation of cartilage development	3	0.09		1.05E-04	4.48E-01
	9.15	0032330	regulation of chondrocyte differentiation	3	0.14		3.55E-04	1.00E+00
	9.15	0001503	ossification	6	1.12		8.36E-04	1.00E+00
	9.15	0098656	anion transmembrane transport	5	0.78		1.09E-03	1.00E+00
ProsBicSta06	2.12	0000278	mitotic cell cycle	32	3.65		0.00E+00	0.00E+00
	2.12	0000280	nuclear division	26	2.15		0.00E+00	0.00E+00
	2.12	0000819	sister chromatid segregation	17	0.81		0.00E+00	0.00E+00
	2.12	0007049	cell cycle	40	6.32		0.00E+00	0.00E+00
	2.12	0007059	chromosome segregation	22	1.23		0.00E+00	0.00E+00

Table 5.4: ORA results for five bicluster stacks with length of 7

Table 5.6 below summarizes the comparison of the ORA results.

	Length of the stack	Number of matched genes	Enrichment score	ORA p-value	FDR
Random gene lists		2.36	0.168	3.84E-03	9.83E-01
Bicluster stacks	3	8.2	2.296	4.04E-04	7.39E-01

	5	7.88	1.7224	5.24E-04	5.04E-01
	6	10.6	3.5912	3.06E-04	5.28E-01
	7	9.24	1.0632	2.74E-04	5.34E-01

Table 5.5: Summary of the ORA results for the stacks and the random gene lists

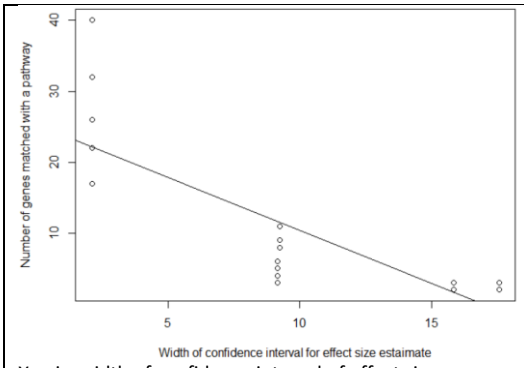
Table 5.5 above summarizes the ORA results for the random gene lists and the bicluster stacks with lengths 3, 5, 6, and 7. The ORA results are measured by four outcomes (described in Section 5.1.1): the number of genes matched with a target pathway, the enrichment score which is the ratio of actual matched gene count over the expected matched gene count based on a reference gene set, the ORA p-value based on a hypergeometric test, and FDR from the Benjamini-Hochberg procedure. Based on the numbers in the table, the four outcomes are correlated. Specifically, higher degree of enrichment tends to be associated with higher number of matched genes, higher enrichment score, lower ORA p-value, and lower FDR.

Furthermore, compared to the random gene lists, the bicluster stacks clearly have higher numbers of matched genes with the target pathways, higher enrichment scores, lower ORA p-values, and lower FDRs. Furthermore, the ORA outcomes tend to improve as the length of the stacks increases. These results show that increasing the length of a bicluster stack has a positive effect on the ORA enrichment results, suggesting that the process of stacking itself can lead to discovery of biologically relevant gene sets. The next question to ask is: what is the value of the MVMA procedure, given that bicluster stacking is already useful? Does MVMA produce results that are predictive of the biological relevance of the gene sets? More precisely, if we have high certainty about the effect sizes of the genes according to the meta-analysis, does it correspond to high biological significance for the gene sets? This question is addressed in the next section.

5.2.2 Relationship between effect size estimates and results of pathway analysis

To answer the question of whether the results from the MVMA and the pathway analysis are statistically correlated, a linear regression approach is taken as described below.

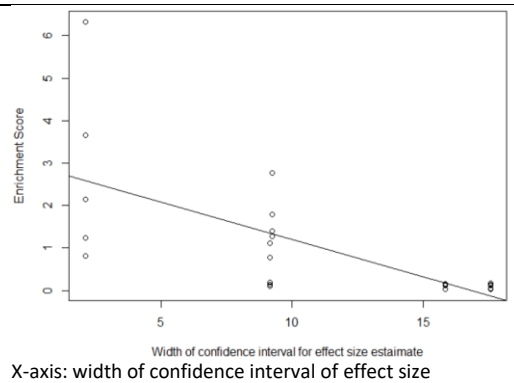
As a reminder, certainty of the effect size estimates is measured by the widths of their confidence intervals. To fit the linear regression model, the confidence interval width is used as a predictor variable while several aforementioned ORA outcomes are used as the response variables. Table 5.6 below lists the linear regression results for the five bicluster stacks with length of 7 (listed in Table 5.5).



X-axis: width of confidence interval of effect size
Y-axis: number of genes matched with a pathway

Coefficients:

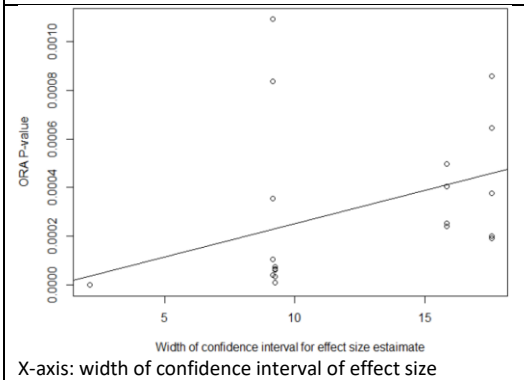
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	25.36	2.66	9.53	0.0000000019
CI_Width	-1.50	0.22	-6.80	**0.000000616**



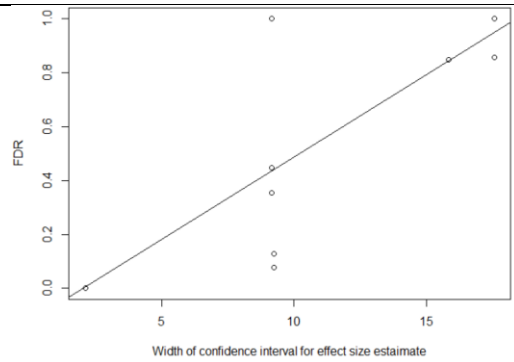
X-axis: width of confidence interval of effect size
Y-axis: enrichment score (or ratio)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.9706	0.4831	6.15	0.0000028
CI_Width	-0.1770	0.0399	-4.43	**0.00019**



X-axis: width of confidence interval of effect size



X-axis: width of confidence interval of effect size

Y-axis: ORA p-value (by hypergeometric test)				
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.00002	0.0001214	-0.16	0.871
CI_Width	0.0000272	0.0000100	2.71	**0.012**

Y-axis: FDR (from Benjamini–Hochberg procedure)				
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.1248	0.1126	-1.11	0.28
CI_Width	0.0611	0.0093	6.57	**0.0000011**

Table 5.6: linear relationships between confidence intervals of the effect size estimates and four ORA outcome measures for bicluster stacks of length = 7

From Table 5.6 above, the linear coefficients for the CI width significantly deviate from zero for all four ORA outcomes when the stack length is 7, according to the p-values (bold highlighted in Table 5.7) for the coefficient estimates. This results indicate the correlation between the predictor and the response variables. In other words, when the stack length is 7, the CI width is predictive of ORA outcomes, implying the effect size estimates can be used to predict the biological relevance of the gene sets as judged by ORA.

Next, the same regression analysis is repeated for stacks with shorter lengths to see whether similar observations can be made. The results are combined and presented in the Table 5.7 below.

ORA outcomes	Number of matched genes		Enrichment score		ORA p-value		FDR	
Length of stack	coefficient estimate	**P-value**	coefficient estimate	**P-value**	coefficient estimate	**P-value**	coefficient estimate	**P-value**
3	-0.305	**0.663**	-0.0589	**0.83**	0.0000672	**0.02**	0.098	**0.0071**
5	-0.101	**0.842**	-0.0212	**0.91**	0.0000353	**0.6**	0.0234	**0.57**
6	-0.684	**0.0964**	-0.294	**0.175**	0.00004187	**0.000013**	0.04779	**1.2E-08**
7	-1.5	**0.000000616**	-0.177	**0.00019**	0.0000272	**0.012**	0.0611	**1.1E-06**

Table 5.7: summary of linear relationship between CI width of effect size estimate and the ORA outcomes in bicluster stacks with varying lengths

From Table 5.7, an important observation is that the p-values for the coefficient estimates tend to decrease as the lengths of bicluster stacks increase, indicating growing significance for the coefficient estimates. In other words, as the stacks become longer, the effect size estimates become more predictive of the ORA outcomes. This observation is not surprising because increasing the stack length implies more data are added to the meta-analysis, which should lead to more predictive power for the pathway analysis results.

When the stack length is reduced to 3, the predictive power measured by the ORA outcomes significantly diminishes. However, this may be compensated by increasing the widths of the stacks (numbers of genes shared among all the member biclusters), according to a preliminary observation not presented here.

5.3 Biological interpretation for selected gene sets.

The goal of this section is to provide biological interpretation for some of the gene sets found to be significant according the pathway analyses described above.

ProsBicSta06 is found to enrich cell cycle pathways according to all three pathway analyses. The relationship between cell cycle control and cancers has been extensively studied. Progression of cancers is largely attributed to deviation of normal cell cycle [162]. Two types of genetic alterations have been shown to be involved in this process. They are gain-of-function and loss-of-function mutations. In gain-of-function mutations, it is believed that the products of the mutated genes participate in signal transduction pathways that promote cell proliferation [163]. In loss-of-function mutations, the normal checkpoints of cell cycle progression are disrupted, resulting in unchecked cell divisions [164]. Thus, the discovery of cell cycle related gene set in

this study is not surprising, and it partially validates the data mining method developed in the current study.

In addition, the stack ProsBicSta13 contains a gene set that enriches the androgen receptor pathway, which plays a key role in prostate normal and cancer development. Androgen deprivation therapy has become a standard treatment for advanced prostate cancer [165].

Finally, the gene set in ProsBicSta19 is found to significantly enrich a number of pathways related to lipid biosynthetic and metabolic process. Previous studies have shown that dysregulated lipid metabolism is often associated with prostate cancers [166]. More interestingly, ProsBicSta19 comes from these three datasets: GSE17044, GSE44905, and GSE7868. The original studies [167][168][169] that gave rise to these datasets did not bring up a discovery of any lipid related pathway. Thus, the stack appears to offer a new finding not mentioned in the original studies. Furthermore, since the stack has high confidence estimates of effect sizes derived from MVMA, as well as significant enrichment results according to ORA, it increases the chances that the discovery of the lipid-related pathways is not only novel but also real.

To summarize, the current study uses prostate cancer related expression data as an example to illustrate a method aimed to identifying gene sets. Several statistically significant gene sets have been found to be biologically significant as well according to pathway analyses. The enriched pathways have been shown by previous studies to be closely related to prostate cancer, which provides further evidence about the effectiveness of the developed method.

5.4 Summary of pathway analyses on bicluster stacks

In this chapter, three different pathway analyses are performed to assess the biological relevance of six gene sets embedded in their corresponding biclusters stacks. They are Over-Representation Analysis (ORA), Gene Set Enrichment Analysis (GSEA), and Network Topology-based Analysis (NTA).

ORA statistically evaluates the fraction of genes from a pre-known pathway found among the set of genes to be validated. GSEA is an improvement over ORA. It takes a gene list and the ranks of the genes in the list as inputs, and produces an enrichment score with a pre-known pathway. NTA utilizes network topology information to assess the significance of a candidate gene list. These three analyses have been widely used to assess the biological relevance of gene sets, and are categorized as three generations of pathway analyses by some authors.

Six bicluster stacks are used to demonstrate the three analyses. In all three cases, ProsBicSta06 exhibits very significant enrichment result, while ProsBicSta01 fails to match any pathway to a meaningful degree. The other four stacks perform better than ProsBicSta01, but not as well as ProsBicSta06 (Section 5.1). An initial observation of these six stacks was made with regard to the relationship between the effect size estimates and the pathway analysis results. The higher the confidence in the estimates of the effect sizes, the more significant the pathway analysis results are.

The correlation between MVMA-derived effect sizes and outcomes of the pathway analysis is further investigated by using more bicluster stacks with varying lengths. The results are twofold. First, longer stacks are more likely to lead to biologically relevant gene sets, compared to shorter stacks and random gene lists. Second, the effect size estimates appear to be a good predictor for the gene set's biological relevance. These results are based on ORA, and not on GSEA or NTA, due to infeasibility of the latter two analyses (Section 5.2).

In other words, stacking of overlapped biclusters alone can result in discovery of gene sets that are likely real, and the candidate gene sets can be further filtered by estimating the effect sizes of the genes through MVMA. The predictive power of the effect size estimates increase as the lengths of the bicluster stacks increase.

Finally, a number of enriched pathways are analyzed in terms of biological functions in relationship with prostate cancers. Literature search reveals that they are closely related to prostate cancer.

Chapter 6 Conclusion and discussion

In this final chapter, I will give an overall summary of the current study, point out some contributions and limitations, and discuss possible future directions that may expand this study.

6.1 Overall summary

The current study tries to answer these two overall questions: (1) what is the data mining method best suited for finding gene sets? (2) how to utilize multiple datasets in order to increase statistical strength? . The motivation is primarily informatics. Massive amount of research data, including gene expression data, have been accumulated over the years due to the technical advancements in comprehensive molecular-level measurements. Gene expression data is the focus of this study.

In Chapter 2, the reasons for identifying gene sets were discussed. Genes do not work alone. They often form functional units or pathways when executing biological tasks. Their functional relationships often translate into statistical correlations when gene expression experiments are carried out. Therefore, an attempt to find gene sets can lead to detection of change of pathway activities and thus may lead to disease prognosis. In addition, the data mining effort may lead to finding of previously unknown pathways.

The first aim of the study (discussed in Chapter 3) tries to address the first overall question: what is the data mining method best suited for finding gene sets. A proven strategy for uncovering functional gene sets is biclustering. Compared to traditional clustering methods such as hierarchical clustering, k-means clustering, and sell-organized feature map, biclustering allows simultaneous clustering on both dimensions of a data matrix. The variety of existing

biclustering algorithms make it possible to identify different gene sets with distinct statistical features. The CCS algorithm adopted in this study aims to identify genes that show correlated expressions across a subset of the samples. The results presented in this study verify the effectiveness of the CCS algorithm in identifying biologically relevant gene sets.

Given the CCS biclusters identified in Aim 1, Aim 2 tries to answer the second overall question by utilizing the biclusters. Since individual biclusters carry limited amount of statistical evidence, making use of multiple patterns represents a necessary strategy. Meta-analysis provides a well-established framework for combining evidences from multiple related sources. Despite the fact that meta-analysis has been widely used in many domains including the clinical and educational fields, adopting it to mining of public gene expression data is not straightforward. A number of challenges arise from the data, including high dimensionality, small sample sizes, and data heterogeneity. To tackle the heterogeneity issue, the random-effects model is adopted.

Similar to the process of selecting participating studies in a traditional meta-analysis, here biclusters are selected based on their levels of overlap (number of genes shared) to form bicluster stacks. Since the genes in a stack are modeled as individual endpoints, and the correlations among the genes are to be taken in account, multivariate random-effects meta-analysis (MVMA) is employed to statistically investigate the stack.

The next issue that appears is calculation of the effect sizes. I propose and validate a method used to estimate the effect sizes of biclusters in small datasets.

There are two main approaches for utilizing multiple datasets, namely integrative data analysis (IDA) and synthesis of summary statistic (SOSS). The result from an experiment shows

that SOSS performs better than IDA in terms of recovery of biclusters. This is because SOSS allows the use of parameters that are individually tuned for the datasets. It is important to point out that SOSS is not always advantageous over IDA. The conclusion here applies to context of biclustering.

To tackle the challenge of high dimensionality, a two-step MVMA method is adopted. It is based on the original formulation proposed by Jackson and Riley [1] with a key improvement. The original two-step method involves estimating the between-study covariance matrix using method of moment (step 1), followed by calculation of the overall effect sizes based on multivariate t distribution (step 2). The improvement proposed is to use weighted sample covariance matrix, subject to matrix regularization, to approximate the between-study covariance in step 1. Compared to the use of method of moments, the alternative step 1 method leads to a significant improvement in classifying real genes from background genes according to a simulation study.

The new two-step method is later applied to analysis of two real bicluster stacks with dimensions of 83 and 74, respectively. The results reveal a sharp contrast in the effect size estimates in terms of confidence intervals between the two stacks. ProsBicSta06 has much narrower estimated CIs than ProsBicSta01.

A narrower estimate on the effect sizes give us relatively high confidence about the underlying effect. However, it does not tell whether the gene set embedded in the stack is actually real, or in other words, biologically relevant. This leads to knowledge-based validation of the gene sets, which is the goal of the Aim 3.

Thanks to a growing number of available knowledge bases that store curated information on biological pathways, it is now possible to conduct various pathway enrichment analyses in an attempt to shed light on the possible biological functions of the gene set. The level of enrichment depends on the specific method and statistic used. Currently, three classes of pathway analyses are widely used, including Over-Representation Analysis (ORA) [102], Gene Set Enrichment Analysis (GSEA) [32], and Network Topology-based Analysis (NTA) [103][104]. All three of these analyses are carried out to assess the gene sets found in ProsBicSta01 and ProsBicSta06. The results are drastically different between the two stacks. ProsBicSta06 shows highly significant enrichment results related to cell-cycle pathways in all three analyses, while ProsBicSta01 exhibits none or little enrichment with any pathway. Thus, the statistical confidence derived from the MVMA is consistent with the pathway analysis results for the two stacks. This finding is further tested using additional stacks with varying lengths. The summarized results show that the MVMA-derived effect sizes for a gene set can be used to predict its biological relevance. The predictive power increases as the length of the bicluster stack increases. This is perhaps the most significant finding of the current study.

A repeated theme is the results is that confidence levels of the effect size estimates appear to be more meaningful than the magnitudes of the effect sizes. Since changes of expression in biological pathways are often small or moderate. The ability in recognizing consistent effect sizes, regardless of how smaller their magnitudes are, is significant because it would allow detection of pathway activity changes.

6.2 Contributions

The current study potentially brings forward a number of contributions to informatics and data mining, as summarized below:

1. High dimensionalities have been a prevalent issue with genome scale data. For example, a dataset derived from the Affymetrix platform contains expressions of over 27,000 probe sets. As a result, dimension reduction has been an active research topic in data mining. In the current study, biclustering on a dataset allows change of focus from the dataset to the biclusters, thus significantly reducing the dimension. Stacking of the biclusters from different datasets further reduces the dimension. This approach is perhaps the first bicluster-based attempt in dimension reduction by utilizing multiple datasets.

2. Biclustering is a powerful technique for finding submatrix patterns. However, it is less useful when applied to individual datasets separately due to high probabilities of getting type 1 or 2 errors as a result of small sample sizes. Meta-analyzing stacks of biclusters allows the potential of the technique to be further realized.

3. Mining multiple datasets is not new, but to the best of my knowledge, the current study is the first effort of casting stacks of biclusters into a MVMA problem. Applying MVMA to multiple biclusters augments the evidence of the gene set, and thus increases the chance of detecting change of pathway activity in the corresponding datasets. The method developed may have potential contribution to the area of knowledge discovery.

4. The two-step method proposed by Jackson and Riley allows MVMA to be applied to small number of datasets. However, it is computationally demanding and impractical

when applied to high-dimensional data. The improvement proposed in this study overcomes the limitation by greatly shortening the run time, at a cost of reduced recall and false positive rate.

5. Gene Set Enrichment Analysis (GSEA) has been widely recognized as a powerful technique for knowledge-based evaluation of gene sets. A required input for GSEA is a list of ranked genes. The ranking is mostly done based on the p-values of differential expressions. To the best of my knowledge, the current study is the first attempt of ranking the gene list based on effect sizes.

6. Most previous studies that aim to identify gene sets often produce candidate gene sets that contain hundreds or even more genes. The MVMA method presented here allows identification of much smaller candidate gene sets, which opens the possibility of uncovering more "concentrated" or more "targeted" gene sets in the data.

7. Utilizing heterogeneous data may maximize the chance of finding gene sets that are active not just in specific sample types (e.g. cell lines), but also in a variety of samples (e.g. cell lines, xenografts, human tissues). Thus, the methodology presented here may allow discovery of "robust" gene sets that are more biological relevant.

6.3 Current limitations of the proposed method

Currently, the proposed framework of applying MVMA to biclusters has a number of limitations:

1. Ideally, the quantitative analyses in Section 5.2 can be done with both ORA and GSEA, which would make results more convincing. Unfortunately, the gene sets identified from the MVMA show fewer enrichment than expected in the GSEA analysis, leaving ORA as the only available option for the quantitative studies.

2. A typical microarray dataset includes multiple conditions, and each condition contains multiple samples. Currently, when the CCS biclustering algorithm is applied to the data, the conditions are ignored and all the samples are treated equally. Disregarding the conditions may represent an information loss.

3. MVMA relies on the assumed hierarchical model involving two levels of multivariate normal distributions as discussed in 4.2.1. If the normality assumption is violated, the results from MVMA may be misleading.

4. Currently, selecting biclusters to form stacks is done solely based on the number of genes shared, without considering other merits such as how similar the biclusters are to each other. Ideally, similarity is measured between the biclusters, and outliers are identified and excluded from the meta-analysis. Including outliers in the meta-analysis may create bias and lower the chance of finding real gene sets.

5. MVMA may not be directly applicable for certain types of biclusters. For example, the Plaid Model algorithm [68] identifies layers of expressions, each representing a biclusters. It is not clear how to apply MVMA to such biclusters.

6.4 Possible directions for future research

To expand the current research, future studies can be done in a number of different ways:

1. The gene sets identified by the MVMA procedure show low level of GSEA enrichment. This may be due to the fact the gene sets contain fewer genes (all less than 200) than an average input gene set in most GSEA studies. Thus, a future research could investigate how to generate longer gene sets, or how to make GSEA work with shorter gene sets.

2. As mentioned earlier, the size of a bicluster stack is specified by the length (the number of member biclusters) and the width (the number of genes shared by ALL the member biclusters). So far the investigation has been focused on exploring the impact of length, not enough on width. The stack ProsBicSta19 is interesting in this regard. It has a length of 3 and width of 30 (see Table 3.2) and a very small CI width (3.94, second smallest after ProsBicSta06, see Table 5.1). More importantly, it shows significant enrichment in both ORA and NTA analyses, and a barely significant enrichment according to GSEA. These results suggest that a stack as short as 3 can still be significant, which opens the possibility of tackling the data scarcity issue in which only a small number datasets are available.

3. The relationship between effect size estimates for the genes and the pathway analysis result for the gene sets can be better established by including more bicluster stacks and using data from different cancer types.

4. Try different biclustering algorithms. Different algorithms recognize different patterns, which may lead to discovery of different gene sets.

5. Try different data types such as RNA-Seq data. This may lead to uncovering of gene sets that are not possible to find in microarray data.

6. The current study focuses on estimation of overall effect sizes for a bicluster stack. Besides the effect sizes, the conditional dependence among the genes can be potentially derived from the data. This information can lead to construction of gene network, which may be useful by shedding light on how the genes are regulated.

7. The GSEA analysis presented in this study uses effect sizes to rank the genes, which may represent an interesting addition to the traditional GSEA that typically analyzes input gene sets ranked by differential expressions. Future studies could be done to explore the impact of ranking the gene sets by effect sizes on the GSEA results.

8. As shown in the simulation studies in Sections 4.3.3 and 4.3.4, the two-step procedure usually leads to lower recalls in identifying the real genes compared to the one-step methods. This can be amended by adjusting the threshold of significance used in the t test in the second step. Further research using more simulated data can investigate how the adjustment can be done.

9. So far, weighted sample covariance matrix has been shown to be valid and efficient to approximate the between-study covariance. However, further statistical characterization may be needed, especially in the context of MVMA. For example, how sensitive it is to outliers, etc.

10. As mentioned in the Section 4.3.2, there is an unsolved discrepancy in how the classification recall responds to the increase of bicluster stack length between the Bayesian method and the two-step procedure. It may be due to the fact that the two methods estimate the parameters using very different approaches and assumptions. Further investigation using different MCMA settings and t distribution parameters should help explain the discrepancy.

References

[1] D. Jackson and R. D. Riley, “A refined method for multivariate meta-analysis and meta-regression,” *Stat. Med.*, vol. 33, no. 4, pp. 541–554, 2014.

[2] D. D. Dalma-Weiszhausz, J. Warrington, E. Y. Tanimoto, and C. G. Miyada, “The Affymetrix GeneChip Platform: An Overview,” in *Methods in Enzymology*, vol. 410, no. 06, 2006, pp. 3–28.

[3] K. Wang, M. Li, and H. Hakonarson, “Analysing biological pathways in genome-wide association studies,” *Nat. Rev. Genet.*, vol. 11, no. 12, pp. 843–854, 2010.

[4] E. E. Schadt, “Molecular networks as sensors and drivers of common human diseases,” *Nature*, vol. 461, no. 7261, pp. 218–223, 2009.

[5] S. Song and M. A. Black, “Microarray-based gene set analysis: A comparison of current methods,” *BMC Bioinformatics*, vol. 9, pp. 1–14, 2008.

[6] K. Wang, M. Li, and M. Bucan, “Pathway-Based Approaches for Analysis of Genomewide Association Studies Kai,” *Am. J. Hum. Genet.*, vol. 81, pp. 1278–1283, 2007.

[7] J. Hedegaard *et al.*, “Methods for interpreting lists of affected genes obtained in a DNA microarray experiment,” *BMC Proc.*, vol. 3, no. Suppl 4, p. S5, 2009.

[8] S. Draghici, *Statistics and Data Analysis for Microarrays Using R and Bioconductor, Second Edition*. Chapman and Hall/CRC, 2016.

[9] K. Eren, M. Deveci, O. Küçüktunç, and Ü. V Çatalyürek, “A comparative analysis of biclustering algorithms for gene expression data.,” *Brief. Bioinform.*, vol. 14, no. 3, pp. 279–92, May 2013.

[10] A. Prelić *et al.*, “A systematic comparison and evaluation of biclustering methods for gene expression data.,” *Bioinformatics*, vol. 22, no. 9, pp. 1122–9, May 2006.

[11] D. Mavridis and G. Salanti, “A practical introduction to multivariate meta-analysis,” *Stat. Methods Med. Res.*, vol. 22, no. 2, pp. 133–158, 2013.

[12] D. B. Kell, “Feature Article The Virtual Human : Towards a Global Systems Biology of Multiscale , Distributed Biochemical Network Models,” vol. 59, no. November, pp. 689–695, 2007.

[13] H. V Westerhoff and B. O. Palsson, “The evolution of molecular biology into systems biology,” vol. 22, no. 10, pp. 1249–1252, 2004.

[14] S. Behjati and P. S. Tarpey, “What is next generation sequencing?,” *Arch. Dis. Child. Educ. Pract. Ed.*, vol. 98, no. 6, pp. 236–238, 2013.

[15] D. C. Koboldt, K. M. Steinberg, D. E. Larson, R. K. Wilson, and E. Mardis, “The Next-Generation Sequencing Revolution and Its Impact on Genomics,” vol. 155, no. 1, pp. 27–38, 2014.

[16] R. Bumgarner, “DNA microarrays: Types, Applications and their future,” *Curr Protoc Mol Biol*, vol. 6137, no. 206, pp. 1–17, 2014.

[17] V. Trevino, F. Falciani, and H. A. Barrera-saldaña, “DNA Microarrays : a Powerful Genomic Tool for Biomedical and Clinical Research,” vol. 13, no. October, pp. 527–541, 2007.

[18] V. Vidova and Z. Spacil, “Analytica Chimica Acta A review on mass spectrometry-based

quantitative proteomics : Targeted and data independent acquisition," *Anal. Chim. Acta*, vol. 964, pp. 7–23, 2017.

[19] M. Mann, "Quantitative , High-Resolution Proteomics for Data-Driven Systems Biology," 2011.

[20] D. K. Trivedi, K. A. Hollywood, and R. Goodacre, "Metabolomics for the masses : The future of metabolomics in a personalized world," *New Horizons Transl. Med.*, vol. 3, no. 6, pp. 294–305, 2017.

[21] D. B. Kell and S. G. Oliver, "Here is the evidence , now what is the hypothesis ? The complementary roles of inductive and hypothesis-driven science in the post-genomic era," no. i, pp. 99–105, 2003.

[22] M. Civelek, A. J. Lusis, M. Genetics, and L. Angeles, "Integrative omics for health and disease," vol. 15, no. 1, pp. 34–48, 2014.

[23] Y. Hasin, M. Seldin, and A. Lusis, "Multi-omics approaches to disease," *Genome Biol.*, vol. 18, no. 1, pp. 1–15, 2017.

[24] D. L. Gerhold, R. V. Jensen, and S. R. Gullans, "Better therapeutics through microarrays," *Nat. Genet.*, vol. 32, no. 4S, pp. 547–552, 2002.

[25] D. B. Kell, "Systems biology, metabolic modelling and metabolomics in drug discovery and development," *Drug Discov. Today*, vol. 11, no. 23–24, pp. 1085–1092, 2006.

[26] K. Bystrom, "Moving towards individualized medicine with pharmacogenomics," *Nature*, vol. 429, no. May, 2004.

[27] G. Russo, C. Zegar, and A. Giordano, "Advantages and limitations of microarray technology in human cancer," *Oncogene*, vol. 22, no. 42, pp. 6497–6507, 2003.

[28] C. Virtanen and W. J, "Clinical Uses of Microarrays in Cancer Research," *Methods Mol. Med.*, vol. 141, no. 5, pp. 87–113, 2007.

[29] M. Rooman, J. Albert, Y. Dehouck, and A. Haye, "Detection of perturbation phases and developmental stages in organisms from DNA microarray time series data," *PLoS One*, vol. 6, no. 12, 2011.

[30] J. D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis, "Significance analysis of time course microarray experiments," *Proc Natl Acad Sci USA*, vol. 102, no. 36, pp. 12837–12842, 2005.

[31] V. K. Mootha *et al.*, "PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes," *Nat. Genet.*, vol. 34, no. 3, pp. 267–273, 2003.

[32] A. Subramanian, P. Tamayo, V. Mootha, S. Mukherjee, and B. Ebert, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proc Natl Acad Sci USA*, vol. 102, pp. 15545–15550, 2005.

[33] M. Kanehisa, "The KEGG resource for deciphering the genome," *Nucleic Acids Res.*, vol. 32, no. 90001, p. 277D–280, 2004.

[34] D. Croft, A. Mundo, R. Haw, and M. Milacic, “The Reactome pathway knowledgebase,” *Nucleic acids*, vol. 42, no. D1, pp. D472–D477, 2014.

[35] A. Fabregat *et al.*, “The Reactome Pathway Knowledgebase,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D649–D655, 2018.

[36] M. Ashburner *et al.*, “Gene ontology: Tool for the unification of biology,” *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, 2000.

[37] S. Carbon *et al.*, “Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D331–D338, 2017.

[38] C. Stark, B.-J. Breitkreutz1, T. Reguly1, L. Boucher, A. Breitkreutz1, and M. Tyers, “BioGRID: a general repository for interaction datasets,” *Nucleic Acids Res.*, vol. 34, no. 90001, pp. D535–D539, 2006.

[39] M. Mramor, G. Leban, J. Demšar, and B. Zupan, “Conquering the curse of dimensionality in gene expression cancer diagnosis: tough problem, simple models,” *Artif. Intell. Med.*, vol. 3581, pp. 514–523, 2005.

[40] Y. Wang, D. J. Miller, and R. Clarke, “Approaches to working in high-dimensional data spaces: Gene expression microarrays,” *Br. J. Cancer*, vol. 98, no. 6, pp. 1023–1028, 2008.

[41] N. Kolesnikov *et al.*, “ArrayExpress update-simplifying data submissions,” *Nucleic Acids Res.*, vol. 43, no. D1, pp. D1113–D1116, 2015.

[42] J. Demeter *et al.*, “The Stanford Microarray Database,” *Nucleic Acids Res.*, vol. 29, no. 1, pp. 152–155, 2001.

[43] C. Stretch *et al.*, “Effects of Sample Size on Differential Gene Expression , Rank Order and Prediction Accuracy of a Gene Signature,” vol. 8, no. 6, pp. 6–11, 2013.

[44] W.-J. J. Lin, H.-M. M. Hsueh, and J. J. Chen, “Power and sample size estimation in microarray studies.,” *BMC Bioinformatics*, vol. 11, no. 1, p. 48+, 2010.

[45] A. K. Dubes and R. C. Jain, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall, 1988.

[46] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston, MA: Kluwer Academic, 1992.

[47] D. Jiang, C. Tang, and A. Zhang, “Cluster analysis for gene expression data: A survey,” *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 11, pp. 1370–1386, 2004.

[48] H. Pirlim, B. Ekşioğlu, A. Perkins, and Ç. Yüceer, “Clustering of High Throughput Gene Expression Data,” *Comput. Oper. Res.*, vol. 39, no. 12, pp. 3046–3061, 2012.

[49] D. Pelleg and A. Moore, “X-means: Extending K-means with Efficient Estimation of the Number of Clusters,” *CEUR Workshop Proc.*, vol. 1542, pp. 33–36, 2000.

[50] C. Goutte, L. K. Hansen, M. G. Liptrot, and E. Rostrup, “Feature-space clustering for fMRI meta-analysis,” *Hum. Brain Mapp.*, vol. 13, no. 3, pp. 165–183, 2001.

[51] L. Rokach and O. Maimon, “Data mining and knowledge discovery handbook.,” Springer US, 2005, pp. 321–352.

[52] D. Bryant and V. Berry, “A structured family of clustering and tree construction methods,” *Adv. Appl. Math.*, 2001.

[53] D. Cheng, R. Kannan, S. Vempala, and Wang .G, “A divide-and-merge methodology for clustering,” *ACM Trans. Database Syst.*, 2006.

[54] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: a review,” *ACM Comput. Surv.*, 1999.

[55] R. Irizarry and M. Love, “Clustering, Basic Machine Learning.” [Online]. Available: http://genomicsclass.github.io/book/pages/clustering_and_heatmaps.html.

[56] T. Kohonen, “Learning vector quantization,” *Neural Networds*, vol. 1, no. suppl. 1, p. 303, 1988.

[57] T. Kohonen, *Self-Organizing Maps*. Berlin: Springer, 1995.

[58] P. Tamayo *et al.*, “Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation,” *Proc. Natl. Acad. Sci.*, vol. 96, no. 6, pp. 2907–2912, 1999.

[59] J. Wang, J. Delabie, H. C. Aasheim, E. Smeland, and O. Myklebost, “Clustering of the SOM easily reveals distinct gene expression patterns: Results of a reanalysis of lymphoma study,” *BMC Bioinformatics*, vol. 3, pp. 1–9, 2002.

[60] H. Wang, W. Wang, J. Yang, and P. S. Yu, “Clustering by Pattern Similarity in Large Data Sets,” *Proc. 2002 ACM SIGMOD Int. Conf. Manag. Data, ACM*, pp. 394–405, 2002.

[61] A. P. Gasch and M. B. Eisen, “Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering,” *Genome Biol.*, vol. 3, no. 11, p. research0059.1–0059.22, 2002.

[62] J. Hartigan, “Direct Clustering of a data matrix,” *J. Am. Stat. Assoc.*, vol. 67, no. 337, pp. 123–129, 1972.

[63] Y. Cheng and G. Church, “Biclustering of Expression Data,” *Proc Int Conf Intell Syst Mol Biol*, vol. 8, pp. 93–103, 2000.

[64] G. Bisson and F. Hussain, “χ -Sim : A New Similarity Measure for the Co-clustering Task,” *Seventh Int. Conf. Mach. Learn. Appl.*, pp. 211–217, 2008.

[65] J. S. Aguilar-Ruiz, “Shifting and scaling patterns from gene expression data,” vol. 21, no. 20, pp. 3840–3845, 2005.

[66] D. Bozdağ, A. S. Kumar, and U. V. Catalyurek, “Comparative analysis of biclustering algorithms,” *Proc. First ACM Int. Conf. Bioinforma. Comput. Biol. - BCB ’10*, p. 265, 2010.

[67] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini, “Discovering local structure in gene expression data: the order-preserving submatrix problem.,” *J. Comput. Biol.*, vol. 10, no. 3–4, pp. 373–84, Jan. 2003.

[68] L. Lazzeroni and A. Owen, “Plaid Models for Gene Expression Data,” *Stat. Sin.*, vol. 12, no. 1, pp. 61–86, 2002.

[69] A. Oghabian, S. Kilpinen, S. Hautaniemi, and E. Czeizler, “Biclustering methods: biological relevance and application in gene expression analysis.,” *PLoS One*, vol. 9, no. 3, p. e90801, Jan. 2014.

[70] D. Bozda, J. D. Parvin, and U. V Catalyurek, “A Biclustering Method to Discover Co-regulated Genes Using Diverse Gene Expression Datasets,” pp. 151–163, 2009.

[71] Q. Sheng, Y. Moreau, and B. De Moor, “Biclustering microarray data by Gibbs sampling,” *Bioinformatics*, vol. 19, no. Suppl 2, pp. ii196-ii205, Oct. 2003.

[72] J. Gu and J. S. Liu, “Bayesian biclustering of gene expression data.,” *BMC Genomics*, vol. 9 Suppl 1, p. S4, Jan. 2008.

[73] B. Pontes, R. Giráldez, and J. S. Aguilar-Ruiz, “Biclustering on expression data: A review,” *J. Biomed. Inform.*, vol. 57, pp. 163–180, 2015.

[74] B. Pontes, R. Girldez, and J. S. Aguilar-Ruiz, “Quality Measures for Gene Expression Biclusters,” *PLoS One*, vol. 10, no. 3, p. e0115497, 2015.

[75] A. Tanay, R. Sharan, and R. Shamir, “Discovering statistically significant biclusters in gene expression data,” vol. 18, 2002.

[76] Berlin J.A., J. Santanna, C. H. Schmid, L. A. Szczech, and H. I. Feldman, “Individual patient-versus group-level data meta-regressions for the investigation of treatment effect modifiers: Ecological bias rears its ugly head,” *Stat. Med.*, vol. 21, pp. 371–387, 2002.

[77] P. J. Curran and A. M. Hussong, “Integrative Data Analysis: The Simultaneous Analysis of Multiple Data Sets,” *Psychol Methods.*, vol. 14, no. 2, pp. 81–100, 2009.

[78] G. V. Glass, “Primary, secondary, and meta-analysis,” vol. 5, pp. 3–8, 1976.

[79] H. R. Rothstein, A. J. Sutton, and M. Borenstein, *Publication bias in meta-analysis: Prevention, assessment and adjustments*. West Sussex, England: Wiley & Sons, 2005.

[80] M. L. Smith and G. V. Glass, “Meta-analysis of psychotherapy outcome studies,” *Am. Psychol.*, vol. 32, pp. 752–760, 1977.

[81] R. Dersimonian and N. Laird, “Meta-Analysis in Clinical Trials,” *Stat. Med.*, vol. 188, pp. 177–188, 1986.

[82] Y. Zhang and Q. Yang, “An overview of multi-task learning,” *Natl. Sci. Rev.*, vol. 5, no. 1, pp. 30–43, 2018.

[83] P. Gong, J. Ye, and C. Zhang, “Robust Multi-Task Feature Learning,” pp. 895–903, 2013.

[84] N. Sarwar *et al.*, “Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: A collaborative meta-analysis of 102 prospective studies,” *Lancet*, vol. 375, no. 9733, pp. 2215–2222, 2010.

[85] R. DerSimonian and N. Laird, “Meta-analysis in clinical trials,” *Control. Clin. Trials*, vol. 7, no. 3, pp. 177–188, 1986.

[86] A. Karthikesalingam *et al.*, “A systematic review and meta-analysis indicates underreporting of renal dysfunction following endovascular aneurysm repair,” *Kidney Int.*, vol. 87, no. 2, pp. 442–451, 2015.

[87] M. Borenstein, L. Hedges, and H. Rothstein, “Meta-Analysis Fixed effect vs . random effects,” *Test*, p. 162, 2007.

[88] Y. Chung, S. Rabe-Hesketh, and I. H. Choi, “Avoiding zero between-study variance estimates in random-effects meta-analysis,” *Stat. Med.*, vol. 32, no. 23, pp. 4071–4089, 2013.

[89] R. D. Riley *et al.*, “Multivariate and network meta-analysis of multiple outcomes and multiple treatments : rationale , concepts , and examples Institute for Health and Care Excellence,” *BMJ*, vol. 358, p. 3932, 2017.

[90] K. J. Ishak, R. W. Platt, L. Joseph, and J. A. Hanley, “Impact of approximating or ignoring within-study covariances in multivariate meta-analyses,” *Stat Med*, no. March 2007, pp. 670–686, 2008.

[91] R. D. Riley, “Multivariate meta-analysis: The effect of ignoring within-study correlation,” *J. R. Stat. Soc. Ser. A Stat. Soc.*, vol. 172, no. 4, pp. 789–811, 2009.

[92] R. Bender *et al.*, “Attention should be given to multiplicity issues in systematic reviews,” *J Clin Epidemiol*, vol. 61, no. 857–865, 2008.

[93] R. D. Riley, K. R. Abrams, P. C. Lambert, A. J. Sutton, and J. R. Thompson, “An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes,” *Stat. Med.*, vol. 26, no. 1, pp. 78–97, 2007.

[94] D. Jackson and R. Riley, “Multivariate meta-analysis: Potential and promise,” *Stat. Med.*, no. January, 2010.

[95] J. B. Reitsma, A. S. Glas, A. W. S. Rutjes, R. J. P. M. Scholten, P. M. Bossuyt, and A. H. Zwinderman, “Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews,” *J. Clin. Epidemiol.*, vol. 58, no. 10, pp. 982–990, 2005.

[96] H. Chu and S. R. Cole, “Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach,” *J. Clin. Epidemiol.*, vol. 59, no. 12, pp. 1331–1332, 2006.

[97] M. Kertai, E. Boersma, J. Bax, M. Heijenbrok-Kal, and M. Hunink, “A meta-analysis comparing the prognostic accuracy of six diagnostic tests for predicting perioperative cardiac risk in patients undergoing major vascular surgery,” *Med. Decis. Mak.*, vol. 89, no. 11, pp. 1327–1334, 2003.

[98] R. DerSimonian and Nan Laird, “Meta-Analysis in Clinical Trials Revisited,” *Contemp Clin Trials*, vol. 45, no. 0 0, pp. 139–145, 2015.

[99] D. R. Rhodes, T. R. Barrette, M. A. Rubin, D. Ghosh, and A. M. Chinnaiyan, “Meta-Analysis of Microarrays : Interstudy Validation of Gene Expression Profiles Reveals Pathway Dysregulation in Prostate Cancer 1,” *CANCER Res.*, vol. 62, pp. 4427–4433, 2002.

[100] J. K. Choi, U. Yu, S. Kim, and O. J. Joon, “Combining multiple microarray studies and modeling interstudy variation,” *BIOINFORMATICS*, vol. Vol. 19 Su, pp. i84–i90, 2003.

[101] J. Wang, S. Vasaikar, Z. Shi, M. Greer, and B. Zhang, “WebGestalt 2017: A more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit,” *Nucleic Acids Res.*, vol. 45, no. W1, pp. W130–W137, 2017.

[102] P. Khatri, M. Sirota, and A. J. Butte, “Ten years of pathway analysis: current approaches and outstanding challenges.,” *PLoS Comput. Biol.*, vol. 8, no. 2, p. e1002375, Jan. 2012.

[103] J. Wang *et al.*, “Integrative genomics analysis identifies candidate drivers at 3q26-29 amplicon in squamous cell carcinoma of the lung,” *Clin Cancer Res*, vol. 19, no. 20, pp. 5580–5590, 2013.

[104] J. Wang *et al.*, “Proteome Profiling Outperforms Transcriptome Profiling for Coexpression Based Gene Function Prediction,” *Mol. Cell. Proteomics*, vol. 16, pp. 121–134, 2017.

[105] J. J. Goeman and P. Bühlmann, “Analyzing gene expression data in terms of gene sets: Methodological issues,” *Bioinformatics*, vol. 23, no. 8, pp. 980–987, 2007.

[106] Pavlidis P, Qin J, Arango V, Mann J, and Sibille E, “Using the Gene Ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex,” *Neurochem Res*, vol. 29, pp. 1213–1222, 2004.

[107] Al-Shahrour F, Diaz-Uriarte R, and Dopazo J, “Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information,” *Bioinformatics*, vol. 21, pp. 2988–2993, 2005.

[108] G. JJ, V. SA, de K. F, and van H. HC, “A global test for groups of genes: testing association with a clinical outcome,” *Bioinformatics*, vol. 20, pp. 93–99, 2004.

[109] L. Tian, S. a Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park, “Discovering statistically significant pathways in expression profiling studies.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 38, pp. 13544–9, Sep. 2005.

[110] K. SY and V. DJ, “PAGE: parametric analysis of gene set enrichment,” *BMC Bioinformatics*, vol. 6, p. 144, 2005.

[111] J. Wang, D. Duncan, Z. Shi, and B. Zhang, “WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013,” vol. 41, pp. 77–83, 2013.

[112] M. Pons, P Latapy, “Computing communities in large networks using random walks,” *Comput. Inform. Sci.*, vol. 3733, pp. 284–293, 2005.

[113] M. Newman and M. Girvan, “Finding and evaluating community structure in networks,” *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, vol. 69, p. 026113, 2004.

[114] A. Barczak *et al.*, “Spotted long oligonucleotide arrays for human gene expression analysis,” *Genome Res.*, vol. 13, no. 7, pp. 1775–1785, 2003.

[115] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-barclay, K. J. Antonellis, and T. P. Speed, “Exploration , Normalization , and Summaries of High Density Oligonucleotide Array Probe Level Data,” no. June, pp. 249–264, 2018.

[116] V. Bewick, L. Cheek, and J. Ball, “Statistics review 7: Correlation and regression,” *Crit. Care*, vol. 7, no. 6, pp. 451–459, 2003.

[117] A. N. Tegge, C. W. Caldwell, and D. Xu, “Pathway Correlation Profile of Gene-Gene Co-Expression for Identifying Pathway Perturbation,” *PLoS One*, vol. 7, no. 12, 2012.

[118] Y. Pita-Juárez *et al.*, “The Pathway Coexpression Network: Revealing pathway relationships,” *PLoS Comput. Biol.*, vol. 14, no. 3, pp. 1–28, 2018.

[119] T. Yun and G.-S. Yi, “Biclustering for the comprehensive search of correlated gene expression patterns using clustered seed expansion,” *BMC Genomics*, vol. 14, no. 1, p. 144, 2013.

[120] A. Bhattacharya and Y. Cui, “A GPU-accelerated algorithm for biclustering analysis and detection of condition-dependent coexpression network modules,” *Sci. Rep.*, vol. 7, no. 1, p. 4162, 2017.

[121] H. Liu, I. Bebu, and X. Li, “Microarray probes and probe sets,” *Front Biosci (Elite Ed)*, vol. 2, pp. 325–338, 2010.

[122] J. D. Allen *et al.*, “Probe mapping across multiple microarray platforms,” *Brief. Bioinform.*, vol. 13, no. 5, pp. 547–554, 2012.

[123] R. Coe, “It’s the effect size, stupid. What effect size is and why it is important,” *Br. Educ. Res. Assoc. Annu. Conf.*, pp. 1–18, 2002.

[124] T. Vacha-haase and B. Thompson, “How to Estimate and Interpret Various Effect Sizes,” vol. 51, no. 4, pp. 473–481, 2004.

[125] J. Cohen, *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Earlbaum Associates, 1988.

[126] L. V. Hedges, “Distribution theory for Glass’ estimator of effect size and related estimators,” *J. Educ. Stat.*, vol. 6, no. 2, pp. 107–128, 1981.

[127] S. Nakagawa and I. C. Cuthill, “Effect size, confidence interval and statistical significance: A practical guide for biologists,” *Biol. Rev.*, vol. 82, no. 4, pp. 591–605, 2007.

[128] S. Kosub, “A note on the triangle inequality for the Jaccard distance,” *arXiv Prepr. arXiv*, vol. 1612.02696, 2016.

[129] S. R. Eliason, *Maximum Likelihood Estimation: Logic and Practice*. SAGE Publications, Inc, 1993.

[130] H. D. Patterson and R. Thompson, “Recovery of inter-block information when block sizes are unequal,” vol. 58, no. 3, pp. 545–554, 2018.

[131] I. R. White, “Multivariate random-effects meta-regression: Updates to mvmeta,” *Stata J.*, vol. 11, no. 2, pp. 255–270, 2011.

[132] A. Gasparrini, B. Armstrong, and M. G. Kenward, “Multivariate meta-analysis for non-linear and other multi-parameter associations,” *Stat. Med.*, vol. 31, no. 29, pp. 3821–3839, 2012.

[133] J. Kruschke, *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*, 2nd Editio. Academic Press, 2014.

[134] D. Jackson, W. IR, and T. SG, “Extending DerSimonian and Laird’s methodology to perform multivariate random,” - *Stat Med. 2010 May 30;29(12)1282-97. doi 10.1002/sim.3602.*, no. 1097-0258 (Electronic), p. 1282-97, 2010.

[135] A. Gelman, J. Carlin, H. Stern, and D. Rubin, “Bayesian Data Analysis. 2nd.,” in *Bayesian Data Analysis. 2nd.*, CRC Press; Boca Raton, 2003.

[136] M. Plummer, “JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling,” *Proc. 3rd Int. Work. Distrib. Stat. Comput. (DSC 2003)*, pp. 20–22, 2003.

[137] J. Higgins and S. (editors) Green, “Cochrane Handbook for Systematic Reviews of Interventions,” Version 5., 2011.

[138] I. J. Myung, “Tutorial on maximum likelihood estimation,” *J. Math. Psychol.*, vol. 47, no. 1, pp. 90–100, 2003.

[139] E. H. Newman and K. Kingsley, “An introduction to the method of moments,” *Comput. Phys.*

Commun., vol. 68, no. 1–3, pp. 1–18, 1991.

[140] R. R. Corbeil and S. R. Searl, “Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model,” *TECHNOMETRICS*, vol. 18, no. 1, pp. 31–38, 1976.

[141] G. M. Allenby and P. E. Rossi, “Hierarchical Bayes Models,” *Handb. Mark. Res. Uses, Misuses, Futur. Adv.*, pp. 1–59, 2006.

[142] J. Hartung and G. Knapp, “On tests of the overall treatment effect in meta-analysis with normally distributed responses,” *Stat. Med.*, vol. 20, no. 12, pp. 1771–1782, 2001.

[143] J. Hartung and G. Knapp, “A refined method for the meta-analysis of controlled clinical trials with binary outcome,” *Stat. Med.*, vol. 20, no. 24, pp. 3875–3889, 2001.

[144] K. Sidik and J. N. Jonkman, “A simple confidence interval for meta-analysis,” *Stat. Med.*, vol. 21, no. 21, pp. 3153–3159, 2002.

[145] S. Nadarajah and S. Kotz, “Mathematical properties of the multivariate t distribution,” *Appl. Math.*, vol. 89, pp. 53–84, 2005.

[146] S. Kotz and S. Nadarajah, *Multivariate t Distributions and Their Applications*. Cambridge, UK: Cambridge University Press, 2004.

[147] A. Genz *et al.*, “Multivariate Normal and t Distributions,” *R Packag.*, vol. version 0., 2018.

[148] C. ROBERTSON and J. FRYER, “The Bias and Accuracy of Moment Estimators,” vol. 57, no. 1, pp. 57–65, 1970.

[149] P. J. Bickel and E. Levina, “REGULARIZED ESTIMATION OF LARGE COVARIANCE MATRICES,” *Ann. Stat.*, vol. 36, no. 1, pp. 199–227, 2008.

[150] M. Pourahmadi, “Covariance Estimation : The GLM and Regularization Perspectives,” vol. 26, no. 3, pp. 369–387, 2011.

[151] Q. Mai, “A review of discriminant analysis in high dimensions,” *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 5, no. 3, pp. 190–197, 2013.

[152] I. Johnstone, “On the distribution of the largest eigenvalue in principal components analysis,” *Ann. Stat.*, vol. 29, pp. 295–327, 2001.

[153] L. Hedges and I. Olkin, *Statistical methods for meta-analysis*. New York: Academic Press, 1985.

[154] J. Friedman and R. Tibshirani, “Sparse inverse covariance estimation with the graphical lasso,” pp. 1–10, 2007.

[155] L. El Ghaoui and A. Aspremont, “Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data.”

[156] D. Nishimura, “A view from the Web, BioCarta,” *Biotech Softw. Internet Rep.*, vol. 2, no. 3, pp. 117–120, 2001.

[157] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo, “The Molecular Signatures Database Hallmark Gene Set Collection,” *Cell Syst.*, vol. 1, no. 6, pp. 417–425, 2015.

[158] P. D. Karp, “The MetaCyc Database,” *Nucleic Acids Res.*, vol. 30, no. 1, pp. 59–61, 2002.

[159] a M. Huerta, H. Salgado, D. Thieffry, and J. Collado-Vides, “RegulonDB: a database on transcriptional regulation in Escherichia coli,” *Nucleic Acids Res*, vol. 26, no. 1, pp. 55–59, 1998.

[160] P. D. Thomas *et al.*, “PANTHER: A library of protein families and subfamilies indexed by function,” *Genome Res.*, vol. 13, no. 9, pp. 2129–2141, 2003.

[161] P. Pons and M. Latapy, “Computing Communities in Large Networks Using Random Walks,” *J. Graph Algorithms Appl.*, vol. 10, no. 2, pp. 191–218, 2006.

[162] N. P. P. Kathleen Collins, Tyler Jacks, “The cell cycle and cancer.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 94, no. April, pp. 2776–2778, 1997.

[163] S. A. Aaronson, “Growth Factors and Cancer,” *Science (80-.).*, vol. 254, pp. 1146–1153, 1991.

[164] C. J. Sherr, “Cancer cell cycles. (Cover story),” *Science (80-.).*, vol. 274, no. 5293, p. 1672, 1996.

[165] P. E. Lonergan and D. J. Tindall, “Androgen receptor signaling in prostate cancer development and progression,” *J. Carcinog.*, vol. 10, no. 20, 2011.

[166] N. Poulose, F. Amoroso, R. E. Steele, R. Singh, C. W. Ong, and I. G. Mills, “Genetics of lipid metabolism in prostate cancer,” *Nat. Genet.*, vol. 50, no. February, pp. 169–171, 2018.

[167] A. Vellaichamy, A. Sreekumar, J. R. Strahler, and T. Rajendiran, “Proteomic Interrogation of Androgen Action in Prostate Cancer Cells Reveals Roles of Aminoacyl tRNA Synthetases,” *PLoS ONE |*, vol. 4, no. 9, 2009.

[168] E. Alfaro, G. Francisco, and A. A. Protter, “Enzalutamide , an Androgen Receptor Signaling Inhibitor , InducesTumor Regression in a Mouse Model of Castration-Resistant Prostate Cancer,” *Prostate*, vol. 73, no. June, pp. 1291–1305, 2013.

[169] E. K. Keeton *et al.*, “A Hierarchical Network of Transcription Factors Governs Androgen Receptor-Dependent Prostate Cancer Growth,” *Mol. Cell*, vol. 27, pp. 380–392, 2007.

[170] K. Y. Yip, D. W. Cheung, I. C. Society, and M. K. Ng, “HARP : A Practical Projected Clustering Algorithm,” *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 11, pp. 1387–1397, 2004.

[171] X. Liu and L. Wang, “Gene expression Computing the maximum similarity bi-clusters of gene expression data,” *BIOINFORMATICS*, vol. 23, no. 1, pp. 50–56, 2007.

[172] G. Li, Q. Ma, H. Tang, A. Paterson, and Y. Xu, “QUBIC: a qualitative biclustering algorithm for analyses of gene expression data,” *Nucleic Acids Res*, vol. 37, 2009.

[173] S. Hochreiter *et al.*, “FABIA: factor analysis for bicluster acquisition.,” *Bioinformatics*, vol. 26, no. 12, pp. 1520–7, Jun. 2010.

[174] T. Murali and S. Kasif, “Extracting conserved gene expression motifs from gene expression data,” *Pac. Symp. Biocomput.*, vol. 88, pp. 77–88, 2003.

[175] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein, “Spectral Biclustering of Microarray Data : Coclustering Genes and Conditions,” *Genome Res.*, vol. 13, pp. 703–716, 2003.

[176] S. Bergmann, J. Ihmels, and N. Barkai, “Iterative signature algorithm for the analysis of large-scale gene expression data.,” *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.*, vol. 67, no. 3 Pt 1, p. 031902,

2003.

Appendix I: List of commonly used biclustering algorithms

Metric-based biclustering algorithms:

Measure	Mathematical representation	Description	Comments
Variance (VAR) [62]	$VAR(\mathcal{B}) = \sum_{i=1}^{\lvert I\rvert}\sum_{j=1}^{\lvert J\rvert}(b_{ij} - b_{IJ})^2$	Bicluster variance is used as a coherence measure, where the goal of his algorithm was to minimize the sum of bicluster variances.	The variance only detects constant biclusters.
Mean Squared Residue (MSR) [63]	$MSR(\mathcal{B}) = \frac{1}{\lvert I\rvert \cdot \lvert J\rvert}\sum_{i=1}^{\lvert I\rvert}\sum_{j=1}^{\lvert J\rvert}(b_{ij} - b_{Ij} - b_{iJ} + b_{IJ})^2$	The lower the mean squared residue, the stronger the coherence exhibited by the bicluster, and the better its quality.	Inefficient for finding those biclusters with strong scaling tendencies.
Relevance Index (RI) [170]	*Relevance index* R_{Ij} for column j∈J is defined as: $R_{Ij} = 1 - \frac{\sigma_{Ij}^2}{\sigma j^2}$ where σ_{Ij}^2(local variance) and σ_j^2 ((global variance) are the variance of the values in column j for the bicluster and the whole data set, respectively	The index gives a high value when the local variance is small compared to the global variance.	The only bicluster patterns that maximize the quality are constant biclusters (either on rows or on columns).
Similarity Score for a Bicluster (SS) [171]	The similarity scores for each row i∈I, and for each column j∈J are defined as: $s(i,J) = \sum_{j\in J} s_{ij}$, and $s(I,j) = \sum_{i\in I} s_{ij}$, respectively. The similarity score between two genes (gene i and a reference gene i*) under condition j is computed as: $s_{ij} = \begin{cases} 0 & \text{if } d_{ij} > \alpha \cdot d_{avg} \\ 1 - \frac{d_{ij}}{\alpha \cdot d_{avg}} + \beta & \text{otherwise} \end{cases}$ where d_{avg} is defined as the average distance value of all the elements in the expression matrix: $d_{avg} = \frac{\sum_{i\in I, j\in J} d_{ij}}{\lvert I\rvert\lvert J\rvert}$ and d_{ij} is the absolute value of the expression difference between the gene i and the reference gene *i** for condition *j* in the expression matrix *a*: $d_{ij} = \lvert a_{ij} - a_{i*j}\rvert$ a^*d_{avg} is used as a threshold to ignore elements with a large d_{ij}, in order to find constant biclusters, and *β* is the bonus for small d_{ij}. This way, *β* enlarges the similarity score for small d_{ij} and ignores d_{ij} greater than the threshold.	Using these equations, the similarity score for a bicluster is computed as the minimum value of the similarity scores of both genes and conditions in the bicluster: $s(\mathcal{B}) = s(I,J) = \min\{\min_{i\in I} s(i,J), \min$ The goal when looking for biclusters is to find sub-matrices with higher values for the similarity score.	Although the type of bicluster found using the similarity score depends on the values for the different thresholds (*α*, *β*, and *γ* for the average), only constant and additive biclusters are recognized.

Pearson's Correlation Coefficient (PCC)[70]	$$PCC(i_1, i_2) = \frac{\sum_{j=1}^{|J|}(b_{i_1 j} - b_{i_1 J})(b_{i_2 j} - b_{i_2 J})}{\sqrt{\sum_{j=1}^{|J|}(b_{i_1 j} - b_{i_1 J})^2 \sum_{j=1}^{|J|}(b_{i_2 j} - b_{i_2 J})^2}}$$ where b_{i1j} and b_{i2j} denote the elements in rows i_1, i_2 and column j, and b_{i1J}, b_{i2J} represent the means of rows i_1 and i_2, respectively	PCC quantifies coherences between pairs of genes. Therefore, in order to measure bicluster coherence, one has to compute all pairwise PCC values between the rows in the same bicluster.	PCC is a very effective metric to quantify co-regulation between pairs of genes [20], and it allows both shifting and scaling patterns to be captured that would be separately identified by additive and multiplicative models, respectively. Nevertheless, PCC is not effective for recognizing constant biclusters or constant row patterns, since these kinds of patterns would make the denominator zero.

Non metric-based biclustering algorithms:

Category	Algorithm	Description	Comments
Graph-based approaches	Statistical-Algorithmic Method for Bicluster Analysis (SAMBA)[75]	(1) It models the input expression data as a bipartite graph whose two parts correspond to conditions and genes. (2) The edges refer to significant expression changes. The vertex pairs in the graph are assigned weights according to a probabilistic model, so that heavy sub-graphs correspond to biclusters with high likelihood. (3) Discovering the most significant biclusters means finding the heaviest sub-graphs in the bipartite graph model, where the weight of a sub-graph is the sum of the weights of the gene-condition pairs in it.	It can detect either up or down regulation.
	Qualitative Biclustering algorithm (QUBIC)[172]	(1) The input data matrix is first represented as a matrix of integer values. (2) A weighted graph is constructed from the qualitative or semi-qualitative matrix, with genes represented as vertices, and edges connecting every pair of genes. (3) Two genes are considered to be correlated under a subset of conditions if the corresponding integer values along the two corresponding rows of the matrix are identical.	It can find both positively and negatively correlated expression patterns,
Probabilistic models	Plaid Models (PM)[68]	(1) The gene-condition matrix is represented as a superposition of layers, corresponding to biclusters.	It can discover overlapped biclusters, but it has a major limitation: the initial choice of

		$Y_{ij} \doteq \sum_{k=0}^{K} \theta_{ijk} \rho_{ik} \kappa_{jk}$, where Y_{ij} refers to the expression level of gene *i* under sample *j* in the input matrix, *K* is the number of biclusters, θ_{ij0} describes the background layer and θ_{ijk}represents four different types of models, depending on the types of biclusters (overlapped, exclusive .. .) (2) The process seeks for a plaid model minimizing the sum of squared errors when approximating the data matrix to the model.	model parameters has a strong influence to the biclustering result.
	Bayesian Biclustering model (BBC) [72]	It uses Gibbs sampling to fit a hierarchical Bayesian version of the plaid model.	It only allows the biclusters overlapped on the gene dimension, not on the condition dimension.
	Factor analysis for bicluster acquisition (FABIA) [173]	It models the data matrix X as the sum of *p* biclusters plus additive noise ϒ, where each bicluster is the outer product of two sparse vectors: a row vector λ and a column vector *z*: $\mathbf{X} = \sum_{i=1}^{p} \lambda_i \mathbf{z}_i^T + \Upsilon = \Lambda Z + \Upsilon.$	The initial choice of model parameters has a strong influence to the biclustering result.
	Conserved gene expression Motifs (xMOTIFs)[174]		This search strategy allows gene overlap and also sample overlap
	Gibbs Sampling (GS) [71]		
Linear algebra	Spectral Biclustering (SB)[175]		
	Iterative Signature Algorithm (ISA) Iterative[176]		

Appendix II: Information on the member genes in the bicluster stacks

ProsBicSta01*

Gene Probe	Gene Symbol	Gene Name	Entrez Gene
227492_at	OCLN	occludin	100506658
209925_at	OCLN	occludin	100506658
235445_at	LOC100508046	uncharacterized LOC100508046	100508046
226154_at	DNM1L	dynamin 1 like	10059
232397_at	LOC101927482	uncharacterized LOC101927482	101927482
1558369_at	MPHOSPH9	M-phase phosphoprotein 9	10198
203196_at	ABCC4	ATP binding cassette subfamily C member 4	10257
243762_at	LINC01297	long intergenic non-protein coding RNA 1297	106146148
212252_at	CAMKK2	calcium/calmodulin dependent protein kinase kinase 2	10645
1558692_at	GLMP	glycosylated lysosomal membrane protein	112770
226726_at	MBOAT2	membrane bound O-acyltransferase domain containing 2	129642
213288_at	MBOAT2	membrane bound O-acyltransferase domain containing 2	129642
225344_at	NCOA7	nuclear receptor coactivator 7	135112
235085_at	PRAG1	PEAK1 related kinase activating pseudokinase 1	157285
205311_at	DDC	dopa decarboxylase	1644
201660_at	ACSL3	acyl-CoA synthetase long-chain family member 3	2181
242726_at	ACSL3	acyl-CoA synthetase long-chain family member 3	2181
204560_at	FKBP5	FK506 binding protein 5	2289
224840_at	FKBP5	FK506 binding protein 5	2289
224856_at	FKBP5	FK506 binding protein 5	2289
226982_at	ELL2	elongation factor for RNA polymerase II 2	22936
214446_at	ELL2	elongation factor for RNA polymerase II 2	22936
226099_at	ELL2	elongation factor for RNA polymerase II 2	22936
212350_at	TBC1D1	TBC1 domain family member 1	23216
41644_at	SASH1	SAM and SH3 domain containing 1	23328
227669_at	MPC2	mitochondrial pyruvate carrier 2	25874
1553645_at	CCDC141	coiled-coil domain containing 141	285025
1563571_at	CTBP1-AS	CTBP1 antisense RNA	285463
209409_at	GRB10	growth factor receptor bound protein 10	2887
231015_at	KLF15	Kruppel like factor 15	28999
222108_at	AMIGO2	adhesion molecule with Ig like domain 2	347902
225330_at	IGF1R	insulin like growth factor 1 receptor	3480
203628_at	IGF1R	insulin like growth factor 1 receptor	3480
225571_at	LIFR	leukemia inhibitory factor receptor alpha	3977
51158_at	FAM174B	family with sequence similarity 174 member B	400451
209706_at	NKX3-1	NK3 homeobox 1	4824
204957_at	ORC5	origin recognition complex subunit 5	5001
205040_at	ORM1	orosomucoid 1	5004
219933_at	GLRX2	glutaredoxin 2	51022
223204_at	FAM198B	family with sequence similarity 198 member B	51313
219872_at	FAM198B	family with sequence similarity 198 member B	51313
209481_at	SNRK	SNF related kinase	54861
218692_at	SYBU	syntabulin	55638
218764_at	PRKCH	protein kinase C eta	5583
223093_at	ANKH	ANKH inorganic pyrophosphate transport regulator	56172
223092_at	ANKH	ANKH inorganic pyrophosphate transport regulator	56172
222449_at	PMEPA1	prostate transmembrane protein, androgen induced 1	56937
222450_at	PMEPA1	prostate transmembrane protein, androgen induced 1	56937
223401_at	ADPRM	ADP-ribose/CDP-alcohol diphosphatase, manganese dependent	56985
212977_at	ACKR3	atypical chemokine receptor 3	57007
225295_at	SLC39A10	solute carrier family 39 member 10	57181
203329_at	PTPRM	protein tyrosine phosphatase, receptor type M	5797
223168_at	RHOU	ras homolog family member U	58480
239202_at	RAB3B	RAB3B, member RAS oncogene family	5865
213139_at	SNAI2	snail family transcriptional repressor 2	6591
228562_at	ZBTB10	zinc finger and BTB domain containing 10	65986

201563_at	SORD	sorbitol dehydrogenase	6652
230782_at	SORD	sorbitol dehydrogenase	6652
209340_at	UAP1	UDP-N-acetylglucosamine pyrophosphorylase 1	6675
202363_at	SPOCK1	SPARC/osteonectin, cwcv and kazal like domains proteoglycan 1	6695
205102_at	TMPRSS2	transmembrane protease, serine 2	7113
226553_at	TMPRSS2	transmembrane protease, serine 2	7113
235888_at	GUSBP1	glucuronidase, beta pseudogene 1	728411
226005_at	UBE2G1	ubiquitin conjugating enzyme E2 G1	7326
205883_at	ZBTB16	zinc finger and BTB domain containing 16	7704
225987_at	STEAP4	STEAP4 metalloreductase	79689
201675_at	AKAP1	A-kinase anchoring protein 1	8165
223544_at	TMEM79	transmembrane protein 79	84283
225819_at	TBRG1	transforming growth factor beta regulator 1	84897
228696_at	SLC45A3	solute carrier family 45 member 3	85414
209250_at	DEGS1	delta 4-desaturase, sphingolipid 1	8560
210946_at	PLPP1	phospholipid phosphatase 1	8611
1554290_at	HERC3	HECT and RLD domain containing E3 ubiquitin protein ligase 3	8916
232639_at	EFCAB12	EF-hand calcium binding domain 12	90288
203910_at	ARHGAP29	Rho GTPase activating protein 29	9411

*Note: Gene probes mapped to multiple Entrtez IDs or not mapped: 206272_at, 215248_at, 226489_at, 227762_at, 228854_at, 229163_at, 230082_at, 236774_at.

ProsBicSta06*

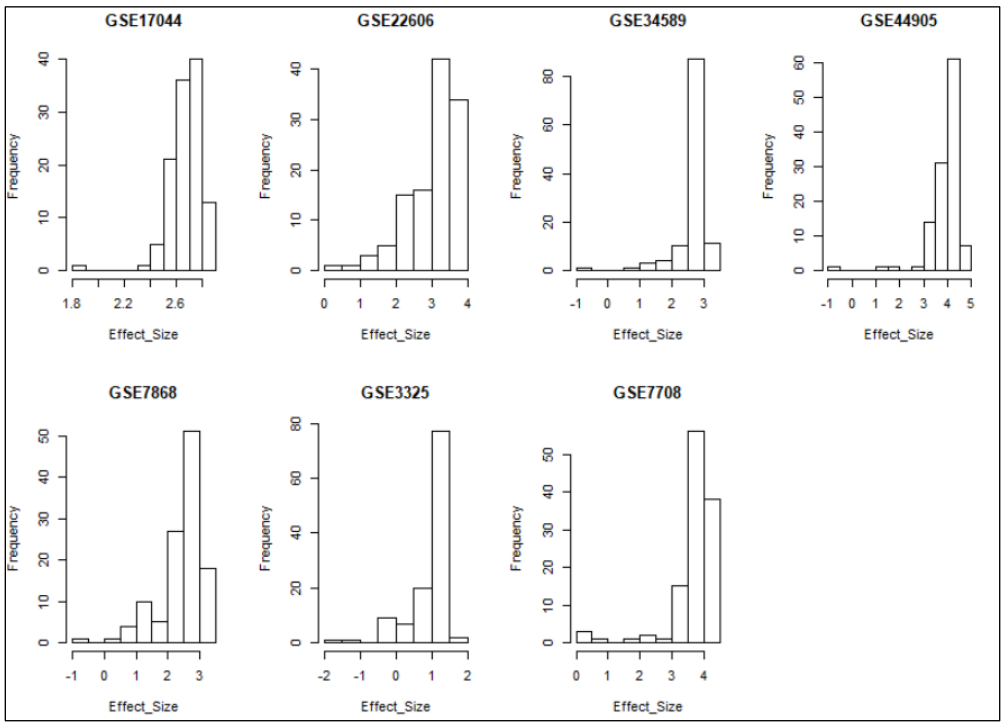
Gene Probe	Gene Symbol	Gene Name	Entrez Gene
218755_at	KIF20A	kinesin family member 20A	10112
204162_at	NDC80	NDC80, kinetochore complex component	10403
203145_at	SPAG5	sperm associated antigen 5	10615
204146_at	RAD51AP1	RAD51 associated protein 1	10635
206023_at	NMU	neuromedin U	10874
202954_at	UBE2C	ubiquitin conjugating enzyme E2 C	11065
224753_at	CDCA5	cell division cycle associated 5	113130
213599_at	OIP5	Opa interacting protein 5	11339
227295_at	IKBIP	IKBKB interacting protein	121457
235572_at	SPC24	SPC24, NDC80 kinetochore complex component	147841
226661_at	CDCA2	cell division cycle associated 2	157313
212805_at	PRUNE2	prune homolog 2	158471
212806_at	PRUNE2	prune homolog 2	158471
216307_at	DGKB	diacylglycerol kinase beta	1607
226610_at	CENPV	centromere protein V	201161
242560_at	FANCD2	Fanconi anemia complementation group D2	2177
212621_at	NEMP1	nuclear envelope integral membrane protein 1	23306
228785_at	ZNF281	zinc finger protein 281	23528
209921_at	SLC7A11	solute carrier family 7 member 11	23657
217678_at	SLC7A11	solute carrier family 7 member 11	23657
218355_at	KIF4A	kinesin family member 4A	24137
232238_at	ASPM	abnormal spindle microtubule assembly	259266
228391_at	CYP4V2	cytochrome P450 family 4 subfamily V member 2	285440
209398_at	HIST1H1C	histone cluster 1 H1 family member c	3006
227350_at	HELLS	helicase, lymphoid-specific	3070
207165_at	HMMR	hyaluronan mediated motility receptor	3161
202094_at	BIRC5	baculoviral IAP repeat containing 5	332
201555_at	MCM3	minichromosome maintenance complex component 3	4172
204058_at	ME1	malic enzyme 1	4199
201710_at	MYBL2	MYB proto-oncogene like 2	4605
219258_at	TIPIN	TIMELESS interacting protein	54962
213008_at	FANCI	Fanconi anemia complementation group I	55215
213007_at	FANCI	Fanconi anemia complementation group I	55215
219502_at	NEIL3	nei like DNA glycosylase 3	55247
219703_at	MNS1	meiosis specific nuclear structural 1	55329
218726_at	HJURP	Holliday junction recognition protein	55355
205053_at	PRIM1	primase (DNA) subunit 1	5557
218820_at	C14orf132	chromosome 14 open reading frame 132	56967
228323_at	KNL1	kinetochore scaffold 1	57082
219099_at	TIGAR	TP53 induced glycolysis regulatory phosphatase	57103
231855_at	KIAA1524	KIAA1524	57650
217995_at	SQRDL	sulfide quinone reductase-like (yeast)	58472
204023_at	RFC4	replication factor C subunit 4	5984
203209_at	RFC5	replication factor C subunit 5	5985
201890_at	RRM2	ribonucleotide reductase regulatory subunit M2	6241
205733_at	BLM	Bloom syndrome RecQ like helicase	641
222848_at	CENPK	centromere protein K	64105
218663_at	NCAPG	non-SMC condensin I complex subunit G	64151
227034_at	SOWAHC	sosondowah ankyrin repeat domain family member C	65124
203755_at	BUB1B	BUB1 mitotic checkpoint serine/threonine kinase B	701
201292_at	TOP2A	topoisomerase (DNA) II alpha	7153
204822_at	TTK	TTK protein kinase	7272
218741_at	CENPM	centromere protein M	79019
200934_at	DEK	DEK proto-oncogene	7913
229305_at	CENPU	centromere protein U	79682
219990_at	E2F8	E2F transcription factor 8	79733
235609_at	BRIP1	BRCA1 interacting protein C-terminal helicase 1	83990

223700_at	MND1	meiotic nuclear divisions 1	84057
209529_at	PLPP2	phospholipid phosphatase 2	8612
203560_at	GGH	gamma-glutamyl hydrolase	8836
203418_at	CCNA2	cyclin A2	890
213226_at	CCNA2	cyclin A2	890
202705_at	CCNB2	cyclin B2	9133
209406_at	BAG2	BCL2 associated athanogene 2	9532
226016_at	CD47	CD47 molecule	961
203764_at	DLGAP5	DLG associated protein 5	9787
204825_at	MELK	maternal embryonic leucine zipper kinase	9833
236641_at	KIF14	kinesin family member 14	9928

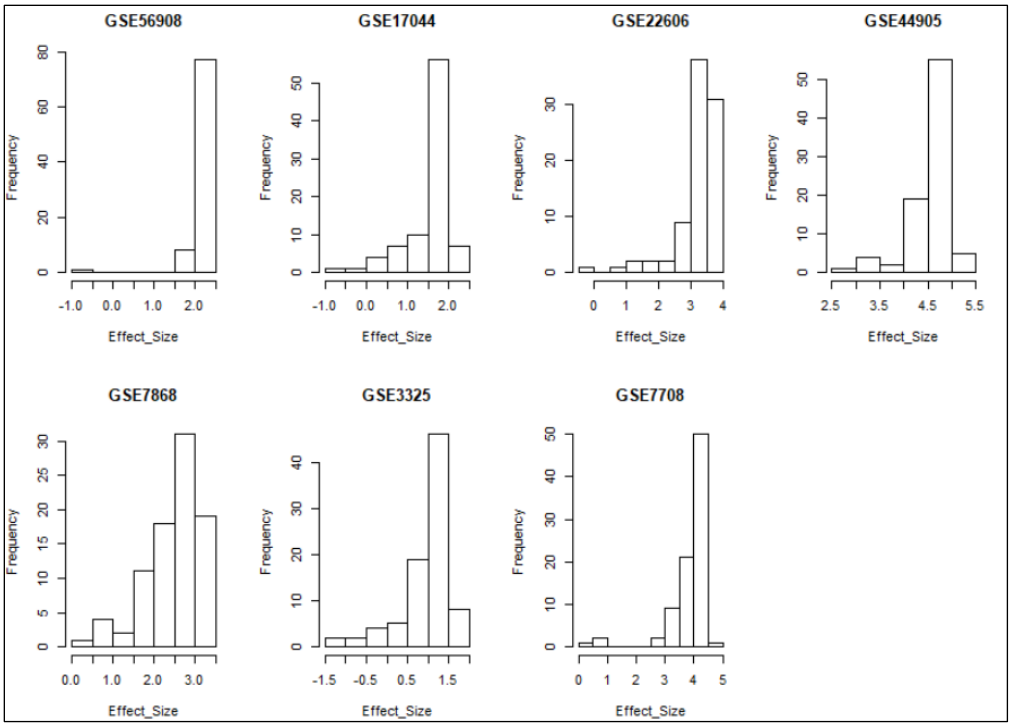
* Note: Gene probes mapped to multiple Entrtez IDs or not mapped: 210187_at, 212126_at, 225834_at, 235363_at, 240478_at, 243063_at.

Appendix III: Distributions of individual effect sizes within four bicluster stacks

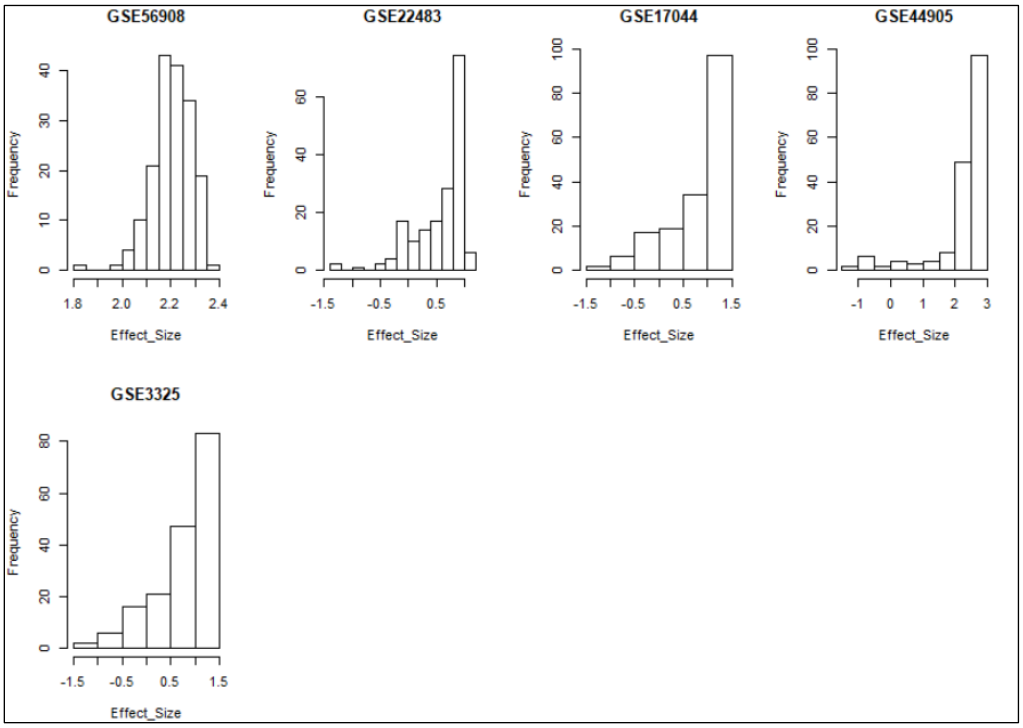
ProsBicSta02:



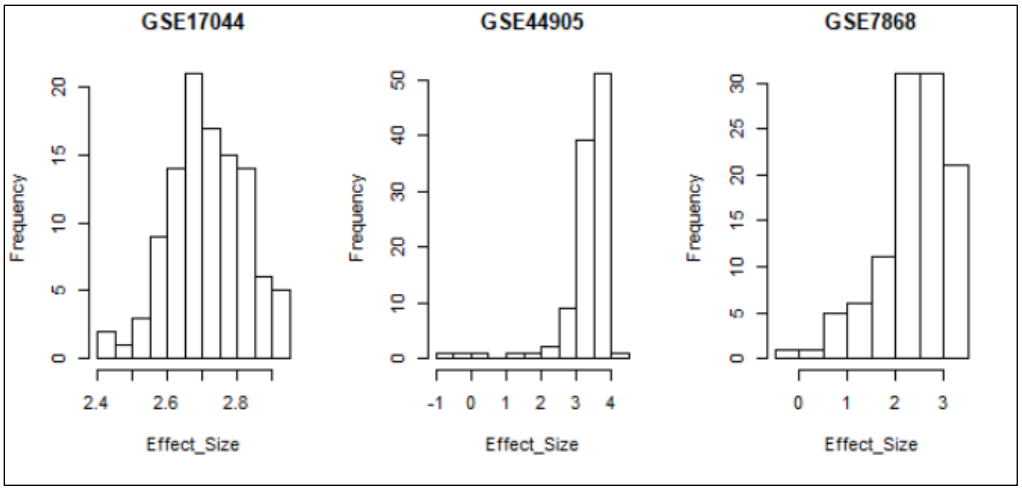
ProsBicSta03:



ProsBicSta12:



ProsBicSta19:


GSE17044
Frequency
0
5
10
15
20
2.4
2.6
2.8
Effect_Size
GSE44905
Frequency
0
10
20
30
40
50
-1
0
1
2
3
4
Effect_Size
GSE7868
Frequency
0
5
10
15
20
25
30
0
1
2
3
Effect_Size

Appendix IV: Ranked seeds in NTA for ProsBicSta06

Gene Symbol	Random Walk Probability	Gene Symbol	Random Walk Probability	Gene Symbol	Random Walk Probability
CENPU	8.83E-03	UBE2C	8.07E-03	HELLS	8.00E-03
CENPM	8.50E-03	RFC5	8.07E-03	CYP4V2	8.00E-03
CENPK	8.49E-03	TIPIN	8.04E-03	E2F8	7.99E-03
BLM	8.43E-03	TOP2A	8.04E-03	KIAA1524	7.99E-03
BUB1B	8.42E-03	KIF20A	8.04E-03	PRUNE2	7.99E-03
MCM3	8.41E-03	NEIL3	8.04E-03	ME1	7.99E-03
NDC80	8.38E-03	OIP5	8.03E-03	CDCA2	7.99E-03
SPC24	8.36E-03	MYBL2	8.03E-03	MND1	7.99E-03
KNL1	8.35E-03	SPAG5	8.03E-03	DEK	7.98E-03
DGKB	8.29E-03	CCNB2	8.02E-03	CD47	7.98E-03
NMU	8.22E-03	HIST1H1C	8.02E-03	PRIM1	7.98E-03
FANCD2	8.18E-03	KIF4A	8.02E-03	MELK	7.98E-03
TTK	8.17E-03	BRIP1	8.02E-03	RAD51AP1	7.97E-03
CCNA2	8.17E-03	BIRC5	8.01E-03	DLGAP5	7.97E-03
RFC4	8.14E-03	ZNF281	8.01E-03	SQRDL	7.96E-03
ASPM	8.12E-03	KIF14	8.01E-03	NEMP1	7.96E-03
CDCA5	8.10E-03	GGH	8.01E-03	SOWAHC	7.96E-03
FANCI	8.10E-03	NCAPG	8.01E-03	CENPV	7.96E-03
HJURP	8.08E-03	IKBIP	8.01E-03	MNS1	7.96E-03
PLPP2	8.08E-03	HMMR	8.01E-03	SLC7A11	7.95E-03
BAG2	8.08E-03	RRM2	8.00E-03	TIGAR	7.95E-03

Appendix V: ORA results for five bicluster stacks with length = 3

Stack Label	CI width for effect size estimate	Gene Ontology ID	Pathway Name	Number of genes matched	Enrich ment score	ORA p-value	FDR
ProsBicSta17	4.16	0019695	choline metabolic process	14	4.43	7.32E-05	3.79E-01
	4.16	0034613	cellular protein localization	6	0.77	1.10E-04	3.79E-01
	4.16	0070727	cellular macromolecule localization	5	0.53	1.74E-04	3.79E-01
	4.16	0006401	RNA catabolic process	2	0.02	1.78E-04	3.79E-01
	4.16	0040023	establishment of nucleus localization	15	5.66	2.76E-04	3.84E-01
ProsBicSta18	9.54	0009725	response to hormone	5	0.54	1.92E-04	1.00E+00
	9.54	0032870	cellular response to hormone stimulus	3	0.13	3.19E-04	1.00E+00
	9.54	0071396	cellular response to lipid	2	0.06	1.55E-03	1.00E+00
	9.54	1900426	positive regulation of defense response to bacterium	2	0.06	1.66E-03	1.00E+00
	9.54	1901701	cellular response to oxygen-containing compound	3	0.25	2.03E-03	1.00E+00
ProsBicSta19	3.94	0008610	lipid biosynthetic process	15	3.17	4.84E-07	4.13E-03
	3.94	0006629	lipid metabolic process	21	6.55	1.19E-06	4.39E-03
	3.94	0044255	cellular lipid metabolic process	18	5.16	2.53E-06	4.39E-03
	3.94	0044711	single-organism biosynthetic process	21	7	3.49E-06	4.39E-03
	3.94	0019367	fatty acid elongation, saturated fatty acid	3	0.03	4.11E-06	4.39E-03
ProsBicSta20	6.33	0046951	ketone body biosynthetic process	14	2.59	2.86E-07	2.44E-03
	6.33	0033148	positive regulation of intracellular estrogen receptor signaling pathway	12	2.36	4.11E-06	1.15E-02
	6.33	0035404	histone-serine phosphorylation	14	3.32	5.27E-06	1.15E-02
	6.33	0046950	cellular ketone body metabolic process	14	3.32	5.38E-06	1.15E-02
	6.33	1902224	ketone body metabolic process	9	1.54	2.34E-05	4.01E-02
ProsBicSta21	4.79	0097052	L-kynurenine metabolic process	8	1.09	1.33E-05	1.07E-01
	4.79	1902946	protein localization to early endosome	20	7.26	3.11E-05	1.07E-01
	4.79	0010737	protein kinase A signaling	15	4.46	3.76E-05	1.07E-01
	4.79	0016482	cytosolic transport	6	0.69	6.87E-05	1.46E-01
	4.79	0008219	cell death	8	1.43	9.46E-05	1.46E-01

Appendix VI: ORA results for five bicluster stacks with length = 5

Stack Label	CI width for effect size estimate	Gene Ontology ID	Pathway Name	Number of genes matched	Enrich ment score	ORA p-value	FDR
ProsBicSta12	9.17	0042127	regulation of cell proliferation	2	0.03	2.78E-04	1.00E+00
	9.17	0030522	intracellular receptor signaling pathway	13	4.51	3.40E-04	1.00E+00
	9.17	0030278	regulation of ossification	13	4.54	3.66E-04	1.00E+00
	9.17	0050861	positive regulation of B cell receptor signaling pathway	5	0.7	6.41E-04	1.00E+00
	9.17	0008283	cell proliferation	2	0.04	8.03E-04	1.00E+00
ProsBicSta13	5.13	0030522	intracellular receptor signaling pathway	14	4.45	1.26E-04	1.00E+00
	5.13	0030521	androgen receptor signaling pathway	11	3.25	3.88E-04	1.00E+00
	5.13	0035116	embryonic hindlimb morphogenesis	10	2.77	4.38E-04	1.00E+00
	5.13	0045923	positive regulation of fatty acid metabolic process	2	0.04	7.76E-04	1.00E+00
	5.13	0001838	embryonic epithelial tube formation	13	4.95	1.23E-03	1.00E+00
ProsBicSta14	11.66	0032870	cellular response to hormone stimulus	15	3.17	4.84E-07	4.13E-03
	11.66	0007050	cell cycle arrest	21	6.55	1.19E-06	4.39E-03
	11.66	0071375	cellular response to peptide hormone stimulus	18	5.16	2.53E-06	4.39E-03
	11.66	0048009	insulin-like growth factor receptor signaling pathway	21	7	3.49E-06	4.39E-03
	11.66	1901653	cellular response to peptide	3	0.03	4.11E-06	4.39E-03
ProsBicSta15	8.08	0042326	negative regulation of phosphorylation	2	0.03	3.70E-04	1.00E+00
	8.08	0001933	negative regulation of protein phosphorylation	2	0.04	5.91E-04	1.00E+00
	8.08	0045936	negative regulation of phosphate metabolic process	2	0.04	5.91E-04	1.00E+00
	8.08	0010563	negative regulation of phosphorus metabolic process	2	0.04	5.91E-04	1.00E+00
	8.08	0033673	negative regulation of kinase activity	2	0.04	7.21E-04	1.00E+00
ProsBicSta16	7.10	0006665	sphingolipid metabolic process	2	0.02	2.25E-04	6.93E-01
	7.10	0044255	cellular lipid metabolic process	2	0.02	2.25E-04	6.93E-01
	7.10	0008610	lipid biosynthetic process	3	0.13	3.19E-04	6.93E-01
	7.10	0030148	sphingolipid biosynthetic process	5	0.66	5.26E-04	6.93E-01
	7.10	0006643	membrane lipid metabolic process	20	9.19	5.52E-04	6.93E-01

Appendix VII: ORA results for five bicluster stacks with length = 6

Stack Label	CI width for effect size estimate	Gene Ontology ID	Pathway Name	Number of genes matched	Enrich ment score	ORA p-value	FDR
ProsBicSta07	6.03	0046890	regulation of lipid biosynthetic process	7	0.88	2.88E-05	1.28E-01
	6.03	0051094	positive regulation of developmental process	21	7.87	2.99E-05	1.28E-01
	6.03	0046889	positive regulation of lipid biosynthetic process	5	0.42	5.95E-05	1.69E-01
	6.03	0019216	regulation of lipid metabolic process	9	1.9	1.21E-04	2.59E-01
	6.03	1901701	cellular response to oxygen-containing compound	17	6.32	1.72E-04	2.94E-01
ProsBicSta08	17.28	0009725	response to hormone	13	4.25	2.86E-04	8.01E-01
	17.28	0042326	negative regulation of phosphorylation	9	2.15	2.90E-04	8.01E-01
	17.28	0015837	amine transport	4	0.4	7.09E-04	8.01E-01
	17.28	0001933	negative regulation of protein phosphorylation	8	1.96	7.58E-04	8.01E-01
	17.28	1901701	cellular response to oxygen-containing compound	13	4.72	7.81E-04	8.01E-01
ProsBicSta09	10.60	0098656	anion transmembrane transport	8	1.33	5.57E-05	2.56E-01
	10.60	0034349	glial cell apoptotic process	3	0.08	6.00E-05	2.56E-01
	10.60	0006470	protein dephosphorylation	7	1.3	3.17E-04	6.81E-01
	10.60	0061035	regulation of cartilage development	4	0.33	3.28E-04	6.81E-01
	10.60	0006970	response to osmotic stress	4	0.35	4.15E-04	6.81E-01
ProsBicSta10	15.97	0061035	regulation of cartilage development	4	0.29	2.13E-04	6.46E-01
	15.97	0061036	positive regulation of cartilage development	3	0.14	3.64E-04	6.46E-01
	15.97	0050861	positive regulation of B cell receptor signaling pathway	2	0.03	4.46E-04	6.46E-01
	15.97	0031666	positive regulation of lipopolysaccharide-mediated signaling pathway	2	0.04	5.93E-04	6.46E-01
	15.97	0034350	regulation of glial cell apoptotic process	2	0.04	5.93E-04	6.46E-01
ProsBicSta11	10.56	2000026	regulation of multicellular organismal development	26	11.54	5.54E-05	4.73E-01
	10.56	0042127	regulation of cell proliferation	23	10.34	1.98E-04	4.89E-01
	10.56	0008283	cell proliferation	27	13.21	2.03E-04	4.89E-01
	10.56	0009888	tissue development	25	11.95	2.57E-04	4.89E-01
	10.56	0051094	positive regulation of developmental process	19	7.94	3.13E-04	4.89E-01