Gene Set Enrichment and Projection: A Computational Tool for Knowledge Discovery in Transcriptomes

Karl Douglas Stamm
Marquette University

GENE SET ENRICHMENT AND PROJECTION: A COMPUTATIONAL TOOL FOR KNOWLEDGE DISCOVERY IN TRANSCRIPTOMES

by

KARL D. STAMM, M.S.

A Dissertation submitted to the Faculty of the Graduate School,
Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy

Milwaukee, Wisconsin

August 2016

ABSTRACT
GENE SET ENRICHMENT AND PROJECTION: A COMPUTATIONAL TOOL FOR KNOWLEDGE DISCOVERY IN TRANSCRIPTOMES

KARL D. STAMM, M.S.

Marquette University, 2016

Explaining the mechanism behind a genetic disease involves two phases, collecting and analyzing data associated to the disease, then interpreting those data in the context of biological systems. The objective of this dissertation was to develop a method of integrating complementary datasets surrounding any single biological process, with the goal of presenting the response to a signal in terms of a set of downstream biological effects. This dissertation specifically tests the hypothesis that computational projection methods overlaid with domain expertise can direct research towards relevant systems-level signals underlying complex genetic disease. To this end, I developed a software algorithm named Geneset Enrichment and Projection Displays (GSEPD) that can visualize multidimensional genetic expression to identify the biologically relevant gene sets that are altered in response to a biological process.

This dissertation highlights a problem of data interpretation facing the medical research community, and shows how computational sciences can help. By bringing annotation and expression datasets together, a new analytical and software method was produced that helps unravel complicated experimental and biological data.

The dissertation shows four coauthored studies where the experts in their field have desired to annotate functional significance to a gene-centric experiment. Using GSEPD to show inherently high dimensional data as a simple colored graph, a subspace vector projection directly calculated how each sample behaves like test conditions. The end-user medical researcher understands their data as a series of somewhat-independent subsystems, and GSEPD provides a dimensionality reduction for high throughput experiments of limited sample size. Gene Ontology analyses are accessible on a sample-to-sample level, and this work highlights not just the expected biological systems, but many annotated results available in vast online databases.

ACKNOWLEDGEMENTS

Karl D. Stamm, M.S.

I need to thank everyone who brought me to this point in life. My family and friends have been supporting my academic endeavor for years. My co-workers at the Medical College of Wisconsin have instrumentally guided my path, and my mentors at Marquette University set the bar for excellence.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND OUTLINE

Explaining the mechanism behind any genetic disease involves two phases, collecting and analyzing genetic data, then interpreting those data in the context of biological systems. While advances have come to both phases, few tools facilitate both phases. My objective is to develop a method integrating several datasets surrounding a biological process, with the goal of presenting the response to a test condition in terms of a set of downstream biological effects. I hypothesize that heuristic methods overlaid with automated domain expertise will highlight relevant systems-level signals underlying complex biological conditions that have not been seen by previous methods. To this end, I have developed an algorithm called Geneset Enrichment and Projection Displays (GSEPD), which visualizes multidimensional gene expression data to identify the biologically-relevant gene sets that are altered in response to a test condition. With inspiration from state of the art model organism experiments, *in vitro* experiments, and genome sequencing, I present a tool that can help to explain the cellular mechanisms of a complex test condition.

1.1 Introducing the Problem Context

Medical researchers are interested in the diseases with as-yet unknown causes. An interesting and longstanding research area is human birth defects, broadly defined as a

problem present in a newborn [1]. Early genetic screening indicated that “not less than 4% of all live births” are impacted by one or more genetic diseases [2]. Presuming a medical condition is caused by genetic or environmental factors, or their combination, researchers seek to discover the shared cause among independent patient cases. For some medical conditions a common environmental factor can be found through retrospective studies, but in the situation of a spontaneous/sporadic birth defect where common environmental causes are ruled-out, researchers could search for a mutation impacting a developmental pathway. Organ development has not been understood completely; only some vital genetic components have been identified [3, 4]. Master control points are known from mutagenesis studies of model organisms [5] but subtle effects on complex organs like the human brain or heart are still unknown [6].

Over 100 genes are required to work together to construct a mammalian heart [7] and altering any one will cause some form of a visible final trait [8, 9]. Each separate gene malfunction can be considered a subtype of the visible trait, and as each gene may malfunction in several ways, a few hundred causally-distinct subtypes of the same visible trait are generated. The presence of multiple subtypes causes a problem for statistical analyses, as each portion of genetic-location specific evidence is relevant for only a subset of the cases with a common disease, thereby decimating the power of a case/control study [10, 11].

Genetic analyses usually are performed by examining “case” and “control” groups of subjects. Researchers collect genetic information and analyze the genetic information for commonalities that segregate the disease case group from the healthy control group [12]. The data collection takes several technical forms depending on the technological and

social platforms. Gene usage platforms such as sequencing and microarray assays are briefly discussed below, each having different strengths, weaknesses, and costs. When studying model organisms, one might create controlled animals, but when studying human traits researchers need to work with public health organizations or recruit volunteers. The prevalence of a disease and the affected population determine the costs and difficulties encountered in collecting samples. For instance, where something like high blood pressure affects millions of adults, it is possible to collect tens of thousands of willing participants to give a blood sample. Conversely, to directly study organ development researchers need to obtain fetal tissue, or work backwards from subjects who have completed organ development. Ethical considerations generally make direct study of healthy human tissue impossible, and researchers are limited to incomplete data and must piece together information wherever it is available.

The simplest solution to the problem of hard to collect samples lies in model organisms. Model organisms are the inbred and outbred lab rats, mice and lower animals, with the simplest model organisms being yeasts and bacteria. All of these model organisms share some biological systems with *Homo sapiens* from which we can learn human-relevant insights. For example, below I present a work involving mice that is relevant to heart development, as the mouse heart is very similar to a human's [13]. However, molecular discrepancies exist between mouse and human too, causing difficulties for analysis and interpretation of results [14]. Direct measurements of human tissues are key, and I present some data on the heart development of humans in **Chapter 2**.

There exist decades of statistical research in genetic analysis of single-gene traits [15, 16]. A trait that has not been explained by single-gene analyses can safely be assumed to require multiple interacting factors. Organ development is known to require many genomic factors to proceed correctly [1]. Through natural selection, modern life has developed redundant systems and safeguards against breakdown of genetic systems [17], and modern research identifies redundant systems that have evolved to work around deficiencies [18]. Research focus is shifting to networks of genes because many biochemical pathways depend on one another, and some proteins fill multiple roles [19]. A phenotype may not manifest when a single gene is changed. This makes detecting the genetic cause more difficult because a mutation seen in one gene may not wholly cause the disease in any one subject's particular genetic background [20]. The notion of *penetrance* of a gene variant captures the probability carrying the mutation causes the disease.

As illustrated by breast cancer, often a genetic mutation causes an increase in probability of disease without directly influencing the visible trait 100% of the time [21]. Incomplete penetrance forces researchers to collect probabilistic results across ever-increasing subject counts. Any mutation in an active gene can be assumed to have some measurable effect on some biological systems. Observing the apparent reaction to a mutation should reveal the activity of related systems [22]. To short-circuit the statistical power problems present in human disease studies of single mutations, my work is directed at biological pathways [23]. If a case/control analysis can be reframed as a perturbed system reaction, then medical researchers can learn about impacted components of the systems involved with the case/control contrast.

My work specifically is to facilitate the endeavors of biological research labs mired in copious, complex, and incomplete gene expression data. Computational Sciences must be brought to bear in an intelligent way to improve research efficiency. Computational time efficiency is not the major factor preventing genetic discoveries: biomedical research is in need of intelligently directed and data driven hypothesis generating tools. Here I present how to merge disparate data types towards the overarching goal of explaining systems that have been opaque to traditional research.

As example datasets, I focus on heart development. Congenital heart disease (CHD) is just one of many examples of a situation that has not been solved by traditional experimentation, but computer scientists could help with modern data mining and processing techniques. Instead of directly assaying every genetic variation, I propose a broad systematic search and refinement system that layers the knowledge we do have into a coherent picture. Specifically, I have developed free and open source software that automates the conversion of gene expression data from human samples into a perturbed-systems overview, targeting the interpretation and knowledge extraction problems inherent in low-sample-count studies.

1.2 Dissertation Outline

In this dissertation I present 1) background studies motivating the methods, 2) software for mining useful knowledge out of imperfect experimental data, and 3) applications to published studies where systems-level results were sought. The culmination of these three sections is a new method for determining altered systems in complex experimental data as alluded to in **Figure 1.1**.

In this work three independent data types are brought together and two major tools are developed. Application of the developed software herein highlights systems level expression patterns in genomic datasets, with the objective of directing further research in novel directions.

Starting on the leftmost column of **Figure 1.1** are three major unique data types accompanying genetic association analyses. "Structural Alteration" refers to the data obtained by the analysis of large-scale genomic alterations, and serves as a solid foundation for the idea that a set of genes can collectively be responsible for disease phenotypes [24, 25].

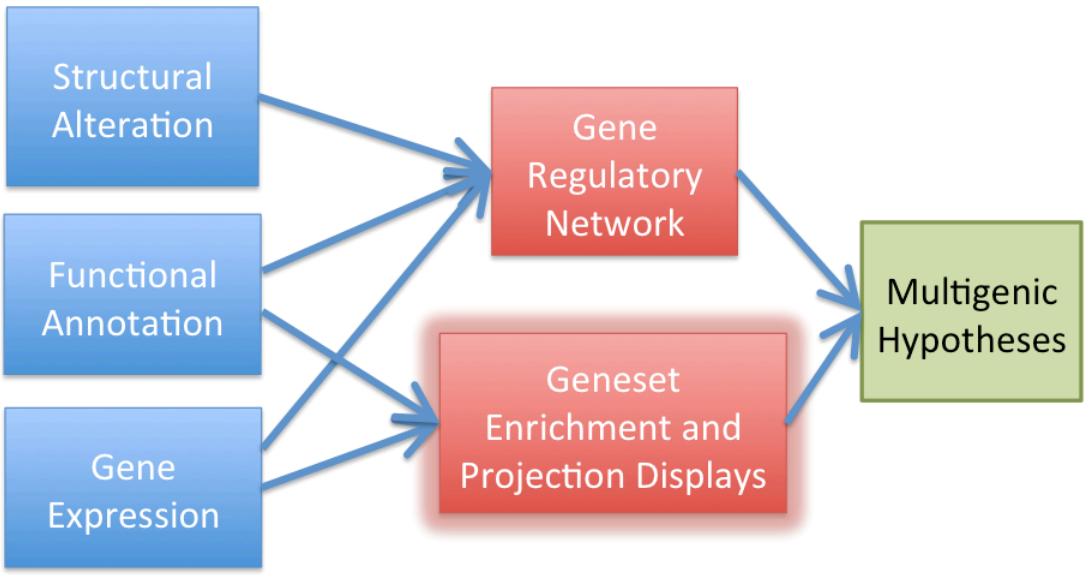

Figure 1.1 Flowchart of Concepts Presented in this Dissertation. Blue boxes represent data or information sources, red boxes are published knowledge-generating tools I and collaborators have produced, with the overall goal is presented in the green box at right. Arrows represent information flow, from primary sources in blue, through toolkits in red, towards the goal in green.

"Functional Annotation" refers to the *apriori* domain specific knowledge obtained from published studies. Primarily Gene Ontology [26], but also the Ingenuity Pathway database [27] are used in my work to mark novel findings in the context of prior information.

"Gene Expression" refers to data sets of mRNA abundance, or transcriptomes. A transcriptome may be measured from any tissue sample or cell culture, and represents a high-dimensional data type. The primary algorithm "Geneset Enrichment and Projection Displays" is developed to help analyze and interpret transcriptomic datasets.

The upper red box, "Regulatory Network," refers to a novel method and application of identifying gene expression dependencies. In **Chapter 2** I describe two studies where we built a set of genes relevant for organ development, show their impact, then build a network of putative gene-gene interactions and refine the interactions listing to testable novel hypotheses.

The lower red box, "Geneset Enrichment and Projection Displays," is a novel algorithm developed herein (abbreviated GSEPD), specifically aiming to aid biological research labs in interpreting their gene expression data. By automating best-practice statistical analyses, then recombining data in a novel way, we can understand genetic-pathway-based sample behavior. GSEPD is the focus of **Chapter 3**.

Both of these tools are brought together on the right of **Figure 1.1** and in **Chapter 4**, where we explore specific case studies in both commercially available stem cells and human CHD. I present how these tools shed more light on more complex situations than was previously possible.

1.3 Motivation to Data Refinement

Observational sciences often work with uncontrolled variability with under-specified hypotheses, and researchers demand computational support to find signal among the noise. The scale of new measurement technologies in biology is overwhelming: a quantitative measure is available associated to every one of the nearly four billion unique locations within a person's DNA. Because no researcher can review each and every finding manually, computationally filtering and prioritizing results is key. If not conducted carefully, genome-wide association analyses may inadvertently lead to errors, biases and misunderstandings. Therefore a very high degree of certainty (such as $p < 10^{-6}$) is required and often no significant results are achieved [28].

When no single factor can be pinned as the global cause of a disease/trait, we need to consider multi-factor hypotheses. Multi-factor association is susceptible to combinatoric explosion [29] where the number of evaluated variables corresponds to the dimensionality of the search space. For example, there are only twenty ways to choose one gene from twenty genes, but there are 1,140 ways to choose three genes from twenty genes. Acknowledging the unfeasibly large scope of searching for pairs or triples of genes in the whole genome, my strategy is a heuristic search, using informed priors to boost the success rate in a restricted *relevant* space.

One may move beyond the DNA to the next level: gene expression, where other kinds of analysis can shed light on what is and is not relevant for whichever process being studied. Recent advances in data acquisition technologies around genomics are starting to shed light on unexplored biological systems, at unprecedented data volumes [30]. New

datasets and analysis methods have formed the field of bioinformatics, wherein computer science meets biology. Modern high-performance computing resources are often required to process and review the results of routine genomic analyses in timely fashion [31, 32]. The genomic approaches mentioned above are used routinely by teams of scientists and healthcare providers, but improved methods are required to better discover knowledge from the volume of data obtained.

1.3.1 Forward Development

Standard genetic analysis methods have focused on controllable experiments in yeasts, mice, or simulation. Human developmental genetic analysis is ripe for computational advancement. Human tissues studies in particular struggle with restricted sample collection opportunities, and restricted experimental designs while supporting high levels of complexity in data and outcome [33].

Low-dimensional outcomes are required for statistical power, but the transcriptomic measurements are inherently high dimensional [34]. Recognizing the multiple testing problem inherent in a high dimensional test, we turn to intelligently biased strategies. Gene sets are the relevant determinant of physiological outcome [35], but searching all possible sets for an association signal guarantees spurious results. Therefore I propose searching restricted spaces with prior knowledge that certain genes are relevant to the organ function [36], or certain gene tuples which are known to interact [37]. Instead of reinventing the wheel, we can proceed from the state of the art by building on curated knowledgebases available online [26, 38].

1.3.2 Relevant Hypothesis Generation

Exploratory or observational studies are considered hypothesis-generating in that they are not designed to test any precise statistical statement, but rather to give an overview and direct more precise future studies. A genome-wide search can produce more results than can be followed up on, so restricting the reported results to those with likelihood of usability would be helpful. Regarding human disease studies, note that every human sample has tens of thousands of unique mutations [39, 40], such that researchers need guidance in the form of curated knowledgebases to sift through the results and find the real cause of disease or tested condition.

1.4 Summary

Bioinformatics covers many disciplines: scaled at its smallest, protein atomic structures, at its largest, writing the tree of life. I see the study of organ development as a high dimensional physical system that is exciting to work with, and an opportunity to give fruitful advances to medicine. The software **rgsepd**, (implementing the method GSEPD) can help make discoveries in human tissues studies such as inborn disease, cancers, drug treatments, stem cells, or trauma and healing situations.

Presently it is not exactly known how organs are formed, or why this process goes wrong sometimes. Our best organ regeneration methods reanimate cellular skeletons with reprogrammed cells [41]. The challenge lies in monitoring ever-smaller and more precious subjects and the destructive measurement process. Only very recently has anyone measured gene usage in time series for a gestational mammal's heart [13], and we subsequently conducted a study to see what kind of gene-gene causality inferences we

could make [7]. To assuage the great cost of bio-specimen acquisition, many studies are designed around the bare minimum number of samples able to be obtained and processed. The cost and sample size problems create a field of independent laboratories working over complex data with simple tools. *My goal is to create automatable intelligence to accelerate knowledge discovery.*

The goal of this dissertation is to highlight the biological research community's need for data analysis support, and show how computational sciences can help by developing automated software tools. Primarily by integrating high-dimensional datasets, I have produced a new analytical software to accelerate knowledge discovery in transcriptomes. An overview of what follows in the remaining chapters is concisely revisited below.

In **Chapter 2** I reinforce the method of searching for clues in complementary data types in both human and mouse. **Chapter 2** explores two biological studies to set the context and motivation of the GSEPD method. In one study, I quantified the genomes of humans with and without congenital heart disease. That study of human genomes solidifies the concept that some experimental conditions are necessarily uncontrolled [25]. In the second study, I build a gene regulatory network from a time series gene expression data set in mouse heart ventricle development. The gene regulatory network study shows the importance of uniquely verifiable results as distinct from the bulk of possible results [7]. These two studies' result set overlap reveals a core of genetic effects necessary for mammalian life, highlighting the utility of independent data types, and the associated needs for computational analysis to integrate such data types, regardless of the studied condition.

In **Chapter 3** I briefly review transcriptomics: the measuring and exploration of the messenger transcripts responsible for gene usage. New technologies are revealing more detail in transcriptomes (such as exon splicing probabilistic effects), and analytical methods have fallen behind the needs of biomedical researchers. The standard method is to compare samples in batches and collect a list of the dimensions with the most difference between two classes, then to query the set of dimensions for meaning. This two-step process of collecting genes and querying the set for biological implication is highly error prone and does not use the full scope of information present in a gene expression data set. I attempt to remedy these concerns by introducing a new software on Bioconductor named **rgsepd** that performs GSEPD: a complete transcriptome analysis from raw sequencing read counts through a novel bio-pathway perturbation detection. GSEPD is novel in that it simplifies the workflow for a biomedical researcher by identifying segregating gene sets between test conditions. I have published **rgsepd** as an open-source toolkit on Bioconductor to let everyone access these automated knowledge discovery methods.

In **Chapter 4** I describe two case studies, their original manual results with limited functional findings, and systematic application of both GSEPD and another popular functional analysis tool to highlight what further functional findings are possible with data-driven tools. The first case study follows stem cell differentiation, tracking the steps nature uses to generate pumping muscle cells. The second case study involves the CHD cohort at the Children's Hospital of Wisconsin and the genomic analyses that have been performed to root out the cause of one type of CHD. In each case study three sets of results are presented, 1) as originally published by the domain expert, 2) as possible with a popular functional analysis tool, and 3) as possible with GSEPD.

In **Chapter 5** I briefly summarize the findings from GSEPD for the application above. Future directions of this avenue of research are identified.

CHAPTER 2

BACKGROUND AND MOTIVATION: COMPLEMENTARY DATASETS

This chapter covers two published studies where disparate data types that complement one another were used to gather a more complete understanding of a shared system. An analysis of the structural alterations in human genomes with known cardiac status is paired with an analysis of gene regulatory networks in mouse cardiac development. The findings of either study are particular to its context, and careful computational integration is required to achieve efficient knowledge discovery [42]. The term "complementary" is used to clarify that each study adds to the other synergistically in information content. The two studies are incorporated here to highlight practical challenges of data processing that guided the development of GSEPD.

The first study, CNV2012, is a study of genomes in CHD patients, resulting in a set of associations between genes and congenital heart defects [25]. Gene associations are moderately useful on their own, reproducing previous studies and evaluating a population for its rate of genetic defect. CNV2012 is included here to show that human populations carry large-scale genomic differences, so any tool development should account for this sample-specific heterogeneity.

The second study, GRN2014, is where we built upon the gene associations from CNV2012. Seeding from the association results in CNV2012, I built a regulatory gene

network in collaboration with The Mayo Clinic and Medical College of Wisconsin Biotechnology and Bioengineering Center. We predicted a set of verifiable gene-gene interactions found in gestational mouse heart. The predicted gene-gene interactions can direct future experimentalists toward hypotheses that are more likely to be true than those in a full genome-wide search. GRN2014 is included in this dissertation to show the complexity possible in a field like mammalian organ development, and to reiterate the need for the more intelligent data usage developed in **Chapter 3**, such as the usage of pre-defined gene sets.

2.1 Candidate Gene Lists and Copy Number Variants

In 2012 we studied a population of healthy and sick children for their prevalence of a class of mutation on a candidate gene list to demonstrate and confirm that class of mutations is important in the development of the disease [25].

Candidate gene listing is a common method to trim the experimental scope such that the new finding is almost guaranteed to have interpretable results [36]. A major drawback of the candidate gene listing method is the limitation in the search space. A study using a candidate gene list can only find results among genes on the list. Building on established science increases the interpretability of new findings while decreasing the ability to discovery truly new effects [43].

Most genome-wide analyses do not find significant associations [44] unless they are restricted to a few dozen genes of relevance, such that the multiple testing correction penalty is not as high [38, 45]. As statistical power is linked to subject counts,

unrestricted searches require large numbers of independent samples [46, 47], which may not be possible with rare diseases or limited budgets.

Even with nearly a thousand subjects, no significant association was found with a genome-wide search. To narrow the scope, a candidate gene list was developed with literature review. "The 100 gene list" of those known to be important for heart development was compiled. Table 1 of [12] indicates 85 cited studies that were reviewed. We were interested in a particular class of genetic aberration that had not yet been thoroughly investigated, the copy number variant. A copy number variant, or CNV, is a structural alteration which is difficult to detect by high throughput analyses [48, 49], so a study had not been performed investigating their incidence among cardiac development cases.

A CNV is a segment of the genome that is non-diploid. Where most of the human genome is expected to have unique sequence, one copy from each parent, non-diploid regions are those that somehow become single copy or 3 or more copies. This is a normal mutational mode and a mechanism of evolution, but like most mutations, a CNV is most often either silent or detrimental. Our study detected CNVs with a microarray technology, described in more detail in the literature [25]. Further discussion of CNV detection methods is outside the scope of this dissertation.

Point mutations that damage the gene are known from other instances [50] but are difficult to assay in a large cohort. The CNV analysis was an affordable way to broadly assay the whole genomes of the cohort. We used Affymetrix GenomeWide SNP 6.0 assays, with over one million probes each.

After collection of samples, they were analyzed as a batch to account for platform biases. CNV segments were called genome-wide and annotated by their underlying gene impacts. The number of CNVs found in both a healthy population and a congenital heart defect cohort are comparable. An interesting takeaway is that both cases and controls show similar counts of CNV losses, the difference being which genes are impacted. Affected individuals have nominally more gain segments, but even unaffected individuals tend to have >5 segments, indicating some level of genomic instability is normal.

2.1.2 Six Subtypes Syndromic – Phenotypic Refinement

In the CNV2012 study, there were 958 subjects with various forms of CHD. If all 958 shared a single causative genetic defect, it would be apparent. However, no single factor was identified, and the implication is that the evidence is spread across several different causes. To address the issue of multiple potential causes, the subjects were labeled with 42 separate diagnoses by domain experts, which were regrouped into six categories. The individual diagnoses are categories summarizing unique individual variation by defined codes (specifically the EPCC 2011 coding scheme). The six categories are “syndromic” in that they represent major named conditions. We hypothesized the categories would show similar genetic burden profiles. We showed the prevalence of various CNVs in various groupings of subjects.

There is reason to believe similar phenotypes have similar causes, i.e. heart valve defects having shared genes. In biological genetics it is known that one or more genes are responsible for various physical systems, and the genes’ perturbations yield various related outcomes [51]. Continuity of causation is key to the belief that gene sets can inform complex outcomes [36].

A Fisher's exact test identified 21 genes as significantly differentially impacted by CNV among cases and controls. This re-verification of candidate genes cements their importance as cardiac development products. Previous indications for each gene came from varied platforms and study populations; no genome-wide analysis had been performed.

A result of CNV2012 was the "spectra" or frequency sets of CNVs at each analyzed gene. **Figure 2.1** notes the rates of gain and loss in each as a frequency with standard error bars and chromosomal location. The results indicate genes may be present in multiple copies or deleted entirely in both sick and healthy individuals.

The major takeaways from the CNV study for us are the use of a literature-review sourced candidate gene list as a technique, and that precisely phenotyped results are a key to precision. Tables in the CNV manuscript list the genes identified by literature review, as well as the CNV findings verifying their association.

The number of gene copies drives the amount of gene-product produced, and is therefore referred to as a gene's *dosage*. Higher or lower than normal dosage results in an amplification or reduction in the available proteins, which may cause disease [52]. The abnormal gene dosage may not always lead to disease, due to the body's regulatory systems that control gene expression [19]. These systems are referred to as "regulatory networks" as proteins, genes, and DNA interact with each other to coordinate development [1, 53].

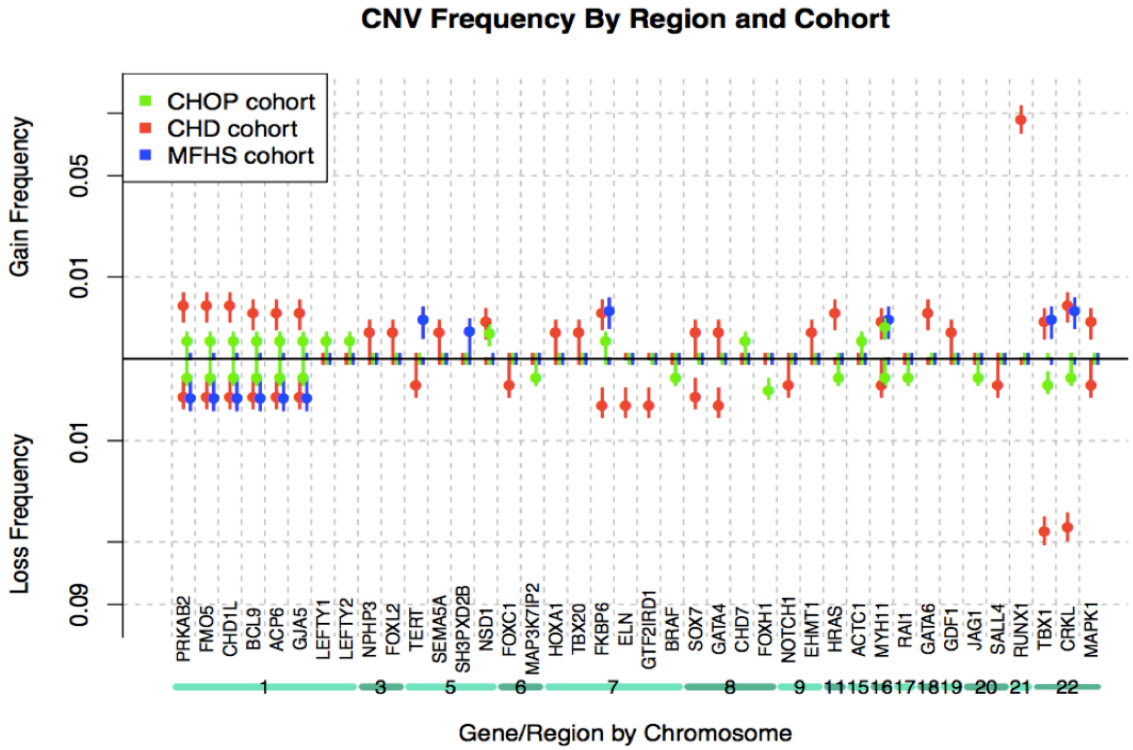

Figure 2.1 CNV Frequency Spectra. Calculated incidence rates of CNV gains and losses in three cohorts. In red, CHD cohort is the Children's Hospital of Wisconsin congenital heart disease group. In green, CHOP cohort is the Children's Hospital of Philadelphia healthy group and can be considered young controls. In blue, MFHS cohort is the Milwaukee Family Heart Study healthy group and can be considered elderly controls. Vertical error bars represent one standard error from the mean in the estimated sampling distribution. For example, gains over gene *FKBP6* (chr7, just left of center) occur in all three cohorts, while losses of the same gene are only seen in the CHD cohort, implying a loss could cause CHD. The red point near the top of the diagram shows an 8.5% rate of copy number gains at gene *RUNX1* in the CHD cohort, mostly indicative of Trisomy 21 patients.

2.2 The GRN2014 Study

The second example source of a complementary dataset comes from our GRN2014 study [7]. There we built and refined a model of the gene regulatory effects necessary to construct a mammalian heart. That study constructed a regulatory network, and involves

model organisms, small sample sizes, tissue specificity, time course dynamics, and uncertainty. Our focus was on providing relevant verifiable results. The first step produced millions of putative gene-gene regulatory interactions, which were later pared down to those that are identifiable, testable, and likely. The result is a shortlist of actionable insights. We estimated correctness via overlap to a gold standard in gene-gene interaction, Ingenuity Pathway Analysis [54].

2.2.1 Mouse Heart Development Dataset

Thanks to a partnership with the Todd and Karen Wanek Foundation for Hypoplastic Left Heart, in 2014 the Mitchell Lab at The Medical College of Wisconsin was part of a data-sharing agreement with the Nelson Lab at Mayo Clinic. We had early access to a unique dataset. Starting from a collection of mouse embryonic tissues in time-course, we had access to the gene expression dynamics of the developing heart for the first time [13].

A time series gene expression dataset is distinct from a case/control study in that we have sequential measurements of transcriptome and a notion of causality. By measuring the same organ's transcriptome at various stages of development, we can see gene expression evolve through the course of the experiment. We know that transcriptomic analyses are sensitive to the cell type peculiarities, so measuring a heart as it grows is key to learning about that process.

The process by which cells grow and divide is regulated by the chemicals present within them [1, 55, 56]. A gene regulatory network (GRN) is a graphical representation of the factors that activate or deactivate each other at the gene level. The network is a graphical representation where genes correspond to vertices, and interactions are

represented by edges. The GRN2014 study was an attempt to deduce the GRN of the developing mouse heart by numerically exploring the dynamics and modeling multidimensional differential equations [7, 57].

2.2.2 Method

Our analysis used a parallel computational technique to fit models to the data, and report which genes seem like they could be regulators of other genes based solely on their time course dynamics [57]. For example, gene A increasing one time step before gene B's rise is mild evidence for A driving B. This is known to be a hard problem computationally because every gene may impact every other. A network (vertex graph) is created by default with every possible edge present, and the edges can be evaluated for significant coefficients. This is computationally intractable in a genome-wide sense, and with limited data many possible network configurations are equivalently supported. Pre-supposing a sparse network with effect propagation leads to a computational shortcut [58]. Each gene is instead evaluated independently and repeatedly as in a random forest technique, and only later overlaid as a probabilistic network. A 'confidence' threshold then filters the network to a sparse collection of the most plausible configuration, by dropping edges that occur with less than a specified frequency. For a detailed description of this method, see the Supplemental Materials of [7, 57].

A major drawback with this method is that many genes may share a time course profile. For example, any two genes that have the same expression dynamics could fill the same role in the network. Although we named the nodes with their originating gene, the software is agnostic to gene name metadata, and represents myriad potential connections. Each node could represent many genes, and each edge represents the product of the node

cardinalities [59]. We needed a way to prevent reporting high-multiplicity edges because they represent gene-gene interactions with a low likelihood of reproducibility when the gene products are physically tested.

2.2.3 Assaying A Gene Dynamics' Uniqueness

To clarify which gene profiles are similar to one another, any distance metric can be used, although which metric is the most appropriate is a difficult question in general. The data set was a branching time course. Each gene was evaluated once for the first few time points, and in two tissues (left and right ventricle) when the fetal development was complete enough to differentiate them. The network was developed only on one heart ventricle, but we know the genes to be differentiable using all data. The gene expression unit of measurement here is an arbitrary product of the microarray technology, a relative fluorescence score, which was normalized to (0-1) at each time point. The normalized (0-1) scores at fifteen measurements produce a data point in a unit hypercube of R^15 as shown in **Figure 2.2**.

To evaluate the expectation that many genes share a similar profile without defining the similarity threshold, we used a clustering technique called the self organizing map (SOM) [60]. The SOM algorithm performs clustering on a predefined topology. An example of the SOM algorithm behavior on a 6x6 hexagonal grid is shown in **Figure 2.2**, where all gene profiles are organized. The self organizing map is so-named because it automatically places data points into identifiable clusters that are locally similar, thus reducing the possibly high dimensional input vector to an element in the finite grid. The SOM algorithm puts elements into a regular grid and iteratively moves each data point to

the grid coordinate that best represents its properties. Thus a finer grid yields more possible clusters with an enforced similarity within each grid cell.

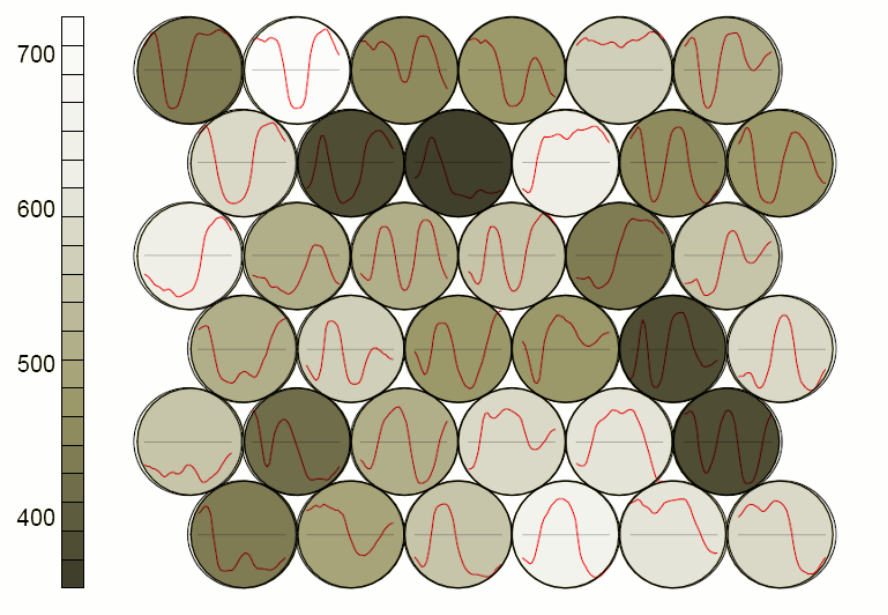

Figure 2.2 Example Self-Organizing Map (SOM) Schematic. A SOM of the GRN2014's mouse time series data is shown on a 6x6 hexagonal grid. The shading shows the number of genes per cell. The canonical profile is drawn with a red curve within each grid cell. Each circle represents a graph of gene expression profile over time, with the vertical axis representing gene expression between 0 and 100%, and the horizontal denoting the time-course. The cell fourth from the left on the bottom row shows a highly populated cell, where more than seven hundred genes are initially turned off, then on, then off again through the time-course.

After ten thousand iterations of randomly sized SOM grids, each gene was clustered together with others at varying degrees of precision (grid resolution). Two dimensional SOM grids were used with random numbers of rows and columns sampled between 4 and 50. Two examples from the 10,000 clusterings are presented in **Figure 2.3.** Resampling the SOM algorithm yields a similarity metric as a percentile of co-clustering

agnostic to the originating R^15 space. Genes that co-clustered more than 80% of the time are said to have the same profile, so similar that they might serve the same role in the numerical network model (Supplemental Methods of [7]). An important caveat to a numerical model is that any number of genes with equivalent input profiles can play the same role in the final output network, and the numerical model cannot differentiation between these genes.

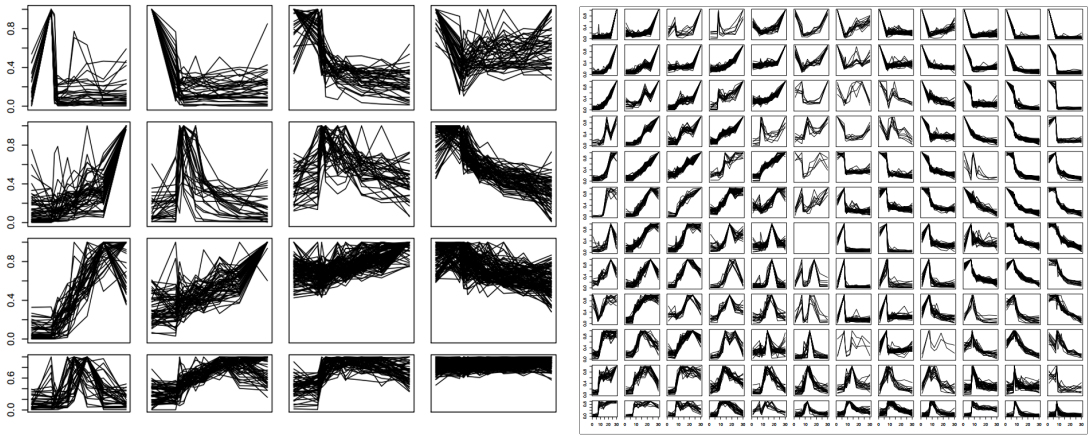

Figure 2.3 SOM Similarity Clustering. Two examples of different resolution SOM grids on the mouse time series gene expression data. At left is a 4x4 grid of all genes, each subplot consists of the time courses that are classified by the SOM into the corresponding cell. The vertical axes are 0-100% expression, and all horizontal axes are time. The 4x4 grid is quite coarse and broadly overlays similar profiles. At right is a 12x12 SOM of the same data clustered more finely.

To convert the collection of SOM-based clusterings to a single canonical cluster, some threshold of co-occurrence is required. Ten example genes' co-occurrence results from the 80% and 90% thresholds are presented in **Table 2.1**, and the 80% level was finally used as the threshold and definition for similarly expressing genes throughout the

time course. Clusters with many genes co-occurring are going to have poor uniqueness with respect to regulatory interaction predictions.

Each node in the network was expanded from one gene to N, encompassing all similarly profiled genes. Each edge then truly represented N_1*N_2 possible edges, where N_1 is the number of genes with similar profile to the source node, and N_2 is the number of genes with similar profile to the destination node. Due to the numeric nature of this analysis, any genes with a shared expression profile were interchangeable in the network.

GeneID	Name	80% Cluster ID	80% Cluster Size	90% Cluster ID	90% Cluster Size
11287	*Pzp*	1	27	1	27
11304	*Abca4*	2	11	2	3
11305	*Abca2*	3	104	3	39
11307	*Abcg1*	4	47	4	1
11352	*Abl2*	5	1	5	1
11363	*Acadl*	6	3	6	1
11364	*Acadm*	7	7	7	6
11370	*Acadvl*	8	65	8	14
11409	*Acads*	9	41	9	22
11419	*Accn2*	10	2	10	1

Table 2.1 Co-Occurrence Clustering Statistics. Top ten genes by ID number of a table of all genes involved in the iterated SOM clustering scheme to identify unique profiles. The 80% and 90% cutoffs show progressively more stringent clusters, for example the second gene *Abca4* is identified as Cluster#2 with 11 genes at the 80% threshold, but just three genes at the 90% threshold. Regulatory network predictions involving *Abl2* can be called unique to that gene, while regulatory network predictions involving *Abca2* are unlikely to be fruitful, with more than 100 other genes sharing its curve at the 80% similar threshold.

Some of the originally proposed edges came to represent thousands of potential configurations, decimating the interaction's verifiability in future experiments. Therefore

we only reported findings on edges with sufficient uniqueness. A z-score filtering step culls low-uniqueness candidate interactions.

A functional annotation was used to evaluate whether predicted interactions involve genes with known similar function. The functional overlap adds independent evidence of the veracity of an interaction. Using the Gene Ontology to note which genes in the network were known to perform which functions: each gene has a series of annotations through the GO hierarchy. The leaf node in the GO hierarchy and the path to the root in the GO hierarchy define a set of increasingly general annotations. All annotated genes are annotated with the root node of the GO hierarchy. Any two genes' paths through the GO hierarchy have an intersection. The cardinality of the intersection is a measure of the degree of known functional similarity between the two genes. We used the Jaccard index to score the similarity of these annotation sets for every pair of genes. The Jaccard index measures similarity between sets by measuring the cardinality of the intersection as a fraction of the cardinality of the union. For example, the sets (A, B, C) and (C, D, E) has Jaccard similarity 1/5 because they share one element from a universe of five. The score was called the GO term overlap (GOTO).

The GOTO score was then used to prioritize gene-gene interactions with a higher-than-zero *apriori* likelihood of biological relevance due to the genes' GO annotations. The GOTO score and the uniqueness z-score were scaled and a weighted average was used as a "Fidelity" score. The weights were determined by evaluation against a gold-standard database of gene-gene interactions [7].

The available gold standard is a commercial database mentioned above, Ingenuity Pathway Analysis. Ingenuity Pathway Analysis is a curated online resource identifying

edges of its graph via text mining and manual review [54, 61]. For each gene in the study we downloaded the participating interactions from Ingenuity Pathway Analysis and screened them for our predicted hits. The details of this analysis are in [7], and by recapitulating known gene-gene interactions, the putative interactions are granted higher confidence. GRN2014's result is a table of newly proposed gene-gene regulatory interactions that take place in the developing heart, filtered to biological relevance and uniqueness (excerpted here as **Table 2.2**). The interactions become pairs of genes for the multi-hit hypothesis of disease etiology. Pairs of genes now strongly believed to be working together on cardiac development and they constitute a resource for knowing which genes are more likely to cause CHD.

Source	**Verb**	**Target**	**IPA**
Myom1	activates	*Myom2*	No
Hbb-bh1	activates	*Hbb-y*	Yes
Hbb-bh1	activates	*Hba-x*	Yes
Hba-a1/2	activates	*Hbb-b2*	Yes
Foxa3	activates	*Nr6a1*	No
Foxa3	activates	*Foxa1*	Yes
Lama3	inhibits	*Lama4*	No
Sox7	activates	*Fli1*	No
Sox18	activates	*Sox7*	No

Table 2.2 Predicted Gene-Gene Interactions. Results from GRN2014 Table 1 indicating the highest fidelity gene-gene interaction predictions [7]. Column **IPA** notes whether the gene pair was found in the Ingenuity Pathway Analysis search.

2.4 Summary

The state of the art in human gene association studies involves a literature review of the condition of interest to generate a list of possibly important genes, creating a candidate gene list. The list could be evaluated for altered dosage (CNV) in developmental patients. The confirmed genes are re-evaluated in a different context, in this case a mouse-based gene network study performed with several outside collaborating laboratories.

The results from the GRN2014 study combined with the CNV2012 give complementary views on heart development. For example, in **Figure 2.1** we see a small percentage of CHD patients carry CNVs (of both gain and loss types) at the gene SOX7. **Table 2.2** reports the top ten interactions ranked by fidelity. *Sox7* is indicated twice, as an activator of *Fli1* and downstream of *Sox18*. *SOX7* is known to be important for heart development but its function was unclear [62]. *FLI1* is known to be important for tissue development in a general sense: it has been associated with sarcomas and leukemias [63]. Bringing these two studies together yields novel approaches to heart development research. Many other refined predictions are available, due to the methodology of integrating complementary datasets.

The techniques shown in GRN2014 could be applied to other data types. Whenever a systems analysis produces a list of putative findings, it is helpful to provide a fidelity ranking so other researchers may pursue only the best results. Computational systems biology studies may report thousands or millions of hypothetical findings [64, 65], but the results can be difficult to integrate with low-throughput experimental

techniques. When presented with many putative results researchers may be more likely to follow up hypotheses that are easier to confirm [66], inflicting undue bias to further research paths. In the GRN study we specifically addressed these issues by measuring how verifiable and how *apriori*-likely each reportable finding is.

Using completely separate experimental protocols, it is possible to refine a view on a core subject. The methods of combining data from these two sources (each of which also draws upon other databases) culminate in a reliable picture of cardiac development, not possible by previous simpler analyses. Cardiac development is just one example of a system where sample collection is difficult, and new data analysis methods are needed. The CNV2012 study serves to highlight how differences between individuals are unavoidable in human tissues analyses, emphasizing the need for better tools. The GRN2014 study serves to highlight how automated searches in a large space often produce too many results to follow up on, and emphasizes the need to collapse results along *apriori* biologically defined gene sets. **Chapter 3** introduces the software **rgsepd**, using these two principles to facilitate systems-level analyses when the data available has more measurement dimensions than samples available.

CHAPTER 3

TRANSCRIPTOMICS AND THE GSEPD

Often considered a synonym to *transcriptomics* is the exploration of the "transcriptome" or the genome of transcripts. A transcript is a gene isoform messenger RNA, or one particular blueprint for a protein. Almost every gene produces several forms of protein through a process called alternative splicing: therefore the transcriptome is larger than the genome. Alternative splicing is a common feature of higher organisms that boosts the complexity possible from the genome [67]. A single genome can produce myriad tissue-types in different contexts, and therefore transcript-level analyses are vital to the understanding of genetic diseases. In this chapter I explain some techniques popular in exploring the human tissue transcriptome, and present a new software algorithm Geneset Enrichment and Projection Displays (GSEPD) designed to facilitate the extraction of important features from a transcriptomic data set.

3.1 Introduction

Why is the transcriptome important? A common belief is that the genome defines the species, but the complexity of various tissue types is reliant on various levels of gene usage at various times. The transcriptome is the set of all transcripts for an organism, and the word is used synonymously with a numerical measure of gene usage: for each gene and in each location. It is a spatio-temporal measure of a high dimensional quantity. A snapshot of the living cell captures the state, and due to the dynamic nature of living cells,

many such snapshots are required to get an average measure of a cell and its variability. Measurement of mixed-cell tissues induces another layer of complexity to transcriptomic analyses.

The modern technique to query the transcriptome is RNA Sequencing, or RNA-Seq for short. In this method, messenger RNA is isolated and loaded into a sequencing machine. The resulting data are sequence reads. With some pre-processing the sequence reads can be converted to a quantitative measure per transcript. Given a reference "transcriptome" we can have a finite set of dimensions measured. The human transcriptome consists of about 35,000 transcripts referred to by their RefSeq ID# in the form NM_123456. A single measure then is a vector of cardinality about 35,000. This would have to be measured for each cell type, sensitive to location, and any other covariates.

Early measures of the transcriptome were limited by the available technologies. For example, before high throughput assays were available, mRNA levels were measured individually and relative to another so-called housekeeping gene, or normalizer [68]. Several genes were found to be "markers" as they were apparently specific to the cell type being studied. When one gene was found active only in one cell type, it became the *de facto* standard marker for that cell type, and is historically used to identify the cell type, and even classify related cell types [69]. Today, cells go through stages defined by their markers. With the advent of higher throughput measures we see more nuance.

A researcher who aims to characterize a cell's state would use RNA-Seq as the proxy for protein production. Single gene measures were not good enough to see higher-level functions and complex developmental orchestration. Canonical pathways developed

over time with databases coming in later to help broaden the scientific corpus and let more researchers access an ever-broadening understanding [70, 71]. Now researchers may focus on one or more pathways and perturb the cells with some experimental condition, then see the results in the perturbation of the transcriptome. The result of a simple experimental condition is a multitude of gene expression changes. Single gene assays cannot explain the whole picture, whereas transcriptome-wide analyses overwhelm and hide the nuance to the researcher.

RNA-Seq showed great promise to finally capture a genome-wide picture of the transcriptome with digital precision [72]. RNA-Seq finds many genes active in a tissue sample. With higher resolution and sensitivity than the older technology, the resulting lists of apparently-active genes can grow considerably [34]. Following this advancement there was a paucity of analysis software to help researchers interpret the mass of data.

Several tools came out in rapid succession to help with gene quantification from the sequencing data. Bowtie was developed to align the reads to a reference genome [73]. Tophat was developed to find new splicings and new isoforms while aligning reads to a reference genome [74]. STAR later was introduced to speed up alignment to a known reference [75]. Cufflinks was developed to compute probabilistic quantification to novel isoforms [76]. Detailed surveys of RNA-Seq data analysis tools have been published [77, 78]. These transcriptome tools helped capture measurements, but not interpretation of the high dimensional results.

Carefully designed experiments are the direct method for simplifying interpretation. A differential gene expression experiment is like subtracting two samples: the difference is apparent and attributable to the experimental condition (subject to

replication). So differential expression is a common term we will be using often. The gene level difference is fundamentally high dimensional, with one measure per gene, and therefore difficult for people to visualize and understand.

Unfortunately pathways can easily be missed with this paradigm: cellular systems are more complex than any one researcher's field of expertise. Each researcher has expertise in a few biological functions and can recognize differences in expected pathways. By looking at the personally understandable subset of gene changes, researchers may miss broad systems level changes [20]. The result is that truly novel findings have the potential to routinely go unseen. The objective of my dissertation, as mentioned in **Chapter 1**, is to develop a tool to aid researchers in exploring gene expression experiments without being biased to the pathways they already are familiar with. The newly developed tool leverages online databases to further expand our perspective in order to extract additional information from annotated gene sets.

3.2 Review of Gene Expression Exploration Techniques

Exploration of the transcriptome started with individual assays of a presence or strength of a known RNA sequence via qPCR or gel technologies. These technologies focus on a given sequence and measure the expression of the gene, but are time consuming. The researcher was limited to assaying sequences that were already well known. As the database of known mRNAs grew, and technologies miniaturized, a fluorescence microarray was developed to assay thousands or hundreds of thousands of sequences in parallel. Finally the researcher can assay a sample's whole gene activity, albeit in a relative way. Later as sequencing technologies matured, direct RNA mass

parallel sequencing was developed to get a real picture of the native sequences without microarray design limiting the view. However, sequencing technology brings technical hurdles to normalization, absolute quantification, and data preprocessing. RNA-Seq is widely recognized as a more accurate method (over microarray), by directly identifying all mRNA molecules of a sample, at a cost of several thousand dollars per sample [79]. Due to the high cost and effort required in sequencing, the microarray technique is not completely obsoleted, since its limited view is sufficient for many experimentalists. Both technologies give a very high dimensional quantitative measurement, leading to challenges in statistical analysis and interpretation. Both technologies suffer the curse of dimensionality, which I am focusing on alleviating.

3.2.1 Differential Gene Expression

The main goal of a transcriptomic analysis is to elucidate the gene usage of a sample. Without contrast, a measurement is meaningless, so usually two or more classes of samples are compared, and what we look for are "differentially expressed genes" (DEG): those genes that are in use differently. Genewise differential expression is an analysis requiring statistical models to determine if each gene is expressing a significantly different level between the sample classes. Simple direct inspection will almost always show a gene varying in usage between samples, so something like a Student's t-test is required. In the case of transcriptomics, the variance is high but parameterizable with respect to other sample properties; so specialized tools have been developed [80]. The computation to find truly DEG is not straightforward, as the numerous technical biases of each experiment must be controlled for [81]. Many tools exist to facilitate these experiments, such as limma, edgeR, and DESeq2 [72, 80, 82-84]. The result of computing

DEGs is always a list of genes or gene transcripts, each associated with a measured change between the sample means, and a p-value from the hypothesis test "this gene was unchanged between sample classes". Depending on the experimental design and control, the resulting DEG list may be a few genes, or hundreds to thousands of genes.

3.2.2 Functional Annotation

After collecting a list of DEGs, the next question would be "What do these genes do?" or "What biological function do these genes represent?" Since the beginning of genome-wide testing, interpretation of results has been a challenge. Often the goal of a transcriptome experiment is not a list of genes altered, but instead knowledge of biological processes impacted. Several interactive tools exist for researchers to enter a set of genes and infer their general function from existing results presented in the peer-reviewed literature [85-88]. **Table 3.1** is an overview of functional analysis tools. Gene Set Enrichment Analysis (GSEA) is one of the original gene set analysis tools, identifying functional commonalities in various gene set data, as in differential expression results [88]. GSEA can cite studies that have found a particular gene list, and thus GSEA gives context to a differential expression result. The contextual result is also a limitation of GSEA, as the previous studies it draws from are likely designed with incomparable source material (for example seeing your gene set has a similar profile to a drug study of a particular tumor is not a particularly usable result.) The next element of **Table 3.1** is DAVID, a website with several analysis tools including functional evaluation of gene sets [89]. DAVID is popular with biologists due in part to its simple user interface, and breadth of results. The broad result set is also a drawback: applicability, reproducibility, and significance of the findings can be questionable. IPA is another web-tool for inferring

meaning from a gene list. The company that owns IPA, QIAGEN, collects both public and manually curated gene associations. The IPA service's knowledgebase is considered one of the most complete and informative, but comes with a paywall, steep learning curve, and can provide hard-to-interpret results like GSEA. The fifth entry of **Table 3.1** is GOSeq [90], an R/Bioconductor [91] toolkit that performs a statistical analysis of a Boolean-flagged gene ID list to provide a set of Gene Ontology (GO) results. GOSeq is robust and highly available. The drawbacks are that it only draws from a single data source, offline installation/maintenance of the GO database can be cumbersome, and any notion of effect size and individual gene behavior is lost. The final entry is GeneMania, available at genemania.org, which is a Flash based application that mines several datasets for an input gene list. Like DAVID and IPA, GeneMania provides interesting results with little regard for reproducibility or significance.

GSEPD uses GO [26] as a functional knowledgebase agnostic to tooling. The GO consists of three tree data structures, the Biological Process, the Molecular Function, and the Cellular Component. Within each are hundreds of thousands of "terms" gaining in specificity as one traverses down from the root node. GO annotations are manually curated by an open-source process, and have become a worldwide free resource for annotating multiple functions to each gene. In this dissertation a gene set generally refers to a GO term, which is a single node in the hierarchy. A broad node is one near the root of the hierarchy, such as GO:0044238 "primary metabolic process" with over 9,000 genes. In contrast, a precise node contains very specific biological processes such as GO:0033188 "sphingomyelin synthase activity" which pertains to just two genes. By

using GOSeq within GSEPD immediately after producing differential expression, significantly upregulated or downregulated GO terms are computed [90].

Tool Name	Online	Input Type	Database	Visualizer	Programming Required	Depends
rgsepd	No	Genome wide read counts	Gene Ontology	Expression heatmaps, PCA, GO scores scatter	Minimal data preparation	R, Bioconductor
GSEA	Website	Significant gene list	Multiple	Gene ranks	Minimal data preparation	None
DAVID	Website	Significant gene list	Multiple	None	No	None
IPA	Website and Java	Significant gene list	Proprietary	Association web graph	No	Paid Account
GOSeq	No	Significant gene list	Gene Ontology	None	Yes	R, Bioconductor
Gene Mania	Website	Significant gene list	Multiple	Association web graph	No	Adobe Flash

Table 3.1 Overview of Functional Analysis Tools. Six methods to calculate functional consequences of an expression dataset are compared in the table. **Online** refers to the installation of the tool, where "No" means the tool is a local installation, available in perpetuity, and "Website" means the tool is accessed remotely, and subject to change or unavailability. **Input Type** refers to the data given by the user: read counts or gene names. **Database** refers to the source of functional data: where most tools aggregate many studies; **rgsepd** and GOSeq rely on the curated Gene Ontology hierarchy. **Visualizer** refers to the figures generated by the tool. **Programming Required** refers to the technical skills required of the user. **Depends** refers to the third-party software installation required of the user.

3.2.3 Enrichment and Projection Display

With functional analysis in hand, researchers could see gene sets X,Y and Z are perturbed, and immediately the question becomes "Which differentially expressed genes are part of those sets?" Noting the genes underlying a set consists of a database join operation. The listing of genes by their GO terms becomes lengthy and therefore difficult

to interpret, as the results are on the order of magnitude of the number of genes times the number of GO terms. A need remains for a way to show that a gene set was not only significantly perturbed, but segregates the samples by gene expression. The simple solution to determining if a set segregates samples is a vector projection.

I developed a software package named **rgsepd** to implement GSEPD in performing automatic differential expression, functional analysis, and the novel gene set projection, creating a score for each sample to each gene set (GO term). With a simple user interface, **rgsepd** allows researchers to produce differential expression lists, GO functional analyses, and the cross-product: a mapping of which genes are perturbed within each gene set. The projection uses simple vector calculus to score samples within the gene set. Additional details of **rgsepd** are presented below.

3.3 Methods

The package **rgsepd** performs several steps in an automated fashion, producing many files for later browsing. The user must first provide a matrix of read counts or integer gene expression values analogous to read counts at RefSeq identified transcripts, as they would to any RNA-Seq statistical analysis package. A second small matrix annotates which samples belong to which class, and which classes the user intends to contrast. The first automation step is gene expression quantification normalization, leading to differential expression. DESeq2 [92] performs the differential expression test at each transcript ID#, and with some cutoff p-value, tens to thousands of genes with linear separation are detected. DESeq2 is a scriptable Bioconductor tool designed specifically for RNA-Seq data. The gene listing produced by DESeq2 is filtered with customizable

parameters for p-value and log-fold-change. Three gene sets (upregulated, downregulated, and the union) are analyzed for functional categories.

Programming using the statistical programming environment R [93] ensures cross-platform availability, and the open-source nature helps ensure longevity and reproducibility. To use **rgsepd**, input should be formatted as a matrix, with RefSeq NM (gene transcript) ID numbers as rows, and sample identifiers as column names. A second matrix of metadata links sample identifiers with test conditions and short sample labels. GSEPD will automatically compute DEG with default parameters to DESeq2, adjusted if necessary for small sample counts [83]. Functional analysis is performed by GOSeq [90], once each for downregulated DEGs, upregulated DEGs, and all DEGs. GOSeq has been shown to have similar accuracy to other enrichment toolkits, and robust to gene type biases [94].

GO terms of certain cardinality are evaluated by GSEPD as clusters of samples in gene-expression space. The default is to search GO terms with more than one gene and less than 31, ensuring precision in the named results, and avoiding the computational costs of permuting and calculating in a many-hundred dimensional space. For a GO term with N genes each sample can be considered a data point in N-dimensional space, of normalized expression (by z-score of log of DESeq2's normalized counts) of the gene. Clustering is performed by k-means to achieve recognizable clusters, which are scored by a validity measure called V-score [95]. Permuting sample condition labels and re-calculating validity computes an empirical significance for each GO term: the segregation p-value. The segregating GO terms are evaluated more thoroughly with scatterplots and subspace principal components analyses (PCAs). Vector projection is performed within each gene

set to score each sample's similarity to either test-condition centroid, highlighting outliers with respect to the subspace.

3.3.1 Systems Architecture

The design of GSEPD is a modular pipeline. **Figure 3.1** gives an overview, with inputs on the left and output on the right. Starting outside GSEPD is data generation at the sequencing facility, and user-driven counts-table construction. Details of generating the counts table are given in **Chapter 4**. Along with sample class labels, the dataset is fed into GSEPD to perform a pipeline. The first stage is differential expression of the desired classes. The results of the differential expression are piped into the functional analysis, which could be any functional annotation tool that takes a list of genes and returns a list of sets. Finally the projection engine takes in the sets found by the functional annotation and the normalized expression table. GSEPD calculates the set-based expression metrics of the samples and the segregation statistics for each set. The output is a table of significantly segregating sets, and metrics between each sample and each set.

The design is driven by the motivation in **Chapter 2** combined with state of the art transcriptome processing. The CNV2012 study revealed the need to accommodate sample heterogeneity. The GRN2014 study revealed the need to bring in database driven set annotations to reveal useful candidate results. Differential expression and functional annotation are commonly used techniques that leave the user with sample-specific heterogeneity unexplored. The current implementation of **rgsepd** takes the sets defined by the Gene Ontology Consortium and further evaluates samples' specific expression by a vector projection and a clustering measure.

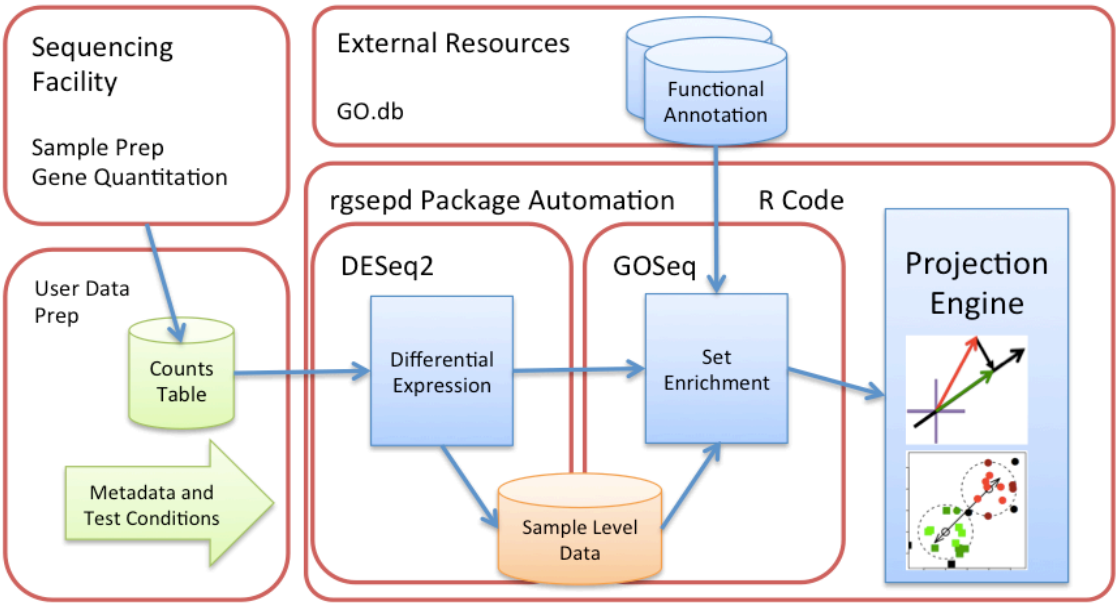

Figure 3.1 Systems Architecture diagram of the components of the GSEPD system, with major sections in red outlines. Blue items indicate automated systems. An experiment starts at the upper left, with the Sequencing Facility where the tissue samples are converted to gene expression quantification through sequencing and processing external to GSEPD. The user then creates a table of count data and defines the sample metadata and conditions to be compared (lower left, green items indicate user inputs). Across the top are External Resources, where functional annotation databases are curated by third parties and plug in to the **rgsepd** software package. The R code wraps subprocesses for differential expression, set enrichment, and set based projection scoring. The orange cylinder of sample data indicates a normalization produced by DESeq2 with useful expression measurements. Within the Projection Engine box are small diagrams of the integral vector projections and clustering analyses.

Specifically the differential expression tool used in the current version of **rgsepd** is DESeq2. The differential expression tool could be replaced in other contexts. The functional annotation stage uses GOSeq. GOSeq requires gene level identifiers from one of several gene name databases. DESeq2 operates on transcripts and is therefore not perfectly compatible with GOSeq. A conversion step in **rgsepd** maps RefSeq transcript IDs into Entrez Gene IDs using the Bioconductor database **org.hs.eg.db** and a hash table.

3.3.2 Projection of Samples onto Differential Axes

The unique final stage of GSEPD is a merging of gene expression values with the gene set enrichment results via a novel spatial projection. Vector projection is a mathematical tool central to GSEPD's ability to report set-based clustering. Each significantly over-represented gene set reported by GOSeq must have some DEGs, so samples of each condition are expected to cluster in the N-dimensional expression space (where N is the number of genes in a GO-Term defined set). If the non-differential genes within the set have high variability or otherwise significantly outnumber the differential genes, the clustering will be poor or not visible. In any case, an N-dimensional line can be drawn through the class centroids, and each sample scored for its position along and distance to the line. Classes are tested by permuted k-means for significant segregation within the gene set. To normalize gene expression between high and low-expressed, genes are scaled to a mean of zero and standard deviation of one, after a log-scale reduces the dispersion common to RNA-Seq. The projection is a linear transformation (illustrated in **Figure 3.2**) of one vector onto another such that we can calculate a proportionality constant and an angle. We need to only define a reference vector, and any point in the space can be projected upon it and the proportionality constant (Alpha) is calculated. The projection action generates a 1-dimensional measure of position along one vector. Therefore with a suitably defined axis, any point can be scored as closer to class A or closer to class B.

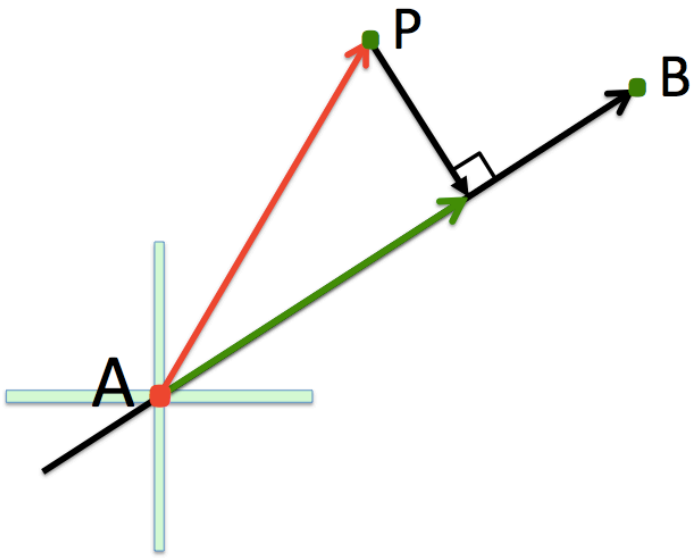

Figure 3.2 Vector Projection Illustration. With the origin at the cross, vector AP is projected onto vector AB, yielding the green projection. In GSEPD, the point A is the centroid of class A, and point B is the centroid of class B. Point P is any one sample.

Consider for example contrasting samples of class A versus class B as shown in **Figure 3.2**. RNA-Seq read count data can be seen as high-dimensional points in Cartesian coordinates, with a dimension for each gene's expression magnitude. An ideal dataset is represented by N_A+N_B points in $R^{(NA+NB)}$. If any genes segregate the samples, they can be detected by testing each individually. While performing the projection, GSEPD shifts the coordinate system so the centroid of sample class A is at the origin (**Figure 3.2**). The end of vector AB is the centroid of the class B samples. This new axis from the origin and along vector AB represents a linear direction that defined class B relative to class A's centroid. One sample represented by point P lies off the axis and represents a sample we want to score along the major axis. In this example, the dimensions are gene expressions of each gene in a given gene set. The vector projection scores a sample along the black

axis. The green vector and its length, here about two-thirds of the way to B, represents the amount sample P is like class B. The ‘Alpha’ score for this sample would be about 0.66. The ‘Beta’ score is contained in the black vector perpendicular to vector AB. Beta values indicate the distance between a sample and the closest point on the axis. If a beta value is high (scaled z-score greater than a configurable threshold) the sample is considered an outlier with respect to the class A-B axis for the given gene set. Each gene set is evaluated separately, so we generate scorings from samples to gene sets, with most of class A nearly alpha=0, and most of class B nearly alpha=1, by construction.

A schematic drawing with several points in each cluster is shown in **Figure 3.3**. Every gene set found significant by GOSeq is evaluated with an axis and sample-projections, creating an alpha and beta score for each sample and GO Term pair. By construction of the mean-centroid, roughly half of the samples in class A will have alpha scores below zero, and roughly half of those in class B will have scores over 1. The distances to the axis are known to scale with dimensionality, and are thus z-score normalized before further processing.

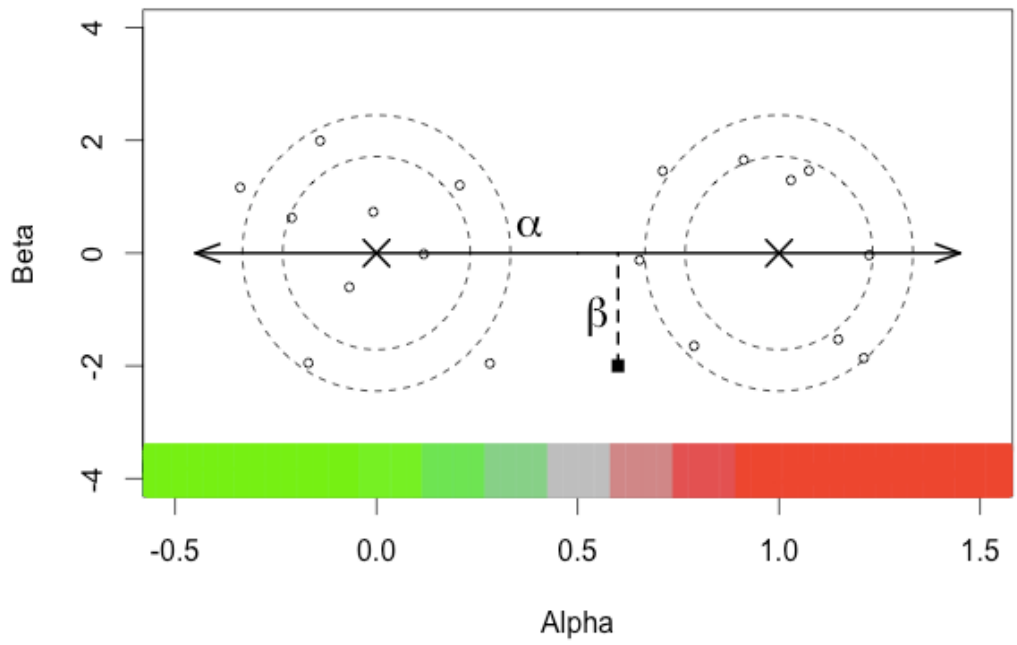

Figure 3.3 Axis Projection Illustration. Small circles represent samples, X represents the samples' centroid for each class. The separation along the condition axis is transformed such that alpha=0(green) at the centroid of the reference condition, and alpha=1(red) at the centroid of the test condition. By extending the colors past the centroids, samples that are hyperactivated will keep the brightest color. The beta score denotes deviation from the axis and is necessary to retain information that a sample may be like one condition in some genes and not others.

Every gene set has a one-dimensional scoring, and the clustering accuracy can be evaluated. Evaluation of clustering accuracy enables prioritized reporting of gene sets that segregate class A and B regardless of dimensionality. As gene sets were chosen from the DEG list, some dimensions (genes) must be segregating. The projection process can identify genes that may not have been on the DEG list, but help segregate the sample classes when taken together with a gene set.

3.3.3 Validity Scores Evaluate Clusters

Many methods exist to measure the goodness of a data clustering, see [96, 97] for two reviews of data clustering metrics. “Validity” (V-score) was chosen to incorporate specificity and sensitivity within a 0-1 scale, where 1 implies a 100% association between class label and cluster label. The V-score is an external index consisting of the harmonic mean of entropic *homogeneity* and *completeness* based on the conditional entropy of each cluster, given a gold-standard [95]. Given the ‘correct’ class labels, it measures how accurate the proposed labeling is. Regardless of the gene set projection score, we can calculate how well the gene set's multidimensional expression values in z-log space cluster with k-means versus their class labels. A perfect score indicates the gene set really does segregate the samples, while noisy or mislabeled samples could induce a lower V-score and still show significantly good validity by permuted class labels.

3.3.4 Clustering Significance by Permutation

To evaluate the significance of a clustering or association in the presence of sample non-independence and measurement noise, careful analysis is required. Significance of each association (differential gene usage and overrepresentation of ontology terms) is computed by their respective tools. The novel subspace projection is a consequence of the multidimensional expression; what we have to show significance of is the segregation of samples. Segregation significance could be computed with linear discriminant analysis if sample count were higher than dimension count. When dimension count is independent of- or larger than sample count, another significance analysis is required. I use a simple Euclidean-distance clustering to find two clusters in the data, then compute its agreement with the given class labels via V-score.

K-means is a method that identifies the centroids and assigns sample to either cluster using a distance metric [98]. The k-means is performed without knowledge of class labels: they are the external gold-standard to measure validity compared to randomized labels given the deterministic expression clustering. The k-means resulting cluster can be evaluated by computing its "validity" with respect to the known sample classes [95]. The V-score ranges from -1 (wrong assignments completely) to +1, (correct assignments completely). With every sample assigned to a cluster, the correlation of this assignment to each sample's class is calculated. To compute an empirical p-value I use random permutation: computing validity of the k-means clustering after repeatedly assigning samples to random classes. The p-value is the proportion of random assignments that achieve a higher V-score. A clustering is thereby evaluated in the context of the natural variability of the gene set.

To estimate how many iterations are necessary for precision of the p-value, GSEPD evaluates a dynamic denominator. Initially 100 runs are explored and more only if there appears a chance for significance (preliminary $p<50/100$.) The minimum calculable p-value is an input parameter P. At least 1/P permutations would be required, and maximally 4/P random permutations are undertaken per GO term to ensure p-value precision. In many instances of complete segregation, this yields a $p=0$ call.

Measuring Euclidean distance from each point to the class centroids provides a final analysis of clustering. Illustrated in **Figure 3.4**, called Gamma scores, as they come after the Alpha and Beta of **Figure 3.3.** An alternative to the axis/projection scoring, another 2xNxM Gamma scores are generated noting how close each of N points are to each of 2 centroids, across M gene sets. The distances are scaled by the dimensionality,

such that the centroids are 1 unit apart in all gene sets. The coloring threshold is calibrated to assign complete indeterminance to just 10% of the center of the axis between centroids, ergo 45% of the line is green and another 45% is red. The absolute value of the distance between a sample and a centroid is scaled with the square root of the dimensionality to match the Euclidean distance formula. These color values go into hierarchical clustering for visual review of the significantly segregating gene sets, while the raw gamma scores are reported in two tables (not shown here, see **Appendix A** for a description of all tables generated by **rgsepd**).

3.4 Results

A run of **rgsepd** produces several principal components analyses figures, heatmaps of gene expression for DEGs, and tables of all output. All thresholds and parameters are configurable before runtime, and configurable output folders and formulaic file naming conventions ensure easy reproducibility or automated parameter sweeps. A tutorial and explanation of all outputs is available within the package vignette/manuals. A full run, from input of a numeric matrix, to completed functional results is achievable in minutes, comparing favorably to weeks of manual processing without an automated pipeline.

I have run **rgsepd** on a time series dataset (five time points with two replicates) along the differentiation of H1ESC cells into cardiomyocytes (heart muscle cells) [99]. We performed RNA-Seq on a time series following duplicate stem cells being developed into heart muscle. This dataset is explored in **Chapter 4**. The raw data have been made publicly available at the NCBI SRA (accession number SRP048993).

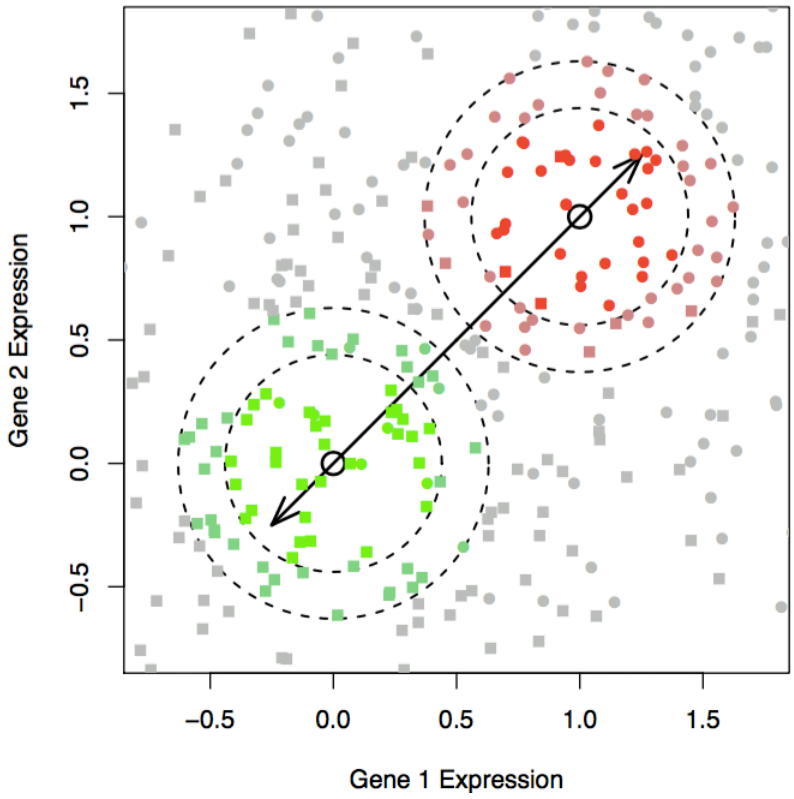

Figure 3.4 Illustration of Gamma Scoring. A simplification of the Alpha/Beta scoring, the Gamma score notes Euclidian distance in gene expression space. The results are colored with the Euclidean distance to each class centroid. Samples of class A are drawn as squares (class B: circles) in this simulated two-gene set. The distance is thresholded to bright and faded red and green, such that points away from either cluster become gray. This scatterplot diagram is not automatically produced in an **rgsepd** run, as the true gene sets are often more than 2-dimensional.

Pairwise comparison of all time points revealed that time points day 3 and day 5 had the fewest differentially expressed genes (3279 HGNC names with $p<0.05$, comprising 2214 GO terms with $p<0.05$, 1073 of which were found to significantly cluster). A feature of GSEPD is the clustering and comparison of non-tested samples (here, time points other than day 3 and day 5) with regard to the GO terms that can

differentiate test samples (such as point P in **Figure 3.2**). Incorporation of non-tested samples can help researchers label unclassified/indeterminate samples by their expression profiles among GO terms relevant to the experiment.

3.4.1 Processing Pipeline

DESeq2 computes differential expression from sequence-count data, precisely modeling the dispersion present in RNA-Seq [83]. While usage of DESeq2 is straightforward for a trained practitioner, this training can be costly or time consuming. **rgsepd** wraps usage of DESeq2 to fully automate the simple case/control comparison method.

Gene Ontology enrichment can be performed in many ways, the most common being a hypergeometric test for over-representation of a set within the DEG. GOSeq is a readily available (Bioconductor [91]) software package in the R environment that performs these tests while controlling for selection biases from sequencing genes of varied length. GOSeq is a tool for calculating GO term presence among a gene set with binary inclusion criteria, particularly differential gene expression results [90]. I incorporate basic usage of GOSeq for the human genome (hg19, refSeq) as part of the pipeline.

Two sets of projection measure are computed for each gene set. The "HMA" heatmaps (example in **Figure 3.5**) show the distribution along an Alpha and Beta axes (defined as **Figure 3.3**), such that samples can be visualized as how much like either class they behave with respect to the direct line between centroids. HMA stands for heatmap-alpha score. The class centroids are always rescaled to be at vector 0 and vector 1.

Figure 3.5 GSEPD Results from the H1ESC Study. The H1ESC dataset of **Chapter 4** is evaluated with GSEPD's Alpha/Beta heatmap (HMA figure). Notes along the bottom are a coded sample identifier ending in the time point name D3 for day 3, D1 for day 1, and so on. This figure shows GO terms with significant segregation between day 3 (green) and day 5 (red). **rgsepd** was instructed via input parameter to display only the top 8 results. The color bar across the top indicates which samples were part of the DESeq2 contrast, here day 3 in green versus day 5 in red, with black denoting non-tested samples.

To demonstrate potential findings, I present an analysis of the H1ESC data. Six time points, named day 0 through day 14 (in duplicate) are seen in the final **rgsepd** result in **Figure 3.5**. This analysis compared samples of day 3 and 5, which is a critical turning point between early tissue development and heart muscle precursors [99]. Each sample has a unique ID noted along the bottom of the figure that is relevant for the user, but not our discussion of the tool. These have suffixes like T4 and I6 corresponding to Tube#4 and Index#6, which were meaningful to the lab performing RNA extraction from the samples.

Many gene sets were found to perfectly segregate due to the sample's cloned nature boosting significance of the differential gene expression. Here **rgsepd**'s parameters were tuned to yield the eight most significant gene sets. An example result is presented in **Figure 3.5.** Regarding the fourth row "cardiac atrium morphogenesis," day 3 is unique (bright green), day 0 and day 1 have expression near the center of the Alpha axis (faded gray) with samples of class day 1 more similar to samples of class day 3, while the samples from later days expressing like the samples of class day 5 (red).

The GSEPD projection Alpha scoring suggests in a data-driven manner that day 3 was the turning point for the stem cells' lineage specification. The next two rows show that day 3 was unique in "mesodermal cell fate specification" and "commitment", suggesting a unique spike of gene activation that deactivated on all other time points. With no biological systems background knowledge, the user of GSEPD can thus extract pathway activation knowledge from RNA-Seq count data.

Within the "cardiac atrium morphogenesis" gene set on row 4 of **Figure 3.5**, 12 of 28 genes were found differentially expressed (GOSeq $p<2x10^{-6}$). GSEPD extracts significant gene sets into multi-page scatterplots, one of which is excerpted into **Figure 3.6**. The example scatterplot consists of 14 pages of two genes each, showing orthogonal views on the 28-dimensional clusters, here *PITX2* is shown downregulated in class day 3 (green) versus class day 5 (red), with gene *NOTCH1* upregulated by 1.5 units of logged, normalized counts. Colored lines (corresponding to cells of the heatmap of **Figure 3.5**) are perpendicular to the thick black axis in the 28-dimensional space, indicating samples of class day 0 and samples of class day 1 fall between the clusters of samples in class day 3

and samples of class day 5, although they will not appear perpendicular in this two-gene subspace.

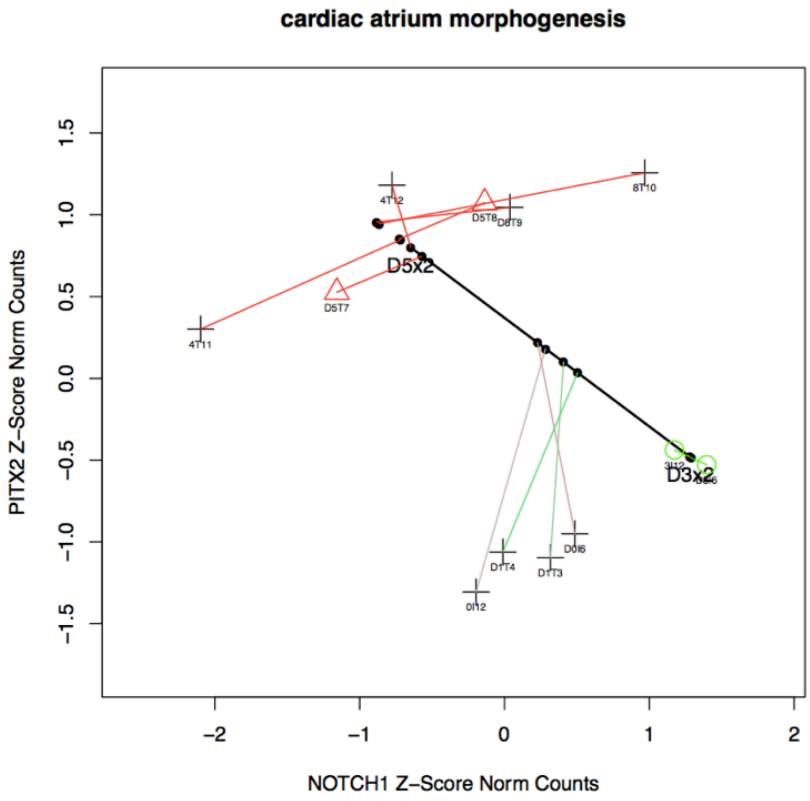

Figure 3.6 Scatterplot of Two Genes. Corresponding to the seventh row of **Figure 3.5**: (gene set 'atrial cardiac muscle tissue development') this diagram is one part of generated file GSEPD.D3x2.D5x2.GO0003209.pdf (first two genes). Points as triangles, circles, and crosses correspond to the input samples. Solid dots indicate the projection coordinate. Labels D5x2 and D3x2 indicate class centroids of the comparison of two samples of day 5 versus two samples of day 3. The small point labels are specified by the user as each sample's "shortname," a parameter given to **rgsepd**.

Samples could be scored by distance to each class centroid (**Figures 3.4** and **3.7**) instead of using location along a constructed axis. The magnitude of the distance between a sample and either cluster centroid can divide the color into five bins: bright and faded

green, medium, and then faded and bright red. Distance is computed as Euclidean distance to each class centroid in the n-gene dimensional subspace (scaled by a factor of the square root of the dimensionality, such that centroids remain one unit apart). The center 10% of the axis is given the medium-gray color, with the ends colored as the sample classes. The 45% closest to class A is green and the 45% closest to class B is red. The first two-thirds of each color section is denoted 'bright' and the latter third 'faded' (**Figure 3.4**).

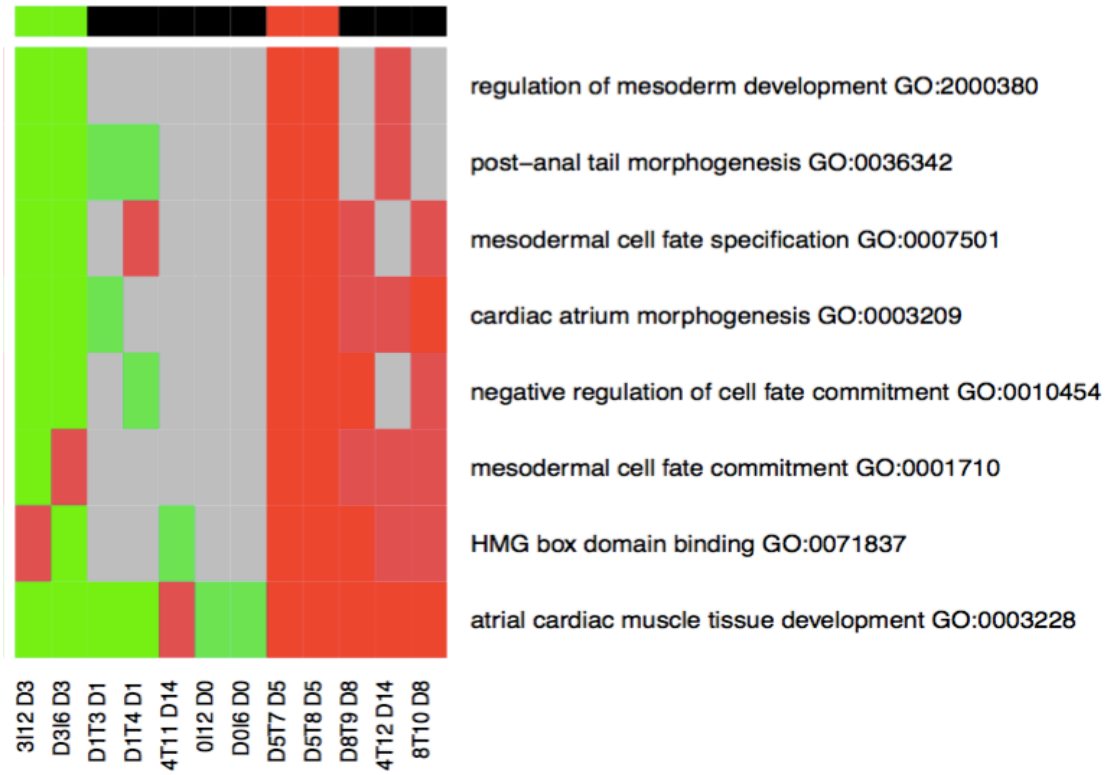

Figure 3.7 Gamma Scores Recolor. Example data from the H1ESC, the same analysis as seen in **Figure 3.5**, alternatively visualized by Gamma scores to simplify axis-scoring into a single dimension from green to red. Less variation is visible, as we here color gray any sample that is more than 0.45 units distant from either class centroid. The coloring is as illustrated in **Figure 3.4**. The color bar across the top indicates which samples were part of the DESeq2 contrast, here day 3 in green versus day 5 in red, with black denoting non-tested samples.

The Gamma color scheme simplifies the interpretation of the result in that green samples can be said to express like the centroid of class A without reservations. As many points may be off-axis in a high dimensional space, the Gamma coloring scheme yields many more gray samples (**Figure 3.7**) than the Alpha/Beta scheme (**Figure 3.5**). While containing the same rows and columns, the hierarchical clustering imposed on the heatmap is based on the color-scores so row and column ordering will differ in the two output figures of each run.

3.4.4 The Bioconductor Package rgsepd

A curated catalog of software accompanied by manuals and tutorials is important for consistent advancement of computational research. The Bioconductor project [91] is a repository of bioinformatics tools for the R programming language with strict guidelines regarding usability and availability. An implementation of GSEPD is submitted as an R package "rgsepd" to Bioconductor. Version 1.5.2, which is discussed in this dissertation, is freely available in Bioconductor 3.3. **rgsepd** version 1.5.2 relies on the following R packages: DESeq2, GOSeq, GO.db, and org.Hs.eg.db. As of July 2016, there have been 2,190 downloads of the various **rgsepd** releases.

3.4.5 Systematic Output Files

All files generated by a run of **rgsepd** are named in the pattern TOOL.YYY.AxNa.BxNb.csv, where YYY represents the file type, A and B the class labels relevant to the table, and N(a/b) the number of samples in each class. This naming convention highlights the conditions and keeps multiple comparisons organized. "TOOL" may be DESEQ, GOSEQ, or GSEPD, depending on the data source. In the

event the user's sample conditions have an "x" in their name, creating ambiguity in this structure, the delimiter is configurable. **Appendix A** lists these tables in chronological order through a run of the tool, as each result builds upon the previous. In total, 15 tables and 11 types of figures are generated in a run. As each interesting GO term may generate several figures, a subfolder is created and files are named with the GO ID number to enable quickly finding the relevant figures. Producing all result files as CSV or PDF permits sharing of results between collaborators.

3.4.6 Software Limitations

The current version of **rgsepd** requires a particular human genome reference data format for input tables. Genes must be quantified against a RefSeq transcriptome definition. This limitation is due to the rigidity of the automatic gene name conversion routines built in to **rgsepd**. Further software development could enable other species such as mouse or plant studies. Additionally, data input is currently restricted to RNA-Seq due to the choice of DESeq2 as the DEG stage (**Figure 3.1**). Accommodation of other input data formats is possible. DEG calculations are restricted by the user interface to a two-class contrast experimental design.

The projection and clustering stages assume absolute valued expression data not compatible with florescence based microarray technologies. Gene set results coming from relative expression platforms will require further experimentation and validation. Count-based data are ideal, such that **rgsepd** would more easily be adapted to sequencing based assays.

3.5 Conclusion

This chapter has outlined transcriptomics as a tool for exploration of biological processes, and has presented a software toolkit to facilitate those analyses. Application to two data sets is shown in **Chapter 4**, highlighting the uniqueness of findings possible with GSEPD. The Bioconductor package has been under development and is publicly available and as per their guidelines includes considerable manuals and tutorials. The package tutorials are included with the software download. It is my sincere hope that this free and open source tool can facilitate many third party research groups that have struggled with transcriptomics and data interpretation in the past.

The **rgsepd** toolkit is available for public use at the website

http://www.bioconductor.org/packages/devel/bioc/html/rgsepd.html

See Appendix A for a complete description of the generated output files.

CHAPTER 4

APPLICATION TO TWO HEART DEVELOPMENT STUDIES

To show their merit, findings and methods from previous chapters are applied to several important clinical situations. Specifically, I present two instances where application of the toolkit from **Chapter 3** implementing the insights of **Chapter 2** to an open problem in cell biology are able to validate and expand upon the findings of an expert in the field of heart development [99, 100]. In each of two case studies I present the original manual findings, then the results that could have been found by GSEA [88], then finally the results that could have been found using GSEPD.

4.1 H1ESC Differentiation Time Series Functional Analyses

The first case study is referred to as H1ESC: pertaining to a stem-cell study (of commercially available type H1E) published in early 2015 in PLoS ONE [99]. In collaboration with the Medical College of Wisconsin Department of Molecular and Cell Biology, we performed a transcriptome analysis of duplicate samples at six time points. The study explored a problem in applied cellular development and refined the best-known recipe for generating living human heart muscle, which is important for later experiments. Following standard protocols, cell culture differentiation has low efficiency: many dead cells and failed experiments [101]. The H1ESC study sought to optimize the

state of the art in stem cell differentiation by evaluating gene expression throughout the differentiation process.

4.1.1 H1ESC Study Methods

The H1ESC study [99] was an exploration of recipes (protocols) for growing commercially available stem cells into beating heart tissue. Pluripotent stem cell differentiation is an expensive task, and a prerequisite to many studies on living human cells [102]. The state of the art methods were inefficient, in the sense that many attempts failed and retrials required. Our collaborator, Dr. John Lough of the Medical College of Wisconsin Department of Cellular and Molecular Biology, had an idea that tweaking the recipe would activate growth patterns in a more precise manner. We took transcriptomic measures across a time-series of the cells' growth and saw pathways activating sequentially. Time-series information had not been available on this model system, so the scientific community did not know the exact timing and level of gene expression on the genomic scale for these processes.

Duplicate samples from each of six time points were collected and RNA was extracted as described previously [99]. Samples were sequenced on an Illumina HiSeq 2000 at the Human and Molecular Genetics Center of The Medical College of Wisconsin, and the Illumina CASAVA pipeline produced FASTQ files from the fluorescence data. Sequence data were processed with RSEM into a measure of transcripts per million mapped (TPM) [103]. RSEM uses Bowtie's unspliced mode [73] onto a reference transcriptome of 38,653 sequences. A data matrix was compiled in R consisting of 6 columns (samples) and 38,653 rows (gene transcripts.) Genes were selected by the

domain expert by colloquial name, and converted to transcript IDs by selecting the highest average expression when more than one transcript were present.

4.1.2 H1ESC Study Results

The H1ESC study explored gene expression on an ad-hoc basis. Genes known to be relevant for the biological process were reviewed. Figure 2 of that study, here excerpted into **Figure 4.1**, shows three biological pathways. From day 1 to day 3 for example we see the precipitous increase of "pan-mesoderm" profiles and "early cardiomyogenesis signaling". We learned a sequential gene activation (*FGFR* series) indicated a possible optimization. The differentiation protocol was modified, and we achieved an increase in cell survival, implying lower cost, and enabling more experiments [99]. With the transcriptome measures at tens of thousands of genes, we relied on experts to identify a handful of important genes, and their categorization.

4.2.3 H1ESC Study GSEA

To go beyond a manual annotation of functional sets, one could use GSEA on the H1ESC dataset. The data matrix originally containing 38,653 transcript identifiers is not compatible with GSEA, and these can be converted to 23,099 named genes for use with GSEA's gene symbol annotations. The Bioconductor package **org.hs.eg.db** was used to convert gene names and included non-gene LINC and MIR elements. GSEA requires a selection of a gene set collection; **c5.all.v5.1.symbols.gmt** was used to select an up to date Gene Ontology set, containing 1,454 gene sets. Analysis was restricted to sets under 31 genes each to ensure comparable results with **rgsepd**'s default parameters. 801 gene sets passed the set size filter. GSEA supports simple two class comparisons and ignores

other samples in the data file. Comparing days 1 and 3 consists of a 2 sample versus 2 sample test, which is below the minimum required for the default gene expression analysis method "Signal2Noise." Processing was completed under the "Diff of classes" scheme, wherein all genes are ranked by the absolute difference in the means between comparison classes.

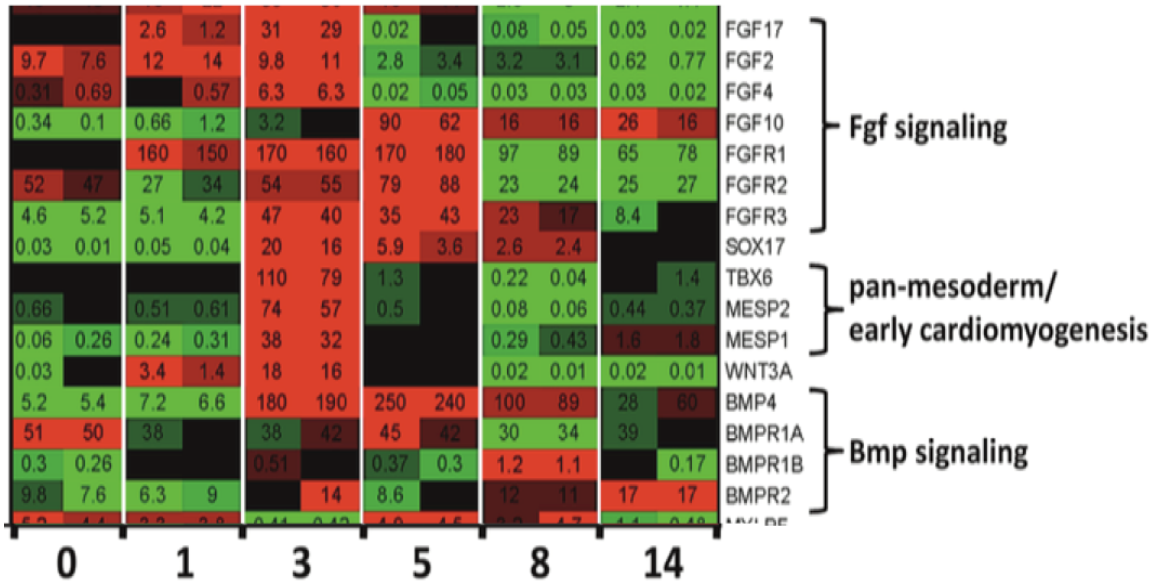

Figure 4.1 Excerpt of Figure 2 of the H1ESC Study, cropped to show annotations of three gene groups. Time points are marked along the bottom, in days post pluripotency induction. The color corresponds to the z-score of the log of the expression, such that high absolute expression for each gene is shown in red, and low in green. Black represents mean (or baseline) expression. Unique to a sequencing based assay, we also annotate absolute gene expression in each cell (black text, units are 'transcripts per million' [103]).

Using the ranked list of 23,099 genes, GSEA reports separate lists of sets found upregulated in each comparison class. Eighty-two gene sets were found significantly enriched and upregulated among day 1 samples, and zero were found among day 3 (both $p<0.05$). The top ten sets from comparing day 1 and 3 are listed in **Table 4.1**. The most significant set was "epithelial to mesenchymal transition," which is expected in this context. The tenth being "muscle cell differentiation" is a good indicator that the samples are doing what we wanted them to. GSEA provides no sample-to-set scores at this high level, but drilling into each category shows a separate gene-level heatmap of expression of the four samples involved in the comparison.

NAME	ES	NOM p-val	FDR q-val
EPITHELIAL_TO_MESENCHYMAL_TRANSITION	0.87	2.08E-03	0.194
CALCIUM_INDEPENDENT_CELL_CELL_ADHESION	0.74	0.00E+00	0.270
INWARD_RECTIFIER_POTASSIUM_CHANNEL_ACTIVITY	0.81	2.04E-03	0.271
MYOBLAST_DIFFERENTIATION	0.78	1.00E-03	0.272
NEUROPEPTIDE_RECEPTOR_ACTIVITY	0.74	0.00E+00	0.287
SARCOMERE	0.76	4.04E-03	0.289
NEUROPEPTIDE_BINDING	0.73	0.00E+00	0.292
MONOVALENT_INORGANIC_CATION_HOMEOSTASIS	0.75	1.11E-02	0.300
INTEGRIN_COMPLEX	0.73	4.00E-03	0.304
MUSCLE_CELL_DIFFERENTIATION	0.73	1.00E-03	0.305

Table 4.1 H1ESC Study GSEA Results. Top ten sets from the comparison of day 1 versus day 3. Four columns extracted from the GSEA report correspond to the set's identifier, the enrichment score, the nominal p-value, and the false discovery rate adjusted significance.

4.2.4 H1ESC Study GSEPD

Running **rgsepd** on the read data generated by RNA-Seq and RSEM, any two classes can be contrasted in depth. Starting from count data at 38,653 gene transcripts, 4,640 were identified as significantly changing between samples of class day 1 and day 3. For this run, **rgsepd** was configured to require $p<0.01$ to consider individual genes differentially expressed, thereby yielding a more precise and conservative result set than if the traditional threshold of $p<0.05$ was used. The functional annotation segment of the pipeline finds 3,720 GO terms enriched $p<0.05$. The projections and segregation analyses evaluated sets passing gene count filters, and found 1,050 sets to significantly segregate samples of day 1 from samples of day 3.

GSEPD produces the GO term based heatmap, seven rows of which are shown in **Figure 4.2** (from the generated file **GSEPD.HMA.D1x2.D3x2.pdf**), with twelve samples displayed. Day 0 being green indicates its similarity to the comparison class day 1 with respect to the gene sets shown. As mentioned in **Chapter 3**, showing the position of non-tested samples like day 5 can be informative to give context to the otherwise high dimensional data. It is validating, informative, and potentially hypothesis-generating for the researchers of this study to see samples of class day 5 being more like samples of class day 3 than samples of class day 1 in the first gene set: mesenchyme morphogenesis, as this was a mesenchymal development protocol. Overall, day 8 and day 14 samples being more like day 3 samples indicates a linear sequential development, wherein, with respect to the functional sets that changed between day 1 to day 3, later stages share more in common with the later time point.

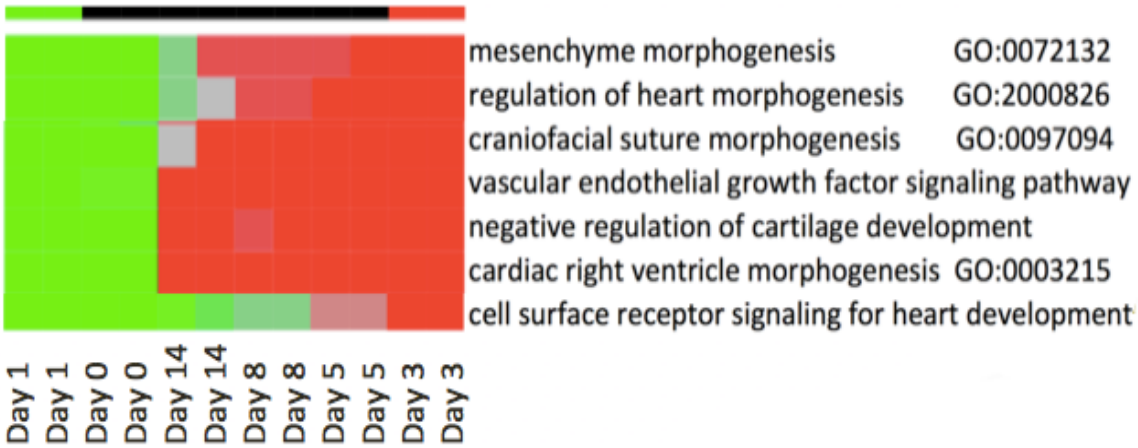

Figure 4.2 Excerpt from the GSEPD Result's HMA comparing day 1 to day 3. The colored bar along the top notes the comparison samples' class labels. Sample classes are listed along the bottom. Rows and Columns are ordered by a hierarchical clustering such that similar samples are adjacent.

To take a broader view, unbiased by an individual's interest in each gene, we can perform GSEPD on sequential time-points to identify the broader processes at work. GSEPD identified "mesenchyme morphogenesis" at the top of the **Figure 4.2**. In the textual result (**Table 4.2**), many pathways were identified as completely segregating the conditions. The third row of **Table 4.2** is an extremely precise "cell surface receptor signaling pathway involved in heart development" with 14 genes differentially expressed (of 24 total). Six of the 16 genes on **Figure 4.1** are "cell surface receptors" (*FGFR1*, *FGFR2*, *FGFR3*, *BMPR1A*, *BMPR1B*, and *BMPR2*). *The GSEPD tool is validated by identifying the same activities an expert would identify.*

GO ID	GOSeq P	Differentially Expressed Genes	Genes in Set	GO Term
GO:0003197	5.49E-11	16	27	endocardial cushion development
GO:0071294	4.99E-10	9	13	cellular response to zinc ion
GO:0061311	1.26E-09	14	24	cell surface receptor signaling pathway involved in heart development
GO:0072132	6.71E-09	15	30	mesenchyme morphogenesis
GO:2000826	6.98E-09	14	27	regulation of heart morphogenesis

Table 4.2 H1ESC Study GSEPD Results. Top five rows by significance, from generated table **GSEPD.HMA.D1x2.D3x2.csv**. These pathways are identified in the comparison of day 1 versus day 3 with GSEPD. The columns correspond to gene set ID, GOSeq significance, number of genes found differentially expressed, the number of genes in the set, and the set's name. These results are sorted by GOSeq p-value, because all sets were found to have equivalent 100% segregation due to low sample variability.

We also find many unexpected gene sets with similar evidence levels. In **Table 4.2**, second after "GO:endocardial cushion development", GSEPD finds "GO:cellular response to zinc ion" with 9 genes differentially expressed of 13 in the GO term (9/13 DE). The cellular sample system seems to be in a known response pattern to zinc ions, which may lead to an extension of this heart development pathway in further research.

Along the right of **Figure 4.1** are manual annotations of gene sets. They are recapitulated by GSEPD in a completely database driven manner, bypassing any individual's preconceptions. GSEPD's ability to quickly identify biologically relevant pathway perturbations from transcriptome data enables faster research and unexpected findings.

Four key genes are highlighted as "Bmp signaling" near the bottom of **Figure 4.1**. GSEPD reports "response to BMP" with 13/22 genes DE (data not shown). The top of

Figure 4.1 is a group of manually annotated genes under the heading "Fgf signaling". GSEPD reports "type 2 fibroblast growth factor receptor binding" with 4/4 genes DE at $p<10^{-4}$. The middle category, "pan-mesoderm/early cardiomyogenesis" is seen as GSEPD's fifth-most significant set, "regulation of heart morphogenesis" at 14/27 genes DE for $p<10^{-8}$ (**Table 4.2**). These 14 identified genes go beyond the three known markers represented in the H1ESC publication. Two other interesting results, at 7/9 genes DE identified are "neurofilament" and "cell migration involved in heart development" in completely disjoint sets. Specifically, between day 1 and day 3 GSEPD identified 1050 GO terms (37 after Bonferroni correction). The segregation scores are maxed-out at 100% due to the small sample of two versus two cloned samples.

With GSEPD I propose we can discover new connections, by harnessing entire databases and presenting systems biology level results in a compact and convenient manner. For example, a novel and unexpected finding is the second row of **Table 4.2**, "cellular response to zinc ion". The word zinc appears nowhere in the publication. Opening the results file GSEPD.RES.D1x2.D3x2.GO2.csv reveals all genes and how each gene's expression changed underlying each GO term. It lists 13 genes under the "cellular response to zinc ion" heading GO:0071294, here included as **Table 4.3**. The **GO2** file (as described in **Appendix A**) collects mean expression in each class (here testing day 1 versus day 3), the test statistics from the differential analysis, and the functional analysis statistics as a cross-product. The **GO2** file allows a simple spreadsheet search function to extract the computed metrics of each gene within the identified GO term. Neither DESeq2 nor GOSeq could provide the combined product.

Quick retrieval of the gene-based evidence underlying each GO term result is a marked functional advancement for the biomedical end-user.

Gene	D1x2	D3x2	PADJ	Gene	D1x2	D3x2	PADJ
CREB1	10.1	10.2	7.13E-01	*HVCN1*	8.81	11.7	4.50E-12
MT2A	12.8	8.43	3.05E-43	*MT1F*	12.3	8.34	1.37E-36
MT1G	12.2	6.85	1.05E-61	*MT1B*	6.86	6.66	8.26E-01
MT1H	10.8	6.58	1.88E-88	*TSPO*	10	9.97	9.88E-01
MT1M	9.5	6.49	1.32E-30	*MT1E*	12.1	7.08	7.51E-95
MT1A	8.7	6.85	5.14E-11	*KCNK3*	6.71	6.69	9.88E-01
MT1X	12.3	8.94	2.22E-43				

Table 4.3 Subsection of GSEPD.RES.D1x2.D3x2.GO2.csv rows 286,317 through 286,329. This file includes every moderately significant GO Term and all genes underlying each, for the purpose of quickly extracting a genomic profile. Column Gene is the HGNC gene name (The Human Genome Organization's Gene Nomenclature Committee is the authority by which genes are given registered identifiers). Column D1x2 is the mean expression of the day 1 class, of which there were two samples. Units are from DESeq2, as a variance-stabilized and log2-normalized read count of the original RNA-Seq, such that 11 is approximately twice the expression of 10. Column D3x2 is the day 3 class mean. Column PADJ is the multiple-testing corrected p-value from the DESeq2 test for differential expression between those classes. This section of the larger listing corresponds to only those genes that are a part of GO:0071294.

The GSEPD results shown in **Table 4.3** note the expression of genes underlying "cellular response to zinc ion." Nine of 13 genes are highly significant with 8 reducing between day 1 to day 3, one gene increased (*HVCN1* from 8.8 units to 11.7), and four unchanged. Metallothioneins 1G, 1H, 1M, 2A, and 1A drop precipitously between the time points, while *HVCN1* activates. A zinc-related finding was unexpected and would have gone unnoticed without GSEPD's merging of gene and gene set based results. A simpler functional annotation tool would bury the interesting results: the GSEPD

intermediary step GOSeq found 1,894 gene sets overrepresented ($p<0.01$) in the day 1 versus day 3 differential expression analysis.

4.2 Genetic/Mechanistic Link in A Congenital Heart Disease

In the second study of this chapter I demonstrate that GSEPD is useful in identifying relevant molecular pathways (gene sets) when sample classes are human disease risk factors.

4.2.1 MYH6 Study Methods

In 2015 we completed a clinical association study known as MYH6, for the gene identified therein, known to be vital for muscle contraction [100]. In the MYH6 study, we started with the genomes of a family with a particular CHD, explored a population with the same disease and made a case for the role of *MYH6* in that CHD subtype. This particular CHD subtype is present in the fetus, and fatal before adulthood [104], precluding high penetrance causes from permeating the population, therefore this mutation has not been thoroughly explored by the medical community.

The genomes of a nuclear family were compared to databases of mutational significance, and a short list of possibly causal genes was developed. We then looked at a population of 190 other patients and evaluated each of the possibly causal genes for occurrence frequency in a case/control analysis. We then contrasted tissue transcriptomes of carriers and non-carriers for *MYH6* variants to find what we call the compensatory pathway: the body's natural response to the mutation.

The MYH6 study explores the transcriptome of several samples of heart tissue. Sequence data collection and processing was as described in section **4.1.1**. Because the

transcriptomes were predominantly divided by their subject's age and tissue type, the more subtle profiles could be lost to noise [105]. To ensure cleaner results, we used a paired test, which limits the available data to those with suitable carrier/non-carrier samples of matched subject age and tissue type. Samples were labeled WV and WR, for those with variant and those with the reference allele. A paired test controls for covariates and yields a short list of genes. We used edgeR [81] with a factor covariate indicating manually identified sample pairs. Criteria for significance was set at $p<0.05$, effect size absolute $\log_2$ difference > 1, and high average expression (defined here as logCPM > 6).

4.2.2 MYH6 Study Results

Twenty-four gene transcripts were both highly expressed and significantly perturbed (**Table 4.4**). *MYH7* was immediately recognized as a gene homologue to *MYH6* (sharing >95% sequence similarity.) Upregulation of *MYH7* is filling the role of a defunct *MYH6*. The compensatory pathway for *MYH6* (a muscle contractile protein) seems to be elevated expression of several other genes related to muscle contraction. Other findings, such as the *EEF* and *RPL* families are possibly the pathway the body uses to select or regulate the proteins used by cardiac cells.

4.2.3 MYH6 Study GSEA

The sample data of the paired analysis can be processed with GSEA. GSEA performs two class contrast analyses, designed for microarray chips, but suitable for other transcriptome measures. Parameters to GSEA were set to match GSEPD as closely as possible: using the Gene Ontology collection (**c5.all.v5.1.symbols.gmt**) restricted to gene sets of size two to 31 genes. GSEA supports several statistical models. Two class comparison was configured to use the "tTest" mode.

Gene	Change Among Carriers	P-Value	Gene	Change Among Carriers	P-Value
ACTA1	315%	0.0051	*MYL2*	366%	0.0039
ALDOA	210%	0.0025	*RPL12*	45%	0.0086
COX6A2	250%	0.0070	*RPL17*	23%	0.0003
DQ668365	265%	0.0007	*RPL21*	15%	0.0001
EEF1A1	36%	0.0033	*RPL41*	21%	0.0017
EEF1B2	50%	0.0053	*RPL9*	28%	0.0012
EF011072	325%	0.0038	*RPS26*	21%	0.0034
ENO3	312%	0.0002	*RPS27A*	36%	0.0003
H3F3AP4	39%	0.0017	*RPS3A*	42%	0.0040
HHATL	241%	0.0026	*RPSA*	22%	0.0007
HNRNPA1	46%	0.0015	*TNNT2*	1231%	0.0028
MYH7	346%	0.0007	*TPM2*	221%	0.0009

Table 4.4 MYH6 Study Paired Test. Twenty-four significant gene level changes were found in the carrier versus non-carrier test [100]. The DEG group into twelve identifiable gene families. Two sets of 3 columns indicate the gene name, effect size, and significance computed with **edgeR** [81].

The dataset originally contained 38,653 transcripts, and was collapsed to 23,099 named genes to be compatible with the GSEA annotation of GO. Significant sets were calculated from the differentially expressed gene list as a ranked object. The output GO collection lists 1,454 sets, 801 of which pass the set size filter. GSEA reports enrichment of 34 sets upregulated in phenotype WV and 2 in WR (with nominal p<1%). Top ten significant sets are shown in **Table 4.5**. GSEA does not score each sample separately for any set's expression profile.

The third row of **Table 4.5** mentions "contractile fiber" and constitutes the agreement with the manual process's indication of a compensatory pathway. The set "mitochondrial respiratory chain," being upregulated in class WV represents an effect

that was unnoticed because the genes that make up this set were not part of the 24 transcripts identified in the MYH6 study.

NAME	ES	NOM p-val	FDR q-val
MITOCHONDRIAL_RESPIRATORY_CHAIN	0.71	0.00E+00	0.037
REGULATION_OF_PROTEIN_AMINO_ACID_PHOSPHORYLATION	0.63	0.00E+00	0.050
CONTRACTILE_FIBER	0.64	1.91E-03	0.056
HYDROGEN_ION_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.64	0.00E+00	0.061
RESPIRATORY_CHAIN_COMPLEX_I	0.73	0.00E+00	0.067
NADH_DEHYDROGENASE_COMPLEX	0.73	2.01E-03	0.073
CONTRACTILE_FIBER_PART	0.66	0.00E+00	0.079
GABA_RECEPTOR_ACTIVITY	0.75	4.00E-03	0.094
PEPTIDYL_TYROSINE_PHOSPHORYLATION	0.61	0.00E+00	0.098
MITOCHONDRIAL_RESPIRATORY_CHAIN_COMPLEX_I	0.73	0.00E+00	0.106

Table 4.5 MYH6 Study GSEA Results. Top ten identified gene sets found by the analysis of the MYH6 carrier versus non-carrier differential expression analysis [100] with GSEA [88] ranked by q-value. Column ES is the enrichment score, NOM p-val is the nominal set significance, FDR q-val is the false-discovery rate adjusted significance.

4.2.4 MYH6 Study GSEPD

GSEPD performs two class contrast analyses where the MYH6 study's manual process used a paired-samples test. To circumvent the experimental design limitation, the user may take advantage of **rgsepd**'s file architecture and insert differential gene expression results into the pipeline. Inserting the results of the pairwise differential expression analysis (Table S3 of [100] consists of **Table 4.4** extended to the whole genome) as the differential expression table of GSEPD lets the functional annotation and projection components operate on the gene results from a more complicated differential expression test.

The GSEPD pipeline found 617 gene transcripts to be significantly differentially expressed, 422 GO terms upregulated in non-mutant, and 254 GO terms upregulated in mutant. Combining the 422 and 254 GO terms and filtering to sets with the requested number of genes (2 to 31 by default), GSEPD evaluated 392 sets (139 consisting of up, 230 down, and 23 with mixed up/down) for segregating ability. Finally GSEPD found 33 gene sets with significant sample segregation (seven of which are represented in **Figure 4.3**). The results include (at rows 4,5, and 6) the muscle contraction pathways found by both the manual process and GSEA. The other sets identified by GSEPD are potentially therapeutic pathways.

Unique to GSEPD is a sample-to-set scoring, represented by the colors of **Figure 4.3**. The WV class defines the red expression profile, and the WR class defines the green. Two samples are seen to be on the wrong side: MAAA1600-WV and VFY-WR, indicating their response to the mutation was inconsistent with the trend discovered in the other six pairs of samples at these gene sets. Some differences are expected, as each subject had a different form of the mutant gene, and a different familial background. The unexpected gene expression measures of samples MAAA1600-WV and VFY-WR open up further avenues for experimentation.

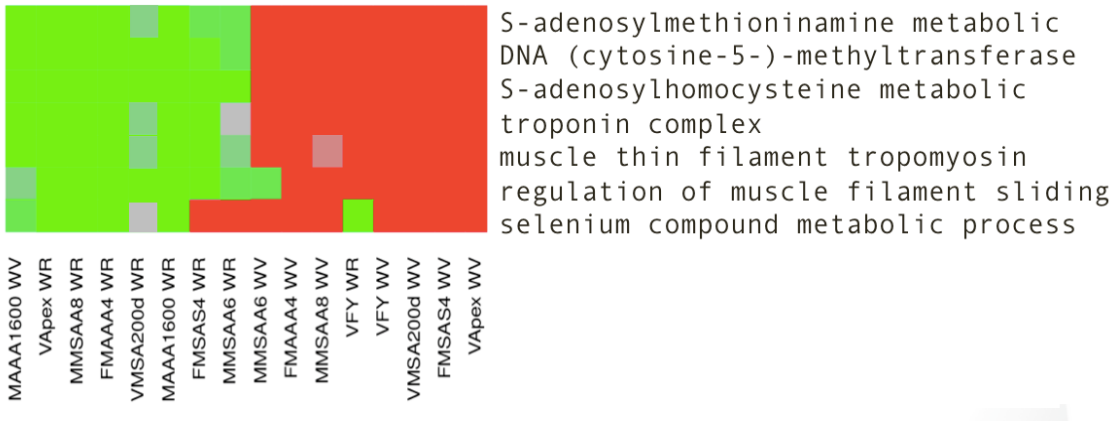

Figure 4.3 MYH6 Paired Analysis with GSEPD, testing 8 *MYH6*-mutant tissue samples (WV: with variant) against 8 non-mutant (WR: with reference), all sourced from the CHD subjects' explanted tissue. Figure is cropped to show only a few rows of a larger result set. Subject pairing identifiers are noted along the bottom and labeled with **WV** for mutant carrier of a *MYH6* variant, or **WR** for wild-type *MYH6*.

4.3 Summary

In Chapter 4 GSEPD and GSEA were applied to two case studies where the domain expert manually annotated function of gene sets found interesting by his or her transcriptome experiment. Finding expression patterns of gene sets via manual review is a major bottleneck of time and effort. The cost potentially prevents complete data exploration. For example, in both case studies the findings were limited to genes recognized by the domain expert, while other findings went unreported. GSEPD provided a window to browse the high dimensional transcriptome data that has *to date* been opaque. I have shown GSEPD can recapitulate the pathway-based findings, while providing a sample-specific gene set analysis. The novel multidimensional projection technique employed by GSEPD highlights systems-level patterns on individual samples.

In the H1ESC study consecutive time points were not carefully analyzed, in part because six time points have fifteen ways to contrast, for which the experimental design was underpowered. With GSEPD fifteen configurations are quickly run, and the user may sort and rank the results together, or quickly browse each comparison's result set. In the MYH6 study we found one weakly defined pathway that was completely novel, as the *MYH6* knockdown genotype is unstudied in humans. GSEPD identifies specific pathways such as "regulation of muscle filament sliding" or "selenium compound metabolic process." which had gone unnoticed, but are avenues of future research for treatment or prevention of human CHD.

CHAPTER 5

CONCLUSION

Computational Sciences is an interdisciplinary research field, aiming to bring the tools of data analytics to other domains. While many scientific domains are in need of computational support, medical bioinformatics is an exciting avenue supporting biomedical research: a field where real advances have the opportunity to save lives.

5.1 Scales Untenable

Biomedical research, specifically human development and genomics, is in a particularly difficult state. Samples are rare or precious, we cannot generate experimental tissues of human fetuses, and the samples we do obtain have all the diversity of mankind. I see this as an opportunity to deploy computational sciences in support of valuable medical research. Human genomes are enormous and so diverse that the next generation reference maps are nonlinear graphical models, challenging the paradigms of sequence alignment and simple gene comparisons. Medicine needs smarter tools. My research work is to create smarter tools that integrate broad knowledgebases such that the culmination can be applied in every forthcoming transcriptome analysis.

5.2 Complementary Datasets

Careful literature review reveals myriad small scale studies where one or a small handful of genes' usages are explained. Collection of these results is a task outside the grasp of any one person. Future tools in AI like IBM Watson are being deployed to scour the medical literature. In that theme, the tools developed and methods applied in the current work aim to bring together large-scale datasets in the form of public databases, which, while ostensibly available, are out of reach for routine annotation on everyday analyses. Researchers deserve to have large scale databases brought to them on their terms, freely and quickly. A wealth of interaction databases are available, but without computational support to annotate their findings, medical researchers are manually perusing interesting hits piecemeal. My work has been to collapse and process data into a meaningful picture. My future work is to bring down the barriers with intelligent automation, amplifying research results.

5.3 Transcriptomics

Moving beyond point mutations and DNA rearrangements to gene usage and gene set usage will enable future researchers to make more effective use of their precious tissue samples. More knowledge could be mined from the same animal experiments or human surgical discards. Other high throughput experiment types are being developed every year, and they too will need intelligent overviews that GSEPD can provide.

The methods of **Chapter 3** can be applied to other data types. Several other sequencing technologies generate genome-wide scores analogous to RNA-Seq's

quantification of transcript activity, such as bisulfite sequencing and methylation specific sequencing [106]. These could also be fed into **rgsepd**. Like with RNA-Seq, epigenetic activation is quantification at the gene level, inducing the same kinds of data deluge we sought to remedy in **Chapter 3**. Collection to the GO Term level and clustering with validity scores could yield more informative results in those research areas [107]. Furthermore, adaptation to non-human experiments is an obvious next step for **rgsepd** version 2. The limited input format of **rgsepd** is RefSeq human transcript identifiers, but more flexible gene name conversion routines would enable any model organism.

5.4 Applications

In this dissertation I have shown two studies in **Chapter 4** where the experts in their field have desired to annotate functional significance to their gene-centric analyses. Although other tools exist, they are limited in interface and applicability. I have created a human transcriptome analyzer in GSEPD that shows inherently high dimensional data as a simple colored heatmap: using vector projections to directly calculate how each sample behaves like each test condition. I believe set-level expression is the way the end-user-researcher understands their data, and projection provides the most understandable dimension reduction. Finally, Gene Ontology analyses are accessible on a sample-to-sample level, and I hope to highlight not just the expected pathways, but the many annotated results that are currently going unseen in vast databases.

In the MYH6 study we showed the effects of a mutation in a heart muscle being compensated by the living host. GSEPD also predicts impacts on the eosinophils and forebrain neurons. These effects would have gone unnoticed, as researchers in neuron

development are unlikely to be interested in a cardiac-defect publication. Bringing together the online databases is the best way to advance biomedical research at the fastest pace possible. The MYH6 study's variant carriers versus non-carriers analysis of **Chapter 4** mentioned a selenium metabolism process. An expression profile exists with statistically significant segregation ability, indicating a real difference related to the *MYH6* mutant status. GSEPD indicates selenium is potentially an avenue for treatment, wherein a drug dose could influence the gene expression, bringing unhealthy patients back in line with the non-carriers.

GSEPD enables a systems-level evaluation of any test treatment on a cell culture. In the H1ESC study we evaluated the effects of Activin-A and *BMP* dosage on heart muscle development. GSEPD could reveal the downstream effects and further streamline the differentiation protocol. A systematic analysis of published data sets like the Illumina Bodymap project or The Cancer Genome Atlas would reveal thousands of subtle systems level effects that have been obscured behind the dimensionality of the transcriptome.

The role Computational Science plays to non-computing domains is a role of analytical and interpretive support. We have a duty to the scientific community to help and advance where the need is greatest. My work has had impacts in the study of organ development [7], treatment of congenital diseases [100], and basic stem cell research [99]. **rgsepd**, as open source software, is freely available to facilitate transcriptome analyses around the world [91]. I hope it has a positive impact on many future studies.

BIBLIOGRAPHY

1. Langman, J. and T.W. Sadler, *Langman's Medical Embryology*. 6th ed. 1990, Baltimore: Williams and Wilkins. xii, 409 p.

2. Day, N. and L.B. Holmes, *The Incidence of Genetic Disease in a University Hospital Population.* American Journal of Human Genetics, 1973. **25**(3): p. 237-46.

3. Roberts-Galbraith, R.H. and P.A. Newmark, *On the Organ Trail: Insights into Organ Regeneration in the Planarian.* Current Opinion in Genetics and Development, 2015. **32**: p. 37-46.

4. Tang, W.W., S. Dietmann, N. Irie, H.G. Leitch, V.I. Floros, C.R. Bradshaw, J.A. Hackett, P.F. Chinnery, and M.A. Surani, *A Unique Gene Regulatory Network Resets the Human Germline Epigenome for Development.* Cell, 2015. **161**(6): p. 1453-67.

5. Menke, C., M. Cionni, T. Siggers, M.L. Bulyk, D.R. Beier, and R.W. Stottmann, *Grhl2 Is Required in Nonneural Tissues for Neural Progenitor Survival and Forebrain Development.* Genesis, 2015. **53**(9): p. 573-582.

6. Arrington, C.B., S.B. Bleyl, N. Matsunami, G.D. Bonnell, B.E. Otterud, D.C. Nielsen, J. Stevens, S. Levy, et al., *Exome Analysis of a Family with Pleiotropic Congenital Heart Disease.* Circulation Cardiovascular Genetics, 2012. **5**(2): p. 175-82.

7. Bazil, J.N., K. Stamm, X. Li, R. Thiagarajan, T.J. Nelson, A. Tomita-Mitchell, and D.A. Beard, *The Inferred Cardiogenic Gene Regulatory Network in the Mammalian Heart.* PLoS ONE, 2014. **9**(6): p. e100842.

8. Posch, M.G., A. Perrot, F. Berger, and C. Ozcelik, *Molecular Genetics of Congenital Atrial Septal Defects.* Clinical Research in Cardiology, 2010. **99**(3): p. 137-47.

9. Priest, J.R., K. Osoegawa, N. Mohammed, V. Nanda, R. Kundu, K. Schultz, E.J. Lammer, S. Girirajan, et al., *De Novo and Rare Variants at Multiple Loci Support the Oligogenic Origins of Atrioventricular Septal Heart Defects.* PLoS Genetics, 2016. **12**(4): p. e1005963.

10. Manolio, T.A., F.S. Collins, N.J. Cox, D.B. Goldstein, L.A. Hindorff, D.J. Hunter, M.I. McCarthy, E.M. Ramos, et al., *Finding the Missing Heritability of Complex Diseases.* Nature, 2009. **461**(7265): p. 747-53.

11. Lescai, F. and C. Franceschi, *The Impact of Phenocopy on the Genetic Analysis of Complex Traits.* PLoS ONE, 2010. **5**(7): p. e11876.

12. Balding, D.J., *A Tutorial on Statistical Methods for Population Association Studies.* Nature Reviews Genetics, 2006. **7**(10): p. 781-91.

13. Li, X., A. Martinez-Fernandez, K.A. Hartjes, J.P. Kocher, T.M. Olson, A. Terzic, and T.J. Nelson, *Transcriptional Atlas of Cardiogenesis Maps Congenital Heart Disease Interactome.* Physiological Genomics, 2014. **46**(13): p. 482-95.

14. Greulich, F., M.O. Trowe, A. Leffler, C. Stoetzer, H.F. Farin, and A. Kispert, *Misexpression of Tbx18 in Cardiac Chambers of Fetal Mice Interferes with Chamber-Specific Developmental Programs but Does Not Induce a Pacemaker-Like Gene Signature.* Journal of Molecular and Cellular Cardiology, 2016. **97**: p. 140-149.

15. Hirschhorn, J.N., K. Lohmueller, E. Byrne, and K. Hirschhorn, *A Comprehensive Review of Genetic Association Studies.* Genetics in Medicine, 2002. **4**(2): p. 45-61.

16. Lewis, C.M., *Genetic Association Studies: Design, Analysis and Interpretation.* Briefings in Bioinformatics, 2002. **3**(2): p. 146-53.

17. Abegglen, L.M., A.F. Caulin, A. Chan, K. Lee, R. Robinson, M.S. Campbell, W.K. Kiso, D.L. Schmitt, et al., *Potential Mechanisms for Cancer Resistance in Elephants and Comparative Cellular Response to DNA Damage in Humans.* The Journal of the American Medical Association, 2015. **314**(17): p. 1850-60.

18. Watson, E., L.T. MacNeil, A.D. Ritter, L.S. Yilmaz, A.P. Rosebrock, A.A. Caudy, and A.J. Walhout, *Interspecies Systems Biology Uncovers Metabolites Affecting C. Elegans Gene Expression and Life History Traits.* Cell, 2014. **156**(4): p. 759-70.

19. Vaquerizas, J.M., S.K. Kummerfeld, S.A. Teichmann, and N.M. Luscombe, *A Census of Human Transcription Factors: Function, Expression and Evolution.* Nature Reviews Genetics, 2009. **10**(4): p. 252-63.

20. Scott, W.K., M.A. Pericak-Vance, and J.L. Haines, *Genetic Analysis of Complex Diseases.* Science, 1997. **275**(5304): p. 1327; author reply 1329-30.

21. Hughes, D.J., *Use of Association Studies to Define Genetic Modifiers of Breast Cancer Risk in Brca1 and Brca2 Mutation Carriers.* Familial Cancer, 2008. **7**(3): p. 233-44.

22. Syed, D.N., M.I. Khan, M. Shabbir, and H. Mukhtar, *MicroRNAs in Skin Response to UV Radiation.* Current Drug Targets, 2013. **14**(10): p. 1128-34.

23. Emmert-Streib, F. and G.V. Glazko, *Pathway Analysis of Expression Data: Deciphering Functional Building Blocks of Complex Diseases.* PLoS Computational Biology, 2011. **7**(5): p. e1002053.

24. Shaikh, T.H., X. Gai, J.C. Perin, J.T. Glessner, H. Xie, K. Murphy, R. O'Hara, T. Casalunovo, et al., *High-Resolution Mapping and Analysis of Copy Number Variations in the Human Genome: A Data Resource for Clinical and Research Applications.* Genome Research, 2009. **19**(9): p. 1682-90.

25. Tomita-Mitchell, A., D.K. Mahnke, C.A. Struble, M.E. Tuffnell, K. Stamm, M. Hidestrand, S.E. Harris, M.A. Goetsch, et al., *Human Gene Copy Number Spectra Analysis in Congenital Heart Malformations.* Physiological Genomics, 2012. **44**(9): p. 518-41.

26. Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, et al., *Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium.* Nature Genetics, 2000. **25**(1): p. 25-9.

27. Kramer, A., J. Green, J. Pollard, Jr., and S. Tugendreich, *Causal Analysis Approaches in Ingenuity Pathway Analysis.* Bioinformatics, 2014. **30**(4): p. 523-30.

28. Ioannidis, J.P., G. Thomas, and M.J. Daly, *Validating, Augmenting and Refining Genome-Wide Association Signals.* Nature Reviews Genetics, 2009. **10**(5): p. 318-29.

29. Edwards, R. and L. Glass, *Combinatorial Explosion in Model Gene Networks.* Chaos, 2000. **10**(3): p. 691-704.

30. Metzker, M.L., *Sequencing Technologies - the Next Generation.* Nature Reviews Genetics, 2010. **11**(1): p. 31-46.

31. Shanahan, H.P., A.M. Owen, and A.P. Harrison, *Bioinformatics on the Cloud Computing Platform Azure.* PLoS ONE, 2014. **9**(7): p. e102642.

32. Zhou, S., R. Liao, and J. Guan, *When Cloud Computing Meets Bioinformatics: A Review.* Journal of Bioinformatics and Computational Biology, 2013. **11**(5): p. 1330002.

33. Roson-Burgo, B., F. Sanchez-Guijo, C. Del Canizo, and J. De Las Rivas, *Transcriptomic Portrait of Human Mesenchymal Stromal/Stem Cells Isolated from Bone Marrow and Placenta.* BMC Genomics, 2014. **15**: p. 1-18.

34. Ramskold, D., E.T. Wang, C.B. Burge, and R. Sandberg, *An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data.* PLoS Computational Biology, 2009. **5**(12): p. e1000598.

35. Thomas, D.C., D.V. Conti, J. Baurley, F. Nijhout, M. Reed, and C.M. Ulrich, *Use of Pathway Information in Molecular Epidemiology.* Human Genomics, 2009. **4**(1): p. 21-42.

36. Tabor, H.K., N.J. Risch, and R.M. Myers, *Candidate-Gene Approaches for Studying Complex Genetic Traits: Practical Considerations.* Nature Reviews Genetics, 2002. **3**(5): p. 391-7.

37. Pattin, K.A. and J.H. Moore, *Role for Protein-Protein Interaction Databases in Human Genetics.* Expert Review of Proteomics, 2009. **6**(6): p. 647-59.

38. Li, M.J., P. Wang, X. Liu, E.L. Lim, Z. Wang, M. Yeager, M.P. Wong, P.C. Sham, et al., *GWASdb: A Database for Human Genetic Variants Identified by Genome-Wide Association Studies.* Nucleic Acids Research, 2012. **40**(Database issue): p. D1047-54.

39. Parla, J.S., I. Iossifov, I. Grabill, M.S. Spector, M. Kramer, and W.R. McCombie, *A Comparative Analysis of Exome Capture.* Genome Biology, 2011. **12**(9): p. 1-17.

40. Worthey, E.A., A.N. Mayer, G.D. Syverson, D. Helbling, B.B. Bonacci, B. Decker, J.M. Serpe, T. Dasu, et al., *Making a Definitive Diagnosis: Successful Clinical Application of Whole Exome Sequencing in a Child with Intractable Inflammatory Bowel Disease.* Genetics in Medicine, 2011. **13**(3): p. 255-62.

41. Uzarski, J.S., Y. Xia, J.C. Belmonte, and J.A. Wertheim, *New Strategies in Kidney Regeneration and Tissue Engineering.* Current Opinion in Nephrology and Hypertension, 2014. **23**(4): p. 399-405.

42. O'Malley, M.A. and J. Dupre, *Fundamental Issues in Systems Biology.* Bioessays, 2005. **27**(12): p. 1270-6.

43. Goldstein, D.B., K.R. Ahmadi, M.E. Weale, and N.W. Wood, *Genome Scans and Candidate Gene Approaches in the Study of Common Diseases and Variable Drug Responses.* Trends in Genetics, 2003. **19**(11): p. 615-22.

44. Johnson, A.D. and C.J. O'Donnell, *An Open Access Database of Genome-Wide Association Results.* BMC Medical Genetics, 2009. **10**: p. 6.

45. Dickson, S.P., K. Wang, I. Krantz, H. Hakonarson, and D.B. Goldstein, *Rare Variants Create Synthetic Genome-Wide Associations.* PLoS Biology, 2010. **8**(1): p. e1000294.

46. Preuss, M., I.R. Konig, J.R. Thompson, J. Erdmann, D. Absher, T.L. Assimes, S. Blankenberg, E. Boerwinkle, et al., *Design of the Coronary Artery Disease Genome-Wide Replication and Meta-Analysis (Cardiogram) Study: A Genome-Wide Association Meta-Analysis Involving More Than 22 000 Cases and 60 000 Controls.* Circulation Cardiovascular Genetics, 2010. **3**(5): p. 475-83.

47. Myocardial Infarction Genetics Consortium, S. Kathiresan, B.F. Voight, S. Purcell, K. Musunuru, D. Ardissino, P.M. Mannucci, S. Anand, et al., *Genome-Wide Association of Early-Onset Myocardial Infarction with Single Nucleotide Polymorphisms and Copy Number Variants.* Nature Genetics, 2009. **41**(3): p. 334-41.

48. Winchester, L., C. Yau, and J. Ragoussis, *Comparing CNV Detection Methods for Snp Arrays.* Briefings in Functional Genomics & Proteomics, 2009. **8**(5): p. 353-66.

49. Cukier, H.N., M.A. Pericak-Vance, J.R. Gilbert, and D.J. Hedges, *Sample Degradation Leads to False-Positive Copy Number Variation Calls in Multiplex Real-Time Polymerase Chain Reaction Assays.* Analytical Biochemistry, 2009. **386**(2): p. 288-90.

50. Amberger, J.S., C.A. Bocchini, F. Schiettecatte, A.F. Scott, and A. Hamosh, *OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an Online Catalog of Human Genes and Genetic Disorders.* Nucleic Acids Research, 2015. **43**(Database issue): p. D789-98.

51. Thomas, D.C., J.W. Baurley, E.E. Brown, J.C. Figueiredo, A. Goldstein, A. Hazra, R.T. Wilson, and N. Rothman, *Approaches to Complex Pathways in Molecular Epidemiology: Summary of a Special Conference of the American Association for Cancer Research.* Cancer Research, 2008. **68**(24): p. 10028-30.

52. Olson, E.N., *Gene Regulatory Networks in the Evolution and Development of the Heart.* Science, 2006. **313**(5795): p. 1922-1927.

53. Hartemink, A.J., D.K. Gifford, T.S. Jaakkola, and R.A. Young, *Combining Location and Expression Data for Principled Discovery of Genetic Regulatory Network Models.* Pacific Symposium on Biocomputing, 2002: p. 437-49.

54. Jimenez-Marin, A., M. Collado-Romero, M. Ramirez-Boo, C. Arce, and J.J. Garrido, *Biological Pathway Analysis by Arrayunlock and Ingenuity Pathway Analysis.* BMC Proceedings, 2009. **3 Suppl 4**: p. S6.

55. Barabasi, A.L. and Z.N. Oltvai, *Network Biology: Understanding the Cell's Functional Organization.* Nature Reviews Genetics, 2004. **5**(2): p. 101-13.

56. Gardner, T.S. and J.J. Faith, *Reverse-Engineering Transcription Control Networks.* Physics of Life Reviews, 2005. **2**(1): p. 65-88.

57. Bazil, J.N., F. Qi, and D.A. Beard, *A Parallel Algorithm for Reverse Engineering of Biological Networks.* Integrative Biology, 2011. **3**(12): p. 1215-23.

58. Bazil, J.N., G.T. Buzzard, and A.E. Rundell, *A Global Parallel Model Based Design of Experiments Method to Minimize Model Output Uncertainty.* Bulletin of Mathematical Biology, 2012. **74**(3): p. 688-716.

59. Langfelder, P. and S. Horvath, *Eigengene Networks for Studying the Relationships between Co-Expression Modules.* BMC Systems Biology, 2007. **1**: p. 54.

60. Kohonen, T., *Cortical Maps.* Nature, 1990. **346**(6279): p. 24.

61. Myslobodsky, M., *Ingenuity Pathway Analysis of Clozapine-Induced Obesity.* Obesity Facts, 2008. **1**(2): p. 93-102.

62. Wat, M.J., O.A. Shchelochkov, A.M. Holder, A.M. Breman, A. Dagli, C. Bacino, F. Scaglia, R.T. Zori, et al., *Chromosome 8p23.1 Deletions as a Cause of Complex Congenital Heart Defects and Diaphragmatic Hernia.* American Journal of Medical Genetics Part A, 2009. **149A**(8): p. 1661-77.

63. Baud, V., M. Lipinski, E. Rassart, L. Poliquin, and D. Bergeron, *The Human Homolog of the Mouse Common Viral Integration Region, Fli1, Maps to 11q23-Q24.* Genomics, 1991. **11**(1): p. 223-4.

64. Jain, P., S. Vig, M. Datta, D. Jindel, A.K. Mathur, S.K. Mathur, and A. Sharma, *Systems Biology Approach Reveals Genome to Phenome Correlation in Type 2 Diabetes.* PLoS ONE, 2013. **8**(1): p. e53522.

65. Newaz, K., K. Sriram, and D. Bera, *Identification of Major Signaling Pathways in Prion Disease Progression Using Network Analysis.* PLoS ONE, 2015. **10**(12): p. e0144389.

66. Hunter, J.E. and F.L. Schmidt, *Methods of Meta-Analysis : Correcting Error and Bias in Research Findings*. Third edition. 2015, Thousand Oaks, California: SAGE. xxxii, 639 pages.

67. Graveley, B.R., *Alternative Splicing: Increasing Diversity in the Proteomic World.* Trends in Genetics, 2001. **17**(2): p. 100-7.

68. de Jonge, H.J., R.S. Fehrmann, E.S. de Bont, R.M. Hofstra, F. Gerbens, W.A. Kamps, E.G. de Vries, A.G. van der Zee, et al., *Evidence Based Selection of Housekeeping Genes.* PLoS ONE, 2007. **2**(9): p. e898.

69. Sumi, T., N. Tsuneyoshi, N. Nakatsuji, and H. Suemori, *Defining Early Lineage Specification of Human Embryonic Stem Cells by the Orchestrated Balance of Canonical Wnt/Beta-Catenin, Activin/Nodal and Bmp Signaling.* Development, 2008. **135**(17): p. 2969-79.

70. Khatri, P., M. Sirota, and A.J. Butte, *Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges.* PLoS Computational Biology, 2012. **8**(2): p. e1002375.

71. Barrett, T., S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, et al., *NCBI GEO: Archive for Functional Genomics Data Sets--Update.* Nucleic Acids Research, 2013. **41**(Database issue): p. D991-5.

72. Trapnell, C., B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. van Baren, S.L. Salzberg, B.J. Wold, and L. Pachter, *Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching During Cell Differentiation.* Nature Biotechnology, 2010. **28**(5): p. 511-5.

73. Langmead, B., C. Trapnell, M. Pop, and S.L. Salzberg, *Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome.* Genome Biology, 2009. **10**(3): p. R25.

74. Trapnell, C., L. Pachter, and S.L. Salzberg, *Tophat: Discovering Splice Junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.

75. Dobin, A. and T.R. Gingeras, *Mapping RNA-Seq Reads with Star.* Current Protocols in Bioinformatics, 2015. **51**.

76. Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D.R. Kelley, H. Pimentel, S.L. Salzberg, et al., *Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with Tophat and Cufflinks.* Nature Protocols, 2012. **7**(3): p. 562-78.

77. Malone, J.H. and B. Oliver, *Microarrays, Deep Sequencing and the True Measure of the Transcriptome.* BMC Biology, 2011. **9**: p. 34.

78. Roy, N.C., E. Altermann, Z.A. Park, and W.C. McNabb, *A Comparison of Analog and Next-Generation Transcriptomic Tools for Mammalian Studies.* Briefings in Functional Genomics, 2011. **10**(3): p. 135-50.

79. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: A Revolutionary Tool for Transcriptomics.* Nature Reviews Genetics, 2009. **10**(1): p. 57-63.

80. Ritchie, M.E., B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, and G.K. Smyth, *Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies.* Nucleic Acids Research, 2015. **43**(7): p. e47.

81. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.* Bioinformatics, 2010. **26**(1): p. 139-40.

82. Law, C.W., Y. Chen, W. Shi, and G.K. Smyth, *Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts.* Genome Biology, 2014. **15**(2): p. R29.

83. Anders, S., D.J. McCarthy, Y. Chen, M. Okoniewski, G.K. Smyth, W. Huber, and M.D. Robinson, *Count-Based Differential Expression Analysis of RNA Sequencing Data Using R and Bioconductor.* Nature Protocols, 2013. **8**(9): p. 1765-86.

84. Ghosh, S. and C.K. Chan, *Analysis of RNA-Seq Data Using Tophat and Cufflinks.* Methods in Molecular Biology, 2016. **1374**: p. 339-61.

85. Mi, H., A. Muruganujan, J.T. Casagrande, and P.D. Thomas, *Large-Scale Gene Function Analysis with the Panther Classification System.* Nature Protocols, 2013. **8**(8): p. 1551-66.

86. Eden, E., R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini, *Gorilla: A Tool for Discovery and Visualization of Enriched Go Terms in Ranked Gene Lists.* BMC Bioinformatics, 2009. **10**: p. 48.

87. Huang, D.W., B.T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, Y. Guo, R. Stephens, et al., *DAVID Bioinformatics Resources: Expanded Annotation Database and Novel Algorithms to Better Extract Biology from Large Gene Lists.* Nucleic Acids Research, 2007. **35**(Web Server issue): p. W169-75.

88. Subramanian, A., P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, et al., *Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.* Proceedings of the National Academy of Sciences, 2005. **102**(43): p. 15545-50.

89. Sherman, B.T., D.W. Huang, Q. Tan, Y. Guo, S. Bour, D. Liu, R. Stephens, M.W. Baseler, et al., *DAVID Knowledgebase: A Gene-Centered Database Integrating Heterogeneous Gene Annotation Resources to Facilitate High-Throughput Gene Functional Analysis.* BMC Bioinformatics, 2007. **8**: p. 426.

90. Young, M.D., M.J. Wakefield, G.K. Smyth, and A. Oshlack, *Gene Ontology Analysis for RNA-Seq: Accounting for Selection Bias.* Genome Biology, 2010. **11**(2): p. R14.

91. Gentleman, R.C., V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, et al., *Bioconductor: Open Software Development for Computational Biology and Bioinformatics.* Genome Biology, 2004. **5**(10): p. R80.

92. Love, M.I., W. Huber, and S. Anders, *Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with Deseq2.* Genome Biology, 2014. **15**(12): p. 550.

93. R Core Team, *R: A Language and Environment for Statistical Computing*. 2011.

94. Mi, G., Y. Di, S. Emerson, J.S. Cumbie, and J.H. Chang, *Length Bias Correction in Gene Ontology Enrichment Analysis Using Logistic Regression.* PLoS ONE, 2012. **7**(10): p. e46128.

95. Hirschberg, J. and A. Rosenberg, *V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure.* Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2007. **June**: p. 410-20.

96. Jain, A.K., M.N. Murty, and P.J. Flynn, *Data Clustering: A Review.* ACM Computing Surveys, 1999. **31**(3): p. 264-323.

97. Yeung, K.Y., D.R. Haynor, and W.L. Ruzzo, *Validating Clustering for Gene Expression Data.* Bioinformatics, 2001. **17**(4): p. 309-18.

98. MacQueen, J., *Some Methods for Classification and Analysis of Multivariate Observations.* Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967. **1**(14): p. 281-297.

99. Kim, M.S., A. Horst, S. Blinka, K. Stamm, D. Mahnke, J. Schuman, R. Gundry, A. Tomita-Mitchell, and J. Lough, *Activin-a and Bmp4 Levels Modulate Cell Type Specification During CHIR-Induced Cardiomyogenesis.* PLoS ONE, 2015. **10**(2): p. e0118670.

100. Tomita-Mitchell, A., K. Stamm, D. Mahnke, P.M. Hidestrand, M.S. Kim, H.L. Liang, M.A. Goetsch, M. Hidestrand, et al., *Impact of Myh6 Genotype on Outcomes in Hypoplastic Left Heart Syndrome.* unpublished, 2016.

101. Löhle, M., A. Hermann, H. Glaß, A. Kempe, S.C. Schwarz, J.B. Kim, C. Poulet, U. Ravens, et al., *Differentiation Efficiency of Induced Pluripotent Stem Cells Depends on the Number of Reprogramming Factors.* Stem Cells, 2012. **30**(3): p. 570-579.

102. Andrews, P.W. and N. Elvasorre, *The Origins of Stem Cells as Tools for Regenerative Medicine.* Biochemical and Biophysical Research Communications, 2016.

103. Li, B. and C.N. Dewey, *RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome.* BMC Bioinformatics, 2011. **12**: p. 323.

104. Noonan, J.A. and A.S. Nadas, *The Hypoplastic Left Heart Syndrome; an Analysis of 101 Cases.* Pediatric Clinics of North America, 1958. **5**(4): p. 1029-56.

105. Patel, A.P., I. Tirosh, J.J. Trombetta, A.K. Shalek, S.M. Gillespie, H. Wakimoto, D.P. Cahill, B.V. Nahed, et al., *Single-Cell RNA-Seq Highlights Intratumoral Heterogeneity in Primary Glioblastoma.* Science, 2014. **344**(6190): p. 1396-401.

106. Fraga, M.F. and M. Esteller, *DNA Methylation: A Profile of Methods and Applications.* Biotechniques, 2002. **33**(3): p. 632, 634, 636-49.

107. Fraga, M.F., R. Rodriguez, and M.J. Canal, *Genomic DNA Methylation-Demethylation During Aging and Reinvigoration of Pinus Radiata.* Tree Physiology, 2002. **22**(11): p. 813-6.

APPENDIX A

Appendix A is a listing of tables and figures generated during a run of **rgsepd**. In each file/figure/table the filenames are constructed to ensure that the user can find and associate the file with the run comparing sample class A with class B.

A.1 Tables Listing

DESEQ.counts.AxN.BxN.csv - post normalization, sample columns. Tables are generated of the normalized counts, produced by DESeq2, but made available in simple CSV format. Similar to the input data table, but library scale normalization has been computed, so these are ready to visualize for ad-hoc analyses. Corresponds to the orange cylinder in **Figure 3.1**.

DESEQ.RES.AxN.BxN.csv – Results file from DESeq2, containing transcript ID numbers from the input data, and computed statistics for fold-change and p-values.

DESEQ.RES.AxN.BxN.Annote.csv – annotated version of the previous table. After some database calls out to Biomart and Ensemble databases, most transcript ID numbers can be associated with a gene name and gene ID number. Here I also annotate the sample class means for expression of each transcript/row. This file makes it possible for the end user to quickly search a gene name and retrieve simplified expression averages.

DESEQ.RES.AxN.BxN.Annote_Filter.csv filtered version of the previous table. Given the constraints of significance presented in advance of a run, a filtered table is produced. An **rgsepd** run culls genes without sufficient change between classes, or

sufficient statistical significance, or sufficient absolute expression. These three values are configurable if the user needs a shorter or longer table of genes for their experimental situation. The filtered list is the basis of the latter stages of functional annotation and projection.

GOSEQ.RES.AxN.BxN.GO.csv – the result of the included GOSeq run. GOSeq computes the functional annotation to the filtered gene set by searching for "overrepresented" gene sets among the space of all GO Terms in a hierarchy. I run it for each of the three major GO trees, then filter the list to nominal significance and annotate the results with the English names for each GO Term identifier. Also included for convenient interpretation are the summary statistics "number of genes in GO Term" vs "number of genes differentially expressed". This list contains many large GO Terms and filtering to the number of genes can quickly refine to precise biological processes found differential between classes "A" and "B".

GOSEQ.RES.AxN.BxN.GO-UP.csv is a table similar to the previous overall GO listing, but here genes are considered only when upregulated in class A versus class B, as the user may be interested in seeing which functional sets are "turned on" in response to the condition.

GOSEQ.RES.AxN.BxN.GO-DOWN.csv is a table similar to the overall GO listing, but here genes are considered only when downregulated in class A versus class B, as the user may be interested in seeing which functional sets are "turned off" in response to the condition.

GOSEQ.RES.AxN.BxN.GO2.csv is a joint table collecting the genewise summary statistics of the DESEQ.RES.AxN.BxN.Annote.csv and copying them onto

each significant row of each GOSEQ table. Broad GO Terms will carry redundant data with more precise terms, so anything greater than 1000 genes is filtered before this table is created. The **GO2** file is the largest single entity, with approximately 347 thousand rows from the run that produced **Figures 3.5-7**. Here a gene set can be searched by name or key word, the genes identified by name and their average and differential expression data readily reviewed or further filtered by the user. This table answers the question "what genes are in this gene set?"

GSEPD.RES.AxN.BxN.MERGE.csv is a filtering and collation of the GOSEQ tables. Cropping the **GOSEQ.RES.AxN.BxN.GO2.csv** to the most significant terms, and those below a specified gene-count yields a manageable list of interesting functional associations. Each is annotated with "Up", "Down", or "Mix" depending on which of the three GOSeq runs found the term.

GSEPD.Alpha.AxN.BxN.csv is a table of Alpha scores, samples on columns by GO Terms on rows. Alpha scores are defined in **Section 3.3.2** and **Figures 3.3** and **3.5**.

GSEPD.Beta.AxN.BxN.csv is a table of Beta scores, samples on columns by GO Terms on rows. Beta scores are defined in **Section 3.3.2** and **Figures 3.3.**

GSEPD.HMG1.AxN.BxN.csv is a table of Gamma1 scores, each sample (on columns) by their distance (N-dimensional Euclidean distance to class A's centroid for a gene set of size N) to each GO Term (on rows). These Gamma1 values are defined in **Section 3.3.4** and used in the coloring scheme of **Figures 3.4** and **3.7**.

GSEPD.HMG2.AxN.BxN.csv is a table of Gamma2 scores, similar to Gamma1 but noting the distance to the class B centroid.

GSEPD.Segregation_P.AxN.BxN.csv lists the clustering Validity scores and associated p-value for each GO Term. Each gene set is evaluated for clustering validity (**Section 2.3.3**) and an empirical p-value is calculated (**Section 2.3.4**). This table lists those results by GO ID.

GSEPD.HMA.AxN.BxN.csv with gene set segregation completed, the most valid GO Terms are represented in this "HMA" file which merges the **GOSEQ.RES.AxN.BxN.GO.csv** with the **GSEPD.Segregation_P.AxN.BxN.csv** such that name-annotated GO Terms can be sorted and selected based on their N-dimensional clustering validity. This table corresponds to the selected rows of **Figure 3.5**.

A.2 Figures

Section A.2 is a listing of figures generated by a successful run of the tool. Like tables, all files are named with the run so they can be kept between runs and explored as needed. All figures are generated without user requesting them individually, and finding detailed results of a given analysis is easy with the operating system's file-search commands. Like the tables, these figures are listed in the order they are generated.

DESEQ.Volcano.AxN.BxN.png A volcano plot is a standard quality assurance check when performing differential expression. The y-axis is negative Log10 p-value, such that insignificant genes are on the bottom, and highly significant on top. The x-axis is the log fold change, indicating whether the gene is up- or down-regulated between conditions. We hope to see a mainly symmetrical response with a uniform distribution of gene differences and significances. This is a scatterplot with black points non-significant, and red with adjusted p below a threshold, giving the appearance of a volcano.

HM.AxN.BxN.XXX.pdf is a heatmap of gene expression for significantly segregating genes: rows correspond to genes and columns correspond to samples. The heatmap is row-normalized, such that each gene's minimum is green and each gene's maximum is red. Here "XXX" in the filename corresponds to the number of rows (genes found significant by the settings of the run). The image size is automatically scaled with a linear function to ensure all row and column labels are legible. This can create a large file if many genes are selected. The file format is kept PDF such that the user can use a search feature to get to a gene of interest. Four versions of this figure are included, as user preferences and publication requirements dictated various renditions and simplifications of the figure. HM is the basic, with row/column annotations, all samples in the input data are kept and annotated absolute expression within cells (as in **Figure 4.1**). A version "HM-.AxN.BxN.XXX.pdf" is stripped of annotations to enable tighter embedding in a publication where zoomed down text is unacceptable but the colors and correlations remain intact. A version "HMS.AxN.BxN.XXX.pdf" is stripped of the non-tested samples, showing only the samples in classes A and B (omitting untested samples, which can perturb the column ordering). Finally "HMS-AxN.BxN.XXX.pdf" is the combination, stripped of extraneous samples and annotations to produce the most compact representation of the DESeq2 results.

GOSEQ.PWF.AxN.BxN.pdf is a byproduct of the GOSeq process. The "parameterized weight function" is a diagnostic of the gene set average length correction inherent to GOSeq. It displays the correlation between gene length and their probability of being found differentially expressed (this is a known bias in RNA-Seq, and the

primary reason for GOSeq rather than a more direct hypergeometric enrichment analysis for GO Term detection) [90].

GSEPD.PCA_AG.AxN.BxN.pdf is a principal components analysis of all genes, scatterplot for the first two principal components. These are standard meta-analyses to evaluate sample clustering and batch effects visually. **rgsepd** annotates which samples belong to which classes with point colors, annotates sample IDs under each point, and annotates the top five driver genes along each axis.

GSEPD.PCA_DEG.AxN.BxN.pdf is a principal components analysis of just differentially expressed genes, scatterplot for the first two principal components. This version enforces visual segregation and is more interesting for the non-tested samples' clustering. **rgsepd** annotates which samples belong to which classes with point colors, annotates sample IDs under each point, and annotates the top five driver genes along each axis.

SCA.GSEPD.AxN.BxN.pdf is a series of scatterplots of the samples' Alpha scores with respect to two GO Terms, indicating cross-term correlations and the direction of divergence among outlier samples. The axes are the scale of Alpha scores, generally 0-1 with the class A samples near (0,0) and the class B samples near (1,1) by construction.

GSEPD.HMA.AxN.BxN.pdf is the projection HeatMap of Alpha scores for significant GO Terms (**Figure 3.5**). Outliers with high Beta values are marked with white dots to indicate the cell color may not be an accurate representation of the sample's expression. The white dot marks are useful when a gene set cannot be reduced to a one dimensional axis. By virtue of the **gplots** package heatmap.2, we have dendrogram correlation by complete linkage clustering, so rows and columns are sorted by their value

profiles. Therefore samples with similar expression profiles are displayed near each other (columns), as are correlated gene sets (rows). The HMA file is the primary result shown in **Chapter 4**.

GSEPD.HMG.AxN.BxN.pdf is the HeatMap of Gamma scores (**Figure 3.7**). The same gene sets as the previous "HMA" figure, but colored with the Euclidean distances as in **Figure 3.4**.

Folder **SCGO** holds figures exploring the significant gene sets' expression profiles. One to three figures per significant GO Term are generated, displaying each sample's behavior with respect to those genes. As a large number of figures is generated, these are expected to be of use only when the user is interested in exploring a select few GO Terms highlighted earlier in the process.

SCGO/GSEPD.AxN.BxN.GO#.pdf is generated for each significant set. The GO Term name is on the title of the figure, and each pair of genes within the set is displayed on subsequent pages (**Figure 3.5** is an example of one page of **GSEPD.D3x2.D5x2.GO0003209.pdf**.) Each sample is annotated by the expression values as z-scored log-scaled normalized counts, derived from the counts table. The Alpha axis is drawn in black between each class centroid, for diagnostic visualization. Non-tested samples are annotated with their scores as colored lines indicating which class they more closely behave like. Importantly, the drawn lines are orthogonal to the axis, indicating the closest possible point, but they will not appear orthogonal due to their high dimensional nature being projected into only two genes. These data make up the content of the GSEPD.HMA heatmap.

SCGO/Pairs.AxN.BxN.GO#.pdf files are generated for significant GO Terms with less than ten genes. Gene sets with few dimensions can be efficiently displayed on a single 'pairs' plot, for immediate visualization of a partially differentially expressed gene set. Each pairing of genes is visualized as a two dimensional scatterplot. All samples are displayed. Axes are labeled with the log-normalized read counts.

SCGO/Scatter.AxN.BxN.GO#.pdf files are generated for significant GO Terms displaying a PCA plot. As in the **GSEPD.PCA_DEG.AxN.BxN.pdf**, major genes are annotated on each axis, and samples are labeled to facilitate visualization of outlier samples. This plot is the PCA across only those genes within the given GO Term, as extracted from the full cross product **GOSEQ.RES.AxN.BxN.GO2.csv.**