

Integrative Genomic Modeling of Complex Traits using Pathway Analysis

by

Brian D. Bennett

Graduate Program in Computational Biology and Bioinformatics
Duke University

Date: _____

Approved:

Sayan Mukherjee, Supervisor

Terrence S. Furey

Elizabeth Hauser

Joseph E. Lucas

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor of Philosophy in the
Graduate Program in Computational Biology and Bioinformatics
in the Graduate School
of Duke University

2012

ABSTRACT

Integrative Genomic Modeling of Complex Traits using Pathway Analysis

by

Brian D. Bennett

Graduate Program in Computational Biology and Bioinformatics
Duke University

Date: _____

Approved:

Sayan Mukherjee, Supervisor

Terrence S. Furey

Elizabeth Hauser

Joseph E. Lucas

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree of Doctor of Philosophy in the
Graduate Program in Computational Biology and Bioinformatics
in the Graduate School
of Duke University

2012

Copyright by
Brian D. Bennett
2012

Abstract

Understanding the root molecular causes driving complex traits is a fundamental challenge in genomics and genetics. Numerous studies have used variation in gene expression to understand complex traits, but the underlying genomic variation that contributes to these expression changes is not well understood. The overall goal of this work is to develop an integrative framework to better understand the genetic and molecular causes of complex traits, including complex diseases. In this work, I present a computational framework that I developed to integrate gene expression and other genomic data to identify biological differences between samples from opposing complex trait classes that are driven by expression changes and genomic variation. This framework combines analysis on the multi-gene biological pathway level with multi-task learning to build predictive models that also uncover pathways potentially relevant to the complex trait of interest. To validate this framework, I first performed a simulation study to test its predictive ability and to measure how well it uncovered pathways that contain genes that are both differentially expressed and genetically associated with a complex trait. The predictive performance of the multi-task model was found to be comparable to other similar methods. Also, multi-task learning, along with other methods that jointly considered pathway scores from both data sets, was able to better identify pathways with both genetic and expression differences related to the

phenotype. I applied this framework to gene expression and genotype data from estrogen receptor (ER) positive and ER negative breast cancer samples. The top 15 predictive pathways from the multi-task model were all related to estrogen, steroids, cell signaling, or the cell cycle. The results from both the simulation studies and the breast cancer analysis suggest that this multi-task framework is useful for both identifying biologically relevant pathways associated with a phenotype across multiple data types while also retaining similar predictive performance as other similar methods.

Contents

Abstract	iv
List of Tables	ix
List of Figures	x
Acknowledgments	xi
1. Introduction	1
1.1 Integrative Analysis	1
1.2 Gene Set Analysis	4
1.3 Multi-Task Learning.....	9
1.4 Summary.....	13
2. Integrative Framework.....	15
2.1 ASSESS	17
2.1.1 Overview	18
2.1.2 Extension for Analysis of Genotype Data	21
2.2 Multi-Task	23
2.2.1 Single-Task SVM.....	24
2.2.2 Multi-Task SVM.....	25
2.2.3 Comparison Methods	27
2.3 Matched Data Methods	29
2.3.1 Summed Prediction Model	30
2.3.2 Summed Enrichment Model.....	30

2.3.3 Merged Model.....	31
2.4 Input Data.....	31
3. Validation Studies.....	34
3.1 Simulation Study	34
3.1.1 Introduction.....	34
3.1.2 Simulating Genotype Data.....	35
3.1.3 Simulating Expression Data.....	37
3.1.4 Generating a Simulated Data Set	38
3.1.5 Varying Similarity of Tasks.....	38
3.1.6 Varying Number of Samples	42
3.1.7 Varying Number of Tasks	44
3.1.8 Finding Pathways Enriched Across Multiple Data Types.....	47
3.1.9 Discussion.....	50
3.2 Breast Cancer Study	50
3.2.1 Introduction.....	50
3.2.2 Analysis	52
3.2.3 Discussion.....	58
4. Additional Gene Set Studies.....	59
4.1 Thirsty Mouse Study	59
4.1.1 Introduction.....	59
4.1.2 Analysis	61
4.1.3 Discussion.....	67

4.2 Sarcoma Mouse Model Study	67
4.2.1 Introduction.....	67
4.2.2 Analysis	68
4.2.3 Discussion.....	70
5. Conclusion	71
References	74
Biography.....	81

List of Tables

Table 1: Gene Set Types.....	39
Table 2: Scenarios with Varying Similarity between Tasks	40
Table 3: Performance of Expression Data with Varying Levels of Similarity	41
Table 4: Performance of Genotype Data with Varying Levels of Similarity	41
Table 5: Performance of Matched Data Models with Varying Levels of Similarity	41
Table 6: Performance of Expression Data with Varying Sample Sizes.....	43
Table 7: Performance of Genotype Data with Varying Sample Sizes.....	43
Table 8: Performance of Matched Data Models with Varying Sample Sizes	44
Table 9: Average Rank of Target Gene Set	49
Table 10: Correct Predictions for Breast Cancer Expression Data.....	53
Table 11: Correct Predictions for Breast Cancer Genotype Data.....	53
Table 12: Correct Predictions using Matched Data Models for Breast Cancer Data	53
Table 13: Top Gene Sets in Breast Cancer Analysis	55
Table 14: Top Gene Sets in Thirsty Mice Analysis	65
Table 15: Addition Gene Sets in Thirsty Mice Analysis.....	65
Table 16: Gene Set Enrichment Analysis for Human Sarcoma.....	70

List of Figures

Figure 1: Overview of the multi-task pipeline	16
Figure 2: Overview of the single-task pipeline	28
Figure 3: Overview of the concatenated data pipeline	29
Figure 4: Performance of Expression Data with Varying Number of Tasks	46
Figure 5: Performance of Genotype Data with Varying Number of Tasks	46
Figure 6: Principle Component Analysis of Thirsty Mice Data.....	63
Figure 7: Enrichment Profile of Opioid Signaling Pathway in Thirsty Mice Analysis	67

Acknowledgments

First, I want to thank my primary advisor, Terry Furey, because your countless hours of help and dedication to making sure that I succeed has allowed me to learn and grow so much while under your mentorship. I also want to thank my co-advisor, Sayan Mukherjee, for your many ideas and assistance with my project. In addition, I want to thank my other committee members, Beth Hauser and Joe Lucas, for your advice and guidance as I progressed. Next, I want to thank my fellow friends and colleagues, both at Duke and UNC, for making my time in the program more enjoyable. I want to thank the Furey lab for your feedback on my project, and Qing Xiong for providing simulated data for my project. Finally, I want to thank my family for your support ever since I was a little child, especially my mother and father. Most of all, I want to thank my wife, Lorena, because without your love and encouragement I would have never been able to achieve what I have.

1. Introduction

A fundamental challenge in genomics is discovering and understanding the molecular and genetic basis of complex traits. A deeper understanding of complex traits will potentially lead to a better diagnosis and treatment of complex diseases. Although variation in gene expression often directly influences phenotypic differences in complex traits [67], additional underlying genomic factors may be driving these expression changes [1,9,52]. Also, many of these genomic differences contribute to phenotypic variation by influencing biological pathways [12-15]. Therefore, an integrative pathway-level analysis may provide a better model of complex traits. The overall goal of this work is to develop an integrative pathway-based framework to better understand the genetic and molecular causes of complex traits, including complex diseases. To achieve this goal, I created a framework to integrate different types of genomic data into a predictive model that utilizes gene set enrichment analysis and multi-task learning.

1.1 Integrative Analysis

Many studies have used gene expression assays to model, at a molecular level, direct influences driving phenotypic variation [67]. However, gene expression differences may be driven by other underlying genomic and environmental factors, including genetic variation [1,10,53,54], copy number alterations [4,8,9,11], and DNA methylation changes [7,11,52]. Therefore, results from a gene expression analysis alone may not provide a complete account of all of the molecular forces driving phenotypic

differences. To overcome this, several previous efforts have explored integrating different genomic data types for samples belonging to different phenotypic classes [2-11]. These studies aimed to uncover genes that influence phenotype with any combination of differential gene expression and other genomic variation. Genes showing both expression and other genomic differences suggest that transcriptional variation may be driven in part by underlying genomic variation. Results from an integrative analysis may provide additional insight into the underlying molecular factors that are driving expression changes and phenotypic differences, and an integrative analysis may strengthen findings if significant differences are discovered across several data types.

Many studies have focused on combining gene expression and DNA copy number data to discover genes that are driven by copy number alterations [2-6]. For example, one study modeled gene expression based on the copy number variation for genes on the same chromosome arm in order to maximize power in finding genes associated with tumorigenesis in breast cancer [2]. Others looked for regions of interest with high copy number alterations that may be driving expression changes and then searched for important genes within these regions [3-6]. In one study focused on finding chromosomal abnormalities that cause colorectal cancer, they calculated fold changes in expression and copy number data between normal and diseased samples, ordered all probes based on chromosomal location, and then searched for large chromosomal segments showing coordinated expression and copy number changes [3]. This analysis

revealed many regions with copy number gain or loss along with differential expression of genes in the region, and they identified several candidate genes within these regions of interest for further study.

Other studies searched for significant differences between phenotypic classes within multiple genomic data types for individual genes [7-9]. For example, one study integrated gene expression, copy number, DNA methylation, and loss of heterozygosity (LOH) data searching for genes associated with breast cancer [7]. They looked for genes that had significant changes in all of these data types when compared to normal. This analysis revealed that *ERBB2*, an important breast cancer gene, simultaneously showed amplification, loss of heterozygosity, loss of methylation, and a drastic increase in gene expression.

Genome-wide association (GWA) studies examine a dense set of single-nucleotide polymorphisms (SNPs) distributed across the entire genome, searching for loci that are associated with disease [1]. However, it is not always clear which gene the locus is functionally affecting [1]. To address this issue, expression quantitative trait loci (eQTL) mapping examines gene expression patterns for samples that have been genotyped to find expression changes driven by genotype differences [68]. Gene expression data and genotype data can be integrated to discover genes with gene expression differences associated with complex traits that are driven by genetic variation. One study selected cancer-associated genes whose expression profiles are

known to predict treatment outcome and looked for genotype patterns within eQTLs associated with these genes [10]. They created a model in which expression profiles and genotype patterns for selected genes were combined and used to predict the treatment success of prostate and breast cancer patients.

All of these studies suggest that an integrative approach may be beneficial for discovering and utilizing gene expression differences that are driven by underlying genomic variation. However, most previous integrative studies performed sequential or independent analyses of each data type. My framework differs from these approaches in that several genome-wide data sets are simultaneously used to build the final predictive model. This eliminates the restriction of using results from one independent analysis to filter results from the other, and instead allows the model to equally and simultaneously consider data from each experiment. One advantage of this approach is that my framework simultaneously considers any combination of gene expression differences and other genomic variation, whereas other methods only consider patterns that fit the model of that particular study. Also, my framework provides a general model that is easily applicable to the integration of many different genomic data types, whereas some of the other methods require a fundamentally different model for each data type.

1.2 Gene Set Analysis

Gene set analysis explores biological data in the context of pathways. This approach examines simultaneous changes in multiple genes in contrast to single-gene

analysis, which searches for differences in individual genes. Pathway analysis explores altered biological pathways where the function of the pathway is disrupted by changes in multiple genes involved with the pathway. Although many integrative studies focus on individual genes [2,4-7,9,10], there are many advantages to performing pathway-level analysis.

One advantage is that a pathway-level model may be more biologically appropriate than a single-gene model. For complex traits, many phenotypic differences are associated with perturbations in specific pathways [12], and there may be several different genes that can be disrupted to alter a pathway [13]. As a result, there are often many different expression patterns or genotype differences that express the same phenotype [15]. When comparing two phenotypes, there may be a significant disturbance in a pathway among many samples, even though the genes causing the disruption may be very different from sample to sample [13]. Also, pathways may be highly altered by consistent but modest changes in many genes involved with the pathway [14]. For example, one study identified a set of genes involved with oxidative phosphorylation, where 89% of the genes were coordinately downregulated in human diabetes, but with a very modest average decrease of about 20% [14]. In these scenarios with phenotypes driven by pathway disruptions, a single-gene analysis may not have the power to discover any individual genes that are associated with phenotype, but a gene set analysis may reveal pathways that are highly altered.

In addition to providing a model that may be more biologically appropriate, pathway analysis also provides other practical advantages. First, gene set enrichment analysis can improve the interpretability of results [23]. Many single-gene analyses produce a large list of phenotypically associated genes. It is often difficult to determine which of these genes are biologically relevant and how these genes relate to each other. A small list of altered pathways is often easier to interpret than a large list of associated genes [23]. Also, pathway analysis provides results that are highly reproducible between studies [16,17]. One study examined three groups that independently assembled lists of genes associated with prostate tumors when compared to normal tissue, and there was a very small overlap of only 6% of genes in common among the three lists [17]. However, there was a surprisingly large overlap of pathways in common among independent gene set analyses for the same phenotype [17]. In addition to this, pathway-based data may be better at predicting phenotype compared to single-gene data when building predictive models for biological data [22].

All of these advantages suggest that a pathway-based analysis may be better than a single-gene analysis, and some studies have performed a pathway-level integrative analysis [3,8,11]. For example, one study integrated gene expression and somatic mutation data to identify pathways frequently altered in prostate cancer [3]. A single tumor was considered to have an altered pathway if one or more genes in the pathway had a somatic mutation or had an expression level that was significantly

different than in normal prostate. Pathways altered in a large percentage of the samples were considered frequently altered. This study identified three well-known cancer pathways as frequently altered: PI3K, RAS/RAF, and RB.

Many methods have been developed to analyze genomic data on the gene set level [16,18-21]. Fisher's Exact test (FE) [18] is one of the oldest methods, and it calculates a significance value for each gene set based on the proportion of differentially expressed genes in the gene set compared to the proportion in the entire data set. Random-Set methods (RS) [19] is a computationally efficient method that calculates an associative statistic for each gene based on the correlation of the expression data for that gene with phenotype and determines enrichment based on the average associative statistic for genes in a gene set. Gene Set Enrichment Analysis (GSEA) [16] is a popular method that also uses an associative statistic for each gene but with a variation of the Kolmogorov-Smirnov statistic [69] to calculate enrichment. Gene List Analysis with Prediction Accuracy (GLAPA) [20] determines enrichment by examining the ability of genes in a gene set to predict phenotype. One study compared the performance of these four methods [51]. This study found that FE performed the worst, with poor sensitivity and a lack of power, whereas the other three methods performed about the same. They found GLAPA to be more conservative and required larger differences between the phenotypes. They observed that RS performed better than GSEA at discovering gene

sets with a mixture of upregulated and downregulated genes. However, they found that GSEA was overall the most consistent.

There are also several methods for pathway analysis of genotype data [21,40]. GSEA-SNP [21] assembles SNP sets for each gene set that contains all SNPs mapped to the genes in a gene set and calculates enrichment scores similar to GSEA using these SNP sets. Gene Set Association Analysis (GSAA) [40] is an integrative method that computes a combined associative statistic for each gene from both gene expression data and genotype data together and calculates enrichment scores similar to GSEA using these combined associative statistics. In general, all of these gene set enrichment methods provide a single measure of enrichment for each gene set across all samples.

All of these findings suggest that an integrative pathway-level analysis may be beneficial. In my framework, I use an integrative supervised learning method to build a pathway-level predictive model, but sample-specific information is required for predictive modeling using supervised learning. ASSESS (Analysis of Sample Set Enrichment ScoreS) [22] is a gene set enrichment analysis method that extends GSEA to provide a measure of pathway enrichment for each sample. To do this, it calculates a separate class-membership likelihood statistic for each sample and gene, instead of an aggregate associative statistic for all samples for that gene. My framework uses ASSESS to produce sample-specific gene set enrichment scores, which can be used to build pathway-level predictive models.

An additional advantage of using ASSESS is providing a method to more easily integrate data types that may have very different structure. For example, expression data consists of gene-based continuous values, whereas genotype data consists of discrete SNP-based genotypes. There is no straightforward method to directly combine data like this that has very different structure. By first obtaining gene set enrichment scores for each data type, this acts as a normalization step to allow each data type to be easily combined with other data types.

1.3 Multi-Task Learning

Machine learning is a type of analysis that computationally learns models from data that can be used for pattern recognition, classification, and prediction [70]. Supervised learning is a type of machine learning that uses data for samples that have been previously classified based on a known characteristic to learn a model that can use new data to predict this characteristic for samples that have not been classified [70]. In biology, one application of supervised learning is building a model using biological data with samples having a known phenotypic class and using this model to predict the phenotypic class of new samples with unknown phenotype. Features in this model that are important for prediction may be biologically relevant to the phenotypic class.

A Support Vector Machine (SVM) [27] is a popular type of supervised learning method and has been shown to be useful in the analysis of genomic data [72]. A basic linear SVM works by mapping training samples with binary class assignments into

high-dimensional Euclidean space based on the data for those samples and finding the linear hyperplane in this space that best separates the samples by class. A new sample can be mapped into this space, and the model can make a class prediction based on which side of the hyperplane the new sample is mapped to. Often, only a subset of training samples that are closest to the hyperplane will be used in the predictive model and are called the support vectors [72]. This is a property that is useful for removing outlier samples in biological data.

An SVM learns the predictive model using a kernel, which is a function that gives a measure for two samples that is related to the Euclidean distance between the samples. This allows the SVM to build the model using only information about the distances between samples, and does not need to make calculations based on the original data. This provides a computational advantage for data sets with a large number of features [71]. This is especially useful for biological data, which often has high dimensionality, such as gene expression data for many thousands of genes. Also, some supervised learning techniques suffer from scenarios where no unique solution exists when the number of features is greater than the number of samples. However, kernel methods allow for a unique solution in these instances by basing the solution on the distance between samples and not on the individual features [71]. An additional advantage of using a kernel is that an SVM can also perform non-linear classification by modifying the kernel in what is known as the kernel trick. The kernel can be modified to

transform the original data into an alternate feature space before learning the predictive model. The SVM then finds the linear hyperplane that best separates the samples in this transformed space, which results in a non-linear separation of the data in the original space. This allows data that may not have meaningful linear structure in the original space to be transformed into a space where the data has linear structure.

Multi-task learning [24] is a supervised learning approach to building predictive models from data that contain complementary information, and regularized multi-task learning (RML) [25] is an SVM implementation of multi-task learning. While other supervised learning methods perform well when there is a single data type, studies have shown an improved performance in predictive accuracy in some instances when simultaneously building multiple models from data with related information [24-26]. My framework uses multi-task learning to build predictive models using the sample-specific enrichment scores from ASSESS for different data types. Multi-task learning provides a way to integrate these data types as different tasks in the model. Multi-task learning aims to take advantage of data with similar information between tasks while also incorporating information unique to each task. In the context of my pathway-based multi-task framework, similar information means similar pathway enrichment among data types, whereas different information means pathway enrichment that is unique to a data type. My multi-task framework aims to identify and take advantage of pathways

with similar enrichment among data types, while also preserving properties that are unique to each data type.

In order to measure the ability of multi-task learning to simultaneously utilize similar and different pathway enrichment properties, I compared multi-task learning to single-task learning and a concatenated data learning model. Single-task learning independently builds separate models for each task and does not consider whether there is similar or different information between tasks. A concatenated data model combines all data together into a single data set containing all samples from all tasks to take advantage of all information together, but it does not distinguish which task the information originated from. Multi-task learning builds a model that attempts to take advantage of the strengths of both single-task learning and concatenated data models. It does this by calculating a common effect shared among all tasks, similar to a concatenated data model. At the same time, it determines a task-specific effect that is unique to each task, similar to a single-task model. Successful multi-task learning models should show an improvement in predictive performance when compared to a single-task model and a concatenated data model.

The model produced by my multi-task framework can also be explored to discover relevant pathways. Pathways that are important for prediction in the model may also be biologically relevant to the phenotype. Pathways with a large common effect indicate that there is a large amount of enrichment in both data types combined.

When integrating gene expression data with other genomic data types, these pathways suggest that underlying genomic variation may be driving expression changes in the pathway. Also, examining the task-specific effects for these pathways can provide a measure of the contribution that each data type has in the significance of the pathway.

1.4 Summary

I have developed a novel pathway-based framework that integrates different types of genomic data and produces a predictive model that can also identify biologically relevant pathways. The overall goal of this framework is to use the predictive model to better understand the root molecular causes of complex traits. This framework consists of two key steps, sample-specific pathway analysis using ASSESS and supervised learning to build a predictive model using regularized multi-task learning. In chapter 2, I discuss this framework in more detail. First, I describe ASSESS and an extension to ASSESS for processing genotype data. I then describe the multi-task SVM and other comparison methods. Next, I discuss extensions to my framework for analysis of matched data. Finally, I discuss the required input data that must be provided to my framework. In chapter 3, I explore the usefulness of my multi-task framework. I first describe several simulation analyses that compared multi-task learning with other methods. Next, I discuss a breast cancer analysis that examined the ability of my framework to discover relevant pathways. In chapter 4, I describe two gene set analyses that explore specific biological questions. First, I explain a study that

explored whether addiction-relation gene sets that are known to be associated with sodium depletion in mice are also associated with thirst in mice. I next describe a study that examined whether a new mouse model of rhabdomyosarcoma (RMS) is similar to human RMS. In chapter 5, I summarize my findings and discuss some future directions for my framework.

2. Integrative Framework

In order to explore the underlying genomic variation driving phenotypic differences and better understand the causes of complex traits, I developed an integrative pathway-based framework. An overview of the analysis pipeline for my framework, using gene expression and genotype data as an example, is presented in Figure 1. The two key steps are first a sample-based analysis on the pathway level using ASSESS (Figure 1b-1c), and second the integration of genomic data into a predictive model using a multi-task SVM (Figure 1d). In this chapter, I first describe ASSESS and discuss an extension that I created for processing genotype data. I then describe the multi-task SVM along with other comparison methods. Next, I describe some extensions to my framework for utilizing matched data. Finally, I discuss the required data that must be provided as input to my framework.

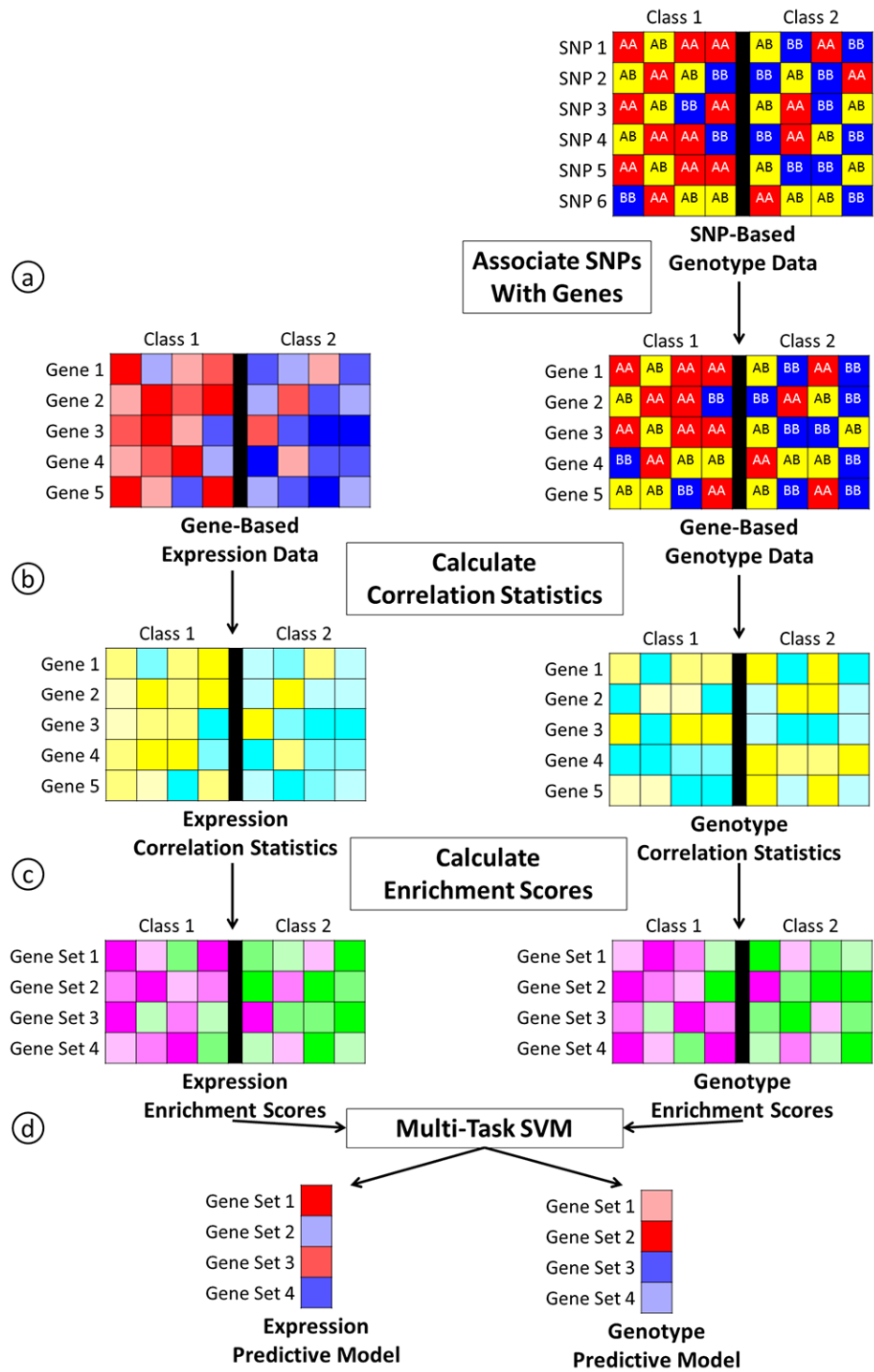


Figure 1: Overview of the multi-task pipeline

2.1 ASSESS

ASSESS (Analysis of Sample Set Enrichment Scores) [22] is a sample-specific gene set enrichment analysis software package. Most gene set enrichment methods provide an aggregate measure of enrichment across all samples for a gene set. However, obtaining pathways enrichment measurements within individual samples may be useful for many biological analyses, such as clustering to find subtypes of cancer or building a predictive model using machine learning. To address this, ASSESS was developed to extend Gene Set Enrichment Analysis (GSEA) [16] and obtain sample-specific gene set enrichment scores for gene expression data. Studies have successfully used ASSESS for several gene expression analyses, such as improved clustering of leukemia subtypes [22] and modeling the progression of prostate cancer on the pathway level [73]. In my framework, I used ASSESS to obtain sample-specific pathway-level information for building a predictive model using supervised learning. One novel application of ASSESS in my framework is obtaining sample-specific enrichment scores for data that is not gene expression, and processing genotype data required an extension to ASSESS.

In step one of my framework, ASSESS takes as input gene-based genomic data for samples belonging to one of two phenotypic classes. It then produces sample-specific enrichment scores for a collection of gene sets. To do this, it first calculates a correlation statistic for each sample and gene (Figure 1b). Then, it ranks all genes based on this

correlation statistic for each sample and uses gene set enrichment analysis to determine the enrichment of pathways within samples (Figure 1c).

2.1.1 Overview

ASSESS takes gene-based genomic data along with phenotype information and provides a measure of the variation of gene set enrichment over all samples for a given gene set. First, ASSESS computes a correlation statistic for each sample and gene that represents the degree to which the gene-based data matches the summary profile of one phenotypic class compared to the other for that gene. It calculates this statistic by determining the log likelihood ratio of the data belonging to one class over the other given the background profile of data for both classes as

$$c_j^i = \log \left(\frac{P(x_j^i \in C_1 | x_j^i, \{x_j^1, \dots, x_j^n\})}{P(x_j^i \in C_2 | x_j^i, \{x_j^1, \dots, x_j^n\})} \right)$$

where x_j^i is data for the i -th sample and j -th gene, and C_1 and C_2 are the two phenotypic classes. The metric for calculating this statistic will vary based on the input data.

The original implementation of ASSESS was designed for use with gene expression data and provides two metrics. The first uses a parametric model to model the data as two normal distributions, one for each phenotypic class. This metric is appropriate when the data is normally distributed. However, gene expression data is typically not normally distributed, so ASSESS provides a second metric that does not require normal data. This metric uses a nonparametric model that estimates the densities

of each class using a Parzen estimator [27] with a Gaussian kernel. This metric is more appropriate for continuous data that is not normally distributed, but it is not appropriate for discrete data, such as genotype data. To address this, I created a third metric for analysis of genotype data that is described in section 2.1.2.

Next, ASSESS independently sorts the correlation statistics for each sample and computes enrichment scores. These enrichment scores are interpreted as the degree to which the genes in a gene set associate the sample with one phenotype class compared to the other. Large enrichment scores indicate that many of the genes in the gene set strongly associate the sample with one class, indicated by the sign of the enrichment score. For each sample and gene set, it uses a tied-down Brownian bridge process [50] to perform a random walk and calculates a running sum statistic as

$$v_i^k(\ell) = \frac{\sum_{j=1}^{\ell} c_{(j)}^i I(g_{(j)}^i \in \gamma_k) - \sum_{j=1}^{\ell} I(g_{(j)}^i \notin \gamma_k)}{\sum_{j=1}^p c_{(j)}^i I(g_{(j)}^i \in \gamma_k) - p - |\gamma_k|}$$

where ℓ is the position along the walk, $g_{(j)}^i$ is the gene in the j -th position in the sorted list for the i -th sample, $c_{(j)}^i$ is the correlation statistic corresponding to this gene, $I(g_{(j)}^i \in \gamma_k)$ is an indicator function for whether gene $g_{(j)}^i$ is in gene set γ_k , and $|\gamma_k|$ is the number of genes in the k -th gene set. It computes the final enrichment as the maximum deviation from zero along the walk as

$$ES_i^k = v_i^k \left[\arg \max_{\ell=1, \dots, p} |v_i^k(\ell)| \right]$$

This deviation from zero is similar to the classical Kolmogorov-Smirnov static [50].

The original enrichment scores for a gene set cannot be directly compared to other enrichment scores for different gene sets, because differences in gene set size may alter the magnitude of the scores. ASSESS can normalize the enrichment scores to allow them to be directly comparable among all gene sets. To do this, it randomly shuffles the class assignments among all samples for many permutations, $\pi = 1, \dots, \Pi$, computes new correlation statistics for each permutation, $c_j^i(\pi)$, and uses these to obtain random

enrichment scores, $\{ES_i^k(\pi)\}_{\pi=1}^{\Pi}$. The normalized enrichment score is calculated as

$$NES_i^k = \frac{ES_i^k}{\left| \text{mean} \{ES_i^k(\pi)\}_{\pi=1}^{\Pi} \right|}$$

To measure significance, I compared the original enrichment scores to the corresponding empirical distribution of random enrichment scores. I calculated a nominal p-value as

$$PVAL(ES_i^k) = \text{Percentage of } [ES_i^k(\pi) \geq ES_i^k] \text{ for } \pi = 1, \dots, \Pi$$

To adjust for multiple comparisons, I computed a FWER adjusted p-value as

$$FWER(ES_i^k) = \text{Percentage of } \left\{ \max_{j=1, \dots, m} [ES_i^j(\pi)] \geq ES_i^k \right\} \text{ for } \pi = 1, \dots, \Pi$$

and a FDR q-value as

$$FDR(ES_i^k) = \frac{\text{Percentage of } [ES_i^j(\pi) \geq ES_i^k] \text{ for } \pi = 1, \dots, \Pi \text{ and } j = 1, \dots, m}{\text{Percentage of } [ES_i^j \geq ES_i^k] \text{ for } j = 1, \dots, m}$$

A notable difference between the original implementation of ASSESS and my implementation is in the calculation of the sign associated with each enrichment score. In the original implementation, the sign of each enrichment score corresponds to the overall direction of differential expression for the genes in the gene set, similar to GSEA. However, both GSEA and the original implementation of ASSESS perform poorly with gene sets that contain a mixture of significantly over-expressed and under-expressed genes [51]. In my implementation of ASSESS, the sign of each enrichment score corresponds to the phenotype profile that best matches the data for that sample and gene set. This allows my implementation of ASSESS to perform just as well with mixed gene sets, and it provides enrichment scores that are more suitable for building a predictive model.

2.1.2 Extension for Analysis of Genotype Data

There are several methods for calculating gene set enrichment scores for genotype data, such as GSEA-SNP [21], but they only provide a single enrichment score for all samples. It may be advantageous to obtain sample-specific enrichment scores for genotype data to build a predictive model or integrate this data type with other sample-specific pathway-based scores for other data types. To address this, I extended ASSESS to obtain sample-specific gene set enrichment scores for genotype data. The key

challenge in extending ASSESS for other data types is obtaining the correlation statistics for each sample and gene. The metrics included with ASSESS can only process gene-based continuous values, whereas genotype data contains SNP-based discrete values. To extend ASSESS for using genotype data, I first mapped the data to the gene level, and then I used a multinomial model to calculate the correlation statistics.

To map the SNP-based data to the gene level, my framework selects a representative SNP for each gene (Figure 1a). Some methods, such as GSEA-SNP, use all SNPs associated with a gene to calculate each enrichment score. However, enrichment scores calculated this way may be biased by differences in the number of SNPs within a gene and by SNPs that are in linkage disequilibrium with each other. In my framework, I eliminated this bias by selecting a single SNP to represent each gene. For all SNPs located within a predefined distance surrounding and including a gene, I performed a Pearson's chi-square test on each SNP to determine its correlation with phenotype and selected the SNP with the highest correlation. This causes the SNP that is most associated with the phenotype in the region surrounding and including the gene to be selected to represent that gene. I used the genotypes of the samples for this SNP as the gene-based data for that gene.

To obtain correlation statistics for the gene-based genotype data, I developed a novel metric. This metric follows the ASSESS model of determining the log likelihood ratio of the data given the background profiles of each class for that gene. I used a

multinomial model to calculate the ASSESS correlation statistic for each sample and gene as

$$c_j^i = \log \left(\frac{p_j^{g1}}{p_j^{g2}} \right)$$

where p_j^{g1} is the percentage of samples with the genotype of the i -th sample for the j -th gene in class 1, and p_j^{g2} is the percentage of samples with this genotype in class 2. If either class contained zero samples with a given genotype, I added a pseudo-count of 1. I then used these correlation statistics within ASSESS to obtain gene set enrichment scores.

2.2 Multi-Task

Regularized multi-task learning (RML) [25] is a supervised learning method for building predictive models for related data. When building multiple models, many supervised learning methods will learn the models independently. However, simultaneously building the models may be useful if the data is related. This may improve the predictive performance of the models and increase the relevancy of features that are important for prediction. To address this, RML was developed to extend a Support Vector Machine (SVM) [27] and produce models that are built simultaneously. To do this, it determines a common element of similarity among tasks and a task-specific element that retains the uniqueness of each task. Studies have successfully used multi-

task learning to increase predictive performance over single-task learning [24-26]. Also, one study previously used RML with ASSESS to model prostate tumor progression [73]. This study interpreted the common element as pathways important in all stages of progression and the task-specific elements as pathways important in individual stages of progression. In my framework, I used RML to simultaneously build predictive models for different data types.

In step one of my framework, ASSESS independently calculates sample-specific enrichment scores for each data type. In step two, I use the enrichment scores from ASSESS for each data type as the input tasks to a multi-task SVM (Figure 1d). Multi-task learning assumes that the samples among the different tasks are independent, and it does not require that the different data come from the same samples or that there are the same number of samples in each task. For comparison to multi-task, I also use methods based on a single-task SVM. Below, I first describe a single-task SVM, and then I describe how this is extended to create a multi-task SVM.

2.2.1 Single-Task SVM

To perform the predictive modeling step of my framework, I used an SVM software package called SVM-Light [39]. An SVM produces the predictive model using a kernel, which is a function used to obtain a measure related to the Euclidean distance between two samples. The SVM trains a predictive model by maximizing the standard SVM dual problem and calculating nonnegative Lagrange multipliers for each sample as

$$\max_{\alpha_i} \left[\sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i y_i \alpha_j y_j K(x_i, x_j) \right]$$

$$0 \leq \alpha_i \leq C$$

where C is a regularization parameter, n is the number of samples, y_i is the class assignment for the i -th sample, x_i is the data for the i -th sample, and $K(x_i, x_j)$ is the kernel function. For all analyses that used a single-task SVM, I used a standard linear kernel as

$$K(x, z) = x \cdot z$$

I used the sample weights obtained from SVM-Light to derive predictive weights for each gene set as

$$w = C \sum_{i=1}^n \alpha_i y_i x_i$$

2.2.2 Multi-Task SVM

To utilize a multi-task framework, I extended SVM-Light to implement RML. I obtained the source code for SVM-Light and modified the code to implement a multi-task kernel described below. The multi-task SVM trains a predictive model for each task simultaneously by maximizing the standard SVM dual problem and calculating nonnegative Lagrange multipliers for each sample and task as

$$\max_{\alpha_t^i} \left[\sum_{i=1}^n \sum_{t=1}^T \alpha_t^i - \frac{1}{2} \sum_{i=1}^n \sum_{s=1}^T \sum_{j=1}^n \sum_{t=1}^T \alpha_s^i y_s^i \alpha_t^j y_t^j K_{st}(x_s^i, x_t^j) \right]$$

$$0 \leq \alpha_t^i \leq C$$

where C is a regularization parameter, n is the number of samples, T is the number of tasks, y_t^i is the class assignment for the i -th sample in task t , x_t^i is the data for the i -th sample in task t , and $K_{st}(x_s^i, x_t^j)$ is the multi-task kernel function. For all analyses that used a multi-task SVM, I used a linear multi-task kernel [25] as

$$K_{st}(x, z) = \left(\frac{1}{\mu} + \delta_{st} \right) x \cdot z$$

where μ is a positive parameter that controls the relatedness of the models, and $\delta_{st} = 1$ if s and t belong to the same task, $\delta_{st} = 0$ otherwise. For all multi-task analyses in this work, I used a μ parameter value of 1. I used the sample weights obtained from SVM-Light to derive task-specific effects for each gene set and task as

$$v_t = C \sum_{i=1}^n \alpha_t^i y_t^i x_t^i$$

These weights are driven by enrichment that is unique to each task and can be interpreted as the predictive element of the model that maintains properties unique to each task. I used these weights to calculate common weights for each gene set as

$$w_0 = \frac{1}{\mu} \sum_{t=1}^T v_t$$

The common weights are driven by similar enrichment that is common to all tasks and can be interpreted as the predictive element of the model that enhances properties

similar in all tasks. I summed the common weights with the task-specific effects for each gene set and task to compute the final predictive weights as

$$w_t = w_o + v_t$$

2.2.3 Comparison Methods

Multi-task learning takes advantage of data with a mixture of similar and unique information between tasks. Therefore, multi-task learning should have an improved predictive accuracy when compared to a model that only considers similar information or a model that only considers unique information. I developed two additional methods to compare to multi-task and determine the advantage that multi-task learning is providing.

The first is single-task learning that only considers information that is unique to each task. It independently builds the predictive models for each task and cannot take advantage of similar information. I performed single-task learning by independently using the enrichment scores from ASSESS with a single-task SVM to build separate predictive models for each data type (Figure 2). The second is a concatenated data model that only considers information that is similar to all tasks. It concatenates the data into a single data set and cannot differentiate which task the data originally came from. I built a concatenated data model by combining all of the enrichment scores for all data types together into a single data set and used a single-task SVM to build a single predictive model from this concatenated data set (Figure 3). Since multi-task learning builds a

model that has an element based on similar information while also retaining a unique element, multi-task learning may have an increased predictive performance when compared to single-task learning or a concatenated data model if the data for each data type contains related information.

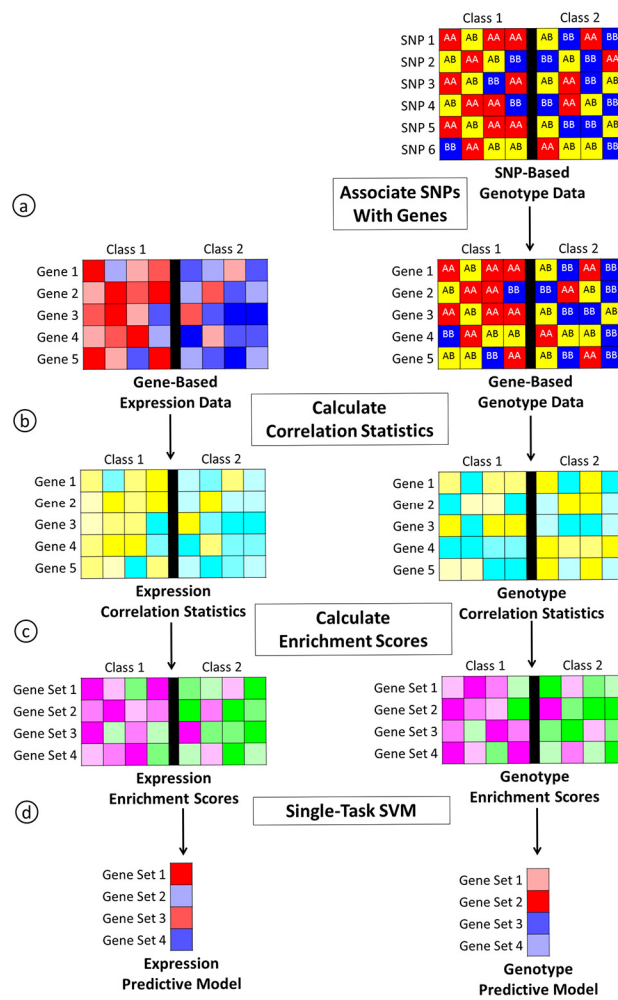


Figure 2: Overview of the single-task pipeline

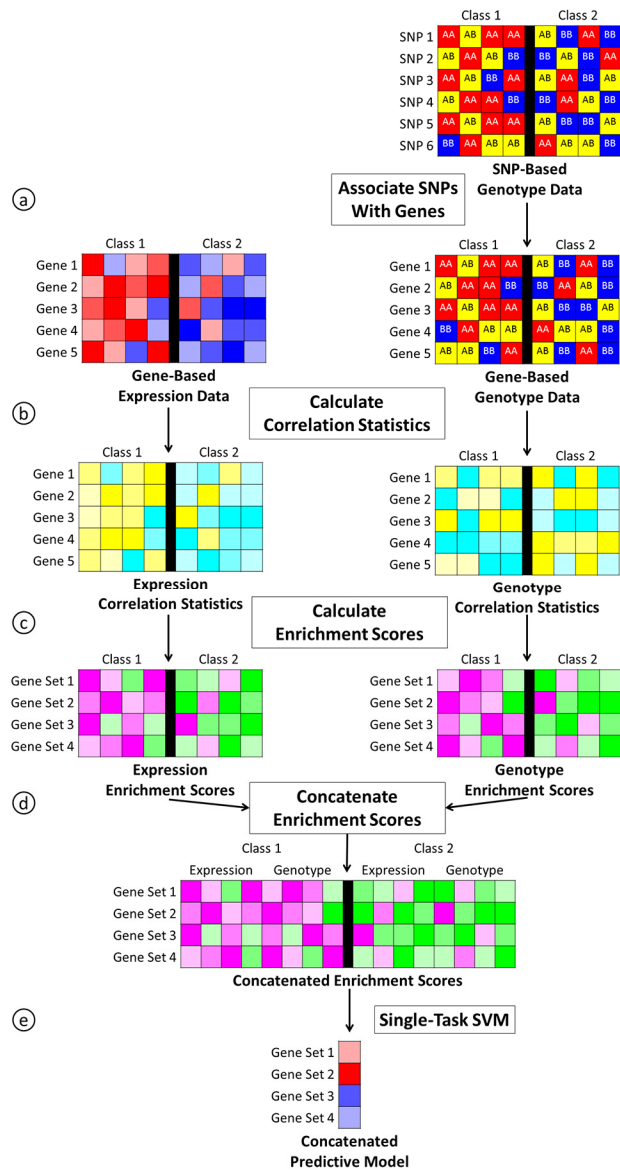


Figure 3: Overview of the concatenated data pipeline

2.3 Matched Data Methods

None of the three previously discussed methods require the data to be matched, meaning that the data from each genomic data type are from the same samples. Since these methods do not treat matched data differently from unmatched data, they may fail

to utilize instances in which pathways show coordinated enrichment across multiple data types in matched samples. Also, the methods that do not require matched data produce separate predictive models for each data type. For new samples with matched data, it may be useful to obtain a predictive model that can assign a single prediction to these samples given all data from all tasks. To address this, I developed models that specially consider multiple data from matched samples that produce a single prediction for each sample.

2.3.1 Summed Prediction Model

The first method I created is a *summed prediction* model that sums the predictions of single-task models. This method independently creates a single-task model for each data type. For a new sample, each model makes a prediction for each data type, p_i . I summed these predictions to make a single prediction as

$$p = p_1 + \dots + p_T$$

2.3.2 Summed Enrichment Model

The second method I developed is a *summed enrichment score* model that sums the ASSESS enrichment scores. This method uses ASSESS to calculate enrichment scores for each data type, $\{ES_i^k(t)\}_{t=1}^T$. For each gene set and sample, I summed the enrichment scores from all tasks to create a single enrichment score as

$$SES_i^k = ES_i^k(1) + \dots + ES_i^k(T)$$

I used these summed enrichment scores within a single-task model to obtain a single predictive model. For a new sample, I calculated summed enrichment scores using the same method. The predictive model uses these summed enrichment scores to make a single prediction.

2.3.3 Merged Model

The third method I created is a *merged* model that merges the ASSESS enrichment scores before building the predictive model. This method uses ASSESS to calculate enrichment scores for each data type. When building the training data set for the predictive model, I merged all of the enrichment scores from all tasks for a given sample into a single feature vector of enrichment scores for that sample. I used these merged enrichment scores with a single-task model to obtain a single predictive model. This model contains weights for every gene set and task. For a new sample, I calculated enrichment scores in ASSESS and merged all of the enrichment scores from all tasks into a single feature vector. I used this feature vector with the predictive model to make a single prediction.

2.4 Input Data

The main input to my framework is a gene-based genomic data set. This data must contain continuous values representing a genomic measurement of each gene for a sample. This makes my framework unsuitable for sequence-based data, such as RNA-Seq, unless continuous measurements for each gene are first derived from this data.

Additional data types that do not fit this description may be used in my framework if a metric is developed to calculate a correlation statistic for each gene and sample that follows the ASSESS correlation statistic model (see section 2.1.1). I have already done this for genotype data (see section 2.1.2).

Although my framework can use any gene-based genomic data that fits the description above, some data types that are especially suitable for analysis in addition to gene expression are copy number variation, DNA methylation, and genotype data. Duplication or deletion of genes often influences their expression levels [8], and many studies have found that copy number variation is associated with phenotypic differences [4,8,11]. Methylation of CpG dinucleotides in the promoters of genes may suppress the expression of the genes [52], and studies have found variation in DNA methylation to be associated with phenotypic differences [7,11,52]. Genotype differences in and around genes, especially in the promoters, may influence expression levels by disrupting the binding of transcriptional machinery [68], and many studies have linked genotype differences to phenotypic variation [10,53,54].

My integrative framework takes as inputs these genomic data sets from multiple data types for which the samples have been assigned to one of two phenotypic classes along with a collection of gene sets. A gene set is a predefined list of genes that have a biological connection to each other, such as functional similarities, close chromosomal proximity, or involvement in the same pathway. The Molecular Signatures Database

(MSigDB) [16], hosted by the Broad Institute, is a popular online gene set database. In order to model phenotypic differences in the context of pathways, I used gene sets associated with biological pathways. Collection C2-CP in the MSigDB is a commonly used collection of manually curated gene sets containing canonical pathways. This collection includes pathway gene sets obtained from online pathway databases, such as BioCarta (<http://biocarta.com>), KEGG [55], and Reactome [56].

3. Validation Studies

In order to determine the usefulness of my framework in discovering and utilizing the underlying genomic variation driving phenotypic differences, I performed several validation studies. In this chapter, I first explain a simulation study that I created to compare multi-task learning with other methods. Next, I describe a breast cancer analysis that I performed to determine the ability of my framework to discover relevant pathways.

3.1 Simulation Study

3.1.1 Introduction

I created a simulation study to compare my multi-task framework to other methods. It would be beneficial to use my multi-task framework if it is better at prediction or discovering relevant pathways in certain conditions. These simulations examine my multi-task framework under a variety conditions and compare its predictive accuracy and its ability to discover relevant pathways to other methods.

Validating my method using real data has disadvantages. First, real data often has too few samples to obtain the power required for an adequate comparison of methods. Second, it is often difficult to control the similarity, structure, and quality of real data. Third, when determining the ability of a method to discover relevant pathways, the correct pathways that should be identified is not always clear. To address this, I created simulated data to examine the performance of my multi-task framework

under a variety of conditions and determine how useful my framework is for genomic data.

Although my framework can be used to integrate many different genomic data types, this study focused on the integration of gene expression and genotype data. I simulated data similar to a previous study that integrated gene expression and genotype data for pathway analysis [40]. For analyses using matched data, the genotype used to generate the expression value for a sample was used as the genotype data for that sample.

3.1.2 Simulating Genotype Data

The genotype data in the simulation study had a genotype of AA, AB, or BB for each sample and SNP. Each genotype data set contained genes that were either genetically associated or had a random genotype. Genetically associated genes are those with a causal SNP within the gene. Causal SNPs are those whose genotype is significantly associated with the phenotype. I obtained the simulated genotype data from a collaborator. They mapped genes that were genetically associated to a single causal SNP, and they mapped genes that had a random genotype to a single random SNP. They simulated the causal SNPs based on parameters described below estimated from genotype information for glioblastoma generated through The Cancer Genome Atlas (TCGA) project [11]. They based these SNPs on the P53PATHWAY, as defined in version 2.5 of the Molecular Signatures Database (MSigDB) [16]. First, they mapped a

single SNP in the glioblastoma data to each of the genes in the P53PATHWAY. To do this, they found the SNPs within the region 1,000 bases upstream of the transcription start site to the end of the transcribed region of each gene. Then, they selected the SNP with minor allele frequency greater than 0.05 that had the highest chi-square association with glioblastoma. They set the allele frequencies of the causal SNPs in the simulated data to that of these selected SNPs in the glioblastoma data. They generated the heterozygote odds ratio for each SNP from $U[1.1,1.3]$ and used an additive disease model with a disease prevalence of 0.02. Using these parameter settings, they generated genotype data using PLINK [41]. They determined the probability that each sample belongs to class 1 based on the following model:

$$\text{logit}\{\Pr(Y_i = C_1)\} = \sum_{j=1}^N g_j^i \beta_j + e_i$$

where N is the number of causal SNPs, g_j^i is the coding of the genotype of the i -th sample for the j -th SNP, β_j is the log of the heterozygote odds ratio for the j -th SNP, and e_i is an error term for the i -th sample drawn from a standard normal distribution. They randomly assigned each sample to either class 1 or class 2, with the probability of being assigned to class 1 equal to the probability calculated in the model above. They also generated random genotype data using PLINK. For the random genotype data, They drew allele frequencies from $\text{Beta}(0.1,0.1)$ and assigned a heterozygote odds ratio of 1.

3.1.3 Simulating Expression Data

The expression data in the simulation study had a continuous expression measurement for each sample and gene. Each gene expression data set contained genes that were either differentially expressed or had random expression. Differentially expressed genes are those whose expression measurements are correlated with the phenotype. I obtained the simulated expression data from a collaborator. They simulated the expression data based on the TCGA glioblastoma study. They based genes that were differentially expressed on the expression of all genes in the P53PATHWAY. They calculated the mean vector μ and the covariance matrix Σ of the genes in the P53PATHWAY. They used these parameters to generate baseline expression levels by drawing from a multivariate normal distribution, $X_0 \sim N(\mu, \Sigma)$. They added a disease effect to these genes by linking each gene to a causal SNP and calculated the final expression level as

$$x_j^i = X_0^{ij} (1 + g_j^i \beta_j^i)$$

where X_0^{ij} is the baseline expression of the i -th sample for the j -th gene, g_j^i is the coding of the genotype of the i -th sample for the j -th SNP, and β_j^i is the effect size of the genotype on gene expression that is drawn from $U[1.0, 1.5]$. They also generated random expression data. They calculated the mean of all genes in the glioblastoma data and took the average of these means as μ_0 and determined the standard deviation of all

genes and the average as σ_0 . They used these parameters to generate random expression levels by drawing from a normal distribution, $X \sim N(\mu_0, \sigma_0^2)$.

3.1.4 Generating a Simulated Data Set

The collaborator simulated a collection of simulated data using the methods described in the previous sections to produce data for 400 samples with 1,600 genes that were differentially expressed and genetically associated and 198,400 genes with random expression levels and random genotype. These samples were evenly divided into two phenotypic classes, such that 200 samples belonged to the positive class and 200 samples belonged to the negative class. To generate each simulated data set for analysis in my simulation studies, I randomly sampled without replacement from this collection of data to obtain a data set with the desired number of samples and genes.

3.1.5 Varying Similarity of Tasks

In the first simulation study, I varied the similarity of the data in two tasks. The two tasks were a simulated expression data set and simulated genotype data set. Multi-task learning may offer an advantage when there is a balance of similar and different pathway enrichment among different data types. Multi-task learning introduces inductive bias to the final predictive model of a task [25]. Inductive bias in a learning algorithm is an assumption about the data that biases the model towards a different result. In the case of multi-task learning, the assumption is that the tasks contain similar information, and this will bias the model towards using signals from the other tasks. If

the enrichment is too different, then a multi-task model may not outperform single-task learning. This is because the inductive bias will mostly add noise to the model and decrease performance. Similarly, multi-task learning may not outperform a concatenated data model if there is too much similar enrichment. This is because the model-task model may not introduce enough inductive bias towards using signals from the other tasks. To test the performance of my multi-task framework, I created a simulation to compare the predictive accuracy of a multi-task model to a single-task and concatenated data model with varying similarity in the tasks.

I simulated gene expression and genotype data with gene sets belonging to one of four gene set types (Table 1). These gene set types either contain similar enrichment across all tasks or unique enrichment only in one task. Different combinations of these gene set types produce data with varying similarity between the tasks. I generated data for 5 experimental scenarios, each with a varied number of gene sets from each gene set type (Table 2). Data with gene sets predominantly from gene set type 1 have similar enrichment across tasks, while data predominantly from gene set type 2 and 3 have different enrichment across tasks.

Table 1: Gene Set Types

Type 1	10 genes differentially expressed and genetically associated
Type 2	10 genes differentially expressed but not genetically associated
Type 3	10 genes not differentially expressed but genetically associated
Type 4	10 genes not differentially expressed nor genetically associated

Table 2: Scenarios with Varying Similarity between Tasks

	Type 1 Gene Sets	Type 2 Gene Sets	Type 3 Gene Sets	Type 4 Gene Sets
Scenario 1	0	20	20	60
Scenario 2	5	15	15	65
Scenario 3	10	10	10	70
Scenario 4	15	5	5	75
Scenario 5	20	0	0	80

For each scenario, I simulated matched expression and genotype data for 50 training samples, which were equally split into 2 phenotypes. The data was matched such that the expression level of a gene for a given sample was generated taking into account the genotype of the SNP associated with that gene for that sample. I used my multi-task framework to train predictive models with these samples. Then, I used these models to obtain predictions for 50 test samples as to which phenotypic class they belong to. I also used the same data and ASSESS-based enrichment scores to evaluate single-task SVMs and an SVM with the expression and genotype enrichment scores concatenated. In addition, I used the same enrichment scores to evaluate the summed prediction, summed enrichment score, and merged models, which utilize matched data (see section 2.3). I repeated this procedure 200 times to obtain 10,000 predictions for each scenario and calculated the percentage of correct predictions for each scenario and model type (Tables 3, 4, and 5).

Table 3: Performance of Expression Data with Varying Levels of Similarity

	Single-Task	Multi-Task	Concatenated
Scenario 1	59.58% ± 0.52%	58.98% ± 0.50%	58.69% ± 0.49%
Scenario 2	59.58% ± 0.52%	59.01% ± 0.48%	59.06% ± 0.50%
Scenario 3	59.58% ± 0.52%	59.17% ± 0.50%	59.20% ± 0.48%
Scenario 4	59.58% ± 0.52%	59.25% ± 0.50%	59.34% ± 0.49%
Scenario 5	59.58% ± 0.52%	59.55% ± 0.50%	59.26% ± 0.49%

Table 4: Performance of Genotype Data with Varying Levels of Similarity

	Single-Task	Multi-Task	Concatenated
Scenario 1	69.26% ± 0.47%	66.00% ± 0.48%	62.33% ± 0.53%
Scenario 2	69.26% ± 0.47%	66.14% ± 0.44%	63.23% ± 0.50%
Scenario 3	69.26% ± 0.47%	66.58% ± 0.45%	63.91% ± 0.50%
Scenario 4	69.26% ± 0.47%	66.87% ± 0.46%	64.53% ± 0.48%
Scenario 5	69.26% ± 0.47%	67.76% ± 0.46%	65.18% ± 0.49%

Table 5: Performance of Matched Data Models with Varying Levels of Similarity

	Summed Prediction	Summed Enrichment Score	Merged
Scenario 1	67.97% ± 0.52%	67.87% ± 0.46%	71.19% ± 0.47%
Scenario 2	67.97% ± 0.52%	67.99% ± 0.44%	71.19% ± 0.47%
Scenario 3	67.97% ± 0.52%	68.11% ± 0.46%	71.19% ± 0.47%
Scenario 4	67.97% ± 0.52%	68.68% ± 0.48%	71.19% ± 0.47%
Scenario 5	67.97% ± 0.52%	68.53% ± 0.49%	71.19% ± 0.47%

For the expression data, the predictive performance was similar for all scenarios and model types (Table 3). For the genotype data, multi-task learning had a significant improvement in predictive accuracy compared to the concatenated model for all scenarios, but failed to perform better than the single-task model (Table 4). Also, accuracy improved for the multi-task and concatenated models as the scenarios

contained more similar enrichment (Table 4). For the models that utilize matched data, the summed prediction and summed enrichment score models failed to perform better than the best unmatched model, but the merged model had a significant improvement in predictive accuracy compared to the best unmatched model (Table 5). Although the difference in predictive accuracy was statistically significant in some cases, the actual predictive performance was similar in these instances.

These results suggest that single-task learning is the best method for obtaining the highest predictive accuracy when building models for unmatched gene expression and genotype data, even when all the gene sets associated with phenotype are similarly enriched in both data types. This could be because the two data types have very different underlying structure, and the single-task model is the best at preserving these differences in structure. However, the multi-task model had a significantly higher predictive accuracy than the concatenated data model, making it the best integrative method for unmatched gene expression and genotype data. The merged model had the best predictive accuracy overall, suggesting that methods that specially consider matched data may be better if the data is matched.

3.1.6 Varying Number of Samples

I next determined the effect that sample size has on my multi-task framework when compared to a single-task or concatenated data model. To do this, I first simulated matched expression and genotype data using gene sets from scenario 3 (see Table 2). In

the previous analysis, I used 50 samples to train the model. In this analysis, I varied the number of training samples from 10 to 200. As above, I used the training samples to build a multi-task, single-task, and concatenated data model, and I simulated an equal number of test samples to generate predictions. I also evaluated the summed prediction, summed enrichment score, and merged models, which utilize matched data. I repeated to obtain 10,000 predictions for each number of samples and calculated the percentage of correct predictions for each number of samples and each type of model (Tables 6, 7, and 8).

Table 6: Performance of Expression Data with Varying Sample Sizes

	Single-Task	Multi-Task	Concatenated
10 Samples	56.47% ± 0.46%	55.85% ± 0.47%	55.72% ± 0.48%
20 Samples	58.15% ± 0.53%	58.33% ± 0.51%	58.26% ± 0.51%
50 Samples	59.10% ± 0.54%	59.02% ± 0.50%	59.16% ± 0.50%
100 Samples	61.00% ± 0.49%	61.26% ± 0.47%	61.20% ± 0.53%
200 Samples	63.67% ± 0.54%	63.10% ± 0.56%	62.74% ± 0.54%

Table 7: Performance of Genotype Data with Varying Sample Sizes

	Single-Task	Multi-Task	Concatenated
10 Samples	55.10% ± 0.49%	56.06% ± 0.47%	56.02% ± 0.48%
20 Samples	62.52% ± 0.49%	62.93% ± 0.47%	62.05% ± 0.48%
50 Samples	70.84% ± 0.47%	69.02% ± 0.45%	65.11% ± 0.51%
100 Samples	76.50% ± 0.45%	74.10% ± 0.48%	70.28% ± 0.54%
200 Samples	82.50% ± 0.53%	80.19% ± 0.50%	74.91% ± 0.56%

Table 8: Performance of Matched Data Models with Varying Sample Sizes

	Summed Prediction	Summed Enrichment Score	Merged
10 Samples	59.19% \pm 0.48%	58.57% \pm 0.49%	59.20% \pm 0.48%
20 Samples	63.45% \pm 0.50%	63.86% \pm 0.49%	64.03% \pm 0.52%
50 Samples	68.60% \pm 0.46%	69.03% \pm 0.50%	72.57% \pm 0.45%
100 Samples	74.34% \pm 0.46%	74.77% \pm 0.49%	79.42% \pm 0.46%
200 Samples	80.44% \pm 0.57%	81.35% \pm 0.50%	86.05% \pm 0.44%

For the expression data, the predictive accuracy was similar among all model types (Table 6). For the genotype data, multi-task learning had a significantly higher predictive performance than the concatenated model for analyses with a sample size of 50 or more (Table 7). However, multi-task learning did not perform better than single-task learning for any of the sample sizes (Table 7). For the models that utilize matched data, the merged model had a significant improvement in predictive accuracy compared to the best unmatched model for all sample sizes (Table 8). The summed prediction and summed enrichment score models also had a significantly higher predictive performance than the best unmatched model for the analysis with a sample size of 10 (Table 8). As expected, the predictive accuracy improved as the number of samples increased for all model types, but multi-task learning did not appear to benefit more than the other model types.

3.1.7 Varying Number of Tasks

I also evaluated the effect of varying the number of tasks. Studies have shown that the predictive accuracy of multi-task models increases as the total number of tasks

increases [75]. I explored whether this is also the case for genomic data. I generated expression data sets, each corresponding to a task, with 20 training samples evenly divided into 2 phenotypes. I simulated phenotype associated gene sets with 10 genes that were differentially expressed between the phenotypes, and background gene sets with 10 genes that represented a null model of random expression. I generated a task by simulating the first 30 gene sets as phenotype associated gene sets and the next 20 gene sets as background. For the last 50 gene sets, 30 were randomly chosen to be phenotype associated and the other 20 background. I generated additional tasks in the same way. As a result, the first 50 gene sets contained similar enrichment among all tasks, and the last 50 gene sets contained enrichment unique to each task. I used this data to build multi-task models with the number of tasks used to build each model varying from 2 to 100. I also used this data one task at a time to build single-task models for comparison. After using the 20 training samples for each task to train the model, I used 20 test samples for each task to obtain predictions. I repeated to obtain 10,000 predictions for each number of tasks and determined the percentage of correct predictions for each number of tasks (Figure 4). I also performed the same analysis with simulated genotype data (Figure 5). For the genotype data, phenotype associated gene sets contained genes that were genetically associated and background gene sets contained genes that were not genetically associated. In these figures, the center dashed line is the accuracy of the single-task model, with the outer dashed lines being the standard error.

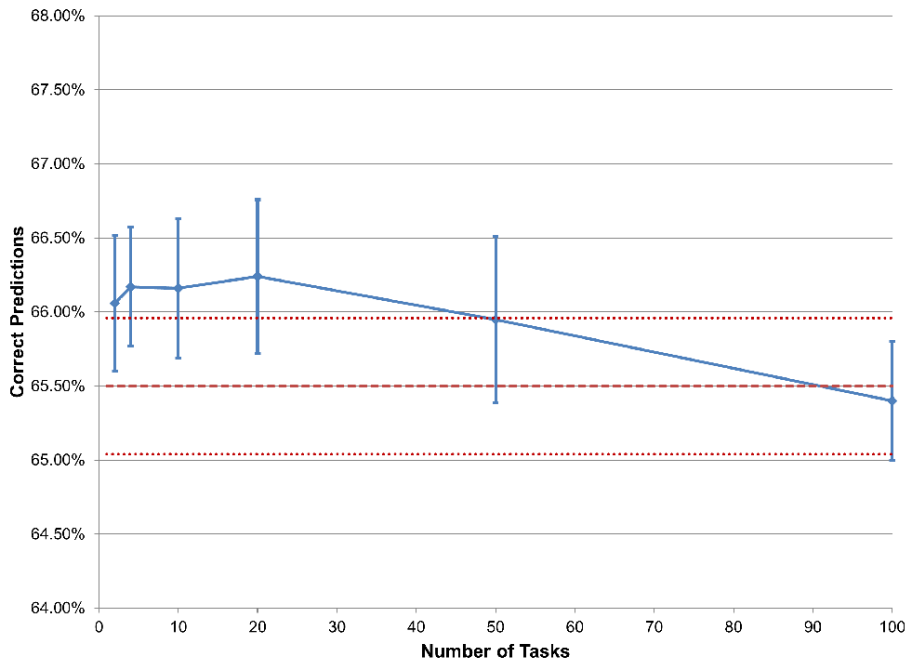


Figure 4: Performance of Expression Data with Varying Number of Tasks

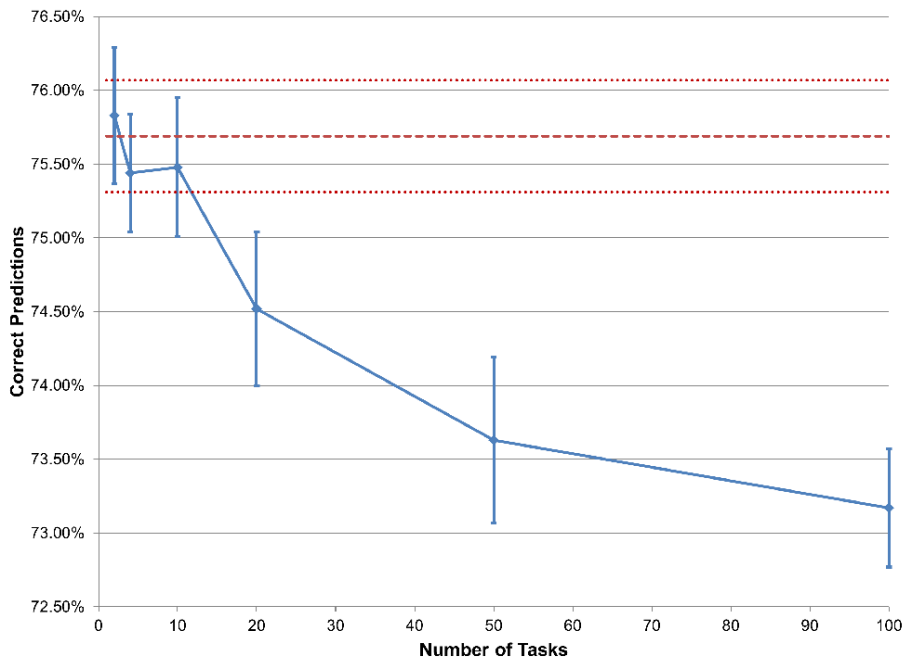


Figure 5: Performance of Genotype Data with Varying Number of Tasks

For the expression data, the predictive accuracies of all multi-task models and of the single-task model showed no significant difference (Figure 4). For the genotype data, the predictive performance of the multi-task experiments with 10 tasks or less was not significantly different than the single-task model (Figure 5). However, multi-task models with 20 tasks or more had a significantly lower predictive accuracy than the single-task model (Figure 5). Although this suggests that analyses with a large number of tasks may have a significant difference in performance between multi-task learning and single-task learning, most analyses of biological data will have a small number of different data types or tasks. For these data, a multi-task model may not be significantly different than single-task in terms of predictive accuracy.

3.1.8 Finding Pathways Enriched Across Multiple Data Types

In addition to class prediction of unknown samples, using my framework to discover pathways with similar enrichment across multiple data types may provide additional insight into the underlying biology influencing the phenotype. For example, pathways with simultaneous gene expression and genotype enrichment may contain genes with genotype differences that are directly influencing gene expression. To determine whether my integrative approach provides an improved ability to discover these types of gene sets enriched across multiple data types, I performed an additional simulation study.

To measure the significance of a pathway, I examined the predictive weights for each gene set that can be derived from the SVM predictive model (see section 2.2.1). Gene sets with higher weights contribute more to prediction, and may also be more important in distinguishing phenotype. After training a multi-task SVM model, a common weight can be derived that is interpreted as a measure of importance for prediction derived from all tasks (see section 2.2.2). Gene sets with larger common weights can be viewed as sharing common information important for prediction across all tasks. These gene sets may represent biological pathways with important factors in multiple data types that are influencing phenotype. I designed the following simulation to determine the ability of my framework to discover this type of gene set.

I simulated matched expression and genotype data for 400 samples that were evenly divided into 2 phenotypes. This data contained the same gene set types as the first simulation study (see section 3.1.5). I created this data set with 1 of gene set type 1 (the target gene set), 5 of gene set type 2, 5 of gene set type 3, and 89 of gene set type 4. I wanted to test the ability of my framework to extract the one target gene set, which contains genes that are both differentially expressed and genetically associated. I used my framework to train multi-task, single-task, and concatenated data predictive models. For multi-task, I determined the rank of the common weight in the predictive model for the target gene set. For single-task, I calculated the rank of the weight in the predictive model for the target gene set in the expression model and the genotype model

separately. For the concatenated data model, I calculated the rank of the target gene set in a single predictive model built by concatenating the enrichment scores of the expression and genotype data. I also took the sum of the weights for both single-task models and determined the rank of the combined weight for the target gene set. In addition, I trained the summed enrichment score and merged predictive models. For the summed enrichment score model, I calculated the rank of the target gene set in a single predictive model built by taking the sum of the enrichment scores from the expression and genotype data. For the merged model, I built a single predictive model by taking the enrichment scores from the expression and genotype data and merging them into a single feature vector for each sample. I then took the sum of the expression and genotype weights for each gene set and determined the rank of the combined weight for the target gene set. I repeated this analysis 1000 times and calculated the average rank for each model type (Table 9).

Table 9: Average Rank of Target Gene Set

	Average Rank
Single-Task Expression Weight	11.01 ± 0.47
Single-Task Genotype Weight	4.17 ± 0.10
Single-Task Weights Summed	3.07 ± 0.16
Multi-Task Common Weight	3.08 ± 0.16
Concatenated Weight	2.89 ± 0.14
Summed Enrichment Score Weight	3.54 ± 0.11
Merged Weights Summed	2.87 ± 0.10

The average rank for the target gene set was significantly lower in all models that considered both tasks (single-task summed, multi-task, concatenated, summed enrichment score, and merged) compared to either of the separate single-task models (Table 9). This suggests that an integrated approach may be beneficial for discovering biological pathways that have an important effect within several genomic data types.

3.1.9 Discussion

The simulation study showed that integrative methods provided an improved ability to discover pathways that are enriched over multiple data types. The predictive accuracy of single-task learning was better than the integrative methods for unmatched data, but multi-task learning performed about the same as single-task learning and performed better than the concatenated data model. This suggests that my multi-task framework provides a useful method for both discovering pathways enriched in multiple data types while also retaining similar predictive accuracy as a single-task model. Also, the merged model had the highest predictive accuracy overall and was one of the best at discovering pathways enriched in multiple data types. This suggests that methods that specially consider matched data may be better if the data is matched.

3.2 Breast Cancer Study

3.2.1 Introduction

To further explore the usefulness of my multi-task framework, I performed an integrative analysis of estrogen receptor (ER) status in breast cancer, which is a well-

studied disease. Many of the top pathways identified in this analysis should be present in the literature as associated with estrogen receptor or breast cancer. Estrogen is a hormone that helps regulate growth in normal mammary tissue and is important in the development of breast cancer [28]. Estrogen regulates gene expression by binding with ER, and this activates downstream targets that regulate growth and differentiation [57]. Breast cancer tumors are often classified as ER-positive or ER-negative, and these two subtypes show very different gene expression patterns [57]. Learning more about the involvement of ER in breast cancer is important, because variation in the expression of ER has important implications in the biology, treatment, and prognosis of breast cancer tumors [28]. I applied my framework to matched expression and genotype data for breast invasive carcinomas (BRCA) that were classified based on ER status, in order to identify pathways relevant to estrogen receptor status in breast cancer and to provide support for the usefulness of my framework.

I obtained breast cancer data generated through The Cancer Genome Atlas (TCGA) project from their data portal (<http://cancergenome.nih.gov>). I selected samples that provided matched gene expression and genotype data. In order to eliminate spurious results due to race, gender, or age, I filtered samples to only include patients who were white, female, 40 to 70 years of age at initial diagnosis, and had a known ER status of positive or negative. I also eliminated the sample with barcode "TCGA-A2-A0CY" because of unreliable genotype data. This resulted in a data set of matched gene

expression and genotype data for 61 ER negative samples and 203 ER positive samples. My collection of gene sets contained curated canonical pathways from collection C2-CP of the Molecular Signatures Database (MSigDB) [16]. In order to eliminate results that are not biologically relevant due to gene set size being too small or too big, I filtered these gene sets to only include those with 15 to 100 mapped genes, resulting in 538 gene sets.

3.2.2 Analysis

First, I wanted to determine the ability of the data to predict ER status. I performed leave-one-out (LOO) cross-validation to calculate predictive accuracy for multi-task, single-task, concatenated, summed prediction, summed enrichment score, and merged models (Tables 10, 11, and 12). For the expression data, the predictive performance was very high for all model types, both with respect to overall accuracy and positive and negative predictive values (Table 10). The phenotypic classification was based on the expression level of ER, and tumors with different ER status are known to show very different gene expression patterns [57]. Therefore, it not surprising that the expression model had a high predictive ability. For the genotype data, the overall predictive performance was moderate, but the negative predictive value (NPV) was low (Table 11). For the models that utilize matched data, the predictive accuracy was moderately better than using the genotype data alone, but not better than using the expression data alone (Table 12). The negative predictive value was greatly improved

for the models that utilize matched data when compared to using the genotype data alone (Table 12). These results suggest that important gene sets in the predictive models may be biologically relevant to ER status.

Table 10: Correct Predictions for Breast Cancer Expression Data

	Single-Task	Multi-Task	Concatenated
Overall Accuracy	92.42% (244/264)	92.05% (243/264)	92.05% (243/264)
Positive Predictive Value	94.63% (194/205)	95.05% (192/202)	95.05% (192/202)
Negative Predictive Value	84.75% (50/59)	82.26% (51/62)	82.26% (51/62)

Table 11: Correct Predictions for Breast Cancer Genotype Data

	Single-Task	Multi-Task	Concatenated
Overall Accuracy	77.65% (205/264)	78.41% (207/264)	78.79% (208/264)
Positive Predictive Value	81.03% (188/232)	81.74% (188/230)	82.10% (188/229)
Negative Predictive Value	53.13% (17/32)	55.88% (19/34)	57.14% (20/35)

Table 12: Correct Predictions using Matched Data Models for Breast Cancer Data

	Summed Prediction	Summed Enrichment Score	Merged
Overall Accuracy	85.98% (227/264)	88.26% (233/264)	83.71% (221/264)
Positive Predictive Value	86.40% (197/228)	89.81% (194/216)	85.40% (193/226)
Negative Predictive Value	83.33% (30/36)	81.25% (39/48)	73.68% (28/38)

I next examined gene sets with the highest weights in the predictive models for their biological relevance to ER status. To allow for a direct comparison of the predictive weights among gene sets, I first normalized the enrichment scores from ASSESS (see section 2.1.1). I calculated the ranks of all gene sets for multi-task, single-task, concatenated, summed enrichment score, and merged models. A complete list of all

ranks and weights for all gene sets and model types is presented in Tables S1 and S2. It is interesting to note that the ranks vary considerably among all models types. This includes significant differences between the multi-task model that considers all data simultaneously and the expression single-task and genotype single-task models which consider only data from one data type. This suggests that using an integrative approach provides results distinct from analyses of either data type alone. This reflects the ability of an integrative model to identify and utilize pathways with common enrichment across multiple data types.

A list of the top 15 gene sets with the highest common weights in the multi-task model is presented in Table 13, along with the corresponding ranks of the expression task-specific weights and the genotype task-specific weights. The common weight from the multi-task model can be interpreted as the importance in distinguishing phenotype drawn from both tasks simultaneously, whereas the task-specific weights provide a way to estimate the contribution that each data type had in the overall integrated rank of the gene set. An analysis of the top 15 gene sets from the multi-task model showed they were related to estrogen, steroids, cell signaling, or the cell cycle, discussed in more detail below. This provides support for the usefulness of my framework in identifying pathways associated with complex traits.

Table 13: Top Gene Sets in Breast Cancer Analysis

	Multi-Task Common Weight Rank	Expression Task-Specific Weight Rank	Genotype Task-Specific Weight Rank
HER2 Pathway	1	1	449
Phase II Conjugation	2	11	12
Steroid Hormone Biosynthesis	3	8	60
FRS2-Mediated Cascade	4	9	59
One Carbon Pool by Folate	5	3	174
Neurotransmitter Release Cycle	6	67	5
Nitrogen Metabolism	7	2	250
Steroid Biosynthesis	8	26	29
Cholesterol Biosynthesis	9	21	53
Apoptotic Signaling in Response to DNA Damage	10	10	138
Riboflavin Metabolism	11	73	10
ECM-Receptor Interaction	12	247	1
Nuclear Receptor Transcription	13	59	17
Mitotic Prometaphase	14	169	6
Steroid Metabolism	15	14	177

Estrogen plays an important role in breast cancer [28]. I found that three of the top 15 gene sets were directly related to estrogen signaling and metabolism: “HER2 Pathway” (rank 1), “Phase II Conjugation” (rank 2), and “Nuclear Receptor Transcription” (rank 13). Human epidermal growth factor receptor 2 (HER2), encoded by the gene *ERBB2*, influences the expression and activity of the estrogen receptor [29]. The “HER2 Pathway” gene set contains the estrogen receptor 1 (*ESR1*) gene. The “Nuclear Receptor Transcription” gene set also contains the *ESR1* gene, and nuclear receptor coactivators are thought to participate with the estrogen receptor pathway [30]. Several phase II conjugating enzymes are involved with the metabolism of estrogen [31].

Tamoxifen is an antiestrogenic drug that is widely used in the treatment of ER positive breast cancer [32]. One study showed that genetic variation in several phase II conjugating enzymes influenced the efficacy of Tamoxifen therapy in breast cancer [33]. Since this study linked genotype differences to Tamoxifen efficacy, it is interesting to note that the Phase II Conjugation gene set has the twelfth highest genotype task-specific weight (Table 13) and is the highest ranked gene set in the multi-task genotype predictive model (w_2 , Table S1). It is also the eleventh highest gene set in the expression task-specific weights (Table 13) and has the fourth highest rank in the multi-task expression predictive model (w_1 , Table S1). This suggests that genotype differences may be directly influencing expression changes. The strong association in both the expression and genotype data resulted in the second highest rank in the multi-task common weights (Table 13), which is higher than the weight in either of the single-task models alone (Table S1). All three of the estrogen-related gene sets contained genes that were generally overexpressed in the ER positive samples.

Estrogen is a steroid hormone, and I found that four of the top 15 gene sets were involved with the synthesis or metabolism of steroids: “Steroid Hormone Biosynthesis” (rank 3), “Steroid Biosynthesis” (rank 8), “Cholesterol Biosynthesis” (rank 9), and “Steroid Metabolism” (rank 15). In addition to estrogen, other steroid hormones, such as progesterone, play an important role in breast cancer [28]. Also, many steroids,

including estrogen, are synthesized from cholesterol, and one study showed that cholesterol levels are linked with breast cancer prognosis [34].

The estrogen receptor participates in cellular signaling initiated by the binding of estrogen and facilitating the activation of downstream processes. In addition to the estrogen-related pathways, three of the top 15 gene sets were similarly involved with other types of cell signaling: “FRS2-Mediated Cascade” (rank 4), “Neurotransmitter Release Cycle” (rank 6), and “ECM-Receptor Interaction” (rank 12). The FRS2-mediated cascade links Fibroblast Growth Factor Receptor (FGFR) to the eventual activation of several important signaling pathways. One study showed that blocking FGFR inhibited breast cancer proliferation and led to downregulation of the MAPK and PI3K pathways [35]. Also, ECM receptors may participate in the control of many stages of breast cancer [36], and neurotransmitters may influence the metastasis of breast tumors [37]. All three of these cell signaling gene sets contained genes that were generally overexpressed in the ER positive samples.

Tumors accumulate genetic damage that results in a perturbed cell cycle which increases the number of tumor cells by stimulating cell birth or inhibiting cell death or cell-cycle arrest [12]. Many of the previously discussed gene sets are involved with the cell cycle or metabolism, and I found that the five remaining gene sets in the top 15 were also involved with the cell cycle and metabolism: “One Carbon Pool by Folate” (rank 5), “Nitrogen Metabolism” (rank 7), “Apoptotic Signaling in Response to DNA Damage”

(rank 10), “Riboflavin Metabolism” (rank 11), and “Mitotic Prometaphase” (rank 14).

Disrupting mitotic prometaphase may influence cell-cycle arrest, and disrupting apoptotic signaling in response to DNA damage may inhibit the cell death of tumor cells. Also, folate, nitrogen, and riboflavin, also known as vitamin B2, are important for cell growth. One study linked increased consumption of folate and B vitamins with reduced risk of breast cancer [38].

3.2.3 Discussion

Results from this study indicate that a pathway-based integrative analysis is a promising approach to identify pathways that are influenced by both gene expression changes and genotype variation. All of the top 15 pathways from the multi-task model built using the breast cancer data have been previously associated with breast cancer. This suggests that an integrative approach may be useful for discovering pathways related to complex diseases, especially diseases that are not as well understood, and for determining the contribution that each data type has for each pathway. The “Phase II Conjugation” gene set is an example that had a strong association in both the expression and genotype data, and this gene set had the second highest multi-task common weight, which was higher than in either of the single-task models alone. This supports the use of an integrative approach in discovering gene sets that may have a direct link between genotype and expression.

4. Additional Gene Set Studies

In order to use gene set analysis to explore specific biological questions, I performed additional studies. In this chapter, I first describe a study that explores whether addiction-related gene sets known for association with sodium depletion in mice are also associated with thirst in mice. Next, I describe a study that determines whether a new mouse model of rhabdomyosarcoma (RMS) is similar to human RMS.

4.1 Thirsty Mouse Study

4.1.1 Introduction

Mammals experience unlearned behavioral patterns with high survival value known as instincts, such as hunger, thirst, and sodium appetite [45]. For these instincts, a physiological strategy has evolved whereby an animal in the wild will develop an intense appetite for food, water, or salt, and then rapidly become satiated once it obtains what it needs [42]. For sodium appetite, gratification and a complete loss of interest in salt will occur before significant absorption of sodium from the gut occurs [43,44]. This behavior has a high survival value because an animal may be very vulnerable while obtaining food, water, or salt, and rapidly exiting the source will decrease chances of predation [45].

One study examined the effects of sodium appetite and gratification in mice [45]. Several methods exist to cause a mouse to experience sodium appetite, such as depriving it of sodium or administering adrenocorticotrophic hormone (ACTH), which

mimics stress [46]. This study measured gene expression in the hypothalamus of mice experiencing sodium appetite driven by sodium depletion or infusion of ACTH. Comparing the mice with sodium appetite to control mice, they performed Gene Set Enrichment Analysis (GSEA) [16] and found that several gene sets containing addiction-related genes [47] were enriched. They next examined several selective antagonists of dopamine receptors and metabotropic glutamate receptor 5 (mGlu5), due to their roles in reward associated with drug abuse [48,49]. They found that SCH23390, a drug that inhibits dopamine receptor D1 (DRD1), Raclopride, a drug that inhibits dopamine receptor D2 (DRD2), and MTEP, a drug that inhibits mGlu5, decreased sodium consumption in sodium-depleted mice.

Thirst is an instinct that is similar to sodium appetite. Although there are many similarities in the way that animals respond to thirst and sodium appetite, these instincts also have a characteristic specificity [45]. This means that thirsty animals do not seek sodium, and sodium-depleted animals do not seek water. The previously discussed study found that SCH23390, Raclopride, and MTEP decreased sodium consumption in sodium-depleted mice. In water-depleted mice, they also found that Raclopride and MTEP decreased water consumption, but SCH23390 did not. This suggests that there are both similarities and differences in the molecular mechanisms driving these instincts. Therefore, researchers are interested in studying what is similar and different about these instincts on the molecular level. In my study, I explored pathways involved with

gene regulation of the instinct of thirst in the hypothalamus, and examined whether there is an association of addiction-related genes, similar to sodium appetite.

4.1.2 Analysis

I obtained gene expression data from the hypothalamus of mice from a collaborator. They used an Agilent whole mouse genome 4x44K monochromatic microarray. The samples included 4 normal mice, 4 thirsty mice that were depleted of water over several days, and 4 satiated mice that were allowed to consume water until satisfied after being depleted. They selected these groups to explore the gene expression changes that occur in the hypothalamus of mice as they become thirsty, and as they become satiated. I first filtered the data to eliminate probes with low signal. To do this, I calculated the signal-to-noise ratio for each probe and sample as

$$S2N = \frac{\text{median signal} - \text{median background signal}}{\text{background standard deviation}}$$

using the green channel median signal (gMedianSignal), the green channel median background signal (gBGMedianSignal), and the green channel background standard deviation (gBGPixSDev). I filtered probes with low signal by eliminating all probes that did not achieve a signal-to-noise ratio of greater than 3 in greater than 50% of all samples. This decreased the total number of probes used in the analysis from 43,020 down to 32,199. To determine the expression measurement of a probe, I used the green channel processed signal (gProcessedSignal). To normalize the expression measurements across samples, I performed Variance Stabilization Normalization (VSN)

[58]. I next mapped this data to the gene level by identifying all probes associated with a gene using the annotation provided with the microarray and taking the median signal across all probes for that sample. The entire filtering process reduced the total number of genes with probes from 24,135 down to a processed expression data set of 18,777 genes.

After preprocessing the data, I first explored whether there are significant gene expression differences in the hypothalamus of normal, thirsty, and satiated mice. Using the gene measurements of all 12 samples from the three groups, I performed Principal Components Analysis (PCA) [59], an unbiased approach that reduces the dimensionality of the data from p genes to k principle components. I used PCA to reduce the dimensionality of the data down to a single principle component (Figure 6). In this analysis, all three groups are distinct, and the samples within these groups cluster together (Figure 6). This shows that each of the three groups exhibits a distinct expression pattern in the hypothalamus. In addition to this, the satiated group clustered farther away from the control group than the thirsty group (Figure 6). This suggests that there are expression changes that are occurring during satiation in addition to those already altered when the mouse became thirsty, and not that the mouse becomes satiated by simply reverting expression levels back to normal.

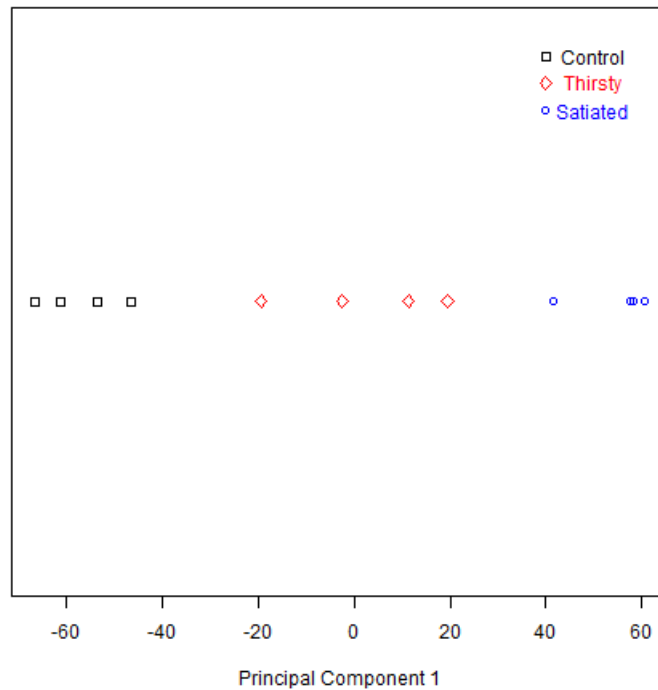


Figure 6: Principle Component Analysis of Thirsty Mice Data

I also looked for expression differences among the three groups at the single-gene level. I performed a Student's t-test for each gene comparing the normal samples and the thirsty samples (Table S3). This test provided a p-value for each gene that measured the amount of differential expression between the phenotypes. I corrected for multiple comparisons by calculating a Bonferroni adjusted p-value for each gene (Table S3). I also performed a t-test and calculated Bonferroni p-values comparing the thirsty samples and the satiated samples (Table S4). There were 56 genes that were significantly different between the normal and thirsty samples with a Bonferroni adjusted p-value of less than 0.05 (Table S3), and there were 50 genes significantly different between the thirsty and satiated samples with a Bonferroni p-value of less than 0.05 (Table S4). This

provides additional evidence that there is substantial regulation of gene expression in the hypothalamus during thirst and satiation.

I next searched for pathways that become altered when a mouse experiences thirst. To do this, I performed GSEA comparing the normal mice to the thirsty mice. I compiled the collection of gene sets that I used in the analysis by first adding the four addiction-related gene sets [47] that were found to be associated with sodium depletion in mice [45]. I then added curated canonical pathways from the C2-CP collection of the Molecular Signatures Database (MSigDB) [16]. In order to compare the addiction-related gene sets to those that were similar in size, I filtered the additional gene sets to only include those with 50 to 300 mapped genes, resulting in 183 gene sets. Results from this analysis revealed 10 gene sets with an FDR q-value of less than 0.25 (Table 14), and none of the addiction-related gene sets had a p-value of less than 0.05 (Table 15). Positive enrichment scores indicate that the genes in the gene set are generally overexpressed in the thirsty mice compared to normal mice.

Table 14: Top Gene Sets in Thirsty Mice Analysis

Gene Set	NES	FDR
Glioma	-1.73	0.01
Regulation of Insulin Secretion by GLP-1	1.55	0.07
Toll Receptor Cascades	1.56	0.09
Metabolism of RNA	1.59	0.13
Processing of Capped Intron-Containing Pre-mRNA	1.45	0.16
Innate Immunity Signaling	1.46	0.18
ErbB Signaling	-1.51	0.18
Opioid Signaling	1.47	0.19
mRNA Splicing	1.46	0.21
TRAF6 Mediated Induction of the Antiviral Cytokine IFN-alphaeta Cascade	1.40	0.24

Table 15: Addiction Gene Sets in Thirsty Mice Analysis

Gene Set	NES	p-value
Cocaine	1.25	0.09
Opioids	-1.09	0.30
Nicotine	-1.00	0.48
Alcohol	-0.93	0.60

An analysis of the top 10 gene sets revealed that they were previously associated with thirst, related to signaling pathways, or involved with RNA processing and metabolism. The “Glioma” gene set is the highest ranked and contains genes associated with glioma, a type of tumor in the brain that arises from glial cells. One study showed that calcium-sensing receptor (CaR) is particularly abundant in the subfornical organ (SFO), which is an important hypothalamic thirst center, and that CaR is expressed in glial cells [60]. The “Regulation of Insulin Secretion by GLP-1” gene set, the second highest ranked, involves Glucagon-Like Peptide-1 (GLP-1), a hormone that is known to

control feeding and drinking behavior [61,62]. One study found that oral water intake of humans decreased under GLP-1 treatment [62]. Five of the other gene sets were related to signaling pathways: “Toll Receptor Cascades”, “Innate Immunity Signaling”, “ErbB Signaling”, “Opioid Signaling”, and “TRAF6 Mediated Induction of the Antiviral Cytokine IFN-alphaeta Cascade”. This suggests that they are many signaling events that occur when a mouse becomes thirsty. The three remaining gene sets are all involved with the processing and metabolism of RNA: “Metabolism of RNA”, “Processing of Capped Intron-Containing Pre-mRNA”, and “mRNA Splicing”.

I also examined whether the association of addiction-related genes with sodium appetite [45] was also present in thirsty mice. None of the four addiction-related gene sets had a significant p-value of less than 0.05 in the thirsty mice analysis (Table 15). However, the cocaine gene set had a relatively low p-value of 0.09 (Table 15). Also, an opioid signaling pathway had an FDR q-value of 0.19, which was the 8th highest in the analysis (Table 14). The enrichment profile of this opioid signaling pathway shows a cluster of genes that are highly overexpressed in the thirsty mice compared to the normal mice (Figure 7). This suggests that addiction-related genes may also be involved with the instinct of thirst.

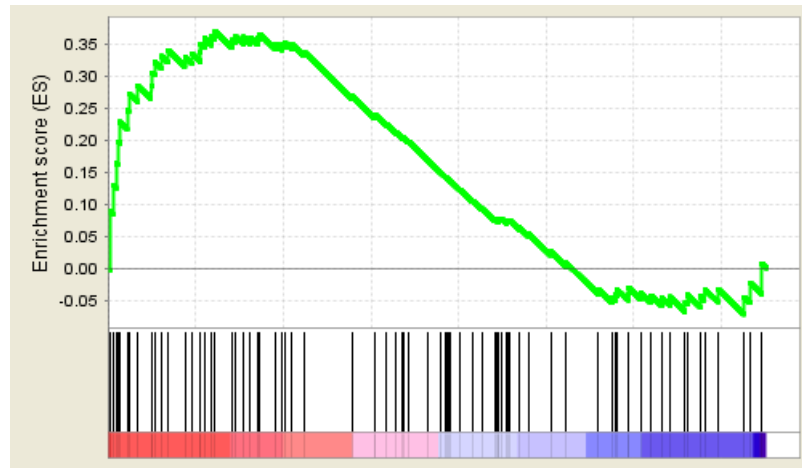


Figure 7: Enrichment Profile of Opioid Signaling Pathway in Thirsty Mice Analysis

4.1.3 Discussion

Results suggest that there is considerable regulation of gene expression in the hypothalamus when a mouse experiences thirst and satiation. Although a study showed several addiction-related gene sets are enriched in mice experiencing sodium appetite, these gene sets were not significantly enriched in thirsty mice. However, the cocaine gene set had a relatively low p-value, and an opioid signaling pathway was among the top 10 gene sets. These results show that there are differences in the biology of sodium appetite and thirst, but suggest that there may also be common components that are shared between the two similar instincts.

4.2 *Sarcoma Mouse Model Study*

4.2.1 Introduction

Rhabdomyosarcoma (RMS) is a type of soft tissue sarcoma that is thought to originate in skeletal muscle progenitors [63]. RMS can be clinically distinguished from

other sarcomas based on the presence of rhabdomyoblasts, and staining for certain myogenic transcription factors such as MyoD and myogenin [63,64]. The rarity of the disease limits the study of RMS in human patients and increases the need for genetically engineered mouse models. One group has developed a novel method for inducing a tumor in a mouse that mimics human RMS at the histological level. In my study, I examined the gene expression of these tumors to provide support for the conclusion that these tumors are more similar to human RMS than other types of human sarcoma at the molecular level.

4.2.2 Analysis

In the first step of my analysis, I assembled a list of RMS-specific genes. To do this, I compared the gene expression of RMS tumors to undifferentiated pleomorphic sarcoma (UPS). UPS is a type of sarcoma that is diagnosed by exclusion in regards to its inability to be classified as any other type of sarcoma by markers of tissue specific differentiation. To identify RMS-specific genes, I searched for genes that were upregulated in RMS compared to UPS. Since both tumor types are sarcomas, genes that are common to all sarcomas will not be differentially expressed between these two groups, but only genes that are specific to RMS will be upregulated. I obtained gene expression data from a collaborator for 18 RMS tumors induced using the novel method and for 6 UPS tumors created using a previously described method [65]. I performed a Student's t-test to calculate a significance value for differential expression for each gene.

I corrected for multiple comparisons by calculating a Bonferroni adjusted p-value for each gene. I selected all genes that were upregulated in RMS compared to UPS with a Bonferroni adjusted p-value of less than 0.05. This resulted in a list of 77 RMS-specific genes (Table S5).

I next determined whether these RMS-specific genes derived from the mouse model could be used to distinguish human RMS from other human sarcomas. I obtained a data set of gene expression for several human sarcoma tumors [66]. Each tumor sample was classified as a particular type of sarcoma. I eliminated samples that belonged to categories containing less than 5 samples. This resulted in a data set of 163 tumors consisting of 12 types of sarcoma, one of which was RMS. Using the list of RMS-specific genes as a gene set, I performed Gene Set Enrichment Analysis (GSEA) [16] for each type of sarcoma, comparing all samples belonging to that type of sarcoma with all other samples in the entire data set (Table 16). Positive enrichment scores indicate that the genes in the gene set are generally overexpressed in the individual type of sarcoma compared to all other sarcoma samples. Only the analysis comparing human RMS to other types of sarcoma had a significant p-value of less than 0.05, and the enrichment score was positive, indicating that the genes were generally overexpressed in the human RMS (Table 16).

Table 16: Gene Set Enrichment Analysis for Human Sarcoma

Tumor Type	Samples	NES	p-value
Ewing's Sarcoma	19	1.09	0.35
Liposarcoma	33	1.40	0.09
Malignant Fibrous Histiocytoma (UPS)	38	-1.31	0.15
Leiomyosarcoma	17	-0.76	0.78
Fibrosarcoma	7	-1.19	0.25
Osteosarcoma	5	1.36	0.12
Synovial Cell Sarcoma	16	-1.02	0.44
Gastrointestinal Stromal Tumor	5	-1.40	0.07
Malignant Hemangiopericytoma	6	-0.41	0.99
Rhabdomyosarcoma (RMS)	6	1.71	0.01
Dermato Tumor Type E	5	0.68	0.89
Malignant Peripheral Nerve Sheath Tumor	6	1.10	0.36

4.2.3 Discussion

The gene set with RMS-specific genes derived from the novel mouse model only had a significant enrichment in the human RMS analysis and was not significantly enriched in any other sarcoma type. This supports the classification of the novel mouse model as more similar to human RMS than other human sarcomas. Further examination of this mouse model may provide additional insight into the biology and treatment of human RMS. It is also interesting to note that there was no significant enrichment in human UPS, even though the RMS-specific genes were derived by comparing mouse RMS to mouse UPS. This suggests that the RMS-specific genes are indeed specific to RMS and may be important in the development and progression of RMS. Further exploration of these genes may reveal more about the biology of RMS tumors and may lead to possible drug targets or treatments for the disease.

5. Conclusion

This work presented a framework for modeling complex traits by integrating different types of genomic data on the pathway level. It accomplishes this by first performing sample-specific gene set analysis using ASSESS, and then building an integrative predictive model using multi-task learning. Results from simulation studies showed that using multi-task learning for analysis of genomic data appears to offer similar predictive performance as other methods. However, results from additional simulation studies and from a breast cancer analysis reveal that an integrative approach, such as multi-task learning, seems to provide an improved ability to discover relevant pathways that are enriched in multiple data types. Therefore, my multi-task framework appears to be a useful method for producing predictive models with accuracy that is comparable to similar methods, while also improving discovery of important pathways and providing information about the contribution of each data type to the enrichment of those pathways. However, future work to improve my framework should be explored, and some possible improvements are discussed below.

There are alternative integrative prediction methods that may be better than multi-task learning. This work revealed that an integrative approach, like multi-task learning, appears to be useful in the discovery of relevant pathways, but multi-task learning may not be necessary. This work presented several additional integrative methods that should be explored, among others, as possible alternatives to the multi-

task learning step in my framework. Also, one disadvantage of many of the integrative methods that were presented in this work is that they apply an equal weight to each of the genomic data types. Some data types may contain more significant information than others, and it would be useful to explore methods that apply different weights to each data type.

There are many additional uses for the enrichment scores beyond building a predictive model using multi-task learning. My framework is used to obtain sample-specific information on the pathway level for multiple data types that is normalized and easy to combine. This information can be used for other powerful integrative sample-level pathway-based analyses. For example, these enrichment scores can be used with a clustering analysis to find subtypes of samples with similar pathway enrichment profiles. Also, a suitable correlation coefficient can be calculated from the enrichment scores to directly determine pathway-level association with a phenotype of interest.

There are also several improvements that can be made with regards to obtaining the ASSESS genotype correlation statistics. My framework obtains correlation statistics for genotype data by first mapping a single representative SNP to each gene, and then uses that one SNP to calculate the correlation statistic. However, this approach removes data for many SNPs that may be useful. Further research should explore methods to calculate a correlation statistic for a gene given multiple SNPs associated with that gene. Also, improved methods for mapping SNPs to genes should be explored.

There are other modifications that can be made to my framework that may improve it. My framework calculates correlation statistics and enrichment scores independently for each data type. It may be useful to explore methods to directly integrate the ASSESS correlation statistics before obtaining enrichment scores. Also, the simulation studies and breast cancer analysis focused on the integration of gene expression and genotype data. However, my framework may also be suitable for integrating other genomic data types, such as copy number variation and DNA methylation data.

References

1. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 6: 95-108.
2. Menezes RX, Boetzer M, Sieswerda M, van Ommen GJ, Boer JM (2009) Integrated analysis of DNA copy number and gene expression microarray data using gene sets. *BMC Bioinformatics* 10: 203.
3. Tsafirir D, Bacolod M, Selvanayagam Z, Tsafirir I, Shia J, et al. (2006) Relationship of gene expression and chromosomal abnormalities in colorectal cancer. *Cancer Res* 66: 2129-2137.
4. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, et al. (2010) Integrative genomic profiling of human prostate cancer. *Cancer Cell* 18: 11-22.
5. Hawthorn L, Luce J, Stein L, Rothschild J (2010) Integration of transcript expression, copy number and LOH analysis of infiltrating ductal carcinoma of the breast. *BMC Cancer* 10: 460.
6. Liu F, Park PJ, Lai W, Maher E, Chakravarti A, et al. (2006) A genome-wide screen reveals functional gene clusters in the cancer genome and identifies EphA2 as a mitogen in glioblastoma. *Cancer Res* 66: 10815-10823.
7. Chari R, Coe BP, Wedseltoft C, Benetti M, Wilson IM, et al. (2008) SIGMA2: a system for the integrative genomic multi-dimensional analysis of cancer genomes, epigenomes, and transcriptomes. *BMC Bioinformatics* 9: 422.
8. Pollack JR, Sørlie T, Perou CM, Rees CA, Jeffrey SS, et al. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A* 99: 12963-12968.
9. Lee H, Kong SW, Park PJ (2008) Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes. *Bioinformatics* 24: 889-896.
10. Glinsky GV (2006) Integration of HapMap-based SNP pattern analysis and gene expression profiling reveals common SNP profiles for cancer therapy outcome predictor genes. *Cell Cycle* 5: 2613-2625.

11. Cancer Genome Atlas Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455: 1061-1068.
12. Vogelstein B, Kinzler KW (2004) Cancer genes and the pathways they control. *Nat Med* 10: 789-799.
13. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, et al. (2007) The genomic landscapes of human breast and colorectal cancers. *Science* 318: 1108-1113.
14. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, et al. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 34: 267-273.
15. Bild AH, Yao G, Chang JT, Wang Q, Potti A, et al. (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439: 353-357.
16. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102: 15545-15550.
17. Maglietta R, Distaso A, Piepoli A, Palumbo O, Carella M, et al. (2010) On the reproducibility of results of pathway analysis in genome-wide expression studies of colorectal cancers. *J Biomed Inform* 43: 397-406.
18. Khatri P, Draghici S, Ostermeier GC, Krawetz SA (2002) Profiling gene expression using onto-express. *Genomics* 79: 266-270.
19. Newton M, Quintana F, den Boon J, Sengupta S, Ahlquist P (2007) Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Ann Appl Stat* 1: 85-106.
20. Maglietta R, Piepoli A, Catalano D, Licciulli F, Carella M, et al. (2007) Statistical assessment of functional categories of genes deregulated in pathological conditions by using microarray data. *Bioinformatics* 23: 2063-2072.
21. Holden M, Deng S, Wojnowski L, Kulle B (2008) GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics* 24: 2784-2785.

22. Edelman E, Porrello A, Guinney J, Balakumaran B, Bild A, et al. (2006) Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. *Bioinformatics* 22: e108-e116.
23. Bild A, Febbo PG (2005) Application of a priori established gene sets to discover biologically important differential expression in microarray data. *Proc Natl Acad Sci U S A* 102:15278-15279.
24. Caruana R (1997) Multitask Learning. *Machine Learning* 28: 41-75.
25. Evgeniou T, Pontil M (2004) Regularized Multi-Task Learning. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* 1: 109-117.
26. Bakker B, Heskes T (2003) Task clustering and gating for bayesian multitask learning. *Journal of Machine Learning Research* 4: 83-99.
27. Vapnik V (1998) *Statistical Learning Theory*. New York: J. Wiley and Sons. 736 p.
28. Osborne CK. (1998) Steroid hormone receptors in breast cancer management. *Breast Cancer Res Treat* 51: 227-238.
29. Stoica GE, Franke TF, Moroni M, Mueller S, Morgan E, et al. (2003) Effect of estradiol on estrogen receptor-alpha gene expression and activity can be modulated by the ErbB2/PI 3-K/Akt pathway. *Oncogene* 22: 7998-8011.
30. List HJ, Lauritsen KJ, Reiter R, Powers C, Wellstein A, et al. (2001) Ribozyme targeting demonstrates that the nuclear receptor coactivator AIB1 is a rate-limiting factor for estrogen-dependent growth of human MCF-7 breast cancer cells. *J Biol Chem* 276: 23763-23768.
31. Shatalova EG, Walther SE, Favorova OO, Rebbeck TR, Blanchard RL (2005) Genetic polymorphisms in human *SULT1A1* and *UGT1A1* genes associate with breast tumor characteristics: a case-series study. *Breast Cancer Res* 7: R909-921.
32. Furr BJ, Jordan VC (1984) The pharmacology and clinical uses of tamoxifen. *Pharmacol Ther* 25: 127-205.
33. Nowell SA, Ahn J, Rae JM, Scheys JO, Trovato A, et al. (2005) Association of genetic variation in tamoxifen-metabolizing enzymes with overall survival and

- recurrence of disease in breast cancer patients. *Breast Cancer Res Treat* 91: 249-258.
34. Tartter PI, Papatestas AE, Ioannovich J, Mulvihill MN, Lesnick G, et al. (1981) Cholesterol and obesity as prognostic factors in breast cancer. *Cancer* 47: 2222-2227.
 35. Koziczak M, Holbro T, Hynes NE (2004) Blocking of FGFR signaling inhibits breast cancer cell proliferation through downregulation of D-type cyclins. *Oncogene* 23: 3501-3508.
 36. Lochter A, Bissell MJ (1995) Involvement of extracellular matrix constituents in breast cancer. *Semin Cancer Biol* 6: 165-173.
 37. Drell TL 4th, Joseph J, Lang K, Niggemann B, Zaenker KS, et al. (2003) Effects of neurotransmitters on the chemokinesis and chemotaxis of MDA-MB-468 human breast carcinoma cells. *Breast Cancer Res Treat* 80: 63-70.
 38. Chen J, Gammon MD, Chan W, Palomeque C, Wetmur JG, et al. (2005) One-carbon metabolism, MTHFR polymorphisms, and risk of breast cancer. *Cancer Res* 65: 1606-1614.
 39. Schölkopf B, Mika S, Burges CC, Knirsch P, Müller KR, et al. (1999) Input space versus feature space in kernel-based methods. *IEEE Trans Neural Netw* 10: 1000-1017.
 40. Xiong Q, Ancona N, Hauser ER, Mukherjee S, Furey TS (2012) Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. *Genome Res* 22: 386-397.
 41. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81: 559-575.
 42. Denton D, McBurnie M, Ong F, Osborne P, Tarjan E (1988) Na deficiency and other physiological influences on voluntary Na intake of BALB/c mice. *Am J Physiol* 255: R1025-1034.
 43. Bott E, Denton DA, Weller S (1965) Water drinking in sheep with oesophageal fistulae. *J Physiol* 176: 323-336.

44. Thrasher TN, Nistal-Herrera JF, Keil LC, Ramsay DJ (1981) Satiety and inhibition of vasopressin secretion after drinking in dehydrated dogs. *Am J Physiol* 240: E394-401.
45. Liedtke WB, McKinley MJ, Walker LL, Zhang H, Pfenning AR, et al. (2011) Relation of addiction genes to hypothalamic gene changes subserving genesis and gratification of a classic instinct, sodium appetite. *Proc Natl Acad Sci U S A* 108: 12509-12514.
46. Morris MJ, Na ES, Johnson AK (2010) Mineralocorticoid receptor antagonism prevents hedonic deficits induced by a chronic sodium appetite. *Behav Neurosci* 124: 211-224.
47. Li CY, Mao X, Wei L (2007) Genes and (common) pathways underlying drug addiction. *PLoS Comput Biol* 4: e2.
48. Santini E, Valjent E, Usiello A, Carta M, Borgkvist A, et al. (2007) Critical involvement of cAMP/DARPP-32 and extracellular signal-regulated protein kinase signaling in L-DOPA-induced dyskinesia. *J Neurosci* 27: 6995-7005.
49. Aston-Jones G, Smith RJ, Moorman DE, Richardson KA (2009) Role of lateral hypothalamic orexin neurons in reward processing and addiction. *Neuropharmacology* 56(Suppl 1): 112-121.
50. Feller W (1971) *An Introduction to Probability Theory and Its Applications, Volume 1*. New York: J. Wiley and Sons. 704 p.
51. Abatangelo L, Maglietta R, Distaso A, D'Addabbo A, Creanza TM, et al. (2009) Comparative study of gene set enrichment methods. *BMC Bioinformatics* 10: 275.
52. Kulis M, Esteller M (2010) DNA methylation and cancer. *Adv Genet* 70: 27-56.
53. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661-678.
54. Shete S, Hosking FJ, Robertson LB, Dobbins SE, Sanson M, et al. (2009) Genome-wide association study identifies five susceptibility loci for glioma. *Nat Genet* 41: 899-904.
55. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, et al. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res* 36: D480-484.

56. Vastrik I, D'Eustachio P, Schmidt E, Gopinath G, Croft D, et al. (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* 8: R39.
57. Gruvberger S, Ringnér M, Chen Y, Panavally S, Saal LH, et al. (2001) Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res* 61: 5979-5984.
58. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 18: S96-104.
59. Mardia KV, Kent JT, Bibby JM (1979) *Multivariate Analysis*. London: Academic Press. 521 p.
60. Yano S, Brown EM, Chattopadhyay N (2004) Calcium-sensing receptor in the brain. *Cell Calcium* 35: 257-264.
61. Nowak A, Bojanowska E (2008) Effects of peripheral or central GLP-1 receptor blockade on leptin-induced suppression of appetite. *J Physiol Pharmacol* 59: 501-510.
62. Gutzwiller JP, Hruz P, Huber AR, Hamel C, Zehnder C, et al. (2006) Glucagon-like peptide-1 is involved in sodium and water homeostasis in humans. *Digestion* 73: 142-150.
63. Linardic CM, Downie DL, Qualman S, Bentley RC, Counter CM (2005) Genetic Modeling of Human Rhabdomyosarcoma. *Cancer Res* 65: 4490-4495.
64. Morotti RA, Nicol KK, Parham DM, Teot LA, Moore J, et al. (2006) An immunohistochemical algorithm to facilitate diagnosis and subtyping of rhabdomyosarcoma: the Children's Oncology Group experience. *Am J Surg Pathol* 30: 962-968.
65. Kirsch DG, Dinulescu DM, Miller JB, Grimm J, Santiago PM, et al. (2007) A spatially and temporally restricted mouse model of soft tissue sarcoma. *Nat Med* 13: 992-997.
66. Baird K, Davis S, Antonescu CR, Harper UL, Walker RL, et al. (2005) Gene expression profiling of human sarcomas: insights into sarcoma biology. *Cancer Res* 65: 9226-9235.

67. Celis JE, Kruhøffer M, Gromova I, Frederiksen C, Ostergaard M, et al. (2000) Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. *FEBS Lett* 480: 2-16.
68. Kendziorski CM, Chen M, Yuan M, Lan H, Attie AD (2006) Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics* 62: 19-27.
69. Hollander M, Wolfe DA (1999) *Nonparametric Statistical Methods*. New York: Wiley. 787 p.
70. Tarca AL, Carey VJ, Chen XW, Romero R, Drăghici S (2007) Machine learning and its applications to biology. *PLoS Comput Biol* 3: e116.
71. Schölkopf B, Tsuda K, Vert JP (2004) *Kernel Methods in Computational Biology*. Cambridge: MIT Press. 410 p.
72. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, et al. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16: 906-914.
73. Edelman EJ, Guinney J, Chi JT, Febbo PG, Mukherjee S (2008) Modeling cancer progression via pathway dependencies. *PLoS Comput Biol* 4: e28.
74. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, et al. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 31: e15.
75. Evgeniou T, Micchelli CA, Pontil M (2005) Learning Multiple Tasks with Kernel Methods. *J Mach Learn Res* 6: 615-637.

Biography

Brian D. Bennett was born on March 14, 1986 and grew up in Westminster, MD. He completed his undergraduate education at the University of Maryland, Baltimore County in May 2008. He received a Bachelor of Science in Bioinformatics and Computational Biology with minors in Computer Science and Statistics, graduating Magna Cum Laude. In August 2008, he entered the Computational Biology and Bioinformatics program at Duke University. He later joined the lab of Terry Furey with Sayan Mukherjee as a co-advisor. His research has resulted in the following publication:

Bennett BD, Xiong Q, Mukherjee S, Furey TS (2012) A Predictive Framework for Integrating Disparate Genomic Data Types Using Sample-Specific Gene Set Enrichment Analysis and Multi-Task Learning. *PLoS One* 7:e44635.