Integrating biclustering techniques with *de novo* gene regulatory network discovery using RNA-seq from skeletal tissues

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Katie Ovens

Permission to Use

Abstract

In order to improve upon stem cell therapy for osteoarthritis, it is necessary to understand the molecular and cellular processes behind bone development and the differences from cartilage formation. To further elucidate these processes would provide a means to analyze the relatedness of bone and cartilage tissue by determining genes that are expressed and regulated for stem cells to differentiate into skeletal tissues. It would also contribute to the classification of differences in normal skeletogenesis and degenerative conditions involving these tissues. The three predominant skeletal tissues of interest are bone, immature cartilage and mature cartilage. Analysis of the transcriptome of these skeletal tissues using RNA-seq technology was performed using differential expression, clustering and biclustering algorithms, to detect similarly expressed genes, which provides evidence for genes potentially interacting together to produce a particular phenotype. Identifying key regulators in the gene regulatory networks (GRNs) driving cartilage and bone development and the differences in the GRNs they drive will facilitate a means to make comparisons between the tissues at the transcriptomic level.

Due to a small number of available samples for gene expression data in bone, immature and mature cartilage, it is necessary to determine how the number of samples influences the ability to make accurate GRN predictions. Machine learning techniques for GRN prediction that can incorporate multiple data types have not been well evaluated for complex organisms, nor has RNA-seq data been used often for evaluating these methods. Therefore, techniques identified to work well with microarray data were applied to RNA-seq data from mouse embryonic stem cells, where more samples are available for evaluation compared to the skeletal tissue RNA-seq samples. The RNA-seq data was combined with ChIP-seq data to determine if the machine learning methods outperform simple, correlation-based methods that have been evaluated using RNA-seq data alone. Two of the best performing GRN prediction algorithms from previous large-scale evaluations, which are incapable of incorporating data beyond expression data, were used as a baseline to determine if the addition of multiple data types could help reduce the number of gene expression samples. It was also necessary to identify a biclustering algorithm that could identify potentially biologically relevant modules. Publicly available ChIP-seq and RNA-seq samples from embryonic stem cells were used to measure the performance and consistency of each method, as there was a well-established network in mouse embryonic stem cells to compare results. The methods were then compared to cMonkey2, a biclustering method used in conjunction with ChIP-seq for two important transcription factors in the embryonic stem cell network. This was done to determine if any of these GRN prediction methods could potentially use the small number of skeletal tissue samples available to determine transcription factors orchestrating the expression of other genes driving cartilage and bone formation.

Using the embryonic stem cell RNA-seq samples, it was found that sample size, if above 10, does not have a significant impact on the number of true positives in the top predicted interactions. Random forest methods outperform correlation-based methods when using RNA-seq, with area under ROC (AUROC) for evaluation, but the number of true positive interactions predicted when compared to a literature network were similar when using a strict cut-off. Using a limited set of ChIP-seq data was found to not improve the confidence in the transcription factor interactions and had no obvious affect on biclustering results. Correlation-based methods are likely the safest option when based on consistency of the results over multiple runs, but there is still the challenge of determining an appropriate cut-off to the predictions. To predict the skeletal tissue GRNs, cMonkey was used as an initial feature selection method to identify important genes in skeletal tissues and compared with other biclustering methods that do not use ChIP-seq. The predicted skeletal tissue GRNs will be utilized in future analyses of skeletal tissues, focussing on the evolutionary relationship between the GRNs driving skeletal tissue development.

Acknowledgements

I would like to thank my supervisors Ian McQuillan and Brian Eames for their constant support and encouragement. Their expectations have pushed me to new heights in my ability to effectively communicate my research to others. I feel the quality of my work has skyrocketed by applying their advice. They have also helped me to identify areas to continue improving, from their pain-staking effort proofreading my written work, watching and commenting on my presentations, and meeting with me weekly to discuss my progress.

Patsy Gómez is also a continuing source of information regarding the biological background of this thesis, and was responsible for collecting the skeletal tissue data. Her and Brian's work have made the motivations behind my part in the research increasingly clear and I am constantly learning new things from them about the biological context of the project.

Furthermore, I would like to thank the other members of my committee, Tony Kusalik, and Kevin Stanley for their time and valuable feedback on my thesis, from the research proposal to the final document. Also a special thanks to Chris Eskiw, who took the time out of his schedule to participate as an external examiner for my defence and provide me with feedback.

Finally, I would like to thank my parents, who have always shown interest (or at least made it appear as such) in anything I have decided to pursue.

Funding for my Masters thesis was provided by the Department of Computer Science at the University of Saskatchewan, as well as Ian and Brian through NSERC.

Contents

List of Tables

List of Figures

List of Abbreviations

ANOVA	Analysis of Variance
ARACNE	Algorithm for the Reconstruction of Accurate Cellular Networks
AUPR	Area Under Precision Recall Curve
AUROC	Area Under ROC
BicAT	Biclustering Analysis Toolbox
ChIP	Chromatin Immuno-precipitation
CTWC	Coupled Two-way Clustering
DREAM	Dialogue for Reverse Engineering Assessments and Methods
EM	Expectation Maximization
ESC	Embryonic Stem Cell
FABIA	Factor Analysis for Bicluster Acquisition
GENIE3	Gene Network Inference with Ensemble of trees
GO	Gene Ontology
GRN	Gene Regulatory Network
HTSeq	High-throughput Sequencing
KEGG	Kyoto Encyclopedia of Genes and Genomes
LASSO	Least Absolute Selection and Shrinkage Operator
PCA	Priciple Component Analysis
ROC	Receiver Operating Characteristic
SAMBA	Statistical-Algorithmic Method for Bicluster Analysis
TMM	Trimmed Means of M Values
TRN	Transcription Regulatory Network
TSS	Translation Start Site
WE	Weighted Enrichment
WGCNA	Weighted Gene Co-expression Network Analysis

Chapter 1

Introduction

Osteoarthritis is caused by the degeneration of articular cartilage and subchondral bone. It is the most prevalent form of arthritis, affecting over 10% of the Canadian population and roughly 50% of people over the age of 60 [2]. This figure is on the rise as the population ages and weight related influences become increasingly common. As the population becomes older as a whole, this issue will place increased financial burden upon healthcare systems as well as having indirect costs from lost wages, and a lower quality of life due to pain and reduced physical functioning [3]. The burden of osteoarthritis is exacerbated by the inadequacies of current therapies. However, recently adult mesenchymal stem cells, which have the ability to differentiate into cartilage or bone, have emerged as a candidate cell type with great potential for cell-based articular cartilage repair technologies [4]. To shed light on the mechanisms behind degeneration of bone and cartilage, it is first necessary to describe normal skeletal tissue development by examining what and how various cellular and molecular components are involved.

The challenge is to determine the genes involved and how they specify differentiation of mesenchymal cells into three main types of skeletal tissue: bone, immature and mature cartilage [5, 6, 7]. Comparing these skeletal tissues may provide insight as to how the process of bone formation differs from the formation of cartilage. One way to approach analysis of these tissues is to look at the transcriptome, which contains the total RNA present inside a cell. The dynamic properties of the transcriptome allows information to be obtained about the gene activity in a particular cell or a number of cells under particular conditions. Important gene activity includes the expression of transcription factors, which regulate the expression of other genes, and ultimately influences the development of each tissue. A gene regulatory network (GRN) consists of genes identified as potential regulators, and the target genes of these regulators. Expression of the regulators influence the expression of the target genes of that particular regulator. The number of genes in a GRN may vary from only two genes to full genomic networks.

A goal of this thesis is to uncover the GRNs underlying skeletal tissue formation. Two main transcription factors are required for skeletal tissue formation. Sox9 is required for immature cartilage development, while Runx2 is required for bone development [5]. Sox9 is hypothesized to be the main transcription factor controlling the GRN active in immature cartilage, and Runx2 is hypothesized to be controlling the GRN active in bone [8]. Since both these transcription factors need to be expressed in order for mature cartilage to form, this thesis hypothesizes that these two GRNs interact in order for mature cartilage to develop. For two

GRNs to interact, genes in one network are also influenced by genes or the transcription factors active in the other network. The alternative to this is that the GRNs are not interacting, but both networks are present in mature cartilage. The GRNs active in mature cartilage could include equal activity of the Sox9 and Runx2 GRNs, more activity in one GRN compared to the other due to the expression level of both transcription factors. The Sox9 and Runx2 GRNs and the level of interaction occurring between them in mature cartilage remains unknown.

A wide variety of technologies are available for constructing GRNs [9]. Genes of importance have been discovered in these skeletal tissue networks using microarrays, RNA-seq and ChIP-seq, which is used to analyze protein interactions with DNA that contribute to regulating gene expression [10, 11]. However, it is of interest to determine whether current knowledge about the genes regulated by Sox9 and Runx2 gives an accurate representation of the GRNs that are active when these three tissues differentiate. It is also of interest to uncover genes whose expression has not been measured and associated with skeletogenesis, the process of skeleton formation. One method of detecting patterns of gene expression in high-dimensional data is to use a clustering technique where genes are grouped together based on similar expression patterns, implying they are more likely to be functionally related [12]. In this thesis, bioinformatic analyses including differential expression and clustering are performed using RNA-seq, which quantitates transcript abundance as a means to measure gene expression, in skeletal tissue. These analyses may contribute to determining the extent that these GRNs may be interacting, if they interact at all, during mature cartilage development. The results of the analyses are compared to what is currently known in the literature about these networks. The comparison is done to determine if what is found agrees with what is currently known in literature about the genes regulated by Sox9 and Runx2, or if there is disagreement with what is in the literature about genes in the Sox9 and Runx2 networks. These results are necessary to determine if it is best to predict new Sox9 and Runx2 GRNs.

A typical GRN construction algorithm predicts GRNs with hundreds of expression samples [13]. A major problem is that the small sample size of typical transcriptome data is a significant limiting factor in gene regulatory network prediction. Expression data tends to have high dimensionality (thousands of genes) versus a limited number (from one to hundreds) of samples implying that there could be many equally good solutions when predicting a GRN. Researchers may not have access to large amounts of data, for *in vivo* studies in particular, due to cost and time constraints or limited resources in publicly accessible gene expression repositories depending on their research area. The number of samples necessary at minimum to form an acceptably accurate network in vertebrates has not been reported in the literature for RNA-seq. For a network to be considered acceptably accurate, it must be useful for further biological predictions and hypothesis testing with minimal false positive interactions. Furthermore, the number of false negative interactions should be low when using a method to discover a new network. It is often necessary to reduce the number of genes used to predict a GRN as well as to combine expression data from multiple experiments [12]. It is not known if the number of samples may be reduced if supplemented with other types of data, including protein-protein interaction, knock-out gene expression or ChIP-seq data. Indeed, more information used for the construction of GRNs is considered best if it is available [14]. Since there are many genes present in GRNs functioning in vertebrates, and the genes in these networks are usually not well-defined, there are a lot more genes that need to be considered for prediction compared to simpler organisms, which will likely also increase the number of samples required. Limiting the genes expressed to a smaller sets of genes of interest is necessary if there is no established group of genes to predict interactions between. This is relevant to this project as the sample size of bone, immature and mature cartilage from mouse is small. In order to successfully construct a GRN from this data, an unsupervised method of categorizing gene expression is necessary to discover underlying GRNs. Furthermore, data from another source is required to test if combining data types increases confidence in the networks.

The algorithms currently available to predict GRNs are increasingly accurate if they are also able to incorporate information from many of these sources including knock-out gene expression, ChIP-seq, data already available for the transcription factors in the pathway or biological annotation [14]. Previous studies have been done with microarray data to show different estimations of the necessary number of samples to generate a GRN that has an accuracy above random [15]. These algorithms are typically evaluated using synthetic data or gold standard networks, usually from *Escherichia coli.* Although there is currently no network considered a gold standard available for mammals, such as mouse, there are small well-established model networks [16]. Therefore, before predicting the GRNs present in cartilage and bone, this thesis will determine how sample size changes the ability for GRN prediction algorithms to accurately construct a GRN using a model network in mouse, where more samples are available for testing. This information will allow us to determine if it could be useful to apply these GRN prediction methods to the skeletal tissue RNA-seq data available for this thesis, which only contains 9 samples in total. This thesis will also attempt to determine if the number of RNA-seq samples required can be small when combining ChIP-seq data with RNA-seq to predict a model GRN in mouse. A random forest method, that has been found to outperform other GRN prediction methods using microarray data, will be compared with methods capable of incorporating ChIP-seq data as well as simple correlation-based methods with no integration capabilities. A clustering technique called biclustering is also evaluated, which can be used to detect gene expression patterns in groups of genes that are unique to a single tissue as well as pattern across all tissues. Biclustering is also used as a means of feature selection to minimize the number of genes potentially in the Sox9 and Runx2 GRNs. Using biclustering to minimize the number of genes will allow the consideration of all genes expressed in RNA-seq gene expression data, which could potentially identify genes that have not been associated with Sox9 or Runx2 before. It could also be used in the future to identify other important transcription factors possible regulating Sox9 and Runx2. How another data type, in particular ChIP-seq, improves GRN prediction accuracy when combined with RNA-seq is not known, nor is whether using a small number of samples is possible if data from other sources are combined.

The interactions present or absent in each predicted GRN for a model mouse network were compared when using different methods to make the predictions. Also, the predicted interactions using only one method were compared to determine the consistency of the results for each method. Each GRN was compared to a well characterized GRN in mouse to determine how many "known" interactions each method identified in their top predicted interactions. It is assumed that the more of these interactions a method is able to identify earlier in their lists of predicted interactions, the better the method is able to perform. However, it is difficult to determine a cut-off for predictions most likely to be true positives without also including almost all possible interactions for particular transcription factors of interest. Furthermore, using biclustering does not allow for the same evaluations that can be done with other machine learning methods since all possible interactions are unlikely to be predicted. Therefore, another means of comparison was to determine the consistency of the top predicted interactions from the different techniques. These evaluations may provide other means of evaluating biclustering methods with other GRN prediction methods. Furthermore, it will provide insight into how integrating data types changes the resulting network and how different approaches to data integration changes results. If a more complete gene network driving skeletal tissue development can be uncovered in this thesis, this may be compared in the future in various organisms at the genomic level using homology-based studies to determine conserved portions of the networks across species in the future. By obtaining an initial estimate of what the Sox9 and Runx2 GRNs in these tissues look like, it will be possible to further evaluate the predicted GRNs from an evolutionary perspective at the transcriptomic and genomic level.

Chapter 2

Research Objectives and Thesis Outline

The first objective of this thesis is to test the hypothesis that two specific GRNs are the main drivers of cartilage and bone development with evidence that the GRNs interact. Furthermore, the GRNs are hypothesized to both be necessary for the development of mature cartilage. How much influence each GRN has in the development of each skeletal tissue is also unknown. For example, since there are genes, such as *Sox9*, required for any type of cartilage development, these genes are likely expressed in both immature and mature cartilage. As such, the GRNs driving development of both tissues may also interact. This can be observed by applying basic bioinformatics techniques including differential expression and global clustering to determine how similar or different these tissues are from each other in terms of gene expression. This analysis tests the hypothesis that there are two transcription factors, Sox9 and Runx2, which are the main drivers of the GRNs controlling differentiation of cartilage and bone. What is currently known about the Sox9 and Runx2 networks from the literature was compared to analyses generated from RNA-seq data from bone, immature and mature cartilage. Based on these comparisons it was determined that it would be necessary to construct a new prediction of these GRNs. Therefore, a second objective was to identify competent methods to predict Sox9 and Runx2 GRNs and whether certain techniques would be more appropriate with few data samples in a complex vertebrate like mouse.

Chapter 3 introduces gene regulatory networks as well as a literature review of current methods used to infer them including their limitations when used with small sample sizes of gene expression data. Then, it provides a background explaining what is currently known about the Sox9 and Runx2 networks, which is important to compare to the initial analyses performed with RNA-seq data from bone, immature and mature cartilage in Chapter 5. Chapter 4 describes the RNA-seq data used in this thesis to predict the GRNs in skeletal tissues and the methods used to test the accuracy of the currently described networks in the literature. Results of differential expression, clustering and comparisons to the current literature networks for Sox9 and Runx2 are presented in Chapter 5. Chapter 6 describes the methods used to evaluate machine learning methods that are able to incorporate multiple data types to infer GRNs and discusses how these results will be used to defend choices made to build preliminary skeletal tissue Sox9 and Runx2 networks. In Chapter 7, preliminary biclustering evaluations are conducted to select a method for feature selection to minimize the genes used for GRN prediction. Chapter 8 presents results of the GRN evaluations for correlation-based methods and machine learning methods with different sample sizes and types by comparing consistency of predicted interactions and accuracy when compared to a model network. From this, a Sox9 and Runx2 network is predicted using ChIP-seq and RNA-seq data currently available from mouse. Chapter 9 discusses results of the evaluation of integrative GRN prediction methods focused on in this thesis. It also includes a discussion of initial network predictions for the skeletal networks, with caveats. Finally, future directions are proposed to further improve the current predictions of the main gene regulatory networks in skeletal tissues.

Chapter 3

Background

3.1 Gene Regulatory Networks

Computationally, gene regulatory networks (GRNs) are generally represented as a (usually undirected) graph, where the nodes of the graph represent genes. The edges connecting nodes of the graph of a GRN indicate interactions or regulatory relationships between the genes, as shown in Figure 3.1. Nodes that have a high number of edges connected to other nodes are referred to as hubs. Hub genes tend to have many edges leading to various nodes of the network and are often transcription factors that directly or indirectly coordinate the expression of a large number of other genes [12, 17]. What genes qualify as hub genes varies, although recent hub gene identification has defined hub genes as the top 5% of the highest-degree nodes in a network []. A transcription factor is responsible for controlling expression of genes by binding to promoters or enhancers to promote or block gene expression. The bound transcription factors are able to collect the genetic machinery necessary for gene transcription, and can increase or decrease the production of mRNA for particular genes depending on where the transcription factor is able to bind [18]. A network with directed edges can also be referred to as a Transcription Regulatory Network (TRN) as opposed to a GRN [12]. When directed edges go in both directions between two vertices (sometimes represented by undirected edges), this may indicate that genes are co-expressed or co-regulated. These types of relationships between genes are predicted using correlation or mutual information, which are discussed later. These edges may also be weighted, depending on the confidence of the interaction [19]. Possible reasons for co-regulation include that they are active in the same pathway, share a common biological function, location or process. It is also possible that their protein products directly bind to one another, or assemble into the same complex, while a directed edge between genes may also be used to represent a step in a metabolic pathway, signal transduction cascade, or stage of development [20]. Therefore, GRNs are important in development, differentiation and for responding to environmental cues, and can provide good evidence for differences between tissues. However, identifying — for each gene — a small number of regulators among thousands of genes using a very limited number of samples in each experiment remains a challenge due to inherent and observational noise in expression data.

Transcription is regarded as a major control mechanism of gene expression [18]. GRN discovery is improving in part by advances made in high-throughput technologies, which enables the measurement of global

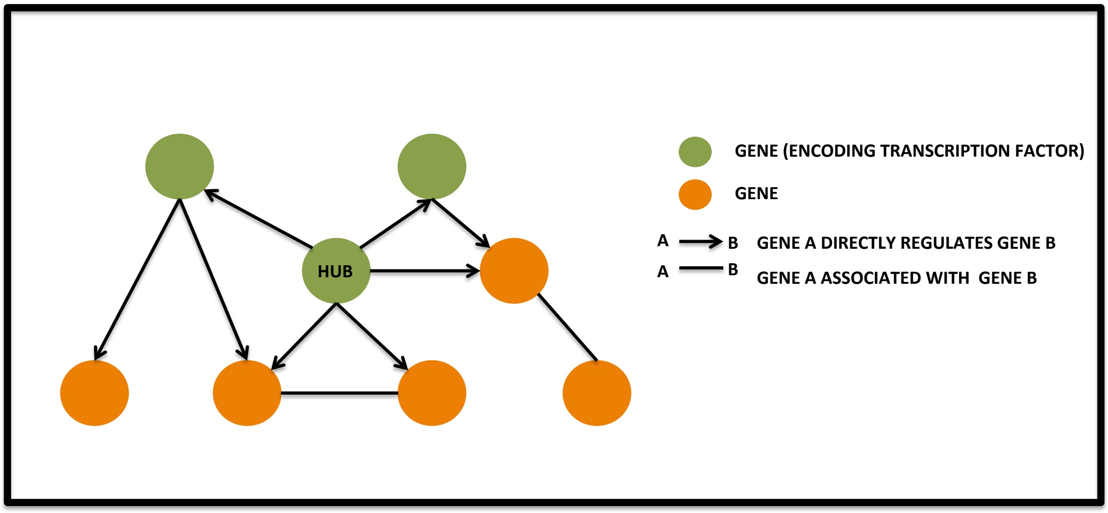

HUB
GENE (ENCODING TRANSCRIPTION FACTOR)
GENE
A → B GENE A DIRECTLY REGULATES GENE B
A — B GENE A ASSOCIATED WITH GENE B

Figure 3.1: Graphical representation of a GRN/TRN

gene expression in biological systems [21]. Using these data alone does not produce a complete or accurate GRN for each skeletal tissue of interest, but integrating different types of "omics" data including genomic, transcriptomic and proteomic data may improve the quality of GRNs reconstructed [21]. Methods currently used to predict GRNs use data including microarrays, RNA-seq, ChIP-chip/ChIP-seq, proteome, metabolome and biological annotations. These data types are discussed in the following subsections.

3.1.1 Microarrays

A DNA microarray is a collection of spots affixed to a solid support, where each spot contains DNA, referred to as probes, representing some feature of interest such as a gene. DNA or cDNA (DNA complementary to RNA) generated from a sample that is able to bind to a particular position on the microarray can be detected [22]. Further, the quantity of bound DNA at each spot can be partially measured to obtain gene expression information. However, microarrays do not include the entire transcriptome (unknown/uncharacterized transcripts etc.) and tend to have higher noise at lower expression levels (limited dynamic range) and so do not provide a complete picture of the transcriptome [23]. This is because the probes present on a microarray have to be designed and therefore all the probes on a microarray must be identified and characterized before being added to a microarray. Microarray technology is also limited largely to well-studied organisms as these are the only species microarrays are available for, which limits evolutionary studies that need to compare many species. Furthermore, splice variants are not taken into consideration with this technology as genes can be transcribed to produce variants of a single gene from combinations of coding regions for a particular gene. Microarray gene expression data still remains the most frequently used type of data for GRN construction even with RNA-seq as a feasible option [24].

3.1.2 RNA-seq

Both RNA-seq and microarray technology follow similar practices for analysis and interpretation of the data they produce, but the technologies have some differences. Next generation sequencing techniques, used for RNA-seq, can be utilized to obtain a more accurate gene expression profile when compared to traditional microarrays, providing increased coverage of DNA sequences and the ability to measure high and low gene expression accurately. RNA-seq is able to provide quantitative approximations of the abundance of target genes in the form of counts for all of the RNA present in a sample, including genes that are novel and would otherwise be excluded from microarrays [23]. Using RNA-seq, a sample of RNA is converted to a library of complementary DNA (cDNA) fragments with identifying adapters attached and sequenced from one (single) or both (paired) ends of each sequence. The resulting sequence reads are aligned with a reference genome or transcriptome (if available) instead of characterized probes on a chip. Since RNA-seq is able to utilize all the RNA in a sample for sequencing, it can detect new transcripts. Furthermore, since RNA-seq does not require probes, it does not have issues with noise due to cross-hybridization where the DNA from a sample pairs with the DNA of a probe that does not match [23]. Microarrays also do not have the dynamic range as high as RNA-seq, since RNA-seq counts correlate with the number of sequences obtained and are not relative amounts as with microarrays [25]. To take advantage of the dynamic range of RNA-seq, read depth is important to consider. If an experiment is performed to discover new transcripts or quatify transcripts that are relatively lowly expressed, than having higher read depth will provide an advantage [26]. It is usually recommended to have about 10M reads, but this may be reduced depending on how well annotated the reference genome is as well as the number of replicates and variation in the data [27]. The number of replicates required depends on the amount of technical or biological variability in the samples [26]. Generally for both microarray and RNA-seq data, there are GRN prediction methods that work best with gene expressions from perturbation and time-series experiments, which often provide more insights on the directionality or the causality of regulatory relationships [24]. There have been recent studies indicating that RNA-seq and microarray *de novo* network discovery tend to complement each other. However, there are genes with “extreme” expression levels, which RNA-seq tends to identify more than microarray, that change the topology of the resulting GRNs [28, 29].

3.1.3 Sequence Data ie. ChIP-chip/seq

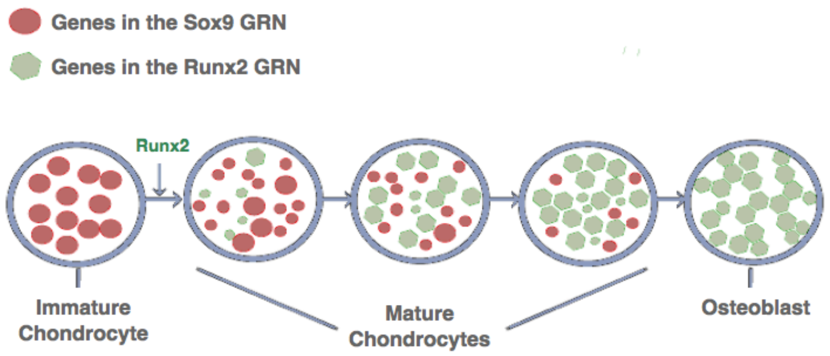
The analysis of sequence data includes the investigation of transcription factor binding motifs, the aim being to detect potential links between sequence motifs and tissue specific gene expression. In ChIP-on-chip (chromatin immunoprecipitation) experiments, DNA fragments that are isolated using a particular protein like a transcription factor are applied to a microarray chip for analysis [30]. This generates a global picture of where the protein binds. However, there are limitations once again by the microarrays available for a genome of interest. ChIP-seq combines ChIP with Next Generation Sequencing such as RNA-seq. With ChIP-seq, analysis assays direct physical interactions between a transcription factor and the DNA to which the transcription factor binds. A sample of DNA is fractionated and an antibody for a particular transcription factor is used to bind to the transcription factors in the sample, which are cross-linked to binding sites on the fractionated DNA. Once these bound fragments are precipitated, the sections of DNA the transcription factor was able to bind to are sequenced using next generation sequencing, which is then analyzed for possible binding sites. Experimentally, transcription factor interactions with DNA are determined by ChIP-seq resulting in p-values of interactions, which are inversely correlated to the probability of an edge being present in a GRN [31]. This data is also used as an evaluation method as they tend to be used on their own to generate many GRNs considered gold standard networks [14]. However, there are limitations depending on the availability of ChIP-seq data for each transcription factor. This type of data has been recently integrated with methods for GRN prediction by enriching results for gene sets, which are expected to include additional evidence for co-regulation [19]. When GRN discovery is transformed into a sparse optimization problem, small transcription factor sets that control the network can be found by solving a least absolute shrinkage and selection operator (LASSO) type problem using transcription factor perturbation sequencing as well as ChIP-chip/seq [32]. It can also be used as the first step to determine potential target genes in the network and calculating correlations between the transcription factor binding data and other gene expression data [33].

3.1.4 Proteome, Metabolome Data and Biological Annotation

Protein interaction and the metabolites produced by protein catalyzing reactions were some of the first commonly used data used to construct networks, but they quickly lose effectiveness when larger, global networks need to be predicted [9, 21]. Protein-protein interaction data can be used to refine gene networks estimated from expression data using Bayesian networks and are particularly useful for predicting the topological structure of a network and the functions of neighbouring genes [34]. It is also possible to integrate functional gene information such as from Gene Ontology, Proteome and KEGG. Gene Ontology (GO), for example, is a controlled vocabulary that describes the attributes of genes and their products including functional characteristics and where they are located in a cell [21]. This type of information alleviates the functional interpretation of genes participating in a GRN.

3.2 Key Transcription Factors in Skeletal Cells

The most abundant tissues in vertebrate skeletal tissues are bone, immature cartilage and mature cartilage. Immature cartilage and mature cartilage differ where immature cartilage will not mineralize, but instead persist over an organism's lifetime and mature cartilage will mineralize and is typically degraded when replaced by bone [6]. Bone is a unique tissue to vertebrates and may develop through two different processes. One of these processes is endochondral ossification, which begins with differentiation of loosely associated cells called mesenchymal cells into chondrocytes. This can persist as cartilage or become gradually replaced by bone. These mesenchymal cell fates are dictated by skeletal cell GRNs. Due to the similarities observed in the functional, embryonic and histological properties of these tissues, it has been hypothesized that the GRNs driving their development share a similar GRN across the tissues [8]. However, bone and cartilage also have properties distinct to each tissue in these categories as wel,l suggesting that there are distinct parts to the GRNs driving cartilage and bone development.



Gómez-Picos and Eames, 2015

Figure 3.2: How the Runx2 network may be related to the Sox9 network present in immature cartilage. Genes in the Sox9 network are indicated by the red objects. Runx2 is hypothesized to be the main regulator of the networks driving mature cartilage and bone formation. Genes in the Runx2 network are represented by the green objects. The introduction of Runx2 and genes regulated by Runx2 to the Sox9 network could allow for the development of mature cartilage. Therefore gene expression in mature cartilage is represented as a mixture of gene expression observed in immature cartilage (driven by Sox9) and bone (driven by Runx2).

Sox9 and Runx2 are candidate transcription factors driving the GRNs responsible for cartilage and bone development respectively. Sox9 is the earliest indicator of mesenchyme differentiating into chondrocytes producing cartilage [5, 6] while Runx2 is considered a master regulator of bone development [35]. Consistently high levels of Sox9 will commit cells to chondrogenesis to produce cartilage, whereas higher levels of Runx2 will push them toward osteogenesis or bone development [5]. The type of tissue that results after immature cartilage development depends upon additional transcriptional control by Sox9 or Runx2. Expression of Runx2 and other transcription factors, such as Sp7, will lead to development of mature cartilage that can be invaded by vasculature, resulting in bone development. Continued action of Sox9 may produce persistent cartilage. In mature cartilage, Sox9 ultimately must become down regulated in order to trigger the maturation of the cartilage. This is required since Runx2 activity is repressed with Sox9 interaction and is hypothesized to be regulated by a wide range of cofactors [36]. Therefore, if both transcription factors are being expressed together it is possible for cells to preferentially differentiate into cartilage.

It is of interest to determine similarities as well as differences in the GRNs of bone and cartilage tissues. If genes in the Runx2 GRN overlap and interact with the genes in the Sox9 network, it will be interesting to determine the extent of the overlap between the GRNs observed in mature cartilage since both Sox9 and Runx2 are required for mature cartilage development. This observation could indicate if the gene expression observed in mature cartilage behaves more like a mixture between the Sox9 and Runx2 networks, if it is more similar to gene expression in one tissue or the other. The combination of the two GRNs also could produce synergistic gene expression where their cooperation leads to gene expression not observed in immature cartilage or bone. It is important to note that Sox9 is dominant to Runx2, so other transcription factors and/or genes may be required for the down-regulation of Sox9 in order for the other skeletal cells to differentiate. This also means that in immature cartilage, the Runx2 GRN is likely to have very little activity or influence. In order for bone development to occur, Sox9 must be down-regulated, which likely means that genes expressed due to Sox9 activity must also become down-regulated or silenced. The alternative to this scenario is that the Runx2 network and the Sox9 network are not both influencing development of mature cartilage, meaning that one of these GRNs could have very little impact in the development of this tissue. This could also mean that, for example, the GRN driving immature cartilage development has very little activity in bone, or no activity at all. The predicted GRNs in this scenario possibly have less overlapping genes and gene expression seen between the skeletal tissues. Mature cartilage could also have gene expression that is a lot more similar to either immature cartilage or bone, depending on the GRN that is most active in the tissue.

Skeletal tissue GRNs have been explored using techniques such as transcriptional profiling and genome wide binding studies [10, 11]. A list of currently known important genes in the Runx2 network has been obtained using microarray data available in the literature, but may still exclude genes participating in the network [10]. The currently known list of genes in the Sox9 network has also been determined from RNA-seq data from the literature using analysis of fold change in expression after Sox9 silencing discussed in Section

3.3 below [11]. The Sox9 study only compares fold change between single replicates of a control and Sox9 silenced sample, which does not allow for statistical measurements of significance. However, it also makes use of ChIP-seq data to make inferences of important genes controlling cartilage development.

3.3 Differential Expression and GRN Prediction

One of the most common uses of transcriptome data is to discover differentially expressed genes that contribute to different phenotypes. When a gene is differentially expressed, it shows differences in expression level between conditions. Since all of the genes in all cells is identical, differential expression of this DNA is one way different cell types develop [37]. For example, different tissue types may have different levels of gene expression or a tissue may have genes expressed that are not expressed in other tissues. The genes that are not utilized still have the potential to be expressed, but may be suppressed by other gene activity and regulatory machinery, or the tissue may lack what is required for the genes to be expressed. Detecting differential expression involves the pairwise comparison of conditions. One of the more simple comparisons tests the null hypothesis that the conditions with a proportion of counts for some gene among two samples is the same as that of the remaining genes. In order to obtain a list of differentially expressed genes with statistical significance it is typically recommended that each condition has at least three replicates, but at least six if preferable to identify differentially expressed genes [38].

Differential expression is one means to establish a prediction for interactions influenced by a single gene, but may not be indication that an accurate GRN can be created from the samples used. It is reliant on a statistical cut-off of confidence of differential expression, which may result in co-regulated modules being left out that could be contributing to a GRN. This also means gene interactions that may be in common among the samples will not be picked up as differentially expressed since a single gene is likely the focus of the study. However, when studying the a GRN in different tissues it is not only imperative to analyze differences, but also similarities between the different networks. Genes in the Sox9 and Runx2 GRNs may be part of both networks and influenced by both transcription factors, so genes may not be differentially expressed between the networks, yet they are important for both networks.

The literature networks available for Sox9 and Runx2 were reported using differential expression and ChIP-seq analysis to identify genes potentially influenced by the transcription factors [10, 11]. Both experiments effectively silenced expression of either Sox9 or Runx2 in chondrocytes and an osteoblast cell line, respectively, and compared to a control. The predicted Sox9 network was generated using expression of Sox9 that was decreased more than 8-fold and compared to a control sample of mouse chondrocytes, as well as focusing on 55% of the genes identified from ChIP-seq data. One limitation of this dataset is that there was only one replicate for each condition, which does not allow for any statistical confidence with the differential expression the authors report, although results are strengthened slightly with the ChIP-seq data. The Runx2 network was predicted using shRNA to silence Runx2, and this identified 159 genes responsive to Runx2 silencing.

Although successful in identifying novel genes potentially regulated by Runx2, the number of probes present on the microarray limits the dataset. Furthermore, it is difficult to measure the quality of this data as it leaves out genes that are known to be regulated by Runx2 such as *Col10a1*. Combining ChIP-seq with differential expression also does not take advantage of the gene expression data as a whole to include a larger portion of genes with correlated expression.

Furthermore, differential expression can help to establish whether a gene is upstream of other genes in the GRN, but will not help to determine if the relationship between this gene and others is likely direct or indirect. An indirect relationship can occur if the expression of one gene influences other genes which are responsible for direct regulation of others and so forth [39]. The genes downstream of this cascade are indirectly influenced by the expression of the first gene. Differential expression will not allow for prediction of other transcription factor influences in a single experiment although predictions can be made with co-expression networks from the gene expression data. One method of obtaining all the genes a transcription factor could be interacting with is collecting the locations where it is able to bind and the gene translation start site (TSS) closest to these binding sites. ChIP-seq is one method to obtain this information.

It is predicted that RNA-seq and ChIP-seq do not influence GRN prediction results in the same way, as ChIP-seq should include the part of the network influence by a transcription factor, not including the type of interaction. The issue with building networks exclusively from ChIP-seq data is the large number of false positive interactions. This is due to many binding events being non-functional [40]. This is where expression data may aid to reduce some of the spurious interactions. It also limits the type of interactions to transcription factor binding events. However, the number of RNA-seq or microarray samples necessary to begin eliminating these spurious interactions from the predicted network is unknown.

Using ChIP-seq may allow for less RNA-seq samples to be used to predict a GRN. However, it is unknown how many samples of expression data are necessary to predict a network when a researcher also has access to other data that can provide an initial hypothesis or prior of what the GRN could look like and reduce the number of genes possibly in the network. This thesis will determine if it allows for less samples of RNA-seq to retrieve the same predictions consistently with current integrative GRN prediction methods.

3.4 Computational Methods for *de novo* GRN Discovery

3.4.1 Clustering

A traditional method of statistical analysis of expression data is to use clustering methods, which relies on the “guilt by association” principle, where genes with similar functional properties tend to interact and exhibit similar expression patterns in a network. Clustering is an unsupervised learning method that can group either the genes or the conditions of an expression matrix, which has a row for each gene, a column for each sample, and has entries that give discrete counts for each gene in each sample. For example, a higher count for a particular gene is seen as a possible indication of higher expression levels of that gene. Clustering is able to group similar patterns of expression across tissue types, conditions, or time steps (the columns), identifying either expression across all tissues with minimal variance or similar changes in expression at different magnitudes. Key features can be explained by grouping these genes or conditions in terms of similar expression patterns across either the genes or conditions being clustered [5]. Genes grouped together based on expression implies they are more likely to be functionally related. Clustering provides a global analysis of the expression data, reflecting expression levels across all conditions, which is an oversimplified view of genes that display expression over select conditions. An example of this type of clustering is discussed further in Section 4.3.1.

Some of the most simple similarity measures used to cluster gene expression data are Euclidean distance and correlation-based methods. Euclidean distance calculates the distances between the expression values of two genes x and y as

$$\sqrt{\sum_{c \in C} (e_{xc} - e_{yc})^2}$$

where e_{xc} is the expression level of gene x under condition c, and C is the set of all conditions [41]. This measure is sensitive to scaling and differences in average expression level, whereas correlation is not. Correlation is an association measure, which is used to estimate the relationships between two variables. Pearson correlation measures the extent of a linear relationship. It is calculated using

$$1 - \frac{\sum_{c \in C}(e_{xc} - \bar{e}_x)(e_{yc} - \bar{e}_y)}{\sqrt{\sum_{c \in C}(e_{xc} - \bar{e}_x)^2 \sum_{c \in C}(e_{yc} - \bar{e}_y)^2}}$$

where $\bar{e}_x$ is the mean expression of gene x [41]. Another measure, Spearman correlation, is based on ranks measuring the extent of a monotonic relationship between x and y. All correlation coefficients take on values between -1 and 1, where negative values indicate an inverse relationship. A correlation coefficient is an attractive association measure since it can be easily calculated, allows for calculating significance levels (p-values), and the sign (+/-) allows one to distinguish between positive and negative relationships. For GRN prediction, close relationships have been found between mutual information and correlation based co-expression networks. Mutual information is discussed further in Sections 3.5 and 3.6. It has been observed that mutual information is often highly related to the absolute value of the correlation coefficient and when they disagree, the correlation findings appear to be more plausible statistically and biologically [42, 43]. It is an attractive method of GRN prediction as well as clustering, since it is possible to estimate correlation with few observations and it does not depend on other parameter choices.

Analyses of RNA-seq data beyond differential expression, such as clustering, are important topics but lack rigorous methodological development with most methods designed with microarray data in mind, which has a different distribution of expression values. Recently, a model-based clustering approach was used to identify co-expressed genes in RNA-seq, which employs either a Poisson or negative binomial mixture model to postulate the over-dispersed gene count data [44]. This algorithm works by alternating between computing probabilities for assignments of each gene to each cluster and updating the cluster means and covariance based on the set of genes predominantly belonging to that cluster [7]. The effectiveness of this clustering method was measured by its ability to cluster genes into clusters with minimal similarities between separate clusters. The method was evaluated in terms of biological significance, as it is required to contribute to elucidating biological processes.

3.4.2 Biclustering Algorithms

There are limitations to GRN prediction using clustering. First, it cannot be presumed that genes that show similar expression profiles are co-regulated as part of the same regulatory pathway. This is because in clustering, all conditions are given equal weights in the computation of gene similarity; thus some conditions may increase the amount of background noise, where there are higher numbers of non-informative variables (genes). Furthermore, each gene can only be assigned to a single cluster even though biologically the gene could be involved in different regulatory pathways depending on the conditions it is acting under. For example, a set of genes could have similar expression levels between two tissues, but they could vary significantly within a third tissue. With clustering, the similarity in the first two tissues would not be identified. Also, it is not possible for a gene, or set of genes, to be present in more than one cluster. To address these concerns, localized clustering methods, or biclustering, was created. The first biclustering algorithms were proposed in 2000 and were called two-way clustering algorithms. This type of algorithm seeks homogeneous subsets of genes and samples by performing a one-way clustering in an iterative manner [45]. To do this, it searches for biclusters with high correlation between the genes by imposing the condition that the mean square residue is below some cut-off value. The Coupled Two-way Clustering (CTWC) algorithm was produced around the same time, which aims to find a set of genes together with a subset of conditions, such that a single cellular process is the main contributor to the expression of the gene subset over the condition subset [46]. This two-way clustering algorithm repeatedly performs one way hierarchical clustering on the rows and columns of the data expression matrix using stable clusters of rows as attributes for column clustering and vice versa. A second type of biclustering algorithm particularly important to this project is probabilistic generation methods, which implement probabilistic techniques in order to discover genes that are similarly expressed across a subset of samples and vice versa [6]. Although biclustering performs better than traditional methods when picking out local gene expression patterns, most biclustering problems have exponential time complexity in the number of rows and columns of the dataset. Consequently, algorithms have to depend on heuristics, so their performance is never optimal. These algorithms are also more often used for feature selection as they are capable of generating lists of related genes, but are unable to infer the types of relationships present in a list of genes without other techniques. In Chapter 7, a performance review and evaluation is performed with the RNA-seq data from skeletal tissues available for this thesis, and a summary of results is provided.

3.4.3 Review of Performance Evaluation of Biclustering Algorithms

Comparative studies have been previously done for both traditional clustering and biclustering methods [47, 48, 49, 50, 51]. A comparison of both clustering and biclustering algorithms is difficult due to most algorithms performing well for the particular tasks assigned yet when further analysis is done, they fail in other areas. Some of the algorithms have been found to be data dependent and so performance relies heavily on the type of data being analyzed [48]. Studies have been done to judge an algorithm's ability to detect biclusters when they do, and do not, overlap using artificial data while others such as Chia and Karuturi used a differential co-expression framework to compare algorithms on real data [52]. Previous performance evaluation of multiple biclustering algorithms tends to involve the introduction of a new biclustering technique in parallel to the evaluation, as a demonstration of the new methods superiority where datasets used for evaluation are either artificially created datasets, or real biological datasets. The synthetic datasets only have the ability to reflect certain aspects of biological reality, but their complexity can be adjusted manually and the solutions are known beforehand making performance analysis a lot easier. Still, biological data tends to hold more sway when judging the performance of a biclustering algorithm.

A common method used to judge biological relevance is the number of Gene Ontology (GO) enriched terms and p-values based on the significance of the GO annotations identified within the data [47]. GO terms are a controlled vocabulary that describe biological properties of gene products. These terms may be used to annotate gene products with various biological processes, cellular components and molecular functions associated with them. In a study evaluating five biclustering algorithms [48], these two methods were argued to be inappropriate, as the number of GO terms and the significance levels of enriched GO terms are dependent on bicluster size. In addition to GO annotations, they considered protein-protein interaction networks. Biclustering algorithms have also been evaluated by defining a scoring method, called gene match score, which uses a clustering method, Bimax, as a reference to test the effects of bicluster overlap and experimental noise [51]. This research suggested that it might be more useful to use multiple algorithms in conjunction, starting with a method to find all possible biclusters before applying another. Other scoring methods that have been used include weighted enrichment (WE) scoring and protein-protein interaction (PPI) network scoring. The algorithms were evaluated by the number of biclusters, ranking of the biclusters generated based on WE scores and ranking of the biclusters based on PPI scores. The results suggested that combining gene expression data with pathway maps within a biclustering framework could be useful to focus on specific gene groups. Identifying particular pathways within gene expression data will play a key role when evaluating the biclustering performance for this project. These studies demonstrate a movement from performance analysis using ideal datasets and using more real data as a means to judge performance. Performance evaluations are discussed further at the beginning of Chapter 7.

3.5 Beyond Feature Selection for GRN Discovery

The Dialogue on Reverse Engineering Assessment and Methods (DREAM) uses crowdsourcing challenges to address fundamental questions in biology including how well current methods are able to describe interacting molecules. One of the more recent projects, DREAM5, performed blind assessments of 35 GRN discovery methods, 29 of which had predicted networks from microarray data submitted by researchers in the community while the other 6 were common ready-to-use methods [14]. The predicted networks were compared against binary gold standard networks. They were assessed by the precision vs. recall curve (AUPR) and the AUROC, which shows the true positive rate vs. 1 minus the false positive rate of interactions between genes. The methods evaluated included combinations and variations of linear regression, correlation, mutual information, Bayesian networks as well as novel techniques put forth by researchers in the community. Regression methods select transcription factors by target gene-specific sparse linear regression or by data resampling techniques. Each gene is considered individually from the others and the expression value for that can be represented as a linear function of all other gene expression levels and of all polymorphisms [53]. The DREAM5 project found that the strategy used for resampling is important in these cases, as the worst performing methods employed no data resampling or bootstrapping technique. These models can also be combined with Bayesian linear regression models or learned using Markov models [14, 54]. Mutual information methods such as context likelihood relatedness (CLC) [13] and Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) [55] have an advantage over correlation-based methods such as Pearson Correlation since they do not assume monotonic relationships and so are able to detect non-linear and irregular dependencies. These methods were outperformed by many independent contributors in the DREAM5 project, but perform well when recovering feed-forward loops. Feed-forward loops have three genes with three interactions between those genes. Gene A influences gene B, which will then influence gene C expression (A $\rightarrow$ B $\rightarrow$ C) and gene A influences C (A $\rightarrow$ C). However, these methods had many false positives for linear cascades [56]. The authors found methods such as Relevance Networks and Bayesian Networks were better at predicting linear cascades as they are more likely to select regulators that independently contribute to target gene expression. However, it is likely these methods are highly dependent on how carefully the data is discretized in order to avoid any loss of information. Furthermore, if data resampling techniques were applied to these techniques, they would likely only be applicable to smaller networks due to performance constraints involved in heuristic searching. It is important to note that these methods were used to measure global dependencies so any local dependencies within subsets of conditions may be missed. Edge detection methods that are able to do this are comparable to other correlation based methods although they may discover slightly more true positive localized relationships between genes [57].

One of the best performing algorithms reported in the DREAM5 project used a non-parametric, non-linear correlation coefficient that is based on Analysis of Variance (ANOVA) [58]. The method performed best when compared to the gold standard *Escherichia coli* network, but was unable to discover a higher proportion of genes in the gold standard network for the eukaryotic species *Saccharomyces cerevisiae* as the way the gold standard network was developed was strictly using ChIP-seq data. Data on physical binding alone can result in false positive interactions unless complemented with a conservation-based motif discovery algorithm [58]. Therefore, the authors speculate that many false positives are present in the gold standard for *Saccharomyces cerevisiae* that GRN discovery methods would never identify based on the expression data used in the evaluations. GENIE3 was another top performing method, which uses tree-based ensemble methods to calculate how important a predictor gene is with respect to a target gene, where greater importance signifies a likely interaction or regulatory link between both genes [59]. GENIE3 decomposes the network discovery task into separate regression problems for each gene in the network. The expression values of a particular target gene are predicted using all other genes as possible predictors. The combination of multiple methods also performed strongly though the quality of the networks was dependent on the information required by each combination. The more limitations with information other than gene expression, the less accurate the networks were. From these results, the project concludes that it is best to exploit direct transcription-factor perturbations, employ strategies like data resampling to avoid overfitting, and develop better approaches to differentiate between direct and indirect regulation.

Compared to GRN discovery using microarrays, little has been done to evaluate GRN discovery using RNA-seq. There are several studies that have been conducted on RNA-seq data for gene network discovery to compare it with generating GRNs using microarray data, but nothing done beyond observing changes in the topology of the GRNs. Pearson's Correlation of the gene expression data has been used as a similarity measure in order to perform hierarchical clustering [60]. Pearson's Correlation has also been applied using a significant correlation threshold, called the Weighted Gene Co-expression Analysis (WGCNA) method, which ranks the edges of a network based on variants of correlation [61]. The results are clustered based on the topological overlap measure, which combines the adjacency of two genes and the connection strengths these two genes share with other genes. It is recommended that 15 samples at a minimum including controls are used to generate significant results, but the authors state more than 20 is ideal [62]. Both studies compared RNA-seq expression data results to similar samples from studies using microarrays, and evaluated the preservation of the network modules across the datasets. This was done by measuring properties of the networks including pairwise relationships between genes, overlap between the networks discovered using each technology and how similar the connectivity was between genes of both networks. It was concluded in both studies that increased dynamic range of expression values and the accuracy of deep sequencing in RNA-seq allowed for better estimation of these network properties. Higher correlation between some genes were found in RNA-seq, which was concluded to be a consequence of genes with relatively low counts, which are not picked up in microarrays due to high background noise. These interactions may result in a more accurate network, although choosing an appropriate cut-off for low counts in RNA-seq studies is also important to minimize false positive correlations.

Recent experiments with RNA-seq data have shown mutual information methods such as CLC and ARACNE are outperformed by even simple correlation strategies such as Pearson or Spearman Correlation [28]. Simple correlation strategies also outperformed methods like WGCNA in these experiments. WGCNA has also been outperformed by regression-based methods like Sparse PArtial Correlation Estimation (SPACE) in evaluations using microarray data [63]. When comparing RNA-seq network results to microarray, hub genes were dissimilar between the two aggregate networks generated. Furthermore, highly correlated genes using one technology were not always well correlated using the other, though Gene Ontology (GO) term results were similar across both networks. This was also the case for the individual networks. Recent consensus measures have provided a cut-off for transcript expression estimates. If expression counts are under the cut-off, they are not reliable in a RNA-seq pipeline, with the bottom one third of transcripts being a major threshold [64]. Using this threshold, the authors determined the genes in the networks that would fall below this threshold. They found that the genes under the threshold tended to have high node degree in GRNs discovered using microarray experiments while RNA-seq experiments resulted in nodes with less edges. These genes contributed many hub genes to the microarray GRNs, which is likely due to lack of sensitivity when faced with noisy expression. One limitation of this evaluation is that machine learning algorithms were not included in this study as a means to generate GRNs, only as an evaluation of the discovered networks using correlation based and mutual information methods. Machine learning algorithms were only used when comparing GRNs generated from RNA-seq to evaluate how similar biological annotations from KEGG, GO and Reactome were related to the connections between genes.

Another recent study used 72 samples of RNA-seq from *Drosophila* using a method based on Pearson Correlation as well as one of the top performing methods in the DREAM5 project, GENIE3, to compare the discovered networks to the gold standard transcription factor motif for eye development [65]. Although comparisons between the networks discovered by each technique are limited, both the correlation and GENIE3 methods yield gene sets that represent candidate transcription factor targets, being a mixture of direct and indirect targets. They recovered many known regulators and *cis*-regulatory elements, but a large part of the predicted network has not yet been explored. This study along with the evaluation of current GRN discovery methods applied to RNA-seq stresses the importance of large sample sizes. From RNA-seq evaluations, it has been concluded that more than 20 experiments each with more than 10 samples of moderate read depth (10M reads for each sample) are required to produce accurate results although this conflicts with suggestions made by the creators of WGCNA, for example [28, 65].

The general consensus of all of these evaluation studies is to construct consensus networks using multiple GRN discovery methods to produce more accurate networks [14, 63, 28, 12]. Given the biological variation among organisms and the experimental variation among gene-expression data sets, it is difficult to determine which methods will perform optimally for reconstructing an unknown regulatory network without testing many strategies. One method proposed, Network Inference using Multiple Ensemble Feature Importance algorithms (NIMEFI), is to weight the results using all the GRN discovery methods to construct a network based on their influence in constructing the final network. Combinations of importance algorithms as used in GENIE3, for example, were also combined [66]. Another option is to combine results from the same methods using various datasets to have more confidence in a GRN discovered. This has been done when comparing multiple species to infer the evolution of a GRN [67]. Pairs of genes whose expression is significantly correlated are identified in multiple organisms indicating co-expression is conserved across evolution. Pearson correlation is traditionally used with microarray analysis for comparing expression profiles between every pair of genes for each organism. All of the genes are ranked according to the Pearson Correlation values to calculate the probability of observing a particular configuration of ranks across different organisms by chance. There are also options of combining information about interaction types present in the data [17]. One of the most successful integrative approaches has been to overlay networks with molecular profiles to identify modules. Molecular profiles include transcriptomic, genomic, proteomic, epigenomic and other cellular information, which are becoming increasingly accessible. However, predicting molecular networks remains under-explored at the systems level, as interaction data are typically measured under single conditions.

Module based inference methods such as clustering and biclustering are an appropriate starting point if the set of expression data is large or heterogeneous compared to more direct query driven methods [12]. These methods are useful when there is no gold standard network or there is little annotation and sequence information available. If a particular section of an already established network needs to be revisited, the already reconstructed network can be used as a starting point to generate a GRN or expand upon a particular piece of it. Biclustering was not involved in any of the method comparisons described previously, but it has been used as a means to infer GRNs using it in combination with other information such as transcription factor binding motif sequences [68]. Although using ChIP-seq data to complement the expression data can allow for a more accurate reconstruction of a GRN, accuracy depends on how much information is available about the transcription factors in the network. If only select transcription factors have this information available, it can bias results to include more interactions involving these transcription factors when trying to derive a GRN [12]. One method using this information is DISTILLER, which uses itemset mining combined with ChIP-on-chip interaction data to search for evidence of co-regulation [69]. Other methods that employ biclustering, such as cMonkey, employs Markov chains to model the biclusters while making use of upstream sequence information as well as association networks and searches for over-represented *de novo*-detected motifs to further support gene co-regulation and report sequence features responsible for the co-regulation [68]. It has been reported that it is possible to identify co-expressed gene-sets in the subgroups of breast tumour samples using this method [47]. Unfortunately, these methods have not been selected for any of the major evaluation studies carried out including the BicAT toolbox, as cMonkey requires sequence binding motif information and was not appropriate to make comparisons to other biclustering methods [70]. Due to this response in the community, an updated version of this method was published earlier last year called cMonkey2 claiming to improve its usability and it can also take other types of information as input such as protein-protein interactions and ChIP-seq [19].

The BicAT toolbox is a means to compare the performance of biclustering algorithms and evaluated methods based on GRN prediction [70]. After obtaining the biclusters, a Bayesian network method was used to learn the subnetworks from the biclusters found and these subnetworks were then combined to make a final GRN. Experiments conducted on datasets using the introduced tool revealed that biclustering algorithms in general have advantages over the conventional clustering ones. To examine whether the performance on the datasets is typical of all network reconstruction methods and is not particular to Bayesian networks with biclustering, the authors compared results with a linear regression method (LASSO). They found the biclustering methods performed consistently regardless of the network reconstruction algorithm used. Current evaluations using this toolbox have found there is no single algorithm that is able to discover all interesting patterns so integrating results based on the enrichment of the output biclusters with gene ontology functional categories is recommended in this case as well. However, they avoided evaluations of many biclustering algorithms due to other information required to run them.

One final integrative biclustering method specifically for GRN prediction is COALESCE (combinatorial algorithm for expression- and sequence-based cluster extraction). COALESCE is a nondeterministic greedy algorithm that seeks biclusters representing regulatory modules in genetics [71]. It finds up-regulated and down-regulated biclusters starting with a pair of correlated genes, updating selected columns by two-population z-test, motifs by a modified z-test, and then selects rows by posterior probability. Although the algorithm was proposed to work on microarray data together with sequence data as well, sequence data has not been used in evaluations [49]. Biclustering methods such as these and other Bayesian network methods that fit a model to the entire dataset are less sensitive to noise, which is identified by a lot of methods that only seek localized patterns. cMonkey2 has been evaluated by the authors against these integrative techniques.

3.6 Limitations of Small Sample Sizes

Methods that are used to predict GRNs tend to be limited in accuracy when only a small number of sampling points are available. When trying to predict interactions in a complex system, it is better to have many more measurements than states, otherwise the system is largely under-constrained and can have many solutions [72]. This is generally referred to as the curse of dimensionality and remains a challenge in GRN prediction. Although integrating data types has been done with some success, there are still challenges associated with it as these various data types do not tend to be directly compatible. Indeed, even combining microarray data across different platforms is difficult. When validating techniques for GRN discovery, researchers tend to utilize samples in the hundreds [13]. One method to combat this may be feature selection, where a much smaller subset of genes is selected from which a GRN can be predicted. Most commonly, feature selection is performed using some method of clustering or using differential expression information [31].

There are select studies that use only a handful of samples from there own research, but either validate or incorporate information external to the studies [73, 74]. One study collected 4 time-series samples at day 4, 8, 11, and 14 with 2 replicates of each in order to infer a network responsible for the differentiation of one type of cell to another in humans [74]. However, they also had access to 52 microarray datasets appropriate for weighting the gene pairs generated in their GRN prediction algorithm in order to determine likely interactions. With prior information of genes more likely to function as transcription factors in humans, the accuracy of the predicted GRN improved further since they were able to restrict the number of possible regulators. Generally, having data in different states allows for sample reduction as opposed to using steady-state samples. Simulation studies artificially generating microarray data from artificial networks also indicate that random perturbations contain more information about gene regulatory interactions compared to single time series with an equivalent data size, even with a higher sampling rate [15].

Also, extensive information is available for the number of samples required depending on the type of GRN an individual wished to predict [72]. However, these numbers are based on how each GRN prediction method behaves theoretically. Currently the number of samples required has been studied for microarrays and the number of data points required is known for simulated time-series data. Experimental performance of ARACNE, SPACE, and WGCNA has been measured in relation to the number of simulated microarray samples provided for each method. With 20 samples and 1344 genes, all of the methods performed better than random, based on area under the ROC curve (AUROC) results, which measures the performance of the algorithm across all sensitivity and specificity ranges [63]. Results continued to improve as more samples were added.

Other research suggests an estimated 64 samples should be enough for researchers to obtain the best possible predictions if considering precision, suggesting that any samples above this is superfluous [75]. These results were observed with networks with sizes ranging from 100 to 1000 on synthesized time series and steady state data as well as one real network from *Escherichia coli* of size 1146 and only with the C3NET algorithm. C3NET works by trying to eliminate nonsignificant connections among gene pairs by testing the statistical significance of pair-wise mutual information values [76]. C3NET can never predict more edges than genes as the maximization step only allows a single edge to another gene so at most the number of edges will be equal to the number of genes used for prediction. Therefore, a connection between two genes will correspond to the maximum mutual information value between a gene and all its neighbours, which will also have the lowest p-value. The author admits these results may not generalize to other methods, one reason possibly being the study was limited to information-theory based algorithms that do not require only the gene expression data with mutual information values and a cut-off for these values in order to eliminate non-significant edges. Also, only one real microarray dataset was tested as well on a relatively simple organism with a well studied network, which means the study may not be applicable to highly complex organisms, such as vertebrates. Precision is used to evaluate C3NET since it is unable to predict more edges than genes present, which increases the number of false negatives. However, this limitation is not a factor in this thesis because only two transcription factors are focussed on, which will require more than a single connection from these transcription factors to two other genes. Precision of a real network may not be more indicative of method performance, as typically with real networks all of the actual interaction taking place within a network are not known, which may inflate the number of false positives. Regardless of the precision of the network (time series alone resulted is poorer performance compared to the steady-state data), the data converged around the same number of samples and increasing this number did not further improve the networks.

The number of samples necessary when combining data from various sources is not well established. There is a question of whether data from other sources each count as a single data point depending on how the data is integrated together (before or after initial network construction). It is also difficult to determine accuracy of large scale GRNs in mammalian systems, as there is no gold standard to compare to presently [77]. Currently for mouse datasets, the smallest found in the literature predicted a GRN with 21 samples, which were only used to compare module conservation with microarray data [61].

Chapter 4

Methodology for Analysis of Gene Expression in Skeletal Tissues

This chapter introduces the methods required for the bioinformatics analysis of RNA-seq data to determine what evidence from gene expression may be observed to suggest that there are two GRNs driving development of bone, immature and mature cartilage. The RNA-seq data from skeletal tissues, referred to throughout the thesis, is introduced as well as how transcript quantitation, normalization and clustering analysis are performed. The results are presented and discussed in Chapter 5.

4.1 Dataset Overview

RNA-seq data provides discrete counts of gene transcripts. There are generally a high number of genes with very low expression counts (in terms of the number of transcripts), and expression levels of fewer transcripts are characteristically high. There have been two underlying distributions proposed to model RNA-seq data [44]. The first is the Poisson distribution, which tends to be used when analyzing technical replicates. Therefore, to decrease the potential false positive rates due to underestimation of sampling error, a negative binomial distribution is generally used to model data containing biological replicates, which tend to contain an overdispersion or more variance in expression levels. In a Poisson distribution, the variance should be similar to the mean, which is too restrictive for data containing biological replicates [1].

Nine samples with three replicates for bone, immature and mature cartilage from mouse, with a total of 13302 genes, will be kept for clustering purposes. The genes selected for clustering were not necessarily differentially expressed in one tissue when compared to the others, but had to be considered expressed over a cut-off in at least one of the tissues. In order to confirm the distribution within the RNA-seq data for this thesis, the samples were sorted by expression levels seen from lowest to highest and the density of expression levels observed are displayed in Figure 4.1. The distribution of average log2 expression across three replicates of each tissue was determined, producing three bimodal distributions. It is common practice to filter out counts that are either close to zero across all samples as well as transcripts expressed at a low level as this may be due to artifacts [78]. It is likely that these genes were expressed without any functional consequence of interest. The initial peak represents genes with a very small number of counts that cannot be attributed to any phenotypic characteristics. The second peak is indicative of a smaller number of genes that have

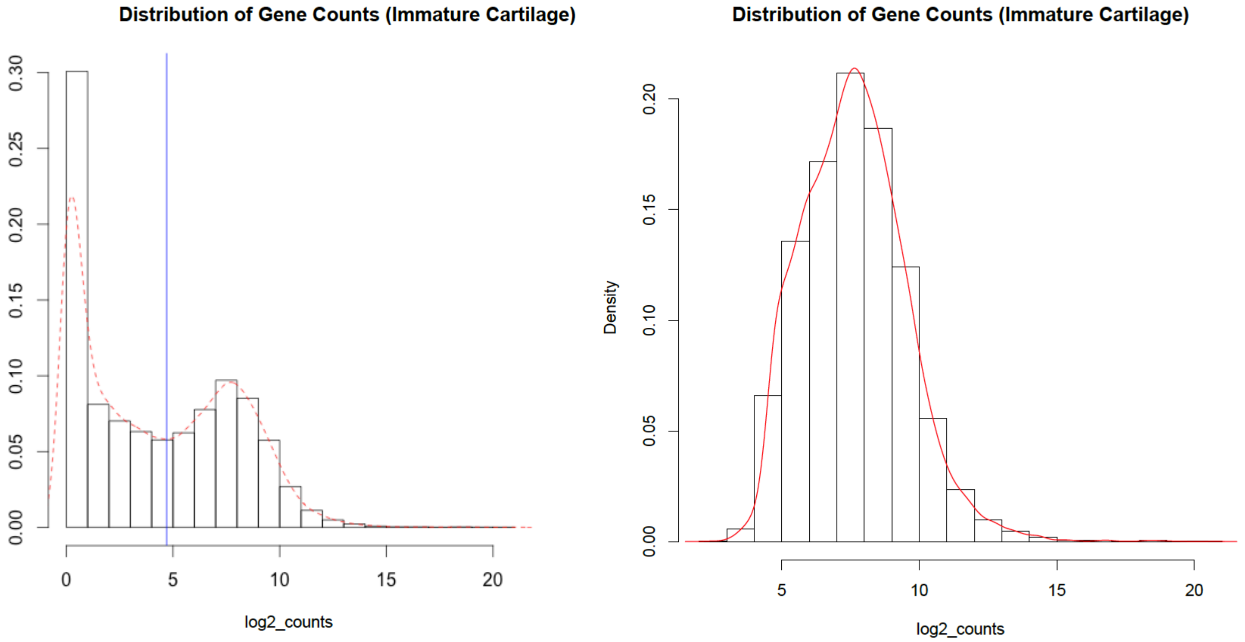
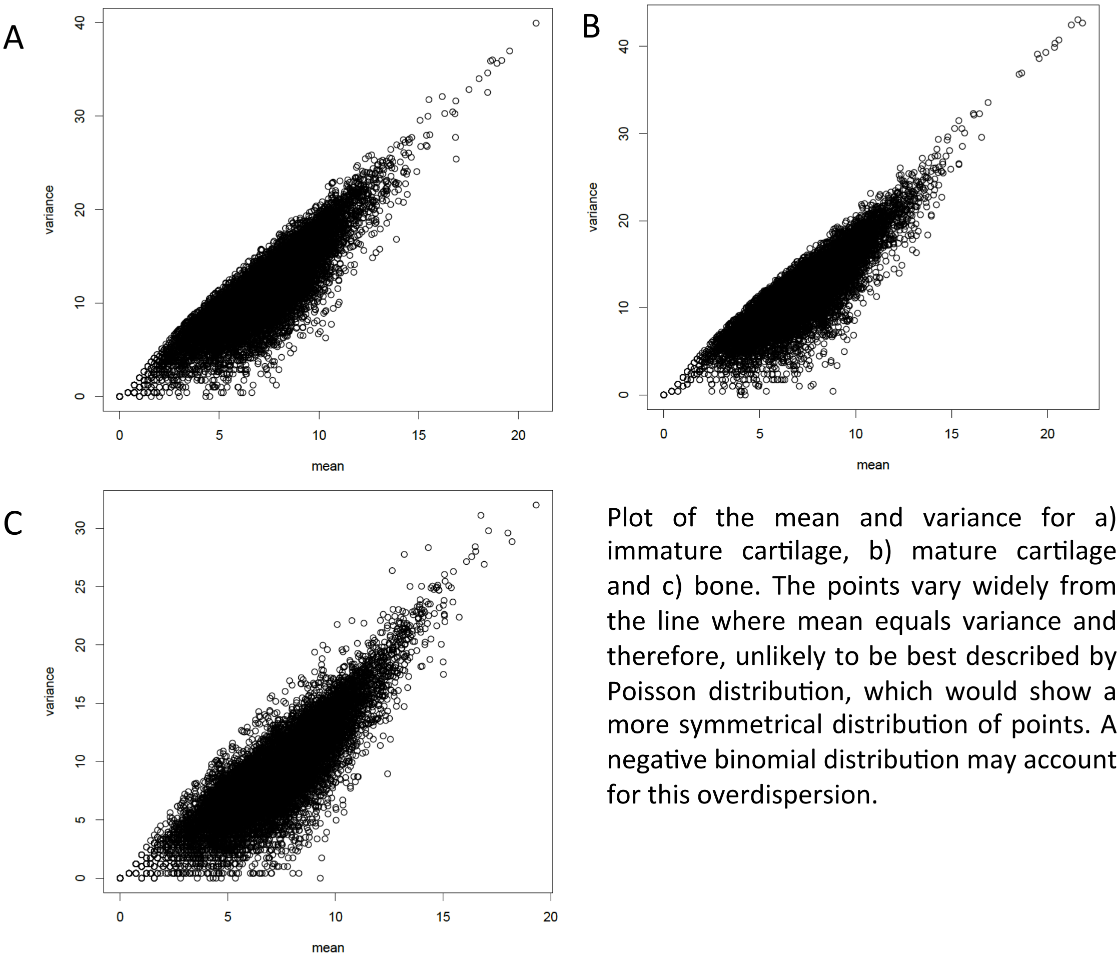


Figure 4.1: Distribution of counts in immature cartilage before and after cut-off of 25 counts was applied. The y-axes show the density of different amounts of gene counts. There is a higher number of low gene counts and a smaller number of genes with high gene counts. Although these genes with low counts could be informative, many are also likely to be un-informative, which could cause problems with downstream analyses. Since the higher gene counts are likely more accurate, the minimum of the bimodal distribution was selected as a cut-off, and the more highly expressed genes were kept for further anlysis. For immature cartilage, the cut-off was 25 counts on average across all three replicates.

expression levels capable of influencing traits of each tissue. An appropriate cut-off to limit biological and technical noise was determined by calculating the minimum value between these peaks before the peak of significantly expressed genes.

Once low counts were filtered out, each expression profile had a distribution that looked closer to either a Poisson or negative binomial distribution. Next, the mean of each gene for all samples was plotted against variance to determine if the data had means similar to variance as in the Poisson distribution or if there is overdispersion in gene expression. From Figure 4.2, a negative binomial distribution appeared best to model the count data, as a negative binomial distribution can account for larger variance [44]. Therefore, the expression levels within each tissue type is likely a mixture of this probability distribution.



Plot of the mean and variance for a) immature cartilage, b) mature cartilage and c) bone. The points vary widely from the line where mean equals variance and therefore, unlikely to be best described by Poisson distribution, which would show a more symmetrical distribution of points. A negative binomial distribution may account for this overdispersion.

Figure 4.2: Plots of the mean expression of each tissue compared to the variation observed between the samples for each tissue. The plots show a larger amount of variation than what would be expected with a Poisson distribution.

4.2 RNA-seq Analysis Pipeline and Comparison to Sox9 and Runx2 Literature Networks

4.2.1 Mapping and Transcript Quantification

Paired-end RNA-seq data from mouse is available as raw transcript sequence reads obtained from Illumina 1.9 sequencing of bone, immature and mature cartilage tissues. These raw sequence reads were assessed using FastQC [79] to check for low quality reads as well as over-represented reads from primer or adapter sequences used in Illumina sequencing. Trimming was applied using the Java application Trimmomatic [80] to filter out low quality sequences as well as possible adapters and primers present. This step resulted in forward and reverse read fastq files. Any unpaired reads were discarded before aligning the reads to a reference genome.

The origin of each read was identified using the mapping and alignment programs called Tophat2 (version

2.0.13) and Bowtie2 (version 2.1.0) respectively [81, 82]. Tophat is a commonly used spliced alignment program that can be used for RNA-seq. Bowtie is the program that acts as the alignment engine for Tophat. Tophat begins by aligning reads using a reference genome in order to construct the transcriptome. The reference genome contains annotation in order to establish the position of the reads along the reference sequence. There is a low tolerance for mismatches, as the reads may not be truncated at the ends if they do not align. Bowtie2 is responsible for extracting the transcript sequences from the annotated reference genome and if there are reads that do not align to this transcriptome construct, they are then mapped to the original genome. The mm10 version of the mouse genome annotation files from Ensembl were used as a reference. Once mapped, the files were sorted by read names to count the reads per gene. Mapping-based assembly is done in order to obtain transcript counts to construct gene expression matrices. The Python program HTSeq [83] was used to obtain discrete transcript counts for each sample, which is a deviation from the traditional workflow using the Tuxedo suite of tools including Bowtie2 and Tophat2. HTSeq produces raw transcript counts, where Cufflinks, the third program in the Tuxedo suite, produces counts that have already been normalized to obtain FPKM (Fragments Per Kilobase per Million mapped reads) values [84]. However, there is some speculation as to how effective this normalization method is for comparisons to be made across samples and it may be best used for within-sample comparisons of genes [85]. FPKM corrects raw counts based on the transcript lengths as well as the sequencing depth and this correction is not affected by results of any other sample. Therefore, it was decided that access to the raw counts would be necessary for other normalization methods as well as to give more flexibility when determining differential and fold change expression levels. In this case, trimmed mean of M-values (TMM) normalization was used to correct for library size as comparison between samples in this case does not require normalization due to different transcript lengths. HTSeq locates the exons where the aligned reads overlap and groups the overlapping counts based on gene ID.

4.2.2 Venn Diagrams of Genes Expressed in Skeletal Tissues

Venn diagrams will be generated using gplots in R with lists of genes considered expressed above background in each tissue. The genes that are unique to each tissue will be determined by selecting the genes that are grouped in the outer portions of the Venn diagram. These are the genes that had counts high enough to be considered expressed in only a single tissue. In the other tissues the counts have to fall below each tissue-specific expression level cut-off. To determine the section of the diagram a gene should be grouped, the cut-off minimum for each tissue was 18, 24 and 25 for bone, mature and immature cartilage respectively from Section 4.1. These are the same cut-offs determined from the bimodal distributions representing each tissue. This is why no gene will be left out of the Venn diagram using these thresholds. Although these genes may be considered expressed in at least one tissue type, they are not necessarily differentially expressed when compared to expression levels of the same genes in the other tissues. For example, some genes may have similar expression, although in one tissue expression of the gene falls just below the cut-off.

4.2.3 Normalization and Differential Expression

Differential expression can be used to identify genes that are expressed in significantly different quantities when comparing groups of samples. The counts were normalized with TMM (trimmed means of M values) using edgeR from Bioconductor, which is a batch normalization technique dependent on the total counts across all samples and is not designed for single sample normalization as with FPKM [86]. It has performed well when compared to other normalization methods that also attempt to resolve isoform expression levels [4]. Pairwise differential expression will also be performed in this thesis using edgeR, a tool containing methods for the normalization of raw count data collected from HTSeq. This tool is capable of handling data that follows a negative binomial distribution such as what is obtained from RNA-seq with biological replicates. It is recommended that genes with small counts across all conditions be removed before performing differential expression [7]. Therefore, genes considered for differential expression analysis were filtered using the cut-offs used to construct the Venn diagrams to remove genes considered unique. The most up-regulated and down-regulated genes in bone, immature and mature cartilage will be determined from the pairwise comparisons by selecting genes in one tissue that were up-regulated compared to both other tissues or down-regulated compared to both. These gene lists will be compared to the genes present in the Sox9 and Runx2 literature networks using set operations in R.

4.3 Model-Based Clustering

One method of detecting patterns of gene expression in high-dimensional data is to use a clustering technique where genes are grouped together based on expression, implying they are more likely to be functionally related. A model-based clustering approach will be used to identify co-expressed genes in RNA-seq datasets, which employs either a Poisson or negative binomial mixture model to postulate the over-dispersed gene count data. Current methods available for model-based clustering for RNA-seq data including biological replicates involve a modified Expectation Maximization (EM) algorithm called MBCluster.Seq [1]. The expectation maximization (EM) algorithm allows for the estimation of probabilistic model parameters when not all data is known. In the context of clustering, the data considered incomplete is the gene assignments to each cluster. The EM algorithm requires one step to compute probabilities for each possible completion of the gene-to-cluster assignments using what is currently known. This creates a weighted training set to provide updated model parameters such as the means and covariance of the genes currently assigned to each cluster.

4.3.1 Algorithm Description for Model-based Clustering

Each number in a RNA-seq expression matrix represents a discrete RNA transcript count representing the gene expression level for every gene. It was necessary to obtain the expression profiles that are the log-fold-change (log-FC) values in order to determine whether a gene is up-regulated, down-regulated or has close to a neutral difference in expression. This measures the expression level of gene g in treatment i relative to the overall mean expression of that gene across all tissues. Due to high dimensionality (# of genes) within gene expression data, grouping genes of interest using clustering algorithms is a useful method of detecting possible patterns in expression across tissue types.

To take advantage of the underlying mixture of distributions in the skeletal tissue data, model-based clustering will be used to detect patterns in the RNA-seq data. The method below uses the EM-algorithm where:

Observed data x: RNA-seq measurements of expression

z: unobserved latent factors which are the assignments of the gene clusters

θ: parameters of means and covariance matrix of the negative binomial distributions representing expression patterns for each cluster.

Responsibility of each cluster $k = \frac{p(z=k|\theta)p(x|z=k,\theta)}{\sum_{k'} p(z=k'|\theta)p(x|z=k',\theta)}$, where k' is over all clusters.

A more detailed version of model-based clustering specific to gene expression is described in [1], is briefly explained here and is utilized from the MBCluster.Seq R package. Currently, this is the only method available for clustering RNA-seq data specifically by taking into account the distribution of the data, which is different than microarray data. In order to cluster using the EM algorithm, presented for this method, with the data for this project, $k = 10$ cluster centers were selected, represented by $\mu_k = (\mu_{k1}, ..., \mu_{kI})$. Each μ_k is a expression profile of a single gene, and I is the number of conditions or treatments represented by the samples.

The negative binomial model the algorithm uses has two parameters. One is the mean, which is calculated as $log\lambda_{gij} = \alpha_g + \beta_{gi}$ where α_g is the geometric mean gene expression of gene g, and β_{gi} is the expression level of gene g in treatment i relative to the overall mean expression. The second parameter estimated by the algorithm is the overdispersion ϕ_g, which will compensate for increased variance in the model compared to the mean where $Var(N_{gij}) = \lambda_{gij} + \phi_g \lambda_{gij}^2$.

The density of the negative binomial distribution or the likelihood of gene g belonging to the kth cluster for all genes being clustered can be represented as $\prod_g \sum_k p_k f(N_{gij}|\alpha_{gk}, \beta_g = \mu_k)$, which can be based on the negative binomial distribution in this case. N_{gij} is the count of reads mapped to gene g for replicate j of treatment i and p_k is the weight of a class ($\frac{1}{K}$) or how likely an observation belongs to cluster k, where K is the total number of clusters. For this algorithm, the authors assumed independence among the genes although this is likely not the case. However, it is not practical to model and estimate the correlation among thousands of variables such as genes with only several replicates and no prior knowledge about the relationship between the variables.

Instead of choosing the cluster center genes uniformly at random from all genes and using their expression profiles as the initial cluster centers, the program only selected one cluster center uniformly at random and then set the additional centers gradually by selecting genes based on the distance between each gene and each of the selected centers. The likelihood $f(N_{gj}|\alpha gk, \mu k^{(1)})$ is maximized with respect to the geometric mean expression (α_{gk}) for each combination of gene g and cluster k.

Once the cluster centers are selected, they are passed manually to the method containing the EM algorithm. The EM algorithm is composed of an E step to calculate the expected complete data log-likelihood that each gene g is in a cluster k. The next step is to maximize f with respect to α_{gk} for each gene g combined with each cluster k. For this algorithm, the responsibility, or Z_{gk}, is the variable indicating if gene g if in cluster k. It equals 0 if g does not belong to the kth cluster and 1 if it does belong. All of the indicator variables in this case are treated as unknown data ($Z = Z_{gk} : g = 1, \ldots, G, k = 1, \ldots, K$). The following portion of the algorithm will then iteratively calculate conditional expectations of Z and update the model parameter estimates.

The EM algorithm consists of the following steps:

1. E step: Calculate the conditional expectation of Z_{gk} given the parameter values from the previous iteration, where m is the iteration number. In other words, given all the current values of the model parameters, determine the cluster k, that gene g will fit best in, to obtain a distribution describing class/cluster k.

$$Z_{gk}^{(m)} = \frac{p_k^{(m)} f(N_{gij}|\alpha_{gk}, \mu_k^{(m)})}{\sum_l p_l^{(m)} f(N_{gij}|\alpha_{gl}, \mu_l^{(m)})}$$

2. M step: Update the parameter estimates of the model.

$$\mu_k^{(m+1)} = argmax \sum_g Z_{gk}^{(m)} log f(N_{gij}|\alpha_{gk}, \mu_k^{(m)})$$

$$p_k^{(m+1)} = \frac{\sum_g Z_{gk}^{(m)}}{G}$$

$$\alpha_{gk}^{(m+1)} = argmax_{\alpha_{gk}} f(N_{gij}|\alpha_{gk}, \mu_k^{(m+1)})$$

3. Repeat until the change in the log likelihood is small.

4.4 Differentially Expressed and Unique Isoforms

In order to generate a transcript count table as opposed to a gene count table, RSEM will be used in order to predict which RNA-seq reads come from each isoform [87]. Genes can be transcribed beginning at different sites, include different coding regions (exons) and different end points, which results in different mRNA sequences, thus potentially changing how the gene functions. These variations of the same gene are referred to as isoforms, as opposed to genes, which encompasses all variations of the gene. Using RNA-seq as opposed to microarray allows the potential to estimate expression of different gene isoforms. A gene could have multiple isoforms with some up-regulated while other isoforms of the same gene are down-regulated. At the gene level, this differential expression could be masked if the up-regulated and down-regulated isoforms cancel each other out or it could result in up-regulation or down-regulation overall for the gene even if the opposite is true for select isoforms, which could lead to misleading results or conclusions.

Using a workflow that has been published for RSEM [87], a transcription reference and index files for Bowtie2 are constructed using the reference genome for mouse from Ensembl. Next, the reads are aligned to the transcriptome using Bowtie2. Since isoforms of a gene normally share a significant portion of their sequences, the read mapping uncertainty increases dramatically. Thus, the first command, rsem-generate-ngvector clusters isoform sequences into 3 clusters according to each isoform's hardness of being mapped uniquely [87]. Then, EBSeq estimates the mean and variance parameters separately for each cluster [88].

It is important to note that estimated counts are not the same as raw counts and therefore common differential expression software such as DeSeq and edgeR are not recommended to calculate differential expression [88]. After the estimated counts, rsem-generate-data-matrix extracts the estimated expected counts from each sample and then generates a count matrix GeneMat.txt that can be used by EBSeq to perform differential expression of isoforms.

After differential expression is performed using EBSeq, controlling false discovery rate at 0.05, the posterior fold changes between tissues were converted to log2 fold change values and significant differential expression was considered to be above 2 or below -2 fold changes. Isoforms will be separated based on up-regulation and down-regulation to determine the isoforms that are only up-regulated in a single tissue. The cut-offs for the isoforms will be set to the original values from Section 4.1 (gene counts: IMM=25, MAT=24, BON=18) using the same strategy as for gene counts to determine if there are particular isoforms most likely dominant to others for a particular gene, although there will also be more genes falling below these cut-offs if the counts are divided among multiple variants.

Chapter 5

Results of Applied Bioinformatics Analysis to RNA-seq Data from Skeletal Tissues

This chapter covers analysis of the similarities and differences between bone, immature and mature cartilage based on the gene expression information obtained using RNA-seq data. The purpose is to test the hypothesis of two GRNs driven by Sox9 and Runx2, and to determine how the extent these GRNs may be interacting with each other in each of the tissues. One method used to analyze the differences between the three tissues will be to analyze genes only expressed in one of the tissues. The reason unique genes are important is that they can help provide evidence for or against the hypothesis of a completely additive GRN. If mature cartilage has gene expression which is a complete mixture of what is found in the other two tissues, then there should not be any genes considered unique. Furthermore, they can help identify genes that are more likely to be under control of Sox9 and Runx2. Each tissue's gene expression profile will also be analyzed using differential expression and clustering to observe to what extent they share similar gene expression in order to determine if it is more likely the GRNs driving development are interacting.

5.1 Comparison of RNA-seq Data to Literature Networks for Sox9 and Runx2

5.1.1 Venn Diagrams of Genes Expressed in Skeletal Tissues

Genes identified in RNA-seq as uniquely expressed in either bone, immature or mature cartilage will be used to determine the number of genes present in the Sox9 and Runx2 networks available in the literature. The literature networks come from two publications discussed earlier in Section 3.3, which use silencing of Sox9 and Runx2 to determine genes that could potentially be up or down-regulated by these transcription factors. A large portion of genes are considered unique in the RNA-seq dataset that do not appear in the differentially expressed gene lists from the publications. The genes in the Runx2 literature network was limited to the genes that appear in the microarray data. How genes were defined as unique to a tissue is further described in Section 5.1.2.

The Venn diagram, showing gene expression overlap of bone, immature and mature cartilage in Figure

5.1, indicates that mature cartilage has a lot fewer genes that are uniquely expressed compared to the other two tissues. Only 321 genes are considered uniquely expressed in mature cartilage compared to 639 in bone and 513 in immature cartilage. Furthermore, the majority of gene expression is above cut-off in all three tissues (10239). Mature cartilage and bone have more overlapping genes expressed above cut-off (857) than bone or mature cartilage have overlapping with immature cartilage (228 and 505, respectively). In particular, bone has the least number of expressed genes in common with immature cartilage, due to the small number of genes above cut-off in only these two tissues (228). There is minimal overlap in the genes present in the literature networks for Sox9 and Runx2 and the genes expressed above cut-off in at least one tissue. More genes overlap with those present in the Sox9 literature network likely due to the number of genes in the network (849) compared to the 200 genes in the Runx2 literature network. However, a higher number of genes considered to be in the Runx2 network are present in the overlapping genes between bone and mature cartilage (17, with 8 up-regulated) compared to bone and immature cartilage (1). The single gene from the Runx2 network in the overlap of genes expressed in bone and immature cartilage is also down-regulated. More up-regulated genes in the Sox9 network are also present in the overlap between immature and mature cartilage (51) compared to bone and mature cartilage (33). There is also a higher number of genes from the Sox9 literature network that are down-regulated, but still expressed above cut-off in both bone and mature cartilage (26) compared to immature and mature cartilage (1).

The unique genes expressed in each tissue also have a small portion of overlapping genes with the literature networks as shown in Figure 5.2. In immature cartilage, there is only 1 down-regulated gene from the Runx2 network and 29 overlapping with the Sox9 literature network. The genes considered unique in mature cartilage and bone samples also have a higher number of genes overlapping with the Sox9 literature network, although with a few more that are down-regulated. However, bone and mature cartilage also has more genes that are considered down-regulated in the Runx2 literature network than up-regulated genes.

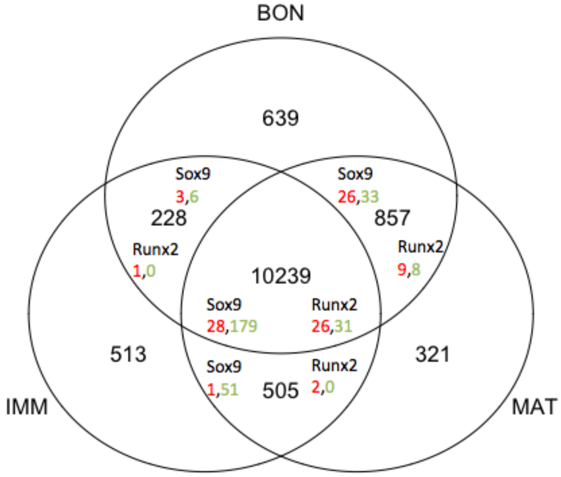


Figure 5.1: Venn diagram of genes expressed in bone (BON), immature (IMA) and mature (MAT) cartilage. Genes are divided into differentially expressed genes also in the literature networks for Sox9 and Runx2. The red indicates genes in the RNA-seq data for bone, immature and mature cartilage also in the literature networks that were reported as down-regulated. The green numbers are genes in the literature network that were repoted as upregulated. The numbers at the center are genes present in both networks as well as whether they are up-regulated or down-regulated in the Sox9 (left) and Runx2 (right) networks.

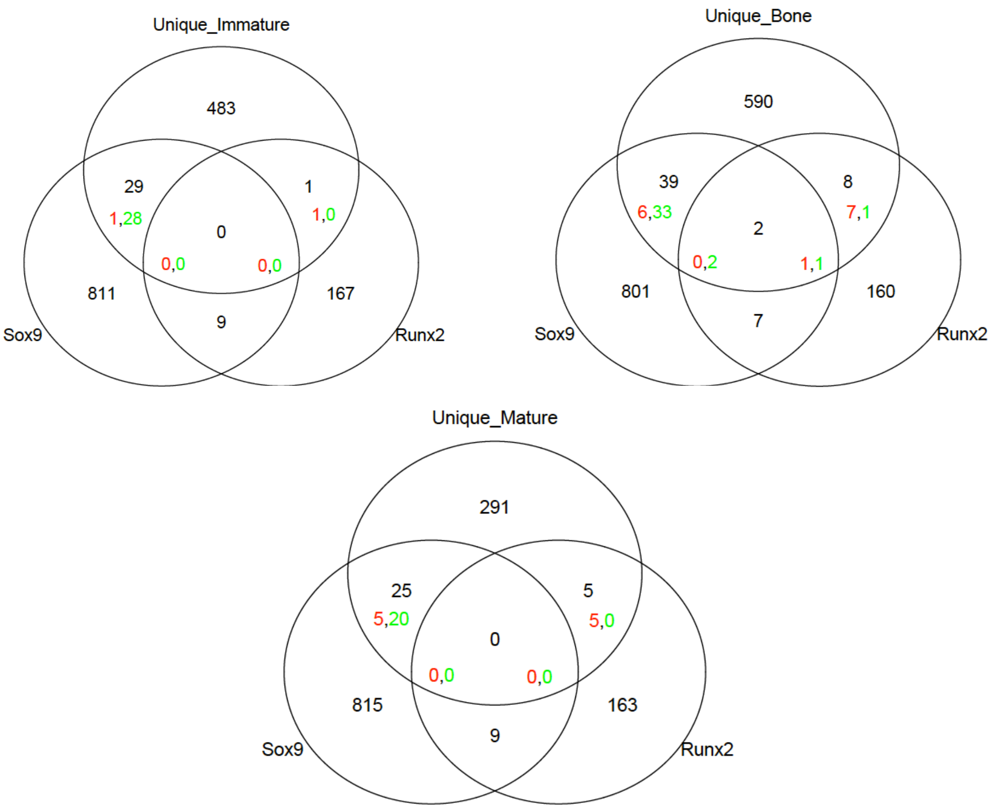
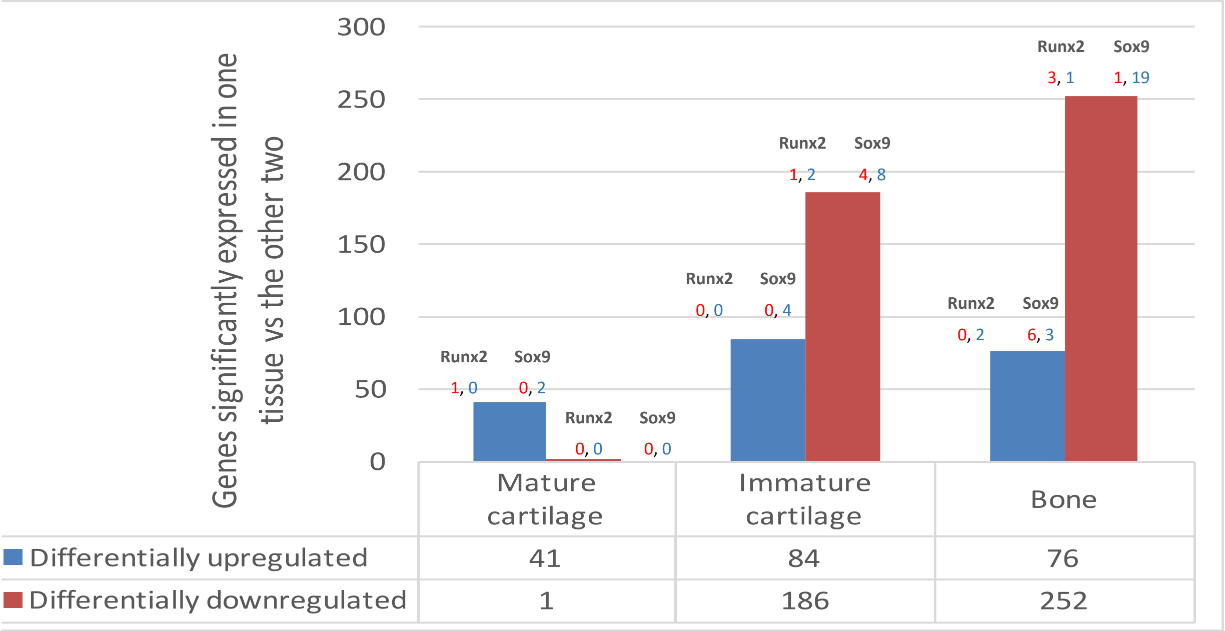


Figure 5.2: Venn diagram of genes uniquely expressed in bone, immature and mature cartilage showing overlapping genes with the literature networks. Genes are divided into differentially expressed genes in the literature networks for Sox9 and Runx2. The red indicates genes down-regulated in the literature networks while green indicates genes up-regulated.

A higher amount of overlap between mature cartilage and the other two tissues may be indication that mature cartilage has genes being expressed in a similar amount to one tissue or another with fewer genes being expressed only in mature cartilage. Also, since bone and immature cartilage have a lot fewer genes expressed that are not expressed in mature cartilage, it is likely their gene expression is the least similar. This suggests that the GRN active in immature cartilage does not have the same influence in bone as there are less genes expressed above threshold in both tissues. From these results it also seems that the current literature networks may not be an accurate depiction of the genes in the Sox9 and Runx2 networks present in these skeletal tissues. It may also be that the RNA-seq data available for bone, immature and mature cartilage is not appropriate for determining potential Sox9 and Runx2 networks accurately. However, the smaller number of unique genes expressed in mature cartilage, as well as the large overlap between genes expressed in all three tissues may be evidence that the GRNs functioning in these tissues share a lot of similarities. Mature cartilage, in particular, has gene expression similar to either bone, immature cartilage or both, more often than having uniquely expressed genes. Further analysis to explore these trends is done using model-based clustering in Section 5.2. Another benefit of using RNA-seq is that, potentially, isoforms can be identified that contributes more or less to the expression of a gene as a whole. An initial exploration of splice variants is performed in Section 5.3.

5.1.2 Differential Expression

Genes that are considered unique, as described in Section 5.1, are not automatically considered differentially expressed in the following analysis. A unique gene is defined as a gene with counts above a cut-off in only one tissue, meaning it is only considered expressed in that tissue. If a gene is considered uniquely expressed, or below cut-off for unique expression in all three tissues, it is also not considered to be differentially expressed. A gene can only be categorized as differentially expressed if it is expressed above cut-off, and considered expressed in more than one tissue. This way, unique genes and differentially expressed genes are separated into distinct groups for analysis. The genes that were differentially expressed in one tissue versus the other two tissues were determined and the number, for each tissue, appears in Figure 5.3. The mature cartilage RNA-seq samples have fewer genes that are considered up or down-regulated compared to immature cartilage and bone with only 41 genes up-regulated compared to both bone and immature cartilage and only 1 down-regulated gene. Of the up-regulated genes, only one gene in the mature cartilage RNA-seq data is in the Runx2 literature network and 2 genes are in the Sox9 network. The down-regulated gene, *Selenbp1* in the RNA-seq data, does not overlap with either network. The tissue with the most overlapping genes with the Runx2 and Sox9 network is bone. The Sox9 literature network contains 19 genes that are up-regulated by Sox9 and down-regulated in the RNA-seq bone samples. Genes that are up-regulated in bone include 6 genes down-regulated by Sox9 in the literature network as well as 3 that are up-regulated by Sox9. In the immature cartilage RNA-seq samples, 8 genes apparently up-regulated by Sox9 in the literature network were down-regulated in the immature cartilage RNA-seq samples. Over all, not many genes in the literature networks overlapped across any of the tissues.



	Mature cartilage	Immature cartilage	Bone
🟦 Differentially upregulated	41	84	76
🟥 Differentially downregulated	1	186	252

Figure 5.3: Number of genes significantly differentially expressed in one skeletal tissue compared to both other tissues. The number of genes overlapping with the literature Sox9 and Runx2 networks are shown above each bar. Red numbers indicate genes that are down-regulated by Sox9 or Runx2 and those in blue indicate genes up-regulated by Sox9 or Runx2. Genes with significantly different expression: up-regulated >2 or down-regulated <-2 log2-fold change was used as a cut-off.

These results suggest that not only is mature cartilage gene expression similar to the gene expression driving immature cartilage formation, but it is also very similar to bone gene expression. Since mature cartilage does not have many genes that are differentially expressed compared to bone and immature cartilage, it suggests that the majority of the gene expression is similar in some way to either immature cartilage or bone. Furthermore, the genes that are differentially expressed in mature cartilage compared to both tissues are almost all up-regulated with only one gene considered down-regulated. Therefore, the GRN driving mature cartilage formation may produce some synergistic effects. It is hypothesized that they would have opposite influence on gene expression where Sox9 down-regulates a gene and Runx2 up-regulates the gene or vice versa. This is because Sox9 is dominant to Runx2 so it has to be suppressed if Runx2 is going to influence gene expression. However, perhaps both transcription factors are able to up-regulate some of the same genes, leading to higher expression in mature cartilage compared to both immature cartilage and bone. This is further discussed in Section 5.2. The tissues where Runx2 should have the most influence on gene expression, bone and mature cartilage, have very few overlapping genes with the Runx2 network (6 and 1 respectively). These results further highlight limitations with either the current network, the RNA-seq data or both due to the lack of overlap between genes in both networks as well.

5.2 Model-based Clustering

Model-based clustering of gene expression patterns across the skeletal tissues is performed to determine if there is evidence that the 2 GRNs driving bone and cartilage development interact with each other in mature cartilage. Principle component analysis (PCA) was performed on the data using prcomp from the stats library in R to determine if the biological replicates of each tissue separated into distinct groups based on gene expression variance. The covariance matrix of the data was also calculated using R in order to determine the eigenvectors and eigenvalues present, which are explained by Abdi et al. [89]. The largest eigenvalues were present in the first two eigenvectors and they explain the majority of the variance in the data. The first component explained 52.3% of the variation while the second component explained 18.8% of the variation. The bone replicates contain a lot less variation compared with the other tissues and overlap with the 95% confidence ellipse of the mature cartilage samples. This suggests that mature cartilage and bone have genes that vary from immature cartilage, but are similar to one another. It may be an indication of the genes that are distinct from the Sox9 GRN in immature cartilage that make mature cartilage and bone distinct tissues with GRNs that possibly include regulatory control by Runx2. The variation seen in bone also varies orthogonally to mature cartilage and immature cartilage in coordinate space.

The algorithm from 4.3.1 was used to perform global clustering on the gene expression profiles across all three tissues. Each cluster was analyzed by comparing average expression levels in each cluster, which clusters had genes from the Sox9 and Runx2 literature networks and where Sox9 and Runx2 are located in the clustering results.

5.2.1 Results

The algorithm from MBCluster.Seq 1.0 package in R, was used to cluster genes based on expression profiles into 10 clusters as specified manually shown in Figure 5.5 [1]. Figure 5.5 shows gene expression that has been clustered according to similar patterns observed in expression across all three skeletal tissue. The clustering was visualized using the hybrid-hierarchical clustering capabilities provided, which begins from an initial partitioning of the genes, then merges the smaller clusters repeatedly to obtain a tree structure. We hypothesize that the majority of mature cartilage gene expression is a mixture of both immature cartilage and bone expression, which is supported upon visual inspection of Figure 5.5 [8]. The average gene counts for each cluster in Figure 5.5 show most clusters have an average expression for mature cartilage, across all genes in the cluster, that is between the average expression of immature cartilage and bone. The exceptions are cluster 1 and cluster 8, which have higher gene count averages in mature cartilage than the other two tissues. The expression in cluster 1 is higher in mature cartilage due to several mitochondrial genes that were placed in this cluster and have much higher expression in mature cartilage compared to all the other genes grouped in this cluster. Therefore, overall, this cluster appears to have higher expression observed across all the genes inside the cluster for immature cartilage, while there is more variability in gene expression observed

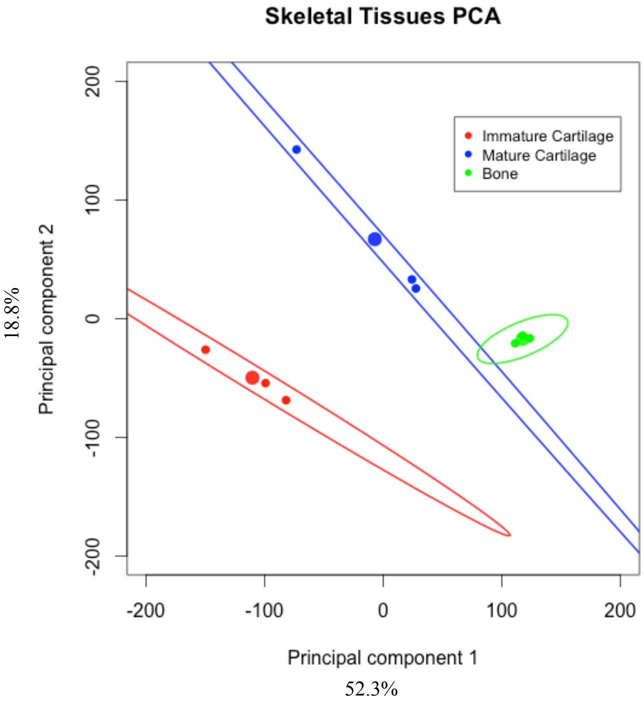


Figure 5.4: PCA of biological replicates for bone, immature and mature cartilage. This was done to visualize any strong patterns of variation within the dataset. It appears that mature cartilage and bone may have more similar gene expression patterns to each other compared to immature cartilage. Bone also appears to have less variation across biological replicates.

in mature cartilage.

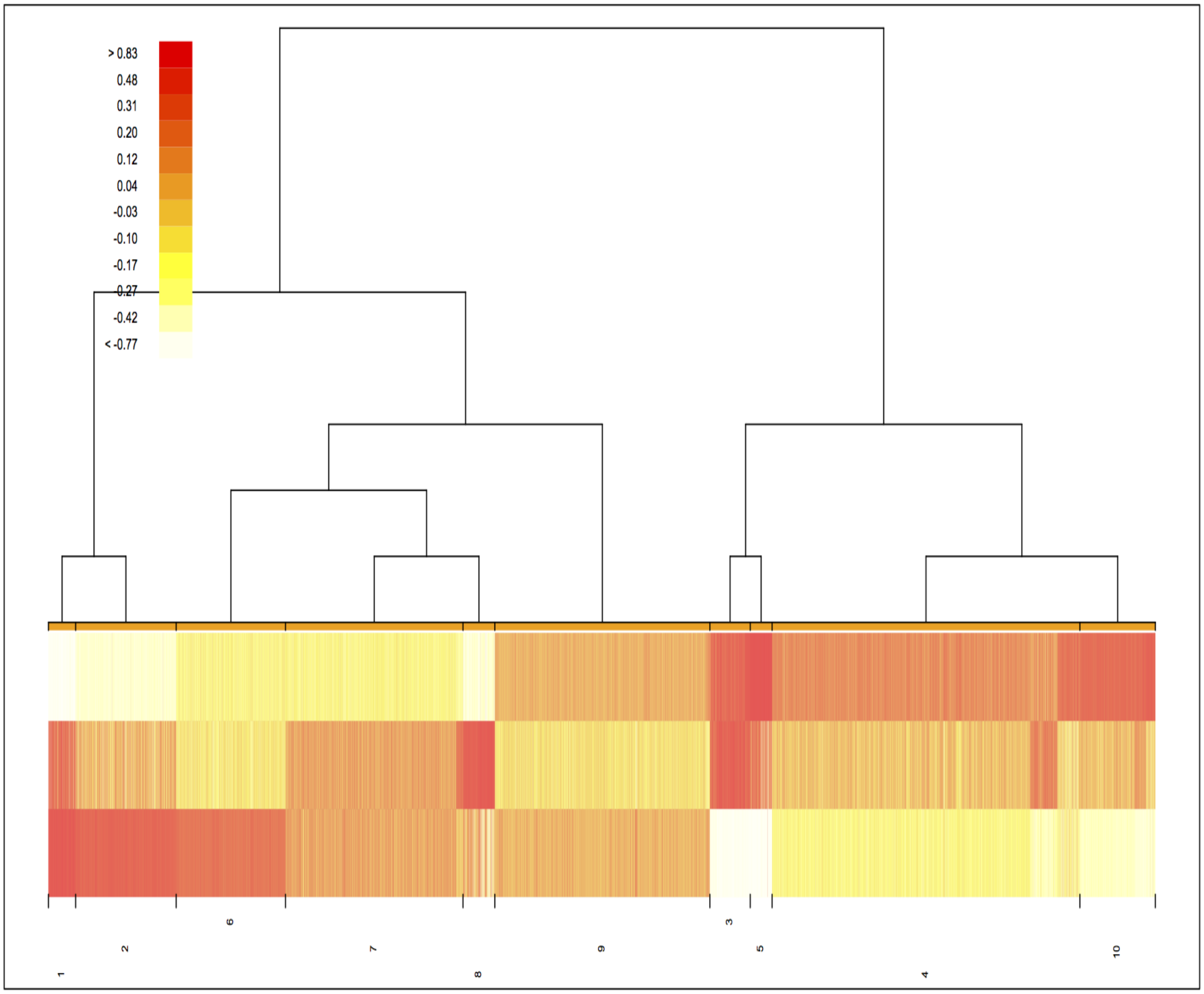


Figure 5.5: Visualization of model-based clustering with bone, immature and mature cartilage using MBCluster.Seq [1]. The gene expression was transformed to have a mean entered at zero. Expression higher than the mean is indicated by red, while yellow and white are lower than the mean expression across all three tissues. Bone expression is indicated by the first row of expression values followed by mature cartilage in the second row and immature cartilage in the final row. Each column (line of colour) indicates expression for a single gene, with the tissue type depending on the row. The numbers from 1 to 10 along the bottom indicate the cluster number.

Table 5.1: Average gene counts for each tissue, in each cluster

Cluster	1	2	3	4	5	6	7	8	9	10
Immature	9192.00	995.53	18.90	444.79	34.96	514.65	502.30	9869.25	579.42	277.52
Mature	14020.66	560.53	332.52	569.25	756.54	282.58	498.84	34719.66	487.39	535.05
Bone	455.35	156.61	382.13	865.47	5539.89	234.91	329.71	1717.56	614.86	1219.93

Sox9 and Runx2 are present in clusters 1 and 3, respectively. Cluster 1 has 55/423 genes from the Sox9 literature network while only 1 gene from the Runx2 network, which is down-regulated. This seems to indicate most genes in this cluster might be associated with the Sox9 network. However, it should also be noted that genes from the network are present in every cluster with cluster 4 having the greatest number from the network, while also being one of the largest clusters. Cluster 3 has 9/86 genes from the Runx2 network, with the highest number of genes from the Runx2 literature network present in cluster 4.

Table 5.2: Number of genes from Sox9 and Runx2 literature networks separated by up and down-regulation. Sox9 is in cluster 1 and Runx2 is in cluster 3.

Cluster	1	2	3	4	5	6	7	8	9	10	Total
# in Sox9 network	55 (0,55)	47 (4,43)	41 (19,22)	82 (14,67)	22 (7,15)	33 (4,29)	55 (6,49)	25 (4,21)	38 (7,31)	25 (5,20)	423
# in Runx2 network	1 (1,0)	9 (4,5)	9 (5,4)	17 (10,7)	7 (1,6)	5 (5,0)	(7,8)	5 (5,0)	6 (1,5)	12 (7,5)	86
# in cluster	327	1208	485	3699	263	1314	2133	382	2584	907	13302
% in Sox9 network	16.8	3.9	8.5	2.2	8.4	2.5	2.6	6.5	1.5	2.8	
% in Runx2 network	0.3	0.7	1.9	0.5	2.7	0.4	0.7	1.3	0.2	1.3	
	*Sox9		*Runx2								

If the proportion of genes is normalized using the total genes in each cluster, shown in Table 5.2, cluster 1 contains the largest proportions of genes from the Sox9 literature networks when compared to the size of each cluster with 16.8% of the genes overlapping. This cluster has similar gene expression in immature and mature cartilage, with lower expression in bone. Therefore, the genes clustered with Sox9 are likely genes from the Sox9 network that do not interact, or are not influenced by, genes in the Runx2 network. The clusters with the most overlapping genes from the Runx2 literature network are in cluster 3 and cluster 5 with 1.9% and 2.7% of overlapping genes, respectively. Cluster 3 contains Runx2 and both of these clusters follow a gene expression pattern of lowest expression in immature cartilage and highest expression in bone and mature cartilage. These genes are more likely to be influenced only by the network driven by Runx2, with little influence due to Sox9 expression. The clusters where Runx2 and Sox9 are clustered support that there are parts of each network that are present and active in mature cartilage, but these parts of the GRN are not interacting with each other to influence the expression of these genes. If genes in mature cartilage are sorted into categories of having expression closer to immature cartilage or bone or if expression is closer to an average between the two, 4667 genes have expression more similar to immature cartilage, 4015 genes are more an average of both tissues while 4620 have expression more similar to bone. Cluster 8 appears to have many genes that are up-regulated only in mature cartilage.

The results of model-based clustering have provided several clusters containing monotonic relationships, where genes differ in expression across all tissues in a by increasing or decreasing if analyzing a gene's expression across immature cartilage, mature cartilage and bone, respectively. This may help to identify genes that distinguish mature cartilage from the other tissues, not necessarily up or down-regulated, but that have different expression in mature cartilage compared to the both immature cartilage and bone. If these genes have not been used as probes in microarray studies to characterize mature cartilage than it could demonstrate the benefits of this RNA-seq method in comparison and provide more genes to classify that tissue type.

One reason for establishing a list of genes possibly in the Sox9 and Runx2 networks is to determine how they overlap and if gene expression in one GRN has an effect on the gene expression in another. From visual inspection of the clusters it looks like mature cartilage is usually an average of the gene expression present in the other two tissues for each gene. However, further inspection shows that there is a large portion of the genes in mature cartilage that either share more similar expression levels with one of the other tissues. This appears to occur almost evenly between immature and bone tissue. This seems to support the idea that mature cartilage, although a distinct tissue, has independent regulation of these genes by one GRN. Some gene expression in mature cartilage is an average of expression in immature cartilage and bone indicates the suggests some interaction between both GRNs in the same tissue. When mature cartilage has gene expression more like immature cartilage, these genes are likely from the Sox9 GRN and those genes that express similarly in bone might be from the Runx2 GRN. For example, cluster 1, which contains Sox9 has genes in mature cartilage that are more similar to expression in immature cartilage as opposed to an overall average. Further, cluster 3, containing Runx2, has genes with expression in mature cartilage more similar to bone. The only genes in cluster 1 showing higher expression in immature cartilage overall compared to immature cartilage are *mt-Rnr1* and *mt-Rnr2*, which skews the overall expression average and is not indicative of the pattern observed with the other genes in the cluster. As both of the networks driven by these two transcription factors drive the formation of bone and immature cartilage when acting independently of each other, mature cartilage shows similar gene expression with one tissue or the other instead of a mixture. Therefore, these clusters may contain many genes that show the most differences in expression between the two GRNs as opposed to the genes that may be present in both networks.

5.3 Preliminary Analysis of Splice Variants

Mature cartilage has 293 isoforms that are considered up-regulated with log2 fold changes greater than 2 compared to bone and immature cartilage. Immature cartilage and bone have 442 and 492 up-regulated isoforms respectively. This mirrors results for genes as well, where mature cartilage has the least number of genes up-regulated only in mature cartilage. The genes that have a dominant isoform or have only particular isoforms differentially expressed between tissues could show genes that are not considered differentially expressed when the sum of counts across all isoforms are considered, but there is differential expression among the isoforms when analyzed individually. Examples of these genes in mature cartilage, for example, would be *Lmo7-002*, which has a log2 fold change of 1.5 (indicating no significant differential expression using our cut-offs) when comparing genes across immature cartilage and bone, but is not picked up as differentially expressed in mature cartilage. There are also genes that do not show up as differentially expressed genes, but have up-regulated isoforms like *Mybph-201.* In the future it will be necessary to determine if this information is due to different isoform expression in the unique genes or if it can be attributed to differences between EBSeq and edgeR methods of detecting differential expression.

5.3.1 Unique Isoforms

This chapter concludes with a preliminary analysis of unique isoforms found in the RNA-seq dataset. The normalized dataset without a set cut-off contains 103,639 isoforms. Using the same cut-offs applied in Section 5.1.1, results in a total of 20,664 isoforms. Of these isoforms, there are 12,746 genes, with 8,681 of these genes only have a single isoform expressed above the cut-off in at least one tissue. Table 5.3 shows the number of genes with a number of isoforms. There are very few genes that have expression level high enough that more than 10 isoforms have expression levels above cut-off. Sox9 only has a single isoform, which is expressed above cut-off where Runx2 has 2 out of 13 expressed above cut-off. Figure 5.6 shows that mature cartilage has the smallest number of uniquely expressed isoforms compared to bone and immature cartilage, much like what is seen in Section 5.1.1. There is also still less isoforms expressed above cut-off between bone and immature cartilage compared to the isoforms in common between bone and mature cartilage as well as mature and immature cartilage.

Table 5.3: Distribution of genes by number of isoforms above cut-off

Number of isoforms	1	2	3	4	5	6	7	8	9	greater than 10
Number of genes	8681	2617	1060	437	161	77	38	20	4	8

5.3.2 Conclusion

The limited knowledge currently available describing the regulation of skeletal development could be further elucidated with the accurate measure of gene expression using RNA-seq technology. In order to utilize this data, the genes expressed in each tissue was plotted as a Venn digram. This showed a large number of genes expressed in all three skeletal tissues suggesting that the tissues require a lot of the same genes to

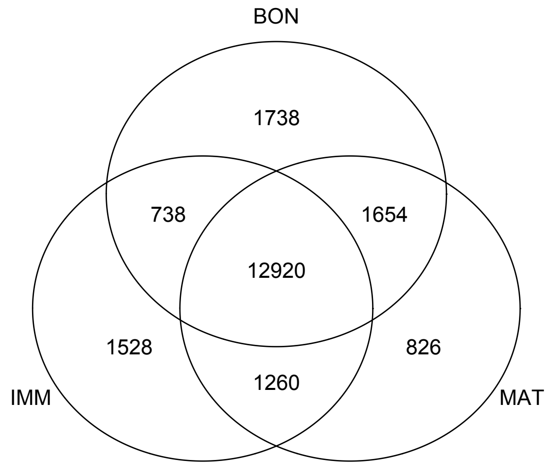


Figure 5.6: Venn diagram of all isoforms expressed above cut-off in bone (BON), immature (IMM) and mature (MAT) cartilage.

be expressed in order to develop. Furthermore, mature cartilage has the least number of uniquely expressed genes, with most gene expression that is above cut-off shared with either bone or immature cartilage. This supports the idea that the GRNs in immature cartilage and bone do not have as much interaction as in mature cartilage. This is also supported by fewer genes being expressed above cut-off that are shared only between immature cartilage and bone. The literature networks for Sox9 and Runx2 that were compared to have little overlap with all three tissues. However, there were more genes from the Runx2 literature network that overlapped with bone and mature cartilage compared to immature cartilage. This suggests that the Runx2 network has more influence on bone and mature cartilage development. The differential expression results also did not have much overlap with the literature networks, but also show that mature cartilage shares a lot more gene expression with the other two tissues, which have more genes that are differentially expressed. Using model-based clustering on RNA-seq data specifically is a relatively new concept that may be capable of grouping expression trends present across different tissue types. RNA-seq data from cartilage and bone tissue in mouse was appropriately modelled using a mixture of negative binomial distributions. Therefore this data appeared appropriate for evaluating the performance of this clustering algorithm and its ability to separate the different molecular processes occurring in each skeletal tissue. The transcription factors of interest were clustered into distinct groups and show evidence of the potential relationship between the Sox9 and Runx2 GRNs. Identifying a proficient means of analyzing expression data from skeletal tissue could contribute to further study of skeletal development using comparison across multiple species and ultimately comparisons being made between the molecular mechanisms of normal tissue development and degenerative skeletal conditions.

It will be of interest in the future to determine gene isoforms that play a dominant role in influence expression compared to the other isoforms of that same gene and if these isoforms are also differentially expressed when comparing bone, immature and mature cartilage. This could help to determine if there is a large portion of isoforms that are not identified as differentially expressed when considering gene expression of all isoforms together. It is unlikely that genes such as these would have been considered in analyses before if they have not been picked up in typical differential expression analysis. This data may also be used in the future to add more detail to GRN prediction using skeletal tissues, but this is currently outside of the scope of this thesis. In order to add this information, a comparison of edgeR and EBSeq differential expression results will have to be done beforehand.

Chapter 6

Methodology for GRN performance Evaluations for RNA-seq Data

This chapter describes the methodology that will be used to compare several biclustering algorithms, which is one method that can be used to predict GRNs, capable of grouping genes and conditions based on gene expression patterns. The biclustering algorithms are described as well as the metrics used to make comparisons. The results for biclustering comparison using RNA-seq data from skeletal tissues is presented in Chapter 7. This chapter also describes the methodology used to compare other machine learning methods for GRN prediction. This will involve using a well-described network in mouse, with available datasets, in order to have a gold-standard network to compare to GRN prediction results produced by each method. The results of GRN prediction methods compared to a literature network available is presented in Chapter 8.

6.1 Comparison of Biclustering Methods for GRN Discovery

In order to choose an algorithm to handle similar, yet functionally distinct tissue types, an analysis of the SAMBA, Plaid and FABIA algorithms handling RNA-seq data from mouse will be performed. These three algorithms were selected because of their accessibility as well as to test algorithms that have differences in performance when handling different tissue samples in previous studies as discussed in Chapter 7. Each algorithm tested can be used with RNA-seq data, though previous studies had only tested their ability to handle microarray results.

A comparison will be made between these three methods, and how they divide the biclusters based on tissue type, to determine if one, or any, provided a better solution to addressing differences between these skeletal tissues. The biclustering algorithms will be judged based on two criteria. First, based on their ability to differentiate various sample types, and second, based on how the groups of genes discovered by the methods are annotated using GO enrichment analysis to measure the biological relevance of the biclusters produced by all of the biclustering methods. The biological relevance of the biclusters will also be measured using what is currently known about the gene networks involved in skeletogenesis, with focus on the transcription factors Sox9 and Runx2 and the biclusters that contain them. It is possible that other genes within the same biclusters as these transcription factors are candidates for further studies on the molecular basis for skeletogenesis.

6.1.1 Biclustering Programs

Plaid The Plaid algorithm uses a series of additive layers over the gene expression matrix to try and explain the underlying structure [90]. Each layer is similar to a two-way Analysis of Variance (ANOVA) model between genes and conditions that represent different biclusters. There is also a background layer containing all the genes not currently in a bicluster. Samples and genes are located within a layer if they have a strong expression pattern that cannot be explained by the background layer. The algorithm fits this model using binary least squares to iteratively update cluster membership parameters of the genes and conditions to minimize the variance of expression levels within the current layer or bicluster.

SAMBA SAMBA [91] models gene expression data as a bipartite graph where each condition and gene is represented as a node of the graph while probabilistically assigned weighted edges connect them if a gene responds under the condition. Genes that have a degree of difference over a certain size, meaning their expression levels differ past the point of a selected threshold, are ignored. The subgraphs with more connectivity than the overall graph correspond to biclusters with a high likelihood.

FABIA FABIA involves Factor Analysis, which will take gene expression data and attempts to explain it with a smaller set of parameters or factors [50]. The program uses a variation of the Expectation Maximization (EM) algorithm in order to iteratively estimate the noise of the observations, in this case expression values, and the most likely weight of the connections between the observations and the factors, or biclusters. This version of the algorithm used a Laplacian prior in order to enforce sparseness, meaning that weak connections between observations and a bicluster will have weights that quickly drop to zero. Once a good estimate of the parameters is found, the biclusters are ranked based on information content or the weights of the connections found in each bicluster. More connections and higher weights suggest high information content within a bicluster.

Plaid and FABIA are available in R in Bioconductor packages biclust and fabia respectively [50, 92]. SAMBA is an open access Java program available in a package called EXPANDER that can be run with the Windows operating system [93]. The parameters required by each algorithm vary from thresholds set for the number and size of biclusters to the number of iterations for each particular algorithm. The parameters are left at the default values except for the number of biclusters generated, which is the number each algorithm was able to create with the highest tissue separation without causing an error on a 2012; Mac with 3.1 GHz dual core processor and 16 GB of RAM based on initial testing. The biclusters each program generates will be selected for analysis of tissue differentiation and biological significance. Each program will be run 20 times with a different number of set biclusters. The number of biclusters with the highest average tissue separation score (defined below) will be selected for analysis and the run that results in the highest tissue separation score will be selected to compare across the biclustering methods.

6.1.2 Evaluation Metrics for Plaid, SAMBA and FABIA

Plaid, SAMBA and FABIA will undergo preliminary evaluations to determine how well each tissue could be identified based on gene expression patterns as well as how distinct both the groupings of functional annotations and the transcription factors are between each bicluster.

Tissue type differentiation A biclustering metric was implemented, described in the recent evaluation performed in [47], in order to determine how well distinct expression patterns in a tissue were grouped together. This metric was used as an indication of how well each algorithm was able to identify each tissue type correctly, which required the tissue type replicates that were present in each bicluster. The tissue replicate names present in each of the biclusters were extracted and the level of overlap was calculated between each bicluster and a list containing all the replicates of each tissue type. The formula is as follows.

$$f(\text{bicluster, tissue}) = 2\frac{\text{Tissues in bicluster} \cap \text{Total replicates tissue}}{\text{Total number of tissues in both lists}}$$

This should give a result of 1 if the tissue types in a bicluster all match the three replicates from a single tissue with no extra samples. The quality measurement was calculated using this matrix by finding the maximum value in the matrix, saving it in a vector and deleting the row and column it was present in. This procedure continued until the matrix was empty and produced a vector of maximum values. The overall quality score is the mean of all the values in this vector.

Gene Ontology Enrichment Analysis The biological relevance of clustering using actual data can also be inferred from GO enrichment analysis, which is one of the most widely used gene-based benchmarks for biclustering methods [94]. This benchmark provides an estimate of the quality of the biclusters by assessing the genes contained in each. It indicates how significantly the sets of genes discovered by a biclustering method are enriched with a similar GO category provided by the Gene Ontology Consortium. Genes are assigned to bins of GO terms, which can be as general as "biological process" to more specific terms such as "apoptosis" or a location based on functional characteristics. Not all genes are annotated with specific terms as their functional characteristics may still be unknown, but it can provide an indication of what types of functional roles these genes may play by reexamining the other genes with which they are grouped together.

GO enrichment analysis will be performed using a web-based program FuncAssociate 2.1, which reports GO terms that appear more frequently than would be expected by chance when examining the set of terms annotated to the input genes [95]. The program has up to date associations available from mouse, downloaded from the Gene Ontology Consortium with 14633 associations available to the 9132 genes clustered in this dataset. In this program, a Fisher's exact test is used to estimate a p-value describing the probability of a term being equally or more frequently observed in another group of genes in the background set. In order to ensure results from this analysis were statistically significant, the genes chosen for the background comparison set includes all of the genes in the RNA-seq dataset and not all genes that could possibly be observed in mouse. If the background set were to contain all genes in mouse, then the significance would be artificially increased for groups of genes associated with skeletal cell development, possibly even for biclusters containing two or three of these genes. With microarray studies, having a background that includes all genes in the genome may increase the number of enriched terms as the microarray dataset is limited to specific probes or genes. With RNA-seq, there is the potential that any gene could be picked up as expressed, so there may be an argument there to keep all genes in the background set, but limiting to just what is expressed means a researcher can be more confident in the enriched terms found. The p-value will be adjusted using 1000 re-samplings of these genes and a p-value cut-off of 0.05 for every GO term. This method can also contribute to the discovery of other pathways of interest depending on the process in which the bulk of genes found in each bicluster are known to be involved.

6.2 Performance Evaluation of GRN Prediction Methods in Mouse

Selected machine learning algorithms using random forest, biclustering techniques and correlation-based methods will also be compared in their ability to retrieve true positive interactions from a complex mammalian GRN using RNA-seq data with varying sample sizes, in comparison to microarrays. There will also be some exploration as to whether the addition of ChIP-seq could improve prediction for parts of the networks. Biclustering will then be applied as a means of feature selection to RNA-seq datasets from skeletal tissue and ChIP-seq datasets for the main transcription factors Sox9 and Runx2 proposed to be the genes driving expression throughout the rest of the GRNs in cartilage and bone respectively. The GRN selected for evaluations using different numbers of samples was the embryonic stem cell (ESC) network with ChIP-seq from two of the main transcription factors characterizing this cell type.

6.2.1 Naïve Embryonic Stem Cell (ESC) Gene Regulatory Network

Currently, no gold standard GRN is available for complex organisms including mammals such as mouse [77]. As such, it remains difficult to evaluate GRN prediction methods for complex organisms. However, cases of well-described networks such as pathways to control pigmentation, tooth, eye and heart development are described [77, 96, 97]. Another commonly studied GRN used for testing GRN prediction methods is the ESC self-renewal and pluripotency network [32, 98, 99]. Mouse ESCs are pluripotent cells derived from the inner cell mass of early blastocysts. They can be maintained *in vitro* for extended periods without loss of their capacity to contribute to all cell lineages when re-implanted back into a blastocyst [100].

The literature-based stem-cell network is a regulatory network extracted from low-throughput studies reported in the stem-cell literature [101]. The network is created by combining data from 271 publications, and it contains cell-signaling and gene-regulatory links that can be direct or indirect. The networks have been updated in the Embryonic Stem Cells Atlas of Pluripotency Evidence (ESCAPE), but have not been used in this case as ESCAPE uses ChIP-seq and RNA-seq/microarray samples and learn the predicted networks

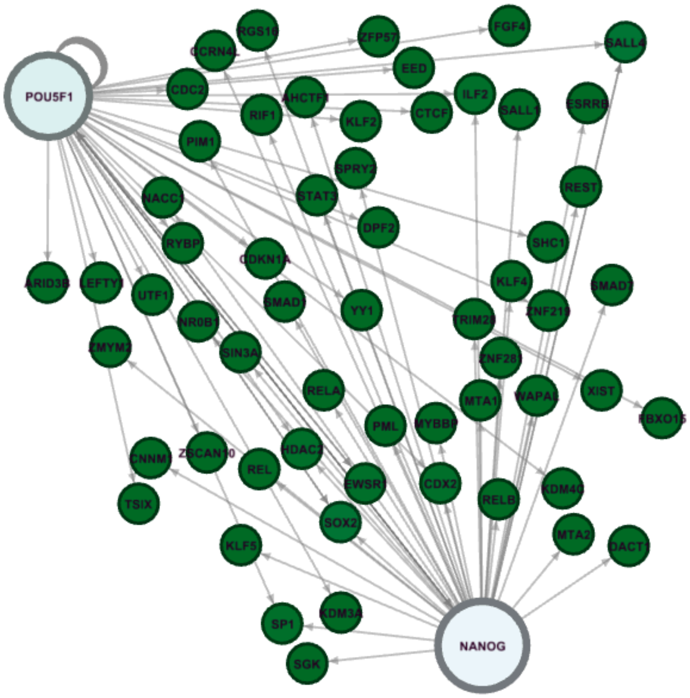


Figure 6.1: Embryonic stem cell transcription factors Pou5f1 (Oct4) and Nanog direct interactions for comparison to ChIP-seq interactions and integration.

to include more interactions. The microarray data used to predict the ESCAPE interactions will also be used for evaluations in this thesis where ESCAPE does not use the RNA-seq being evaluated. Therefore, the more gene expression or ChIP-seq samples similar to those used to predict the ESCAPE networks, the more "accurate" the network may be, which may cause the microarray data to give much better predictions. There are 146 genes, 249 unique interactions in the list and of these, 97 are transcription factor binding interactions that could theoretically be identified using ChIP-seq. 62 of these interactions should be possible to infer from the RNA-seq available alone, and more if ChIP-seq is also considered. For example, only 5 TF-binding events can be inferred for Pou5f1 using the RNA-seq data as the other genes were not in the data. However, there are a total of 17 TF-binding events reported for this gene in the known literature interactions, more of which are contained within the ChIP-seq data.

GRN prediction methods have been evaluated using high-throughput data from ESC [32, 98]. However, the datasets currently used for evaluation are microarray time-course, silencing or overexpression studies although there are several studies available for RNA-seq data with a large enough sample size to attempt GRN prediction tests. Additionally, many data samples are available for single cell RNA-seq, which were not selected to evaluate these methods due to the noise inherent in these datasets, but are an option if hundreds of samples are required for GRN prediction. They require different normalization methods and more samples to obtain accurate gene counts and therefore any conclusion made using this data may not be applicable to our own RNA-seq data if these datasets have a different level of accuracy.

In order to compare to results obtained using microarray datasets, two experiments were selected from Array Express E-MTAB-3234 and E-MTAB-2830, containing 48 samples and 30 samples respectively, to give a total of 78 samples to use for testing GRN prediction methods for consistency and accuracy based on sample size used to predict the GRNs. It should be noted that if single-cell RNA-seq is not used to generate samples, experiments with a large number of samples are usually from *in vitro* studies, which are artificial by nature, as opposed to *in vivo* as it is much more difficult. It can be time consuming, and possibly even impossible depending on the species, to generate a large number of samples. This number was determined to be appropriate based on other simulated measures of sample size generally being between 60-70 samples that allow for performance above random [76]. The researchers originally collected these samples with different miR-142 levels due to Cas9 silencing in embryonic stem cells starting from before differentiation and then through the process of differentiating to an endoderm precursor. Pluripotency-associated genes are in charge of regulating this specification of cells [102]. Therefore, it is a combination of time-series and perturbation data.

Besides forming the regulatory circuit, the three core transcription factors Oct4, Nanog and Sox2 contribute to the hallmark characteristics of ESCs by activation of target genes that encode pluripotency and self-renewal mechanisms and repression of signalling pathways that promote differentiation [102]. Focusing on Oct4 and Nanog, two of the commonly studied transcription factors, will be done to determine the effect on the specificity of this part of the network, and to determine if RNA-seq is able to contribute to information about this part of the network as well as being used to predict interactions without direct influence of these transcription factors. The silencing of Oct4 makes it impossible for cultured embryos to form stable cell lines. Both of these transcription factors are necessary to maintain pluripotency [103]. The other important transcription factors are not included in evaluations due to lack of good quality samples. Furthermore, it mirrors the focus of Sox9 and Runx2 in skeletal tissue. It is highly unlikely these are the only two transcription factors that are important for skeletal tissue development, but these are the only two being utilized for GRN prediction. In most cases of GRN discovery, knowledge of all important transcription factors might not be known. We would like to determine if the top performing algorithm outlined in the DREAM5 project, GENIE3, is improved using ChIP-seq data provided that not all transcription factors in the data have ChIP-seq data available. Therefore, using the output of these programs, the objective is to determine if using ChIP-seq improves the true positive rates of these methods, where true positives are the predicted interactions currently known to occur. It was also necessary to determine which method predicts the most true positives and the least false positives with a network of equal size to the literature ESC network. If no method is able to predict interactions correctly for an already reduced subset of 126 genes, then there is not much hope of current methods to predict accurate GRNs for complex organisms where little may be known about what genes are involved in these networks. Open source programs available with the potential to incorporate ChIP-seq data in different styles were selected for testing against all of GENIE3, Pearson and Spearman correlation. One biclustering method specific to GRN prediction will be tested as well as a random forest method extended from GENIE3, but both are able to make use of RNA-seq and ChIP-seq data without integration of other data types. cMonkey2 has the potential to be compared to the other biclustering algorithms evaluated in Chapter 7, which currently are unable to utilize ChIP-seq data within the biclustering process.

6.3 Random Forest

Random forest is an ensemble method for classification, or regression in this case, where weaker models are combined to create a stronger model [104]. Assume the number of samples in the training set is N. A randomly selected subset of N samples with replacement is used for training to grow a decision tree. Many of these decision trees are made to model a response variable, each based on a randomly selected subset. What decides how the samples are split into children nodes in the tree is chosen randomly from a set of predictors that are available to select based on their ability to decrease the nodes impurity. Individual decision trees tend to overfit the data, so averaging over multiple decision trees is done.

6.3.1 GENIE3

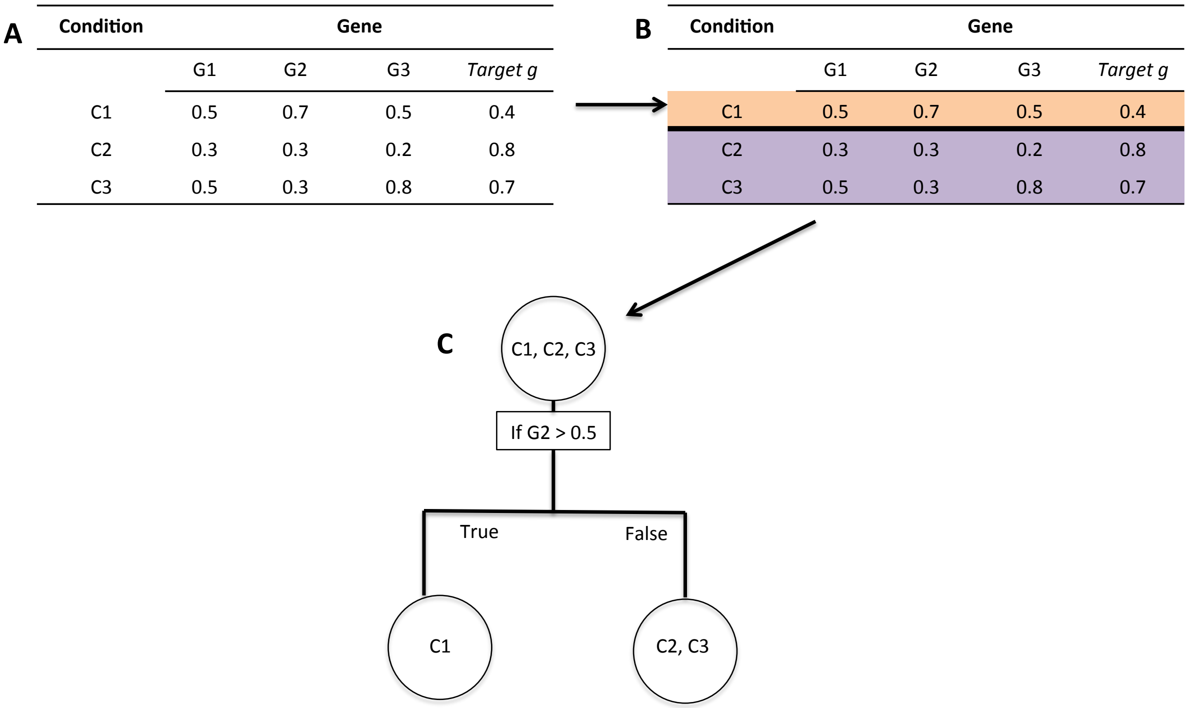
GENIE3 first generates a sample where the expression profile of gene j is the output and the expression of all other genes in the sample is the input. Genes that are strong predictors for gene j expression profiles are considered the genes regulators. A decision tree is constructed for each gene j 1000 times (using bootstrapped samples) where the root of the tree contains all observations which are split into subsets that are more similar than those in the parent node, which is shown in Figure 6.2. These trees are averaged in order to get the most likely genes regulating gene j. The importance score (IM) is described as the total decrease in node impurity due to the splitting based on gene j. If D is a node in a tree, the IM is calculated as follows:

IM(D) = Samples(D) * Var(genes_D) - #Samples(D_leftchild) * Var(genes_leftchild) - #Samples(D_rightchild) * Var(genes_rightchild) .

The importance measure in higher when there is a high number of nodes with low variance in the parent compare to its children once the genes are split again [106]. Using gene k as a predictor for gene j: IM(g_k) is equal to the sum of all nodes with split on g_k importance measure divided by the total number of trees.

6.4 ChIP-seq Data Integration

Integration of ChIP-seq will be done using ChIP-seq available for Oct4 (Pou5f1) and Nanog from Whyte et al. as sequence quality scores after trimming were high without removing a large portion of the sequences for mapping [107]. Alignment and normalization of the RNA-seq data from ESC was performed as described



A

Condition	Gene			
	G1	G2	G3	*Target g*
C1	0.5	0.7	0.5	0.4
C2	0.3	0.3	0.2	0.8
C3	0.5	0.3	0.8	0.7

B

Condition	Gene			
	G1	G2	G3	*Target g*
C1	0.5	0.7	0.5	0.4
C2	0.3	0.3	0.2	0.8
C3	0.5	0.3	0.8	0.7

Figure 6.2: Diagram of random forest method used by GENIE3. Part A shows the target gene g expression ratios as well as three other genes potentially influencing g. There are three experiments, or samples. In order to determine how to split the data to begin constructing a tree, GENIE3 attempts to minimize the variance of gene g expression values, which should group the samples into groups that show similar responses for gene g (shown in Part B). A visualization of the tree after the first split is shown in Part C. In this case, a threshold of 0.5 was selected to split the samples into two groups. This pattern continues to create more splits in the tree until no more splits can be made [105].

in Chapter 7, Section 4.2.1. ChIPseeker 1.8.3 in R will be used to annotate the peaks found with Model-based Analysis of ChIP-Seq (MACS) 1.4.2 and only peaks found within 3000 base pairs of a transcription start site (TSS) of a gene were kept for integration purposes with the methods below. In this thesis, the main GRN prediction methods that are focussed on are correlation-based methods, random forest and biclustering.

6.4.1 iRafnet

One limitation of GENIE3 is that it is unable to incorporate data other than microarray and RNA-seq in order to make interaction predictions. Recently, iRafnet adapted GENIE3 with the potential to adapt and incorporate other sources of heterogeneous data [108]. The method uses a weighted sampling strategy where the gene expression data is considered the main input data to make inferences using the random forest technique as with GENIE3, but also utilizes other data, such as protein-protein interactions (PPI), knockout or ChIP-seq data to derive prior information before incorporating the gene expression data. This prior is an indication of how likely an interaction occurs between two genes. At each node of the random forests that are generated to model the expression value of gene g as a function of potential regulators, a random set of data is selected and N potential regulators are sampled according to the prior information, or weights, for that data.

The authors claim that this integrative method can be adapted for information including transcription factor (TF)-DNA-binding, which can be obtained from ChIP-seq [108]. However, it has not been implemented to work with the program to generate an appropriate weight matrix to calculate weights, nor was a description to generate the weight matrix for this particular data included in the paper. It is also possible a researcher may wish to focus on increasing accuracy of the network for particular transcription factors of interest when they likely do not have access to data on all the predicted transcription factors in the network or the resources to generate the data for all transcription factors. It is also not known if biasing to the random forest algorithm, which averages results across many trees, will pick up a select few transcription factors.

As iRafnet does not describe how to integrate ChIP-seq, an attempt was made here. In order to bias iRafnet to a smaller set of transcription factors, every gene that does not appear in Nanog or Oct4 ChIP-seq data, but has the potential to regulate other genes, are weighted evenly. All genes that do have these transcription factors potentially influencing them gets a different weight, which is dependent on how close the genes are to the binding sites of the transcription factors. A probability is adapted from MACs, which reports a p-value indicating the likelihood a transcription factor is influencing gene expression of a particular gene based on where in the genome it is binding to influence transcription. This allows all genes to have the possibility of being part of the larger network, for future exploratory work or formulating hypotheses, but should give more confidence to the genes regulated by the transcription factors that appear in the ChIP-seq data. A benefit of integrating ChIP-seq in this manner is that every gene can still take part in the network if there is enough evidence from the expression profile that it is a part of it, while focusing on increasing the accuracy of the network for transcription factors of interest. The algorithm used within iRafnet to accomplish this is given in Algorithm 5.1.

Algorithm 6.1: Adaptation of iRafnet step A1 calculations for all target genes

```
r ← number of potential regulators (all transcription factors possibly in the network)
n ← number of samples
imp ← matrix(0,p,p) #p by p matrix to store importance
p ← total number of genes
For(j in 1:p){
        #matrix zj of expression profiles of potential regulators (n by r)
        zj ← expression_matrix[TF_names, ]
        #matrix xj containing expression profile of target gene j
        xj ← expression_matrix[j, ]
        #sampling weights
        sw ← vector(size← nrow(expression_matrix), prior_value)
        sw["Nanog"]← Macs_pval
        sw["Oct4"]← Macs_pval
        #normalize the probabilities
        sw ← sw/sum(sw)
rout ← RF(x= sw(sorted), y= xj, importance= TRUE, mtry= round(sqrt(p)), ntree= 1000, sw= as.
    double(sorted), numsource= 1L)
imp[index, j] ← c(importance(rout)[,2])
#the kth element of this vector will be the importance score placed on gk −> gj
}
```

The program has been adapted in order to include ChIP-seq datasets for as many transcription factors as are available although only two will be used for downstream experimentation. To test iRafnet using ChIP-seq data, different weighting schemes will be applied:

1. using the normalized maximum $-10 * log(\text{p} - value)$, which is the smallest p-value reported for each gene interaction with Oct4 or Nanog,

2. reducing the weight of the interactions reported in the ChIP-seq data,

3. all weights of potential regulators equal (No ChIP-seq data influence).

This is only one potential method to integrate the data. For example, there may also be other strategies integrating the data at different stages of the program or perhaps after the program has run in order to identify key regulators.

6.4.2 cMonkey2

To compare outside biclustering methods to integrated versions specifically for GRN prediction, cMonkey2 is a program available that has its own biclustering method as well as the potential to integrate data from other sources than expression data such as ChIP-seq and PPI networks [19]. Originally, cMonkey did not receive much use in the wider community, perhaps due to it making use of sequence information, which other biclustering algorithms tend not to, thus making it difficult to compare and therefore be established as a good GRN predictor. However, it is now possible to compare results with cMonkey2, which works by enriching clusters for gene sets which are expected to include additional evidence for co-regulation (ie. genes under the same regulatory influence from ChIP-chip/seq) in order to find co-regulated modules. The algorithm works by calculating an enrichment score and, in order to predict the network structure, uses a program called Inferelator (described below) to predict the network.

First, the cMonkey2 pipeline needs to be overridden with a new file including the gene enrichment scoring function. Second, a JSON file is created with groups of genes (gene sets), which is obtained from ChIP-seq data, in this case from the genes potentially regulated by Nanog and those potentially regulated by Oct4 (including the genes themselves in the gene sets). Given these gene lists that possibly overlap, the enrichment scoring function determines the amount of overlap between these genes annotated in each set and the genes in each bicluster for every iteration using Fisher's exact test. The gene set that results with the smallest p-value for each bicluster is used for training, and row scores are generated to increase the probability that genes stay in a bicluster if they are in the enriched set and tries to add more genes from the set if possible. The authors explain the gene scores are computed by a simple heuristic where they multiply the log10 of the p-value by 1.0 for genes which are in the bicluster and are members of the enriched set; by 0.5 for genes which are in the set but are not in the bicluster; and by 0.0 for all other genes.

6.4.3 Inferelator

After biclustering is performed using cMonkey2 to group genes into modules, Inferelator can be used to predict transcription factors that are most likely regulating the genes present in each bicluster. The program uses linear regression LASSO [109]. Since the main focus of this project is to predict interactions of select transcription factors and not necessarily include protein-protein interactions of any gene not also considered a transcription factor, this program was also selected to compare to iRafnet as it may also be used with and without ChIP-seq data. The output of cMonkey2 was modified in order for Inferelator to make predictions, as Inferelator was originally designed for cMonkey and has been minimally updated.

6.5 Evaluation

Testing these programs with ChIP-seq data is an attempt to answer if and how adding samples of RNA-seq with the application of ChIP-seq data allows for improvement of the GRN prediction accuracy. Furthermore, it is desirable to know how much the consistency of the interactions improves during each run of the program. This will determine whether interactions predicted by ChIP-seq alone begin to disappear with the addition of new samples from expression profiles, and at what point this begins to happen.

These methods will be evaluated in terms of the number of true positives compared to false positives to determine how confident a researcher can be in a predicted network for a complex organism. This will be used as opposed to accuracy since with gene regulatory networks, there are a lot of true negative interactions due to the sparse nature of biological networks [76, 110]. A method that makes no prediction will still achieve high accuracy since the number of true negative interactions is large in comparison to true positives, false positives and false negatives. One limitation to the methods is that they make many predictions so without thresholding in some way the number of false positives compared to true positive interactions will be high. For example, GENIE3 will predict interactions multiple times with different importance values so if there are 100 genes to predict interactions there are $(100) * (100 - 1)$ possible interactions, but GENIE3 can produce a result of 100000 interactions or more if no maximum is specified. To compare against the gold standard network the number of predictions will be minimized to 250 in order to compare the 248 "known" interactions. cMonkey2 was selected for generating an initial prediction for the Sox9 and Runx2 networks as a feature selection method to compare to the other evaluated biclustering methods. The resulting network was visualized in Cytoscape [111].

To ensure the RNA-seq data selected is appropriate for evaluations, the same analysis will be done for GENIE3 and Pearson correlation using microarray data that has been used previously to predict the ESC network [98]. Finally, the predicted GRNs from each method will be compared to each other to determine how often these methods are making similar predictions to each other using the same sized GRNs. The methods used to make the comparisons described above are further explained in Chapter 8.

6.6 Microarray and RNA-seq Comparisons

6.6.1 Generation of ROC curves

Rates of true positives, true negatives, false positives and false negatives were calculated as follows:

True positives (TP): True positives are calculated by determining the number of predicted interactions that are in the list of known interactions, which is done by determining overlap between dataframes in R.

False positives (FP): False positives are the number of predicted interaction that are not in the list of known interactions from the literature.

False negatives (FN): False negatives are equal to the number of known interactions that are not in the list of predicted interactions, meaning the program failed to predict this number of interactions.

True negatives (TN): True negatives are calculated by first determining the number of unique genes in the list of interactions and then calculating all possible combinations of these genes not including self-interactions. The number of false positives, true positives and false negatives are subtracted from this number.

Ten predicted GRNs will be generated for each method excluding cMonkey2, which will have five predicted GRNs due to the length of time required to run the program. The true positive rates and false positive rates will be plotted to generate a Receiver Operator Characteristic (ROC) curve. True positive rates and false positive rates will be calculated as follows.

True positive rate: $= \frac{\mathrm{TP}}{(\mathrm{TP+FN})}$

False positive rate: $= 1 - \frac{\mathrm{TN}}{(\mathrm{TN+FP})}$

In order to calculate these values, the number of possible interactions that could be predicted from the ESC literature network is required. Only 126 genes are present in the RNA-seq data so not all 248 interactions from the literature network have the potential to be predicted by any method. This was also the case for microarray data with only 60 genes from the literature network. Therefore, only the interactions that could be produced will be included for comparisons, so that sensitivity (true positive rate) and 1-specificity (false positive rate) could reach 1. ROC curves will be plotted using R and the area under the curve (AUROC) will be calculated using the trapezoidal method in the flux R package. The number of genes used in the RNA-seq dataset will also be minimized to the same set of genes available in the microarray dataset to determine if the number of genes used to predict the GRN in this case changes the performance of GENIE3.

6.7 Measuring Consistency of GRN Prediction Methods

The consistency of each algorithm using RNA-seq data will be determined using the top 250 predicted interactions and determining how many are different on average across all runs of the algorithms. The top 250 interactions were selected for some comparisons for two reasons. The first is that recent evaluations in literature using this network have used this cut-off [98]. Secondly, the importance values begin to plateau after roughly 246 interactions in the random forest methods, where the confidence of interactions does not change as drastically. Therefore, it was assumed that after this point, the consistency of results would change by greater margins since the order of very similar importance values could shuffle. All 78 samples will be used to predict the GRNs to ensure there is no difference in the samples that were used by each algorithm. When comparing an algorithm to itself, each list of predicted interactions will be compared to all other GRNs predicted. When comparing two different algorithms, the same run from each method will be compared to generate an average. The importance measures from GENIE3 will be plotted in R to determine if there is a natural point at which to cut-off the number of possible interactions. Furthermore, using randomly selected subsets of RNA-seq samples, the average number of differences will be calculated between 10 runs of GENIE3 in order to determine if more samples correlates with a decrease in the number of differences between two predicted GRNs. In order to investigate the consistency of results produced by GENIE3 depending on the number of samples used, the 78 samples will first be split into distinct subsets of equal size. Six GRNs will be predicted using sample sizes from 6 to 13 since a maximum of six GRNs can be made using 13 distinct samples. Secondly, the number of overlapping interactions will be plotted for each sample size. Results of these comparisons are presented in Section 8.3.

Chapter 7

Evaluation of Biclustering Methods using RNA-seq Data from Skeletal Tissue

Performance analysis of current biclustering algorithms was recently conducted on microarray data [47, 49]. The first group of researchers measured the performance of 12 biclustering algorithms by evaluating each bicluster on artificial datasets generated from six different models as well as evaluating the genes of biclusters discovered in expression data of rat peripheral and brain regions. The second and most recent study focused on the ability of 15 biclustering or clustering methods to distinguish various sample types rather than their performance in discovering various bicluster patterns in the data. It was found that the groups of genes discovered by CTWC, FABIA, ISA, Plaid, SAMBA and hierarchical clustering were enriched with GO terms and performed acceptably for both distinct tissues and breast tumours. Furthermore, CTWC, Plaid, SAMBA, hierarchical clustering, constant MSBE and FABIA methods best distinguished the sample-types in the expression matrix containing multiple tissues. Overall, Plaid was found to be a robust method when tested on the five heterogeneous tissues used consisting of expression data with bicluster structures with small overlaps on their genes and samples. Plaid was also found to work well with the rat peripheral and brain regions as well as the multi-tissue samples studied by Hochreiter et al., who proposed FABIA as a biclustering algorithm [50]. FABIA uses a similarity measurement in combination with the Munkres algorithm to estimate the sample differentiation and when it was compared to Plaid it was out-performed when handling multiple tissues, but the best option when handling tumours from breast tissue alone. Due to the results of this most recent paper correlating to the performance analysis of other evaluations described, biclustering method evaluation in this thesis has been narrowed down to SAMBA, Plaid and FABIA. These three algorithms were selected for evaluations using skeletal tissue because of their accessibility as well as to test algorithms that have differences in performance rank when handling different tissue samples in previous studies. If skeletal tissues have gene expression typical of tissue subtypes then FABIA would be expected to marginally outperform Plaid and SAMBA due to high tissue gene expression similarities. If gene expression were distinct enough between the skeletal tissues, Plaid or SAMBA would be predicted to outperform FABIA. An appropriate method should not necessarily separate all the tissues, but be able to identify patterns unique or similar across the skeletal tissues. Each algorithm tested was also applicable to RNA-seq data though the previous studies tested their ability to handle microarray results. Therefore, in this chapter, we discuss the results of comparing these three biclustering methods using skeletal tissues to determine if any of the methods are able to produce potentially biologically relevant results.

7.1 Results

Figure 7.1 shows the results of tissue sample differentiation. FABIA outperformed SAMBA according to the tissue separation metric, with an average separation of 90% in the best run, but was unable to group all three replicates of immature cartilage and bone as distinct tissue types. Plaid was able to distinguish between all tissue types in at least one bicluster. This means there were three separate biclusters each containing all three replicates of one tissue to give a tissue separation score of 100%. Mature and immature cartilage were grouped together in the remaining biclusters with only one bicluster containing bone, which did not share an expression pattern across genes in the bicluster with other tissues. Biclusters 4 and 6 only contained select replicates from mature and immature cartilage. SAMBA produced biclusters with an average tissue differentiation of 63% and discovered more localized expression patterns across two or all three tissue types. There were fewer genes contained in each bicluster than those found using Plaid and FABIA.

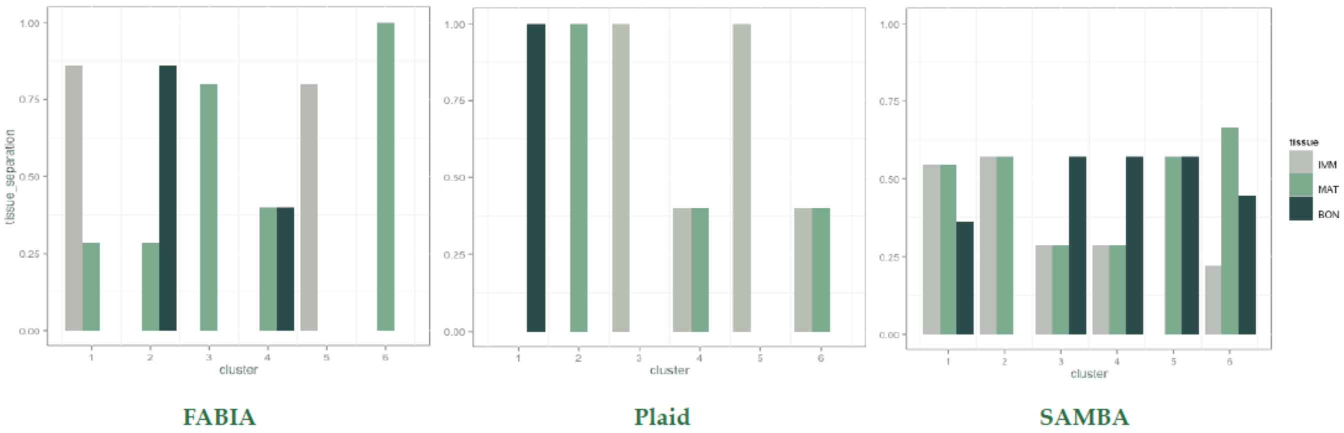


Figure 7.1: Results of tissue differentiation analysis for Plaid, FABIA and SAMBA biclustering algorithms. Plaid was able to detect local gene expression patterns distinct to each tissue (IMM=Immature cartilage, MAT=Mature cartilage, BON=Bone). FABIA was able to distinguish all mature cartilage replicates while SAMBA was unable to discover any localized expression patterns unique to a single tissue.

There was significant enrichment observed in all six biclusters produced by FABIA containing terms particular to bone and cartilage development similar to terms found in biclusters from Plaid and SAMBA. Two biclusters contained terms associated with wound healing such as coagulation, platelet derived growth factor binding and other blood related terms. One bicluster from each of the other two methods also produced terms of this nature mixed with other, more general terms, including wound healing and coagulation. Plaid also produced enriched biclusters although the bicluster containing only mature cartilage produced no enriched terms and many terms that were enriched were not specific to skeletal tissue. Figure 7.2 shows a comparison between the number of enriched terms in biclusters produced by Plaid and FABIA. The complete tables of GO terms for all three methods can be found in Appendix 2.

Plaid was able to produce multiple biclusters containing Runx2 (bicluster 1 and 4), but no cluster contained Sox9 in the run selected with a tissue separation score of 100%. Runx2 was present in biclusters separating bone from both other tissues. Sox9 was sometimes, but not always, present in a bicluster. Sox9 and Runx2 are required in biclusters. This is because in order to make a predicted GRN with these transcription factors as the main drivers of genes in the GRN, they and genes sharing similar expression patterns are required to make predictions for interactions involving Sox9 or Runx2. Therefore, since Plaid does not consistently produce at least one bicluster containing Sox9, the program cannot be used for feature selection. Runx2 was present in two biclusters using FABIA (bicluster 1 and 5) with immature or mature cartilage present in each bicluster. These biclusters could contain genes that are located in both networks driven by these transcription factors that share activity in mature cartilage. FABIA also produced distinct biclusters that contained Sox9 (bicluster 4) and Runx2 (bicluster 1). Both of these clusters were annotated with terms for cartilage and bone development respectively. Sox9 did not appear in any of the biclusters found using SAMBA including those annotated with cartilage development terms. Runx2 was contained in only one bicluster (bicluster 4) including terms associated with bone development in the presence of cartilage tissue including “endochondral ossification”, “replacement ossification” and “biomineral tissue development”. However, all three replicates of bone were not present in this cluster, and they contained replicates of both immature and mature cartilage.

All of the biclusters produced by SAMBA produce significantly enriched terms associated with bone and cartilage tissue. The GO terms enriched within SAMBA biclusters were occasionally more specific then Plaid and FABIA resulting in terms such as “collagen type IX” and “FACIT” collagen - which includes collagen types IX, XII, XIV, XIX and XXI [112] - as well as more general terms such as “limb morphogenesis”. The expression patterns of the genes within the collagen associated biclusters show up-regulation in the clustered mature cartilage tissue. There were, however, other terms consistently present in biclusters using all techniques containing bone tissue including those associated with cell migration, motility and locomotion. The bone samples in the RNA-seq dataset were from neural crest cells, which are migratory cells, which explains these terms. Cartilage samples were from the limb, which is a possible explanation for terms like “limb morphogenesis”.

7.2 Discussion

Since not all replicates for a tissue were present in a single bicluster for some cases across all three methods, this suggests that there is a chance that these patterns are not generalizable for all samples of these tissues and may be due to differences between the biological replicates. Also, more general terms appeared interchangeably between biclusters containing bone and cartilage separately and may not be the best indication of making distinction between the developmental processes. However, the biclusters annotated with GO terms related to wound healing are potentially important for bone and cartilage development. One example would be terms annotated with platelet derived growth factor binding. Not only important for wound healing, this is a potent activator of cells with a mesenchymal origin and differentiation of these cells potentially results in skeletal tissue formation [113]. Vasculature remodelling is also characteristic of bone formation [5]. So perhaps FABIA is more sensitive at picking up particular patterns involving these processes.

If mature cartilage does have expression that is similar to bone or immature cartilage as well as gene expression that is in between the other tissues in the biclusters, it may not be the case that all the tissues would be separated, or separating the tissues into distinct biclusters would not be useful. This is because in these cases, a pattern should be seen in large portions of the genes where expression in immature cartilage is high when expression in bone is low (or vice versa) and expression in mature cartilage is somewhere between. Therefore, it would be expected that biclusters would usually contain at least two tissues. It appears that SAMBA is best at selecting patterns that are observed across all three tissues. If SAMBA had been able to identify Sox9 in at least one bicluster, it would have also been a viable means of feature selection to compare to cMonkey2 in Section 9.1. However, perhaps this suggests that there are other transcription factors playing a important role in skeletal tissues. Since there has to be something controlling expression of Sox9 in mature cartilage and bone to keep Sox9 down-regulated, perhaps some of these transcription factors are present in the biclusters produced by SAMBA or Plaid. FABIA and Plaid, however, do separate at least one of the tissues. Plaid can separate all three, which seems unlikely to be a desirable outcome if gene expression is behaving as it was described above. Using Plaid, Runx2 was grouped in a bicluster separating bone from both other tissues, which would be expected as expression of Runx2 facilitates the development of bone from undifferentiated mesenchymal cells. However, if Runx2 plays a role in mature cartilage formation it would also be expected to appear in a bicluster containing mature cartilage tissue, which is in bicluster 4 along with immature cartilage. This may suggest that Plaid is not useful for biclustering these tissues without minimizing the number of genes prior to biclustering, potentially minimizing noise in gene expression Plaid may be sensitive to. Both methods are able to separate mature cartilage. If the Sox9 and Runx2 GRNs were truly additive, where the Runx2 network formed in the presence of the Sox9 network to create a mixture of gene expression between both networks, then mature cartilage having distinct gene expression from immature cartilage, in particular, would not be expected. The alternative hypothesis, that the Runx2 network in bone is completely separate from the Sox9 network, could explain if mature cartilage had unique expression compared to bone. However, this does not explain the unique expression compared to gene expression observed in immature cartilage. The genes found by FABIA in the bicluster only containing mature cartilage includes genes with high expression in mature cartilage compared to the other two tissues, such as *Col10a1*, which were grouped using model-based clustering as well. Therefore, FABIA separating mature cartilage from the other tissues, in this case, is appropriate. It will be of interest to further explore the annotations unique to biclusters containing a single tissue type to make further judgement of biclustering algorithm performance.

From the model-based clustering presented in Section 5.2, a large portion of gene expression in mature cartilage falls somewhere between immature and bone. It also shows a comparable number of instances where expression in mature cartilage is more similar to bone or more similar to immature cartilage gene expression as opposed to an average between the two. There was also one cluster where mature cartilage showed a group of highly expressed genes not indicative of an additive GRN, and appeared to be the combination of the GRNs driving development, when interacting, producing synergistic changes to gene expression. Therefore, since mature cartilage gene expression is not always similar to immature cartilage or bone, it has unique gene expression patterns, which could be an explanation for Plaid and FABIA being able to separate these tissue types. Another possible reason for this separation could be the variation of the biological replicates. From the PCA in Section 4.1, the variation in the samples of mature cartilage is higher than the other tissues. This could indicate the pattern is more easily identified in mature cartilage. The biclustering algorithms may be able to pick up on this more overt variation when it shows a conserved pattern across mature cartilage while the gene expression does not share the same pattern in the other two tissues. Therefore, the separation of mature cartilage samples can be explained and does not necessarily mean these biclustering methods perform poorly.

7.3 Conclusion

RNA-seq data from cartilage and bone tissue was used to evaluate the performance of three biclustering algorithms and their ability to separate tissue types as well as molecular processes enriched in each bicluster. Based on these metrics, the Plaid biclustering algorithm was able to separate tissue types, but was unable to produce clusters with terms enriched for either cartilage or bone development. It also produces larger biclusters than the other two techniques. The larger bicluster size may explain why there are no significant terms enriched as Plaid may be more sensitive to noise. Therefore, although it has the potential to separate tissue types, there are more genes used to separate the tissues that may either not be well described using GO terms. Furthermore, it did not always produce a bicluster containing Sox9 or Runx2 at least once suggesting the method could be sensitive to noise in real biological data if the number of genes has not been minimized by another method beforehand, so this method alone is likely inappropriate. FABIA was able to find multiple biclusters with local gene expression patterns that contained transcription factors Sox9 and Runx2. FABIA was unable to separate the tissues completely although it was able to separate mature cartilage, which shares a lot of similar gene expression with either immature cartilage or bone. This may be due to gene expression unique to mature cartilage, which also provides evidence that the Sox9 and Runx2 GRNs are not completely additive. Therefore, the results from FABIA will be used to construct preliminary networks for Sox9 and Runx2. It is also important to note that cMonkey2 has not been compared to these methods, so the biclustering results presented will also be compared to cMonkey2 as it was found to perform more proficiently than other popular methods on GRN prediction.

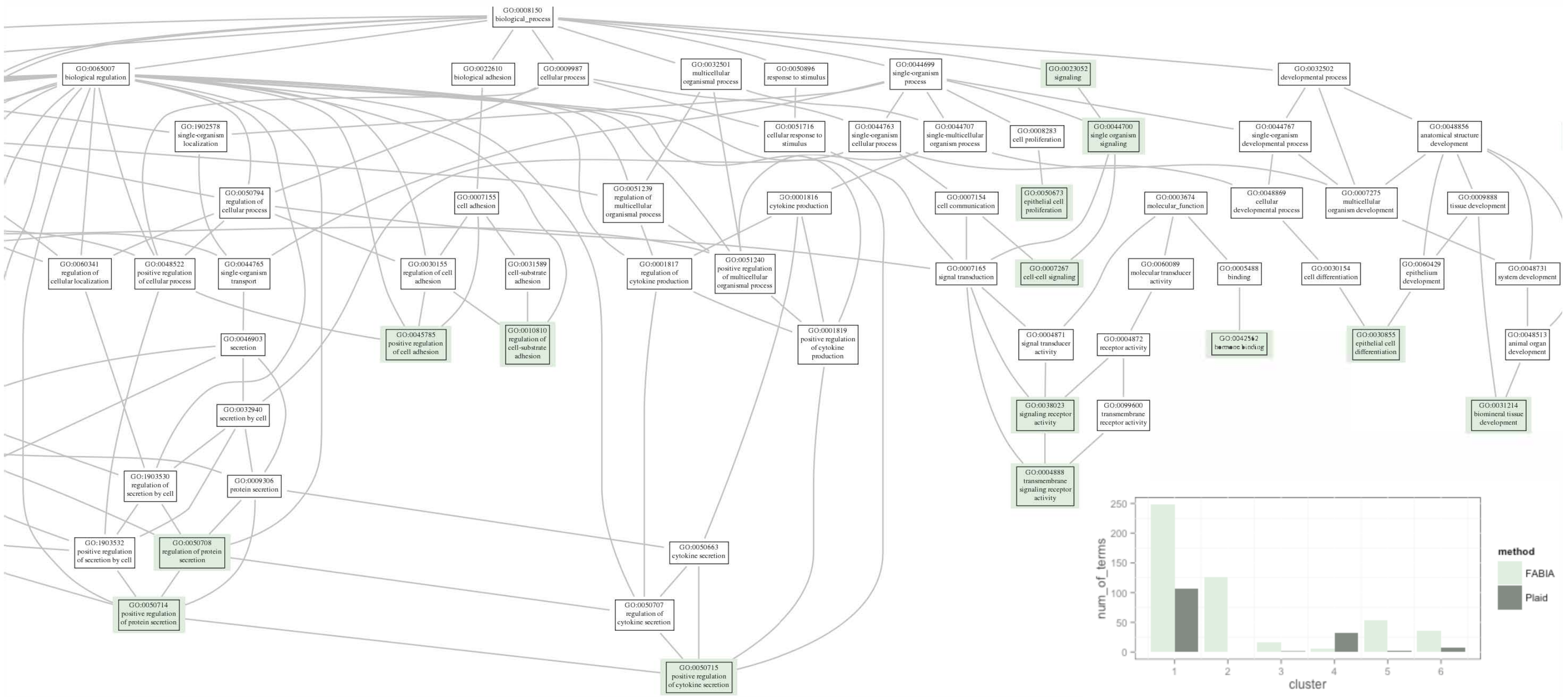


Figure 7.2: Gene Ontology terms shown for FABIA bicluster containing only mature cartilage samples. The number of enriched terms are compared between Plaid and FABIA. In clusters separating mature cartilage, Plaid did not have any enriched terms while FABIA had terms such as "biomineralization".

Chapter 8

Comparisons of GRN Prediction Method Performance

ESC RNA-seq data has not been previously used to perform an evaluation of GRN prediction performance as opposed to microarray data, which has been done in [98]. Therefore, it was necessary to determine if RNA-seq data would behave similarly to microarray data when used to predict GRNs. In the case of the microarray data with 60 genes, there is a potential for 3540 ($60 * 59$) interactions to create a complete network. This is what was used previously to measure GRN inference performance although not with the integration of ChIP-seq [98]. Once duplicate genes were removed from the array, there was a total of 8127 genes left on the array with 60 genes in total from the ESC network in the literature, so the number of genes had to be minimized to these 60 in order to compare to the literature network. When GENIE3 is run with 126 genes, the maximum number of interactions is 15750, which is the number of interactions possible given each of the 126 genes potentially interacting with 125 other genes. This program does not take into consideration that some of the interactions could be self-regulated and so the sensitivity of this algorithm never reaches 100% with the mouse literature network. Then, to generate ROC curves, these interactions were removed when comparing to the GRNs predicted by these programs.

8.1 Microarray and RNA-seq Comparisons

GRN prediction from RNA-seq and microarray was performed. 78 samples were used from each data set selected randomly for each run of GENIE3. A comparison between the first 250 predicted interactions using 126 genes from the RNA-seq dataset and 60 genes from the microarray dataset were used. The number of true positive interactions seems comparable, since the AUROC was 0.818 and 0.816 using the microarray and RNA-seq data respectively, although the microarray data contained 90 samples versus 78 RNA-seq samples, and it only predicted interactions for 60 genes instead of 126. The AUROC was calculated again after the RNA-seq dataset was minimized to the genes intersecting with the genes in the microarray data leaving 58 genes total, to confirm that the RNA-seq data produced similar AUROC values to ESC microarray data. This was done since the RNA-seq datasets have not been used specifically for GRN prediction before. This reduced performance slightly with a AUROC of 0.798. Increasing the number of samples that the microarray could use to all 90 of the samples available did not change the AUROC, which remained at 0.818, shown in Figure 8.1.

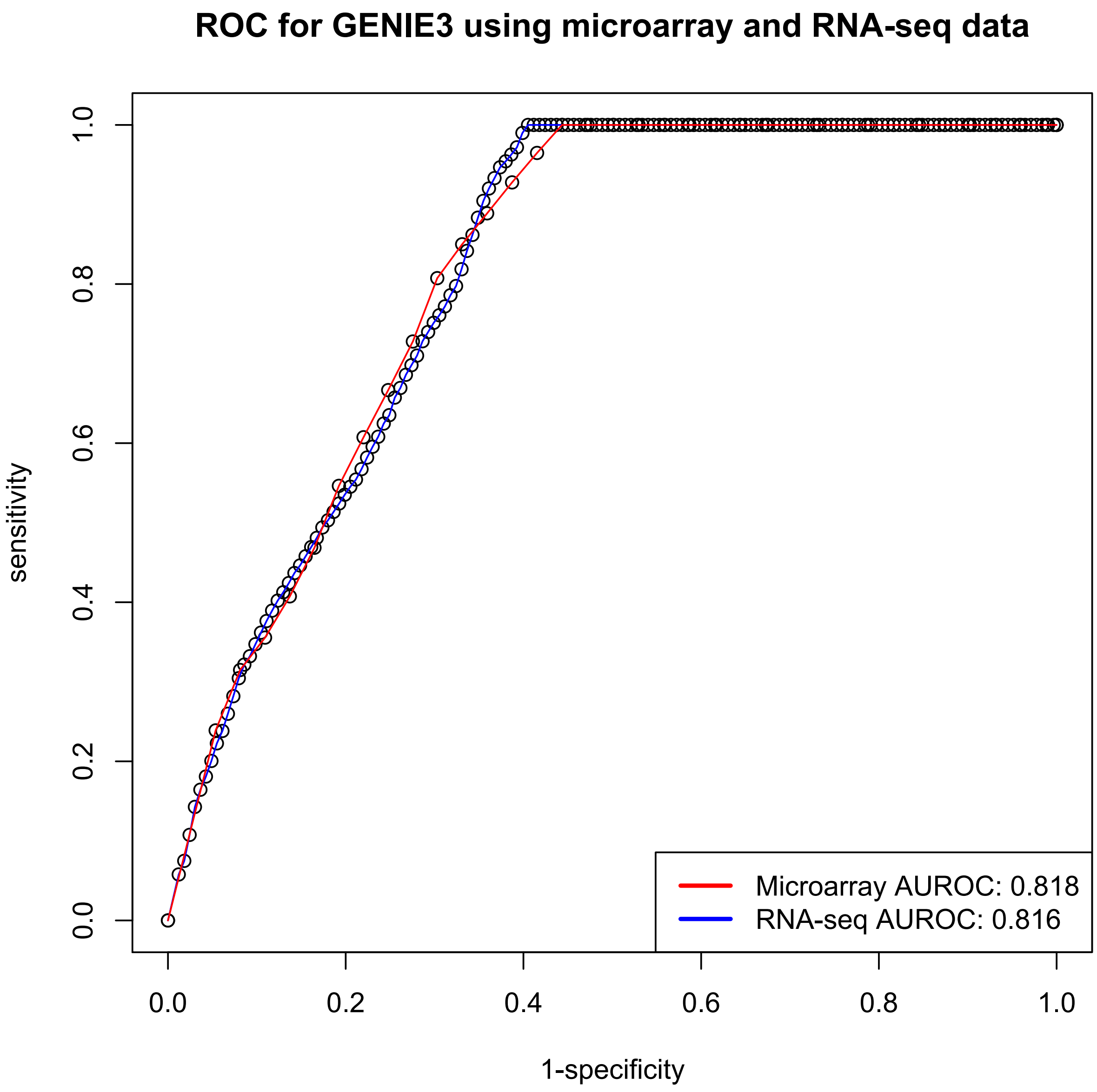


Figure 8.1: Number of true positives retrieved by GENIE3 with different numbers of samples used to predict the GRN. The number of TP in the top 250 predictions are shown.

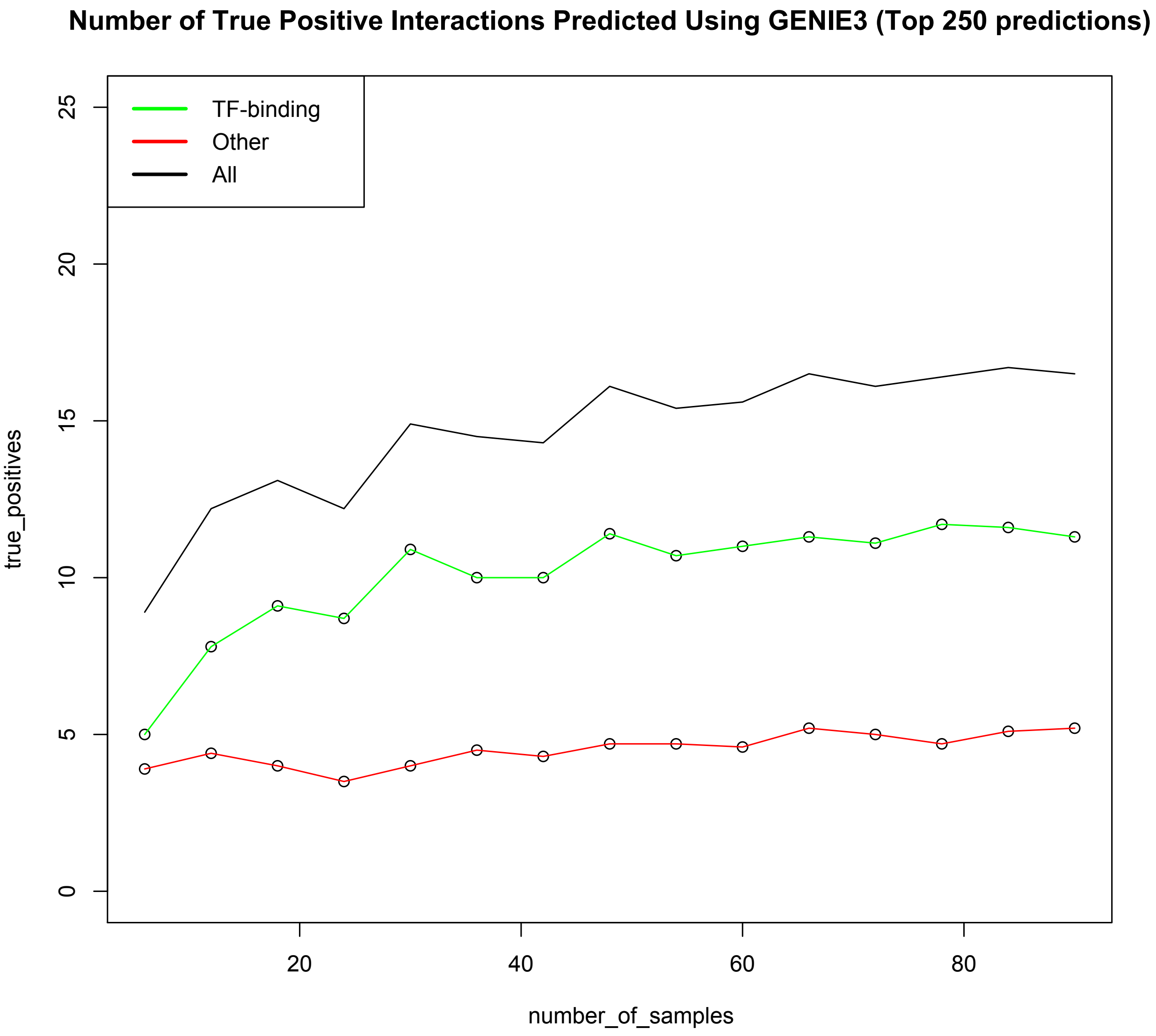


Figure 8.2: True positive results for GENIE3 using microarray data.

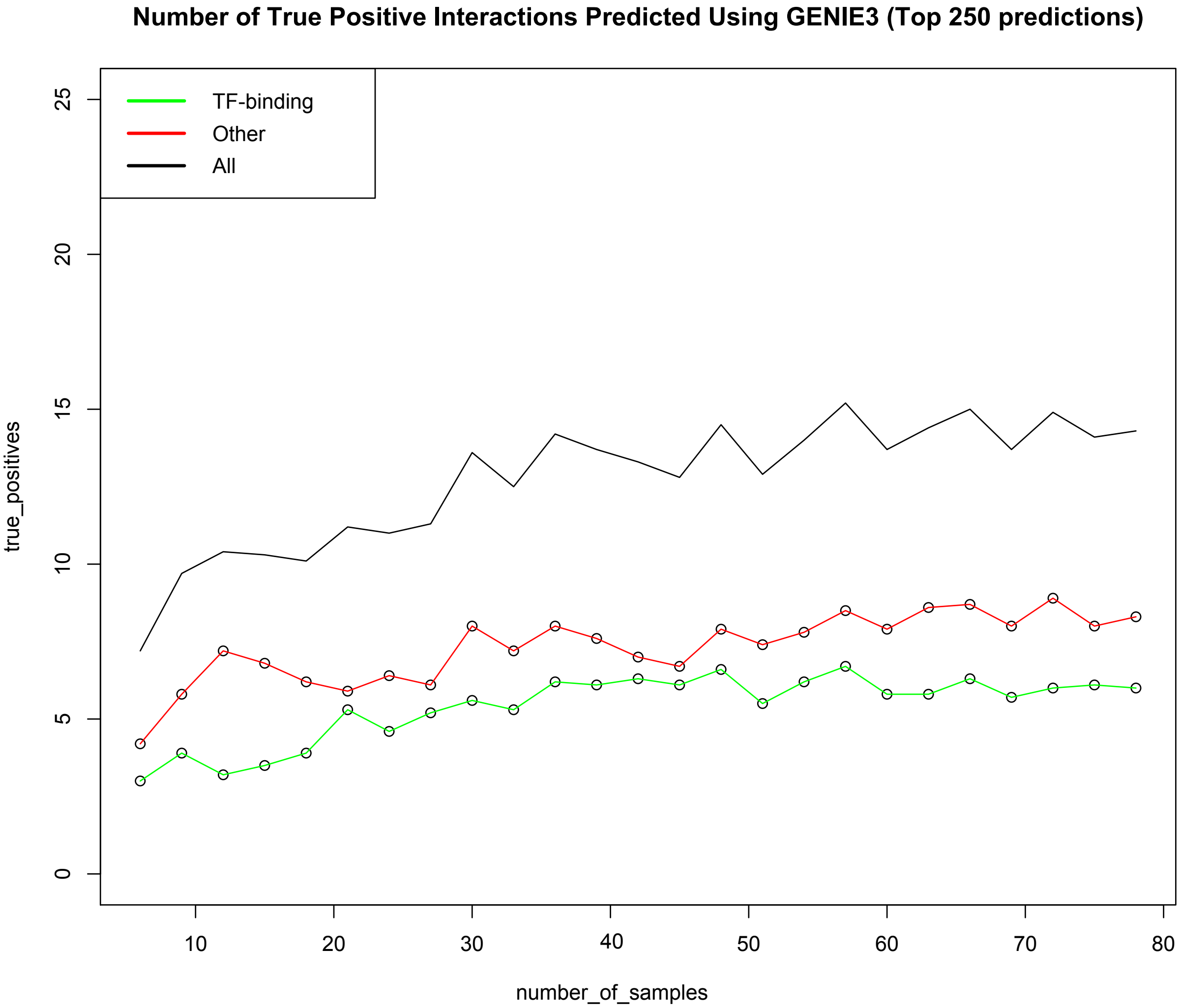


Figure 8.3: Number of true positives retrieved by GENIE3 with different numbers of samples used to predict the GRN. The number of TP in the top 250 predictions are shown.

Figure 8.2 and 8.3 shows the number of true positives predicted in the top 250 interactions using microarray data or RNA-seq resulted in similar numbers overall. However, more predictions for transcription factor binding were discovered using the microarray data compared to the number of other interactions, which was opposite to when the RNA-seq data was used. Using RNA-seq resulted in a lower or equal numbers of predicted transcription factor interactions in almost all the other GRN prediction methods evaluated and presented in Section 8.2.

8.2 Sample Size

When the number of samples was decreased to 6 randomly selected for each run of the methods, GENIE3 achieved a AUROC of 0.801 using the RNA-seq data, which is a drop of 0.015. It is possible that increasing the number of runs would have increased the AUROC even more to make it equivalent to using 78 samples since there are more selections of distinct samples for GRN prediction using only 6 of the 78 samples. When using all 78 samples, there is only one option for GRN prediction, which is including all 78 samples to make a prediction. With 10 runs, a maximum of 60 samples were used to predict each GRN since only six samples are selected randomly for each run. However, it is clear that adding another 72 samples to 6 samples did not increase the performance of the method overall when considering different cut-offs. However, with a cut-off of 250 top interactions, the number of true positives did improve from 6 to 10 samples. The rate of true positives discovered is compared to the number of interactions considered, as the number of false positives remained consistent. In Figure 8.5, with 6 samples, Spearman correlation had a comparable number of true positives to using 78 samples with fluctuation using numbers of samples between these values. iRafnet performed the worst out of all of the methods with a trend that was relatively flat with no improvement from 6 to 78 samples, shown in Figure 8.6. Furthermore, it was not able to detect as many true positive interactions in the top predicted interactions, with few transcription factor interactions making it inappropriate for predicting a GRN controlled by transcription factor activity. Figure 8.4 shows Pearson correlation had more true positive interactions than GENIE3 in the top 250 predicted interactions where GENIE3 predicted 15 true positives on average after reaching 60 samples where Pearson's correlation was able to predict over 20 true positives. Pearson's and Spearman correlation also outperformed GENIE3 when using a smaller number of samples. However, from the AUROC, the performance was better overall than Pearson's and Spearman's correlation with both iRafnet and GENIE3 ultimately able to discover the most true positive interactions compared to the total number of predictions made.

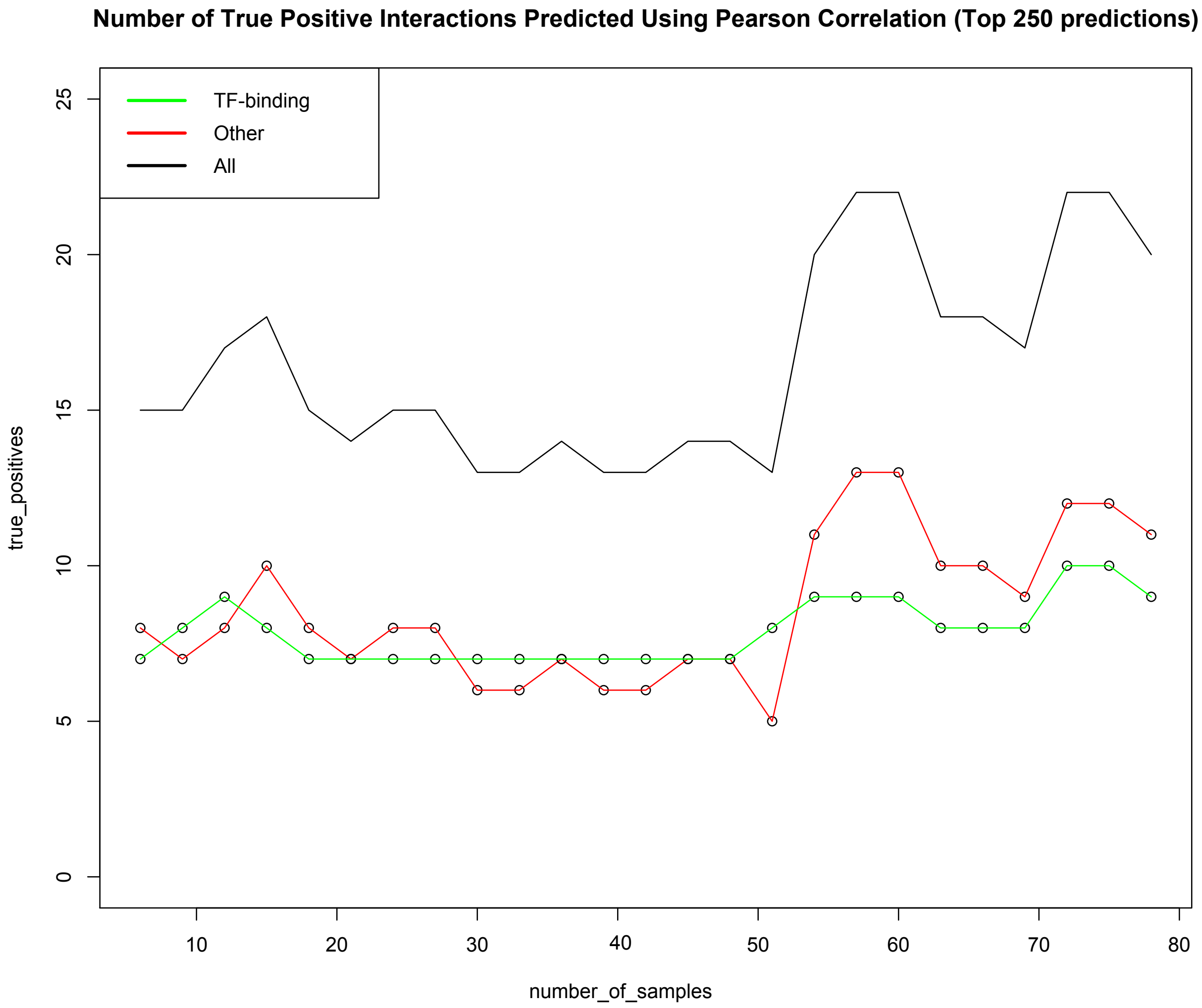


Figure 8.4: Pearson Correlation to predict GRN from RNA-seq

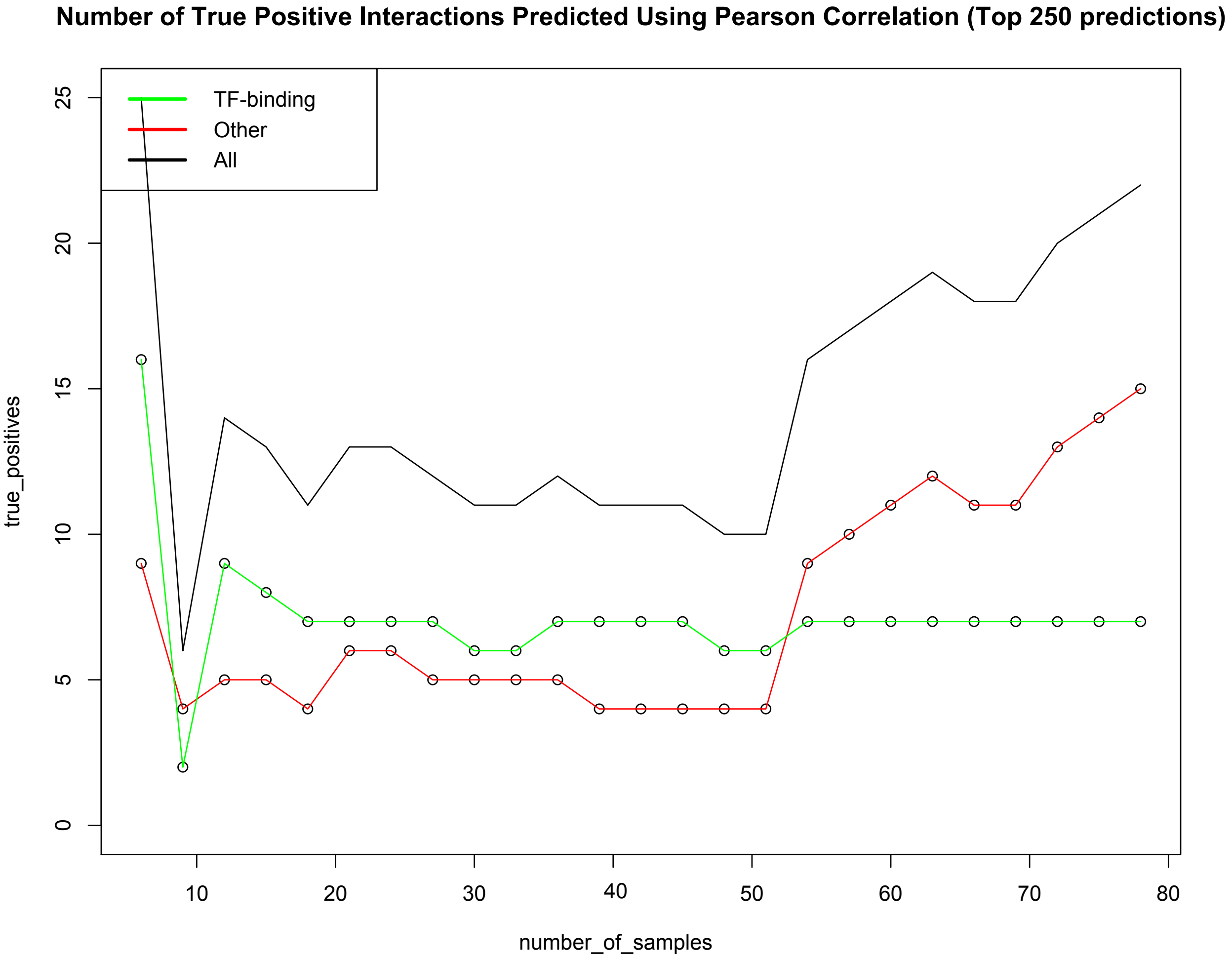


Figure 8.5: Spearman Correlation to predict GRN from RNA-seq

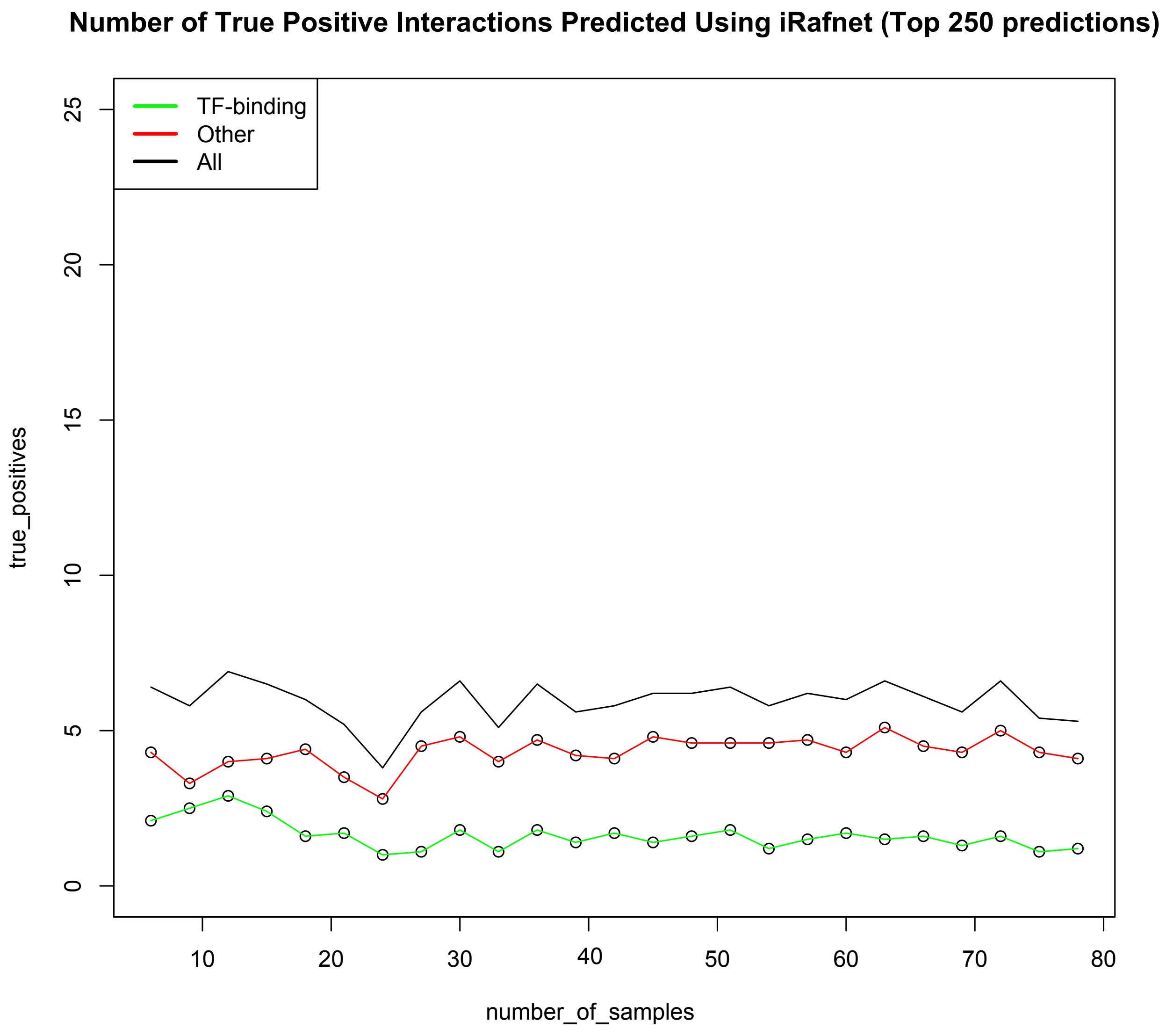


Figure 8.6: iRafnet performance with no influence of ChIP-seq. Since it is not possible for influences between the same gene, performance is lower than GENIE3

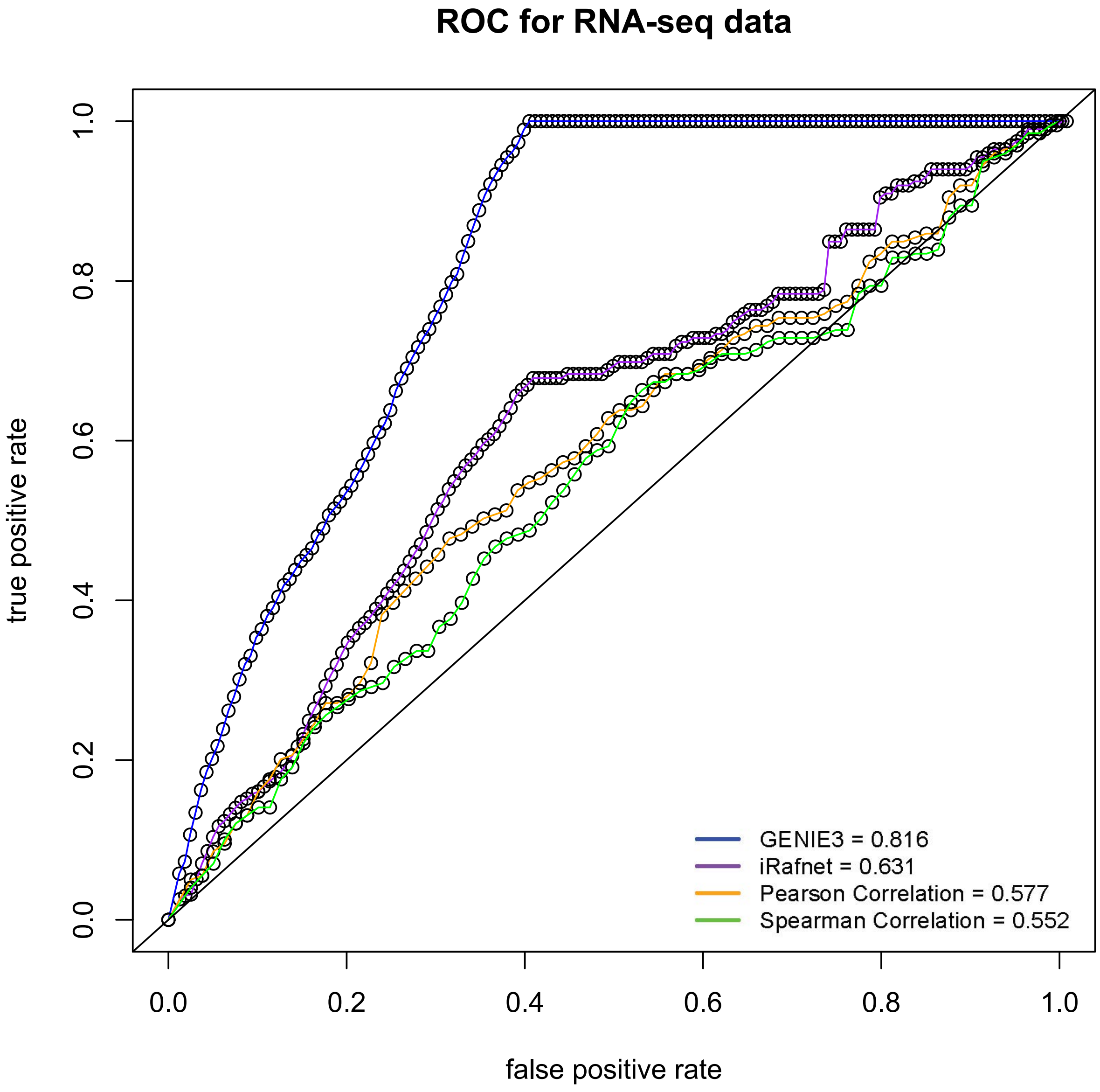


Figure 8.7: ROC for GENIE3, iRafnet, Pearson's and Spearman correlation. GENIE3 outperformed the other methods

8.3 Consistency of Predicted Interactions

A table of the top 250 interactions for each method was constructed and compared across the GRN prediction methods to determine how similar the results of each program were, and is shown if Table 8.1. The most consistent methods when using all 78 samples to predict a GRN are Pearson's and Spearman correlation. Both of these methods result in the same top 250 predictions for every run of the program. They are also the programs that share the most interactions between them. The other methods, even while using the same samples to make the predictions are predicting many different interactions from each other.

Table 8.1: Average Number of Different Interactions Between Predicted GRNs

	Spearman	Pearson	GENIE3	iRafnet	Inferelator (cMonkey2)
Spearman	0	59 (+/- 0)	192.5 (+/-0.45)	248.9 (+/-0.1)	234 (+/-2.70)
Pearson	59 (+/- 0)	0	189.9 (+/-0.46)	247.9 (+/- 0.1)	234.8 (+/-2.31)
GENIE3	192.5 (+/-0.45)	189.9 (+/-0.46)	30.69 (+/-0.32)	247.63 (+/- 0.07)	240.82 (+/-0.44)
iRafnet	248.9 (+/-0.1)	247.9 (+/- 0.1)	247.63 (+/- 0.07)	23.48 (+/-0.26)	248.02 (+/-0.34)
Inferelator (cMonkey2)	234 (+/-2.70)	234.8 (+/-2.31)	240.82 (+/-0.44)	248.02 (+/-0.34)	217 (+/-6.45)

The more samples are used the less variable the results of the program. The average number of differences were plotted in Figure 8.8 for GENIE3 according to sample size with as many GRNs made from distinct samples made. For example, with 6 samples, 26 distinct GRNs could be predicted from the 78 samples of RNA-seq data. This was done up to 39 samples where only 2 GRNs could be predicted with distinct samples for each. As the number of samples increased, the average number of differences decreased until the number of samples was increased from 26 to 39. When 39 samples are used to construct 2 GRNs, the average number of differences between them is within the standard error of the predictions using 26 samples to construct 3 distinct GRNs. This shows that increasing from 26 to 39 distinct samples no longer increases the consistency of the predicted interactions.

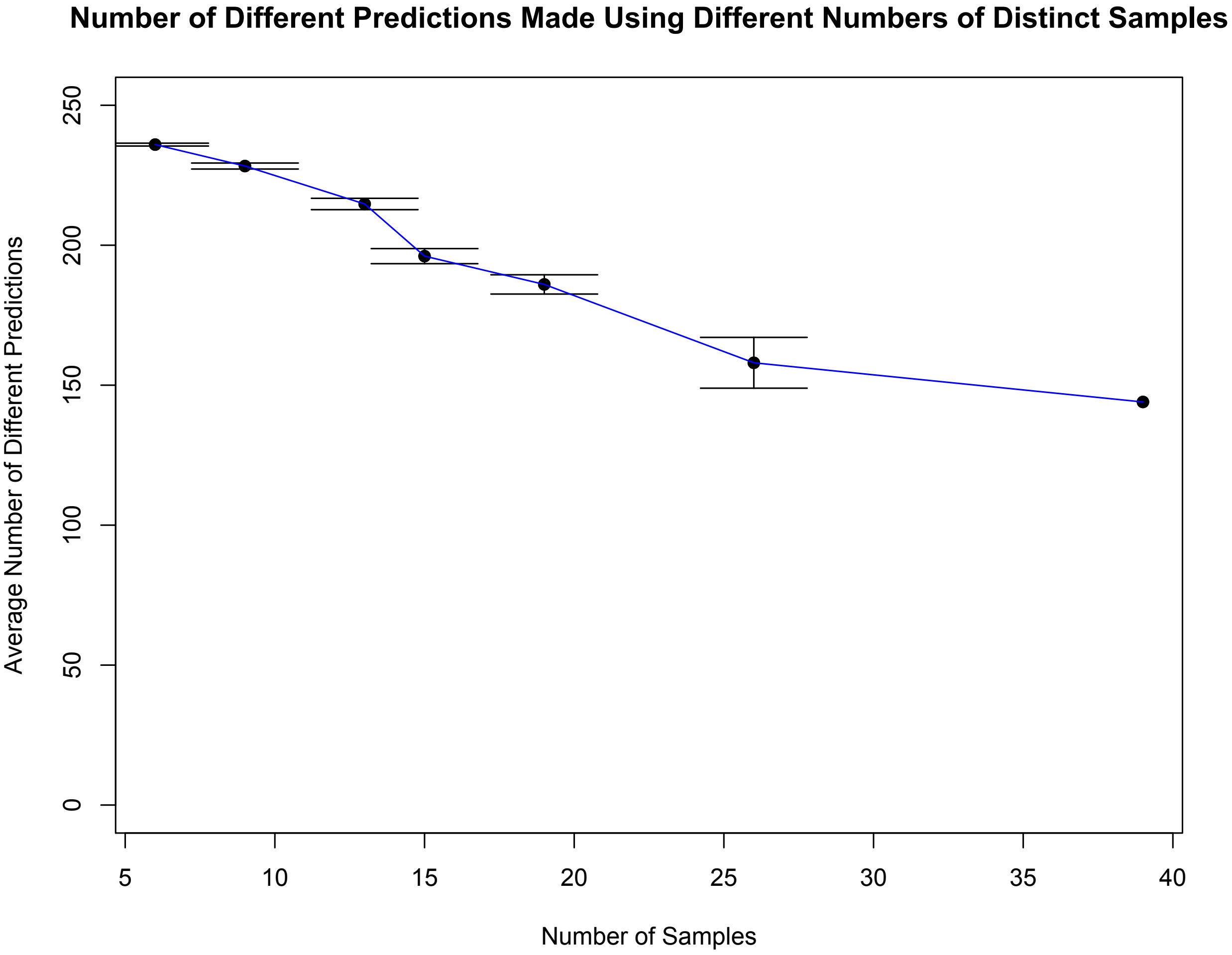


Figure 8.8: Average number of different predictions made with GENIE3

A recommended cut-off for the importance measure is not provided by random forest GRN prediction methods. However, depending on the number of samples used to infer a GRN, the importance measure will plateau quickly, shown in Figure 8.9. There is initially a spike where the importance values are quite high relatively compared to others and this difference gradually plateaus with no obvious value at which to stop considering the predicted interactions accurate. The plateau begins when the importance measure is equal to 0.04. An importance value any less than that means that interactions below the top 250 interactions have very little change in their importance measures.

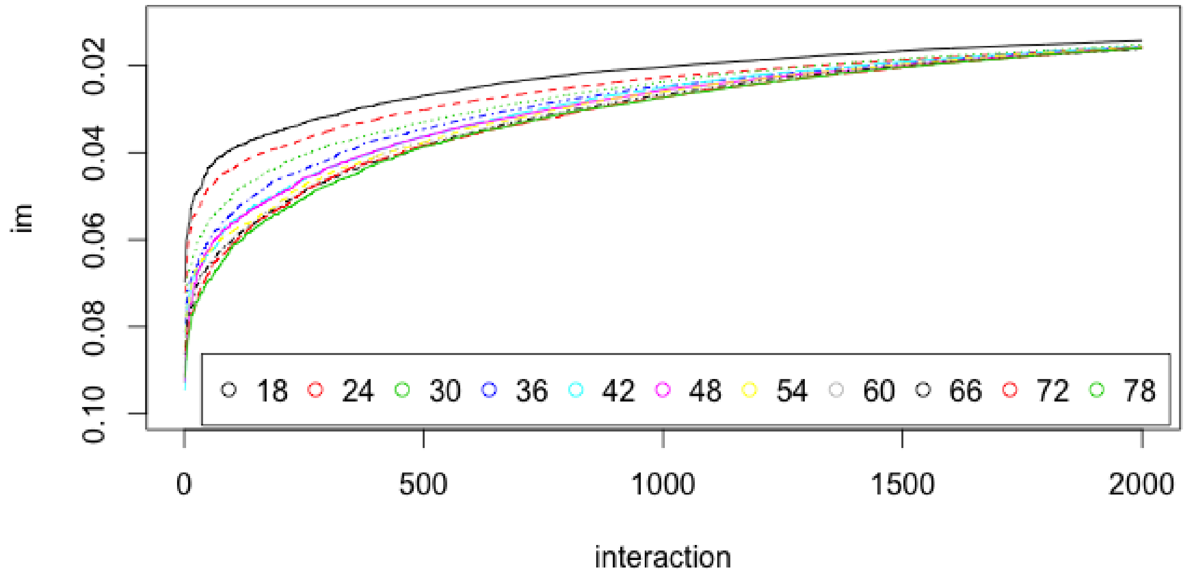


Figure 8.9: The value of importance measures from GENIE3 for different numbers of interactions. The number of samples used to predict the GRN was increased from 18 samples to all 78 samples (indicated by the legend). The more samples used, the more interactions had higher importance measures, but the importance measures quickly plateaued and do not change a significant amount from 1000 to 2000 interactions.

8.4 cMonkey2 and iRafnet Performance using ChIP-seq

8.4.1 iRafnet With and Without ChIP-seq

This method, described in Section 6.4.1, was initially run without any influence of ChIP-seq where all potential regulators were given a weight of 1 including Nanog and Oct4 for each gene j. Performance was below GENIE3 when comparing the rate of true positives and false positives in the top 250 predicted interactions. The method uses a weight of zero for each gene j when it is under consideration as the researchers assume that a gene is not able to influence its own expression, but some of these interactions are reported in the literature network. Although iRafnet is the only method that can take this into account, once the influence of Nanog and Oct4 are included using ChIP-seq data, performance drops instead of improving. The number of promoter binding events, in particular, is less for iRafnet with or without ChIP-seq data as the predicted interactions for Nanog and Oct4 have a lower placement in the list of predicted interactions.

8.4.2 cMonkey2 With and Without ChIP-seq

Without ChIP-seq, cMonkey performance was comparable to GENIE3 results at sample sizes 40 and 78 although the result varied more depending on the number of samples used to predict the GRNs shown in Figure 8.10. One aspect this evaluation does not take into consideration is the transcription factors that are in the same bicluster as other genes, which are not accounted for using Inferelator. It is assumed that the transcription factors in a bicluster could be influencing other genes in that same bicluster, but they are not reported in Inferelator results. As such, many true positives could be missed using this evaluation. However, since the GRN is only limited to 250 interactions, there are only so many interactions that would be picked up, possible only for one or two transcription factors. Since ChIP-seq was used for Nanog and Oct4, the number of true positive interactions identified in biclusters for each number of samples were identified in Figure 8.13 and Figure 8.12.

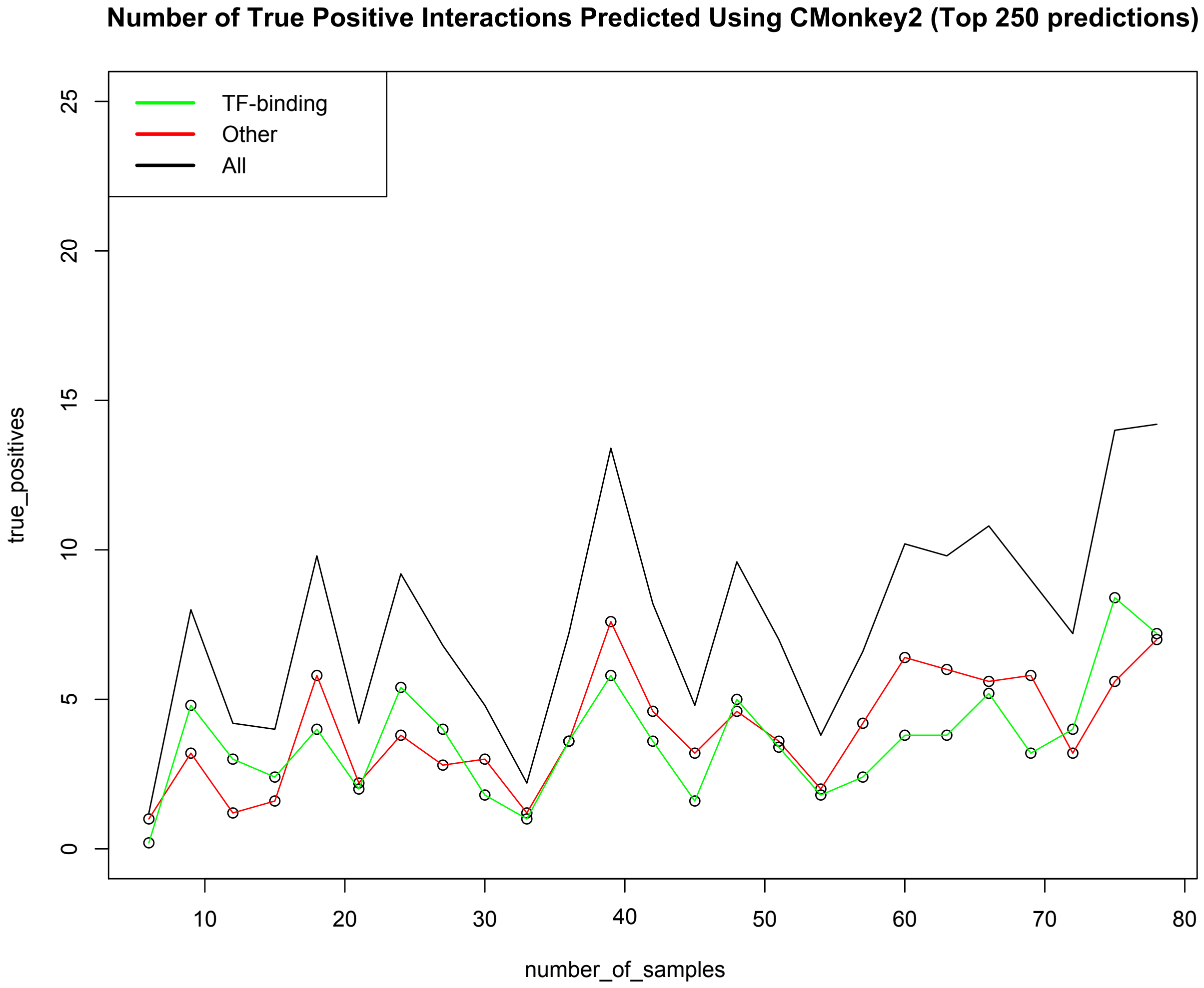


Figure 8.10: cMonkey2 results with no influence of ChIP-seq

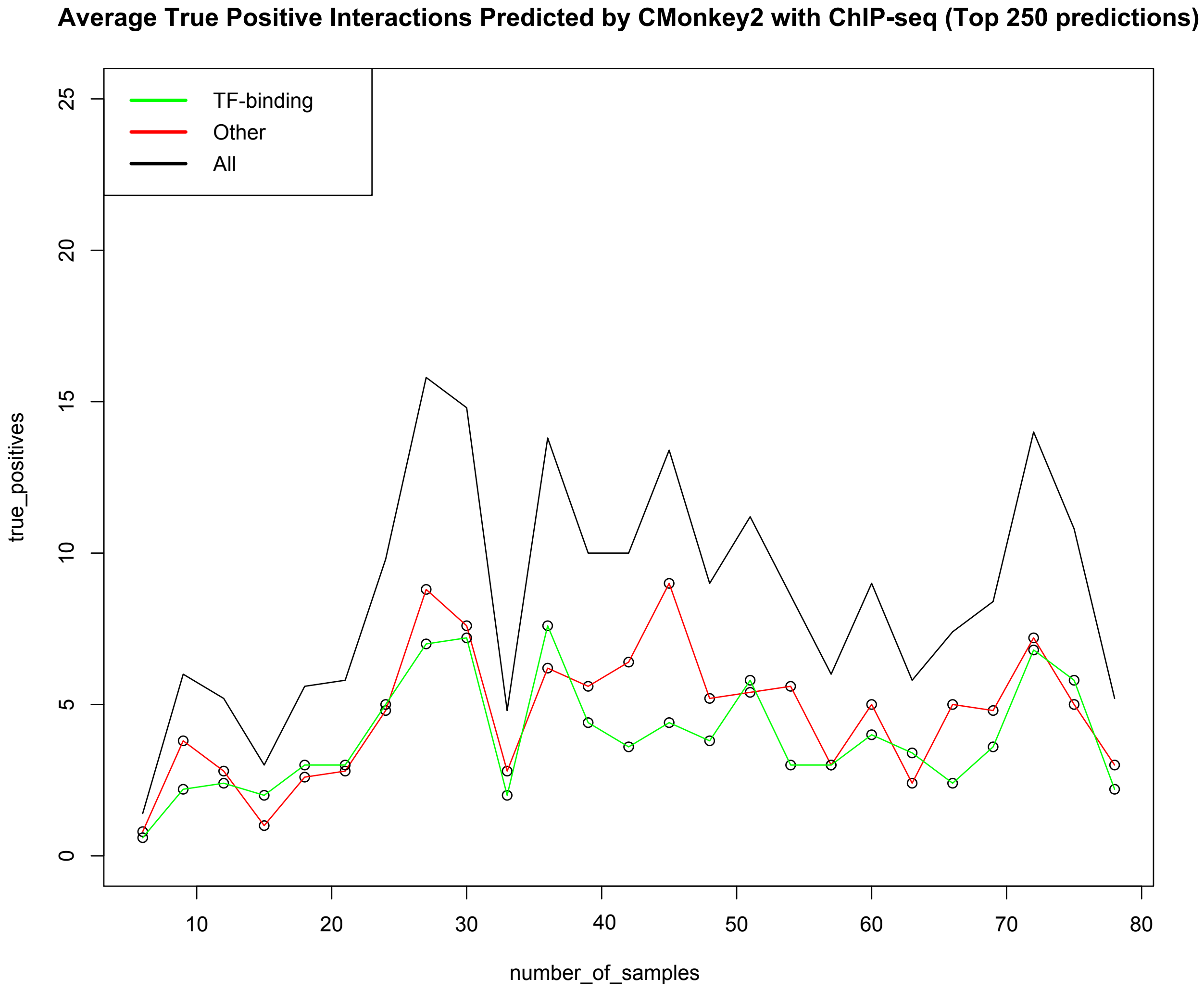


Figure 8.11: cMonkey2 results with ChIP-seq

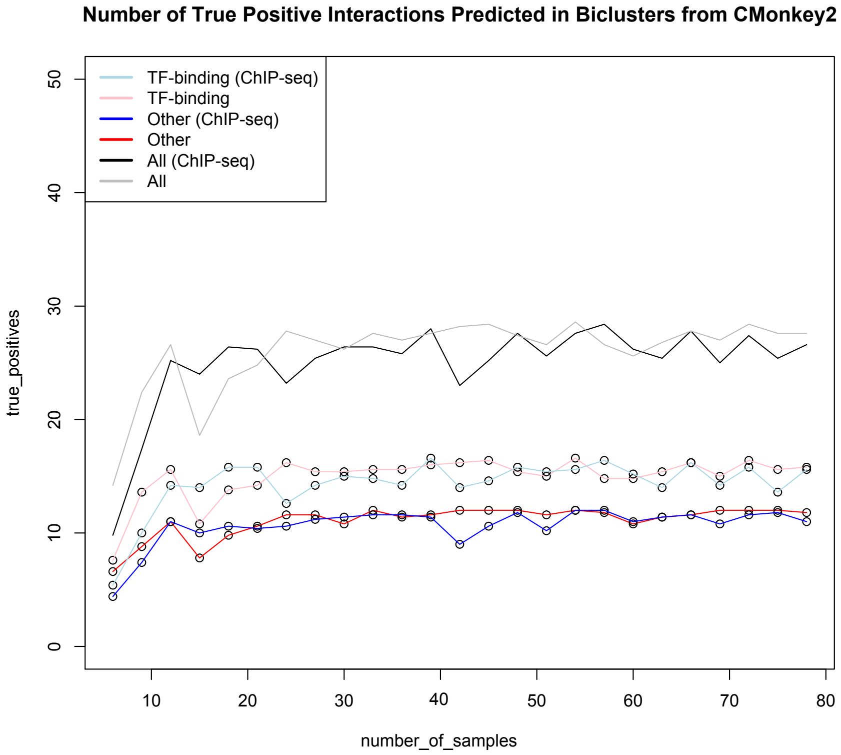


Figure 8.12: Number of true positives for Oct4 in biclusters

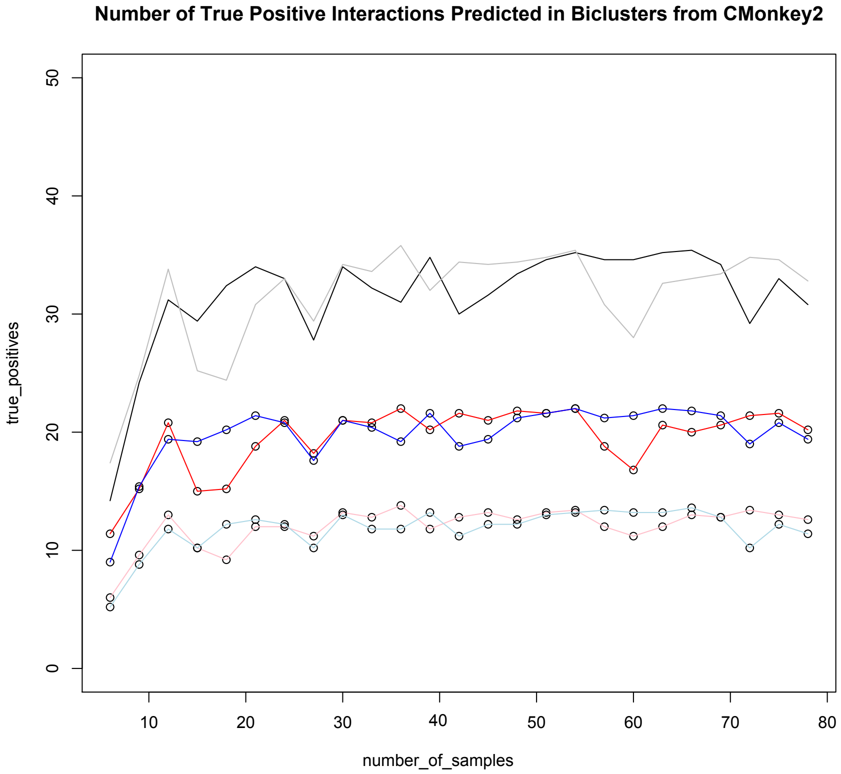


Figure 8.13: Number of true positives for Nanog in biclusters

ChIP-seq was not able to improve the rate of true positive predictions using Inferelator for most sets of samples although significantly higher peaks in Figure 8.11 were observed between 20 and 30 samples as well as at 46 samples and 72 samples. Taking the potential interactions occurring in bicluster modules also did not improve with the use of ChIP-seq. However, using ChIP-seq did not decrease the number of true positive interactions overall. All true positive interactions for Nanog and Oct4 were predicted consistently within the biclusters when using more than 10 samples without ChIP-seq data so it was not necessary for the information provided by ChIP-seq to be utilized.

8.5 Discussion of GRN Prediction Evaluation

It was determined that the ESC RNA-seq data could be used for the purpose of evaluating GRN prediction methods. The AUROC was much better than 0.5 and as such, performed better than randomly predicting interactions with GENIE3, iRafnet and correlation-based methods. The number of samples used did little to change the AUROC, which leads to the conclusion that the RNA-seq data would be appropriate for GRN prediction, but could suggest two other possibilities. It is possible that i) the samples for microarray and RNA-seq ESC data are equally good for predicting the ESC GRN or ii) the samples are not appropriate for predicting the GRN using either microarray or RNA-seq.

It is difficult to determine, using the ROC curves, that all of these programs will never find a large portion of the currently known interactions without taking a lot of possible interactions into consideration. This was a similar case with the microarray data. However, one difference noted between both types of data, was that using the microarray data resulted in more predicted promoter binding events in the top 250 predicted interactions. This might suggest that the topology of the network predicted using microarray data would be different than when using RNA-seq, depending on the cut-off used. One limitation of the true positive rate using correlation based methods is that the methods predict an association, but the direction of the interaction will not be specified. However, this may be applied to some predictions after the fact if there is a database of transcription factor available for the organism as well as PPI information. Although these methods perform above random, given there are $N * (N - 1)$ potential interactions where N is the number of genes, the sparsity of the network means that a low number of true positive interactions will indicate the method predicts interactions better than random guessing. However, having $\frac{10}{248}$ or $\frac{19}{248}$ true positive interactions does not provide much confidence for the quality of the other interactions that have been predicted.

Although all of the methods perform better than random, the number of true positives in 250 predicted interactions is less than 20 in total out of 248, which is likely why AUROC is reported much more in publication than the actual number of correct interactions predicted using different cut-offs. It also appears that increasing the number of samples may not increase true positive predictions in a linear or exponential fashion for real biological data. Therefore, due to this flat trend it may not be best to focus on increasing the number of samples used with these methods, but using and combining different methods. When using iRafnet, it was thought that if more weight was placed on several transcription factors, then the chance of seeing interactions with those transcription factors influencing other genes would increase, but this is not the case. This is likely due to the number of genes selected in order to generate the trees for each gene j using random forest. Instead of selecting a random sample of genes at each node, genes are selected according to sample weights. This means that biasing towards a small subset of genes is not possible unless all other regulators are discounted. The sample size will be greater than 2 in this case (GENIE3 and iRafnet select the square root of the number of potential regulators for each sample of genes). However, it is not practical to take a subset of two (the number of transcription factors where there is ChIP-seq available) in order to construct the trees. Since this might be the case, 9 other potential regulators were given a weight of 1 and the rest were always zero. These others were selected based on the interactions in the literature network, selecting those more commonly in TF-binding relationships. Again, this improved results back to comparable levels with GENIE3, but no more than that. Perhaps this would improve further if at least the number of ChIP-seq data from different transcription factors was equal to the number of samples taken at each node, but this removes the biasing aspect to only a select few transcription factors. However, this still does not completely explain the decrease in performance when ChIP-seq is added. iRafnet results in a total of 6502 interactions being predicted where the importance measure is not equal to zero. Therefore, one benefit of using iRafnet as opposed to GENIE3 is that there is a very obvious place to set a cut-off in the list of predicted interactions. These interactions were sorted according to highest importance measure and the top 250 interactions were used for comparison to GENIE3 and Pearson and Spearman correlation. However, both Nanog and Oct4 have all 125 predicted interactions present with importance measures greater than 0 so a cut-off would still have to be applied in order to produce a useful result if the focus is on select transcription factors.

The challenge remains that there first needs to be some feature selection performed before GENIE3 could be used to predict a complex organism's GRN. Furthermore, an appropriate cut-off to predictions is also necessary. From current results, it seems that in order to achieve good sensitivity, this will result in all possible predictions for some transcription factors predicted, which is not useful if a researcher has those particular transcription factors of interest. They would do no worse by generating all possible combinations of interactions, although in the case of GENIE3 they would be ranked by importance measure. One benefit to biclustering (by using cMonkey2) is it divides the genes into modules. There are less false positives associated with the genes, as they will only be associated with the other genes in the module. As such, since the biclusters contain fewer false positives in comparison to true positives found overall inside the module, there are fewer interactions to narrow down using other means as long as the transcription factors of interest are in a minimal number of biclusters. Therefore, genes of interest can be focused on without eventually predicting all possible interactions one gene has with all the others. However, this also means that if a gene is left out of a module it may never be associated with the genes in a module, where random forest methods will eventually predict an interaction even if overall it results in performance no better than random guessing. Inferelator also adds more information on top of the biclusters if it is likely genes of interest are influencing expression external to their assigned module. Inferelator was designed in 2006 and has been minimally updated in the past three years. Therefore, using other methods to infer interactions within the biclusters may now be feasible. Originally, cMonkey2 did not use Inferelator and only compared to other methods based on similarities of the genes in each module including similar sequence motifs and biological enrichment, but not on predicted interactions as there is no method to predict these interactions other than knowledge of the transcription factors in each module potentially regulating the other genes in the module.

From the true positives predicted using a low number of samples, Spearman correlation achieves better performance initially compared to the other methods. This spike in true positive predicted interactions may be due to the number of runs being limited to 10 or 5 depending on the runtime of each method. During each run different samples are combined at random so perhaps a better selection of samples was run with Spearman correlation. The variation in cMonkey2 results may also show evidence for different subsets of samples resulting in better GRN prediction than others. This may suggest that it is likely that the number of samples is not as important for making many true positive predictions as it is to use data with as much variation as possible in order to pick up patterns in gene expression.

Chapter 9

Application of cMonkey2 to Skeletal Tissues

cMonkey2 was applied to the skeletal tissue RNA-seq dataset due to its ability to find localized patterns, including those that may be in a single tissue for which there are only three samples. The random forest methods utilize techniques that require more than three samples. The Sox9 and Runx2 networks were predicted using cMonkey2 and Inferelator, which resulted in 913 biclusters. Some of these biclusters that were empty or contained only a single gene, so they were removed before Inferelator was applied. These results were minimized to only biclusters that contained either Sox9 or Runx2 or the biclusters that Inferelator predicted could be regulated by Sox9 or Runx2. This left 13 biclusters, which could be expanded to include regulatory interactions with the other transcription factors in the biclusters, but Figure 9.1 shows the biclusters that are directly associated with Sox9 or Runx2 expression.

9.1 Comparison of cMonkey2 Predicted Interactions to FABIA Biclustering Results

cMonkey2 produced biclusters separating mature cartilage from the other skeletal tissues, but was unable to separate the other two tissues, much like the results from FABIA. Since FABIA seemed to produce the most biologically relevant results from analyses performed in Chapter 7, it was selected to compare overlap of genes associated with Runx2 and Sox9 found by each method. Where FABIA was able to produce up to 10 biclusters, with 6 biclusters used to achieve the highest tissue separation score. cMonkey2 produced 13 biclusters in total that either had Sox9 or Runx2 inside them or, using Inferelator, potentially regulating the genes inside another bicluster. The biclusters produced by FABIA that also contain Sox9 or Runx2 had 852, 772 and 253 genes. These are much larger than the number of genes contained in the biclusters produced by cMonkey2 shown in Figure 9.1. As such, only 3 genes overlapped between the results produced by cMonkey and the results from FABIA (*Igf1, Fxyd6* and *Lgmn*). This is not to conclude that cMonkey2 results are not useful. For example, cMonkey2 biclusters Sox9 with genes that provide instructions for making part of type IX collagen, *Col9a2* and *Col9a1*, so the results are not necessarily any less biologically accurate. FABIA was unable to group these genes with Sox9. It will be of interest to further compare the genes present in the biclusters of these programs, potentially also including Plaid and SAMBA results, to determine if results from these programs may be combined or if we may be any more confident in the importance of genes appearing in at least one bicluster using each program.

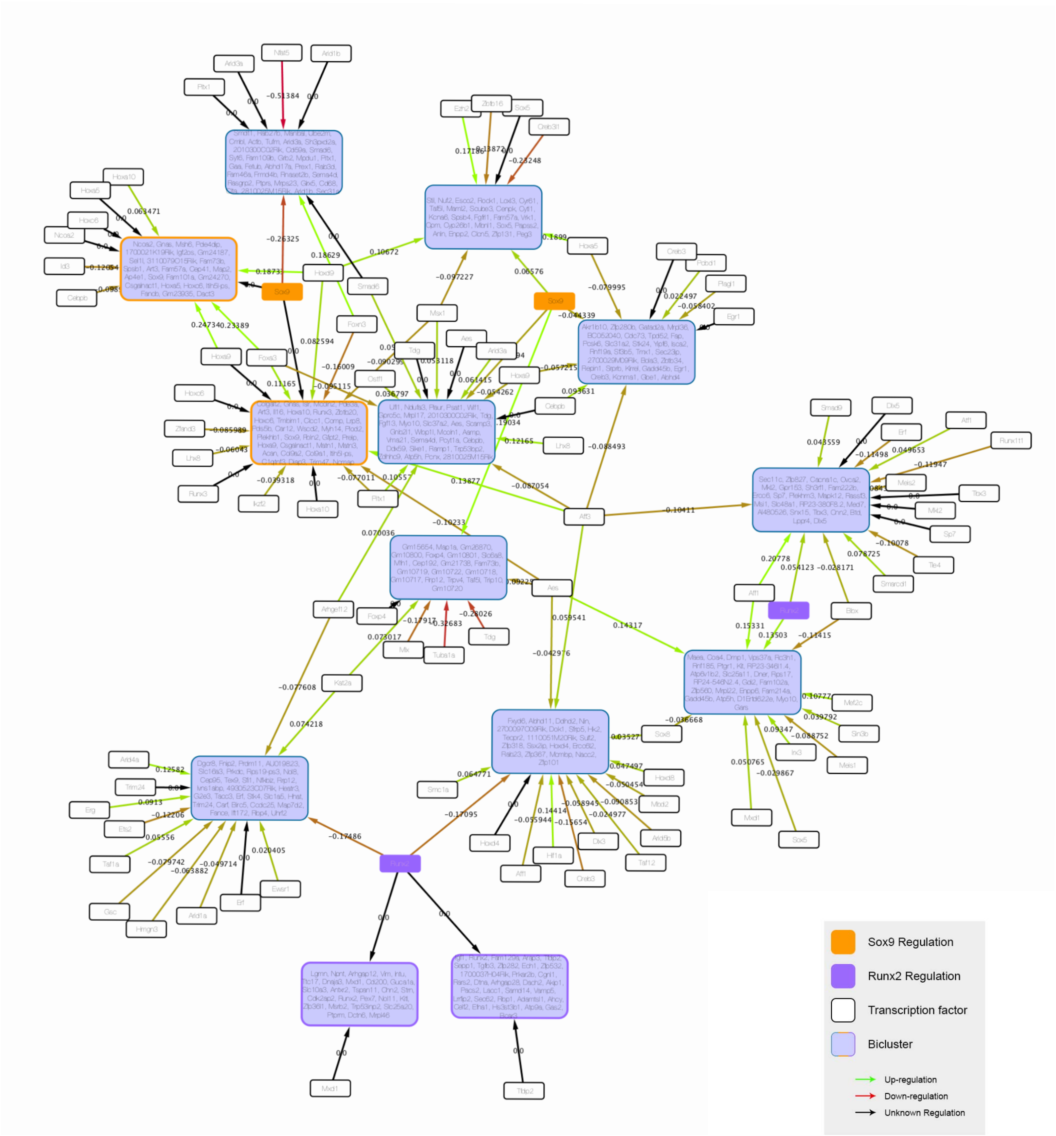


Figure 9.1: Sox9 and Runx2 GRNs visualized with Cytoscape as predicted by cMonkey2. The diagram shows biclusters potentially regulated by transcription factors within the same bicluster as well as transcription factors not placed in the bicluster. Biclusters outlined with orange are biclusters that contain Sox9 while biclusters outlined in purple contain Runx2. The biclusters outlined with neither colour do not contain these transcription factors, but Runx2 or Sox9 may be regulating the genes that are inside the biclusters indicated by directed arrows. The black arrows (unknown regulation) means there is no confidence associated with the interaction since the gene potentially regulating others in the bicluster is also in the same bicluster. Up and down-regulation interactions are predicted using Inferelator. To view the genes within each bicluster, this data is also available in Appendix C

9.2 Limitations of Testing GRN Prediction Methods

Some limitations of predicting GRNs apply to the analysis of both Chapter 8 and Chapter 9. Therefore, it is likely that the limitations observed with ESC data also apply to skeletal tissue data. One limitation in this thesis is the network available for evaluation of the GRN prediction methods in vertebrates may not be complete as GRNs tend to be complex, and interactions that do occur may not alway be identified using low throughput techniques. Therefore, the low true positive rates in the top 250 interactions may not be an indication of a method's ability to make accurate predictions, but the lack of research available to identify more interactions that are occurring in these networks. To try and counteract this limitation as much as possible, the genes used to predict interactions had to be in the literature network. Therefore, interactions could only be predicted between these genes. However, the number of interactions in the literature network (248) versus what could be predicted using the RNA-seq data (15750) is a lot lower so a lot of interactions could be missing. Unfortunately, this may not provide indication of a methods ability to discover new interactions, compared to finding interactions that are already known. However, more true positive interactions discovered with high confidence may also give reason to believe the other genes predicted–currently considered as "false positives"–with high confidence may just be novel interactions.

9.2.1 Predicted Interaction Cut-off

All of the methods do not have a defined cut-off. This may also be a reason for the large rate of inconsistency when attempting to predict a larger network in methods like GENIE3. Since the importance values beyond 250 are similar, it is likely that multiple runs of the program shuffle these predicted interactions around, which is why each run of the program results in different predictions depending on the cut-off. This could explain why other researchers have reported only 3 edges recovered while multiple runs of GENIE3 produced an average of 18 using the same microarray data [98]. This suggests that it is possible the proposed method does not in fact perform any better than GENIE3 unless perhaps it is more consistent in its results. To get consistent interactions predicted for the top 248 interactions of the ESC network with GENIE3, more than 78 samples of RNA-seq data would be required. For novel discoveries this is of particular concern, since even collecting 78 samples or attempting to find appropriate datasets online when the research is looking at something new is likely not possible.

9.2.2 Using AUROC for Measuring Performance

The best performing GRN prediction method proposed by the authors in the most recent evaluation using the ESC network was able to discover 10 true positive edges using ESC microarray data, which is still very low although when the AUROC appears to perform better than random [98]. Due to the difference in performance between 10 runs of GENIE3 in this project and the results reported in literature, there is a question of how good results can be if they are not consistent. Using GENIE3, the best performing GRN prediction method, the last true positive interaction for Nanog is around the 5800th predicted interaction. At this point, if referring to the ROC curves, using a cut-off of 5800 predicted interactions includes all of the interactions in which Nanog could possibly be involved.

9.2.3 Addition of ChIP-seq: Quality and DNA Binding Locations

From the initial exploration of incorporating ChIP-seq data to gene expression data to predict GRNs, ChIP-seq does not largely seem to impact the performance of random forest or biclustering. One caveat to these findings could be the quality of the ChIP-seq data. When using fastQC as a quality check to determine the quality of the sequences, many required trimming. When the sequence length is originally small before trimming (25-36bp) any more trimming can have a impact on the number of uniquely aligned sequences to the genome. The sequences that do not align uniquely are not counted as a gene feature and so cannot contribute any information about gene expression or likely binding sites in downstream analysis. Therefore, binding sites may not achieve counts above background.

Another possible limitation might be using a 3000 base pair (bp) cut-off from the transcription factor start site. This cut-off was used as the ChIPSeeker program labels these as binding events to a promoter while binding sites further away might be a different kind of regulation. As the focus was on promoter binding events for comparisons with the ESC literature network, this cut-off was used. However, this cut-off may exclude possibly important regulatory events. Since the focus is on the 126 genes in ESC, any binding events outside of these 126 genes are not considered when predicting the GRN, since most of these binding events were picked up using the cut-off of 3000 bp. There was still potential to increase the number of predicted interactions, perhaps not all of them, but more than the RNA-seq data alone. If dealing with a larger set of genes to predict a GRN, it may be useful to reevaluate if the 3000 bp cut-off would be appropriate.

Another limitation would be minimizing the number of genes before GRN prediction with the ChIP-seq data. There still may be a benefit to incorporating ChIP-seq data when the number of genes used to predict a network is large. In this thesis, the genes in the ESC RNA-seq samples were minimized to those known to take part in the network, which is not possible with uncharacterized networks, although there is the option of using clustering and biclustering to minimize the number of genes. The way cMonkey operates, for example, is it groups genes that have the potential to interact with the transcription factors for which there is ChIP-seq data. Since, for these evaluations, the genes were already minimized to only include genes known to be in the ESC literature network, the options for grouping particular genes together had already been minimized. If more genes not known to be in the network were included, and less likely to be involved, the ChIP-seq data may have proven more useful.

9.2.4 Auto-Regulation

One limitation across all of the methods tested is their inability to accurately predict auto-regulation, which is the regulation of a gene's expression by itself. For example, both Oct4 and Nanog have binding sites that allow for changes in expression of themselves. These interactions have been confirmed and are in the literature ESC network. This is detected by ChIP-seq, but cannot be picked up in expression data since the expression of a gene is always going to be the most highly correlated with their own expression and so these interactions are left out of the prediction, meaning when evaluating using the ROC, these interactions are never found. Hence, they cannot be taken into consideration if sensitivity is to ever equal 1. With the binding site information, the expression data did increase the confidence of these interactions so this is one benefit of using iRafnet. But in comparison to the other methods, its performance was poorer overall although it outperformed correlation-based methods. However, with Nanog and Oct4, there are only 2 possible occurrences of auto regulation, and it is not obvious that the addition of the ChIP-seq data would be helpful with ChIP-seq data for more transcription factors.

9.2.5 Sox9 and Runx2 Literature Networks

New gene expression and ChIP-seq data for bone and cartilage tissue has recently become available [114, 115]. Therefore, it will be necessary to update the literature networks not only based on our data, but other data available in order to determine if any datasets, including the RNA-seq data used in this thesis, are potential outliers. It is possible that the current datasets compared to the RNA-seq data from skeletal tissues is an outlier or it could be that the RNA-seq data is not appropriate for GRN prediction. Predicting the GRNs with other datasets is another means that could increase the confidence for the current GRN prediction. If multiple datasets are resulting in very different lists of predicted interactions, they may not be appropriate for GRN prediction or it may be necessary to average the predictions across the datasets. The more agreement there is among the predictions using different datasets suggests more confidence can be placed in the predicted network.

9.3 Future Directions

From comparisons between the microarray and RNA-seq data, the two datasets are just as useful as each other for GRN prediction, but whether the resulting predictions made by either are good is less clear. It might be interesting to test a dataset with the same genes from another tissue other than ESC and see if the same top interactions would be predicted. This may provide insight as to whether the interactions the methods are able to identify in the top interactions can be attributed to a GRN that is functioning in one type of cell compared to another tissue that are not pluripotent with no self-renewal. With simulated data it is possible to confirm that a particular GRN is functioning as it was placed in the data artificially. With actual biological data however, it is generally assumed that the GRN is functioning and has the potential to be picked up by the GRN prediction methods. Another method to test this might be to randomly swap gene expression values among the data to get rid of relationships in the gene expression and run all of the methods again. If the number of predictions does not drop significantly in the top interaction or if the AUROC remains high or above random, this may suggest that the datasets are not appropriate for predicting the relationships specifically found in a GRN.

A second objective would be to improve how to compare biclustering techniques to other machine learning techniques for GRN prediction. The genes within biclusters are associated in some way, but there is only one confidence value (the residue), which does not indicate whether some of the transcription factors are more likely to be direct regulators compared to others. This means it is necessary to consider every potential interaction that could occur between the genes in a bicluster, which increases the number of false positives (while possibly increasing the number of true positives as well). The ROC is not a proficient way to measure the performance of these algorithms as it is not necessary that all potential interactions be predicted. Only 97 interactions were ever predicted for Nanog when there could be 126, for example using cMonkey2.

Adding more machine learning methods to this evaluation may provide more or less evidence to support the usefulness of increasing sample size for a complex organism. It may be necessary to have some means of predicting with bicluster interactions by attaching a level of confidence to each possible interaction as opposed to a residual which provides overall confidence based on how correlated all genes are to each other. As long as it is possible for a GRN prediction method to produce a list of predicted interactions, the current methods of comparing performance may be used. If these methods can be integrated to predict interactions most likely in the biclusters, there could be potential for comparing the current biclustering methods more easily to other GRN prediction methods. It is possible that applying ChIP-seq data after biclustering as opposed to it influencing the genes grouped together initially could make the addition of ChIP-seq information more useful as well. One limitation currently for cMonkey2 in higher organisms is the motif database is not in a format that allows the program to automatically run, and it is necessary to curate your own database or minimize genes in some other fashion so that only genes that are in the current database are accounted for [19]. Therefore, it may be necessary to curate a database of sequence data for skeletal dataset in order to incorporate motif finding if cMonkey2 predictions prove to make sense biologically at this stage. It will also be of interest to locate other transcription factors with high connectivity with the Sox9 and Runx2 networks. These transcription factors may also have a significant influence on the expression of Sox9 and Runx2 as there has to be other gene expression that is influencing the down-regulation of Sox9 in order to mature cartilage and bone to develop.

9.3.1 Evolution of Gene Regulatory Networks

Gene regulatory networks tend to have complex structures and it is a current challenge to determine which connections in a GRN are modified and how they are modified in order to produce a novel phenotype. It is thought that the co-option of older GRNs can lead to the development of novel structures [39]. Examples of this phenomenon include beetle horn formation resulting from the co-opting of the appendage formation GRN [39]. Not all genes in the networks required for appendage formation are required for the development of the beetle horn although knockdown of key parts of the network suggest that parts of the network are necessary for beetle horn formation. Another GRN for echinoderm larval skeleton development could have been co-opted from an ancestral GRN that directed the formation of their adult skeleton [116]. The GRN in this case for adult was already well understood so co-expression studies showed genes active in similar manners during both processes. In other, less related species like the sea cucumber, it has been shown that it is likely this GRN underwent further remodelling. The GRNs defining skeletal tissue development in vertebrates may be an example of GRN co-option leading to the generation of new morphologies [8].

9.3.2 Gene Regulatory Networks Evolution in Skeletal Tissues

The distinct characteristics of the Sox9 and Runx2 GRNs have recently been explained from an evolutionary perspective [8]. It is hypothesized that cartilage is a much older tissue than bone, meaning that the GRN(s) characterizing the development of this tissue have been established for a longer period of time before bone appeared in evolutionary history. One possibility is that bone evolved separately from cartilage meaning the gene expression and the GRN that defines bone does not necessarily have any relation to the genes expressed in cartilage. Another option is that bone development evolved gradually through co-option of the Runx2 GRN that was established in mature cartilage [8, 117]. It is further hypothesized that a mixture of the GRNs in immature cartilage and bone characterizes mature cartilage development, meaning the tissue arose somewhere between the process of bone co-opting the Runx2 portion of the Sox9/Runx2 GRN mixture in mature cartilage. Learning more about the topology of these networks will aid in determining regulation and function of the genes in these networks and elucidate the evolution of skeletogenic mechanisms.

Potentially, biclustering may be adapted to include other information. To do so, biclustering algorithms currently used will have to be adapted in order to handle the sequence data contained with the data used to construct the gene expression matrix that is currently used to bicluster. cMonkey2 can use sequence information currently, but only from a database and does not focus on mutation across the same gene, but on potential regulator binding sites [19]. Using synonymous (change of a nucleotide that does not change the amino acid sequence produced) and non-synonymous (change of a nucleotide that does change the amino acid sequence produced) mutations will group genes by different degrees of conservation within a single tissue as well as across tissues. It would be interesting to determine how results correspond to results using gene expression profiles to determine genes potentially in a network. Since evolution of the GRNs will be a main focus of my research in the future, the incorporation of sequence information in terms of synonymous and non-synonymous mutations in genes would be useful.

Chapter 10

Conclusions

The limited knowledge currently available describing the regulation of skeletal tissue development could be further elucidated with the accurate measure of gene expression using RNA-seq technology. Using this gene expression data to predict possible GRNs driving the development of bone and cartilage tissue could potentially identify genes not known to be involved in these processes as well as confirm hypothesized key regulators of the networks as well as others. The first objective of this thesis was to determine if there is evidence of two interacting GRNs in mature cartilage by analyzing gene expression. Furthermore, since there is some information available in literature about genes potentially regulated by transcription factors Sox9 and Runx2, it was necessary to determine if this information agreed with the RNA-seq data from the skeletal tissues since minimal agreement between these data sources could provide justification for predicting new networks. Results of model-based clustering, as well as differential expression and simple comparisons of gene expression across bone, immature and mature cartilage show evidence that if there are two GRNs driving bone and cartilage formation, they are likely both active in mature cartilage. It was determined that there are less uniquely expressed genes in mature cartilage compared to immature cartilage and bone. Gene expression in mature cartilage was usually an average of gene expression in immature cartilage and bone, or had similar gene expression to one of the tissues. This suggested that most genes in the GRN driving mature cartilage development, were under control of both Runx2 and Sox9 GRNs. It also appears that the number of genes in mature cartilage that have expression more like immature cartilage or bone are nearly even, suggesting one GRN is not necessarily the dominant GRN with more activity in mature cartilage. However, since there are genes expressed that are unique to mature cartilage, suggesting that the Sox9 and Runx2 GRNs are not entirely additive. The expression of both Sox9 and Runx2 may influence some genes to increase in expression according to differential expression results as well as clustering results. To confirm these results it will be required that more data from other sources be analyzed in the same manner, or more samples added to the current RNA-seq dataset to confirm if the current set of data is an outlier and unreliable for biological interpretation. It was also determined that the Sox9 and Runx2 networks in the literature available for this project did not contain many overlapping genes with those considered expressed using the RNA-seq data from bone, immature and mature cartilage. Therefore, the analysis of other datasets that have become available for determining a potential Sox9 or Runx2 GRN will be useful for strengthening evidence either for or against a relationship between the Sox9 and Runx2 GRNs.

The second part of the thesis focused on predicting new Sox9 and Runx2 GRNs, given that the genes in literature reported as being regulated by these transcription factors did not agree with the RNA-seq data available for skeletal tissues. This required that the number of samples used to predict the GRNs be enough to predict interactions in the GRNs better than random. Initial predictions of GRNs functioning in skeletal tissues are useful for future research comparing genes conserved and expressed among vertebrates. Challenges associated with predicting a GRN include the large number of samples required to predict interactions. Since the number of samples available is small, it was necessary to determine if these samples would be enough on their own, or if results of GRN prediction using these samples could be improved by either adding more data or using a particular method over others. Random forest methods outperformed correlation-based methods, but increasing sample size did little to improve performance, with a maximum of 90 samples of microarray data and 78 samples of RNA-seq data for ESC in mouse. Furthermore, neither using a strict cut-off nor considering many different cut-offs up to the total possible interactions lead to significantly improved results with any method. Other techniques to improve results such as ensemble techniques combining methods or results from different data sets to have more variable samples may have a greater impact to results. The consistency of results across methods was highly variable with machine learning methods while fairly consistent when comparing correlation-based methods. Increasing the number of samples has the potential to improve consistency within a single method, but using all 78 samples of RNA-seq data still resulted in a large range in predictions between methods.

Furthermore, with the ChIP-seq data used, there was no evidence of improvement using the ESC data although there were limitations due to the number of genes already being minimized. From biclustering results, the addition of ChIP-seq did not do anything to improve the number of interactions predicted, because they are all predicted regardless of the use of ChIP-seq data. Therefore, it was not possible to reduce sample size with the addition of ChIP-seq data using the methods tested. Machine learning methods were found to outperform correlation-based methods although both have limitations such as requiring a cut-off to predicted interactions. The only method that did not require a cut-off was cMonkey2, a biclustering method capable of discovering modules of related genes. As cMonkey2 can find patterns within datasets with small sample sizes as well as not predicting all possible interactions for a transcription factor without the application of a cut-off, it was selected to make an initial prediction for the Sox9 and Runx2 GRNs. Biclustering using other methods was also done, since the number of genes used to predict Sox9 and Runx2 GRNs needs to be minimized before predictions are made. It was also done to observe how the skeletal tissues separated into different biclusters based on patterns in gene expression and discover more evidence of interacting GRNs. The method that had results similar groupings of the tissues to cMonkey, FABIA, was compared to cMonkey2. It seems that biclustering methods that do not separate all the tissues into distinct biclusters may be more biologically relevant since the Sox9 and Runx2 GRNs are likely active to some extent in all three tissues. Therefore, using biclustering methods like cMonkey, FABIA and SAMBA, which do not separate all the tissue may ultimately be more useful moving forward with testing the predicted GRNs in skeletal tissue. Identifying a proficient means of analyzing expression data from skeletal tissue to construct GRNs would contribute to the further study of skeletal development using comparison across multiple species. Ultimately, comparisons can also be made between the molecular mechanisms of normal tissue development and degenerative skeletal conditions. This will allow for properties of skeletal tissue differentiation to be utilized for future therapies.

References

[1] Y. Si, P. Liu, P. Li, and T. P. Brutnell, "Model-based clustering for RNA-seq data," *Bioinformatics*, p. btt632, 2013.

[2] C. Lagacé, A. Perruccio, M. DesMeules, and E. Badley, "The impact of arthritis on canadians," *Arthritis in Canada*, pp. 7–34, 2003.

[3] T. Stafinski and D. Menon, *The Burden of Osteoarthritis in Canada: A Review of Current Literature.* Edmonton: Institute of Health Economics, 2001.

[4] C. C. Wyles, M. T. Houdek, A. Behfar, and R. J. Sierra, "Mesenchymal stem cell therapy for osteoarthritis: current perspectives," *Stem Cells and Cloning: Advances and Applications*, vol. 8, p. 117, 2015.

[5] B. F. Eames, P. T. Sharpe, and J. A. Helms, "Hierarchy revealed in the specification of three skeletal fates by Sox9 and Runx2," *Developmental Biology*, vol. 274, no. 1, pp. 188–200, 2004.

[6] B. F. Eames, L. De La Fuente, and J. A. Helms, "Molecular ontogeny of the skeleton," *Birth Defects Research Part C: Embryo Today: Reviews*, vol. 69, no. 2, pp. 93–101, 2003.

[7] A. Cole, "A review of diversity in the evolution and development of cartilage: the search for the origin of the chondrocyte," *Eur Cell Mater*, vol. 21, pp. 122–129, 2011.

[8] P. Gómez-Picos and B. F. Eames, "On the evolutionary relationship between chondrocytes and osteoblasts," *Frontiers in Genetics*, vol. 6, 2015.

[9] R. Heinrich and S. Schuster, *The Regulation of Cellular Systems.* Springer Science & Business Media, 2012.

[10] H. Wu, T. W. Whitfield, J. Gordon, J. R. Dobson, P. Tai, A. J. van Wijnen, J. L. Stein, G. S. Stein, and J. B. Lian, "Genomic occupancy of Runx2 with global expression profiling identifies a novel dimension to control of osteoblastogenesis," *Genome Biology*, vol. 15, no. 3, p. R52, 2014.

[11] Y. Lu, S. Liang, Y. Mori-Akiyama, D. Chen, B. de Crombrugghe, H. Yasuda, *et al.*, "SOX9 regulates multiple genes in chondrocytes, including genes encoding ECM proteins, ECM modification enzymes, receptors, and transporters," *PloS One*, vol. 10, no. 11, p. e107577, 2014.

[12] R. De Smet and K. Marchal, "Advantages and limitations of current network inference methods," *Nature Reviews Microbiology*, vol. 8, no. 10, pp. 717–729, 2010.

[13] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner, "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles," *PLoS Biology*, vol. 5, no. 1, p. e8, 2007.

[14] D. Marbach, J. C. Costello, R. Küffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, M. Kellis, J. J. Collins, G. Stolovitzky, *et al.*, "Wisdom of crowds for robust gene network inference," *Nature Methods*, vol. 9, no. 8, pp. 796–804, 2012.

[15] F. Geier, J. Timmer, and C. Fleck, "Reconstructing gene-regulatory networks from time series, knockout data, and prior knowledge," *BMC Systems Biology*, vol. 1, no. 1, p. 11, 2007.

[16] D. L. Silver, L. Hou, and W. J. Pavan, "The genetic regulation of pigment cell development," in *Neural Crest Induction and Differentiation*, pp. 155–169, Springer, 2006.

[17] K. Mitra, A.-R. Carvunis, S. K. Ramesh, and T. Ideker, "Integrative approaches for finding modular structure in biological networks," *Nature Reviews Genetics*, vol. 14, no. 10, pp. 719–732, 2013.

[18] A. J. Griffiths, J. H. Miller, D. T. Suzuki, R. C. Lewontin, W. M. Gelbart, *et al.*, *Transcription: An Overview of Gene Regulation in Eukaryotes.* WH Freeman, 2000.

[19] D. J. Reiss, C. L. Plaisier, W.-J. Wu, and N. S. Baliga, "cMonkey2: Automated, systematic, integrated detection of co-regulated gene modules for any organism," *Nucleic Acids Research*, p. gkv300, 2015.

[20] A. J. Hartemink, "Reverse engineering gene regulatory networks," *Nature Biotechnology*, vol. 23, no. 5, pp. 554–555, 2005.

[21] M. Hecker, S. Lambeck, S. Toepfer, E. Van Someren, and R. Guthke, "Gene regulatory network inference: data integration in dynamic models—a review," *Biosystems*, vol. 96, no. 1, pp. 86–103, 2009.

[22] R. Bumgarner, "Overview of DNA microarrays: types, applications, and their future," *Current Protocols in Molecular Biology*, pp. 22–1, 2013.

[23] Z. Wang, M. Gerstein, and M. Snyder, "RNA-seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[24] A. Sîrbu, M. Crane, and H. J. Ruskin, "Data integration for microarrays: Enhanced inference for gene regulatory networks," *Microarrays*, vol. 4, no. 2, pp. 255–269, 2015.

[25] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder, "The transcriptional landscape of the yeast genome defined by RNA sequencing," *Science*, vol. 320, no. 5881, pp. 1344–1349, 2008.

[26] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, *et al.*, "A survey of best practices for RNA-seq data analysis," *Genome Biology*, vol. 17, no. 1, p. 1, 2016.

[27] Y. Liu, J. Zhou, and K. P. White, "RNA-seq differential expression studies: more sequence or more replication?," *Bioinformatics*, vol. 30, no. 3, pp. 301–304, 2014.

[28] S. Ballouz, W. Verleyen, and J. Gillis, "Guidance for RNA-seq co-expression network construction and analysis: safety in numbers," *Bioinformatics*, p. btv118, 2015.

[29] B. Trost, C. A. Moir, Z. E. Gillespie, A. Kusalik, J. A. Mitchell, and C. H. Eskiw, "Concordance between RNA-sequencing data and DNA microarray data in transcriptome analysis of proliferative and quiescent fibroblasts," *Open Science*, vol. 2, no. 9, p. 150402, 2015.

[30] P. J. Park, "ChIP–seq: advantages and challenges of a maturing technology," *Nature Reviews Genetics*, vol. 10, no. 10, pp. 669–680, 2009.

[31] J. Linde, S. Schulze, S. G. Henkel, and R. Guthke, "Data-and knowledge-based modeling of gene regulatory networks: an update," *EXCLI Journal*, 2015.

[32] J. Qin, Y. Hu, F. Xu, H. K. Yalamanchili, and J. Wang, "Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via lasso-type regularization methods," *Methods*, vol. 67, no. 3, pp. 294–303, 2014.

[33] C. Angelini and V. Costa, "Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems," *Frontiers in Cell and Developmental Biology*, vol. 2, 2014.

[34] N. Nariai, Y. Tamada, S. Imoto, and S. Miyano, “Estimating gene regulatory networks and protein–protein interactions of Saccharomyces cerevisiae from multiple genome-wide data,” *Bioinformatics*, vol. 21, no. suppl 2, pp. ii206–ii212, 2005.

[35] G. S. Stein, J. B. Lian, A. J. Van Wijnen, J. L. Stein, M. Montecino, A. Javed, S. K. Zaidi, D. W. Young, J.-Y. Choi, and S. M. Pockwinse, “Runx2 control of organization, assembly and activity of the regulatory machinery for skeletal gene expression,” *Oncogene*, vol. 23, no. 24, pp. 4315–4329, 2004.

[36] G. Zhou, Q. Zheng, F. Engin, E. Munivez, Y. Chen, E. Sebald, D. Krakow, and B. Lee, “Dominance of SOX9 function over RUNX2 during skeletogenesis,” *Proceedings of the National Academy of Sciences*, vol. 103, no. 50, pp. 19004–19009, 2006.

[37] G. SF, *Developmental Biology: Differential Gene Expression.* http://www.ncbi.nlm.nih.gov/books/NBK10061/: Sunderland (MA): Sinauer Associates, 6th edition ed., 2000.

[38] N. J. Schurch, P. Schofield, M. Gierliński, C. Cole, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. G. Simpson, T. Owen-Hughes, *et al.*, “How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?,” *RNA*, vol. 22, no. 6, pp. 839–851, 2016.

[39] M. Rebeiz, N. H. Patel, and V. F. Hinman, “Unraveling the tangled skein: the evolution of transcriptional regulatory networks in development,” *Annual Review of Genomics and Human Genetics*, vol. 16, pp. 103–131, 2015.

[40] C. S. Poultney, A. Greenfield, and R. Bonneau, “Integrated inference and analysis of regulatory networks from multi-level measurements,” *Methods Cell Biology*, vol. 110, pp. 19–56, 2012.

[41] P. D’haeseleer *et al.*, “How does gene expression clustering work?,” *Nature Biotechnology*, vol. 23, no. 12, pp. 1499–1502, 2005.

[42] L. Song, P. Langfelder, and S. Horvath, “Comparison of co-expression measures: mutual information, correlation, and model based indices,” *BMC Bioinformatics*, vol. 13, no. 1, p. 328, 2012.

[43] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, “The mutual information: detecting and evaluating dependencies between variables,” *Bioinformatics*, vol. 18, no. suppl 2, pp. S231–S240, 2002.

[44] S. Anders and W. Huber, “Differential expression analysis for sequence count data,” *Genome Biology*, vol. 11, no. 10, p. 1, 2010.

[45] S. C. Madeira and A. L. Oliveira, “Biclustering algorithms for biological data analysis: a survey,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 1, no. 1, pp. 24–45, 2004.

[46] G. Getz, E. Levine, and E. Domany, “Coupled two-way clustering analysis of gene microarray data,” *Proceedings of the National Academy of Sciences*, vol. 97, no. 22, pp. 12079–12084, 2000.

[47] A. Oghabian, S. Kilpinen, S. Hautaniemi, and E. Czeizler, “Biclustering methods: biological relevance and application in gene expression analysis,” *PloS One*, vol. 9, no. 3, p. 90801, 2014.

[48] L. Li, Y. Guo, W. Wu, Y. Shi, J. Cheng, and S. Tao, “A comparison and evaluation of five biclustering algorithms by quantifying goodness of biclusters for gene expression data,” *BioData Mining*, vol. 5, no. 1, pp. 1–10, 2012.

[49] K. Eren, M. Deveci, O. Küçüktunç, and Ü. V. Çatalyürek, “A comparative analysis of biclustering algorithms for gene expression data,” *Briefings in Bioinformatics*, vol. 14, no. 3, pp. 279–292, 2013.

[50] S. Hochreiter, U. Bodenhofer, M. Heusel, A. Mayr, A. Mitterecker, A. Kasim, T. Khamiakova, S. Van Sanden, D. Lin, W. Talloen, *et al.*, “FABIA: factor analysis for bicluster acquisition,” *Bioinformatics*, vol. 26, no. 12, pp. 1520–1527, 2010.

[51] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, "A systematic comparison and evaluation of biclustering methods for gene expression data," *Bioinformatics*, vol. 22, no. 9, pp. 1122–1129, 2006.

[52] B. K. H. Chia and R. K. M. Karuturi, "Research differential co-expression framework to quantify goodness of biclusters and compare biclustering algorithms," *Algorithms for Molecular Biology*, 2010.

[53] D. Allouche, C. Cierco-Ayrolles, S. de Givry, G. Guillermin, B. Mangin, T. Schiex, J. Vandel, and M. Vignes, "A panel of learning methods for the reconstruction of gene regulatory networks in a systems genetics context," in *Gene Network Inference*, pp. 9–31, Springer, 2013.

[54] H. Li, "Statistical methods for inference of genetic networks and regulatory modules," *UPenn Biostatistics Working Papers*, p. 17, 2007.

[55] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera, and A. Califano, "Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics*, vol. 7, no. Suppl 1, p. S7, 2006.

[56] D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky, "Revealing strengths and weaknesses of methods for gene network inference," *Proceedings of the National Academy of Sciences*, vol. 107, no. 14, pp. 6286–6291, 2010.

[57] A. T. Kwon, H. H. Hoos, and R. Ng, "Inference of transcriptional regulation relationships from gene expression data," *Bioinformatics*, vol. 19, no. 8, pp. 905–912, 2003.

[58] R. Küffner, T. Petri, P. Tavakkolkhah, L. Windhager, and R. Zimmer, "Inferring gene regulatory networks by ANOVA," *Bioinformatics*, vol. 28, no. 10, pp. 1376–1382, 2012.

[59] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, "Inferring regulatory networks from expression data using tree-based methods," *PloS One*, vol. 5, no. 9, p. e12776, 2010.

[60] R. S. Sekhon, R. Briskine, C. N. Hirsch, C. L. Myers, N. M. Springer, C. R. Buell, N. de Leon, and S. M. Kaeppler, "Maize gene atlas developed by RNA sequencing and comparative evaluation of transcriptomes based on RNA sequencing and microarrays," *PLoS One*, vol. 8, no. 4, p. e61005, 2013.

[61] O. D. Iancu, S. Kawane, D. Bottomly, R. Searles, R. Hitzemann, and S. McWeeney, "Utilizing RNA-seq data for de novo coexpression network inference," *Bioinformatics*, vol. 28, no. 12, pp. 1592–1597, 2012.

[62] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[63] J. D. Allen, Y. Xie, M. Chen, L. Girard, and G. Xiao, "Comparing statistical methods for constructing large scale gene networks," *PloS One*, vol. 7, no. 1, p. e29348, 2012.

[64] S.-I. Consortium *et al.*, "A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium," *Nature Biotechnology*, vol. 32, no. 9, pp. 903–914, 2014.

[65] D. Potier, K. Davie, G. Hulselmans, M. N. Sanchez, L. Haagen, D. Koldere, A. Celik, P. Geurts, V. Christiaens, S. Aerts, *et al.*, "Mapping gene regulatory networks in Drosophila eye development by large-scale transcriptome perturbations and motif inference," *Cell Reports*, vol. 9, no. 6, pp. 2290–2303, 2014.

[66] J. Ruyssinck, V. A. Huynh-Thu, P. Geurts, T. Dhaene, P. Demeester, and Y. Saeys, "Nimefi: gene regulatory network inference using multiple ensemble feature importance algorithms," *PloS One*, vol. 9, no. 3, p. e92709, 2014.

[67] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A gene-coexpression network for global discovery of conserved genetic modules," *Science*, vol. 302, no. 5643, pp. 249–255, 2003.

[68] D. J. Reiss, N. S. Baliga, and R. Bonneau, “Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks,” *BMC Bioinformatics*, vol. 7, no. 1, p. 280, 2006.

[69] K. Lemmens, T. De Bie, T. Dhollander, S. C. De Keersmaecker, I. M. Thijs, G. Schoofs, A. De Weerdt, B. De Moor, J. Vanderleyden, J. Collado-Vides, *et al.*, “DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli,” *Genome Biology*, vol. 10, no. 3, p. R27, 2009.

[70] F. M. Alakwaa, N. H. Solouma, and Y. M. Kadah, “Construction of gene regulatory networks using biclustering and bayesian networks,” *Theoretical Biology and Medical Modelling*, vol. 8, no. 1, p. 39, 2011.

[71] C. Huttenhower, K. T. Mutungu, N. Indik, W. Yang, M. Schroeder, J. J. Forman, O. G. Troyanskaya, and H. A. Coller, “Detailing regulatory networks through large scale data integration,” *Bioinformatics*, vol. 25, no. 24, pp. 3267–3274, 2009.

[72] F. M. Alakwaa, “Modeling of gene regulatory networks: A literature review,” *Journal of Computational Systems Biology*, vol. 1, no. 1, p. 1, 2014.

[73] K. Raza and R. Parveen, “Reconstruction of gene regulatory network of colon cancer using information theoretic approach,” *Confluence 2013: The Next Generation Information Technology Summit (4th International Conference)*, pp. 461–466, 2013.

[74] F. Zhu, L. Shi, J. D. Engel, and Y. Guan, “Regulatory network inferred using expression data of small sample size: application and validation in erythroid system,” *Bioinformatics*, p. btv186, 2015.

[75] G. Altay, “Empirically determining the sample size for large-scale gene network inference algorithms,” *IET Systems Biology*, vol. 6, no. 2, pp. 35–43, 2012.

[76] G. Altay and F. Emmert-Streib, “Inferring the conservative causal core of gene regulatory networks,” *BMC Systems Biology*, vol. 4, no. 1, p. 132, 2010.

[77] D. Djordjevic, A. Yang, A. Zadoorian, K. Rungrugeecharoen, and J. W. Ho, “How difficult is inference of mammalian causal gene regulatory networks?,” *PloS One*, 2014.

[78] E. Korpelainen, J. Tuimala, P. Somervuo, M. Huss, and G. Wong, *RNA-seq Data Analysis: A Practical Approach.* CRC Press, 2014.

[79] S. Andrews *et al.*, “FastQC: A quality control tool for high throughput sequence data,” *Reference Source*, 2010.

[80] A. M. Bolger, M. Lohse, and B. Usadel, “Trimmomatic: a flexible trimmer for illumina sequence data,” *Bioinformatics*, p. btu170, 2014.

[81] C. Trapnell, L. Pachter, and S. L. Salzberg, “Tophat: discovering splice junctions with RNA-seq,” *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, 2009.

[82] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome,” *Genome Biology*, vol. 10, no. 3, p. 1, 2009.

[83] S. Anders, P. T. Pyl, and W. Huber, “HTseq–a Python framework to work with high-throughput sequencing data,” *Bioinformatics*, p. btu638, 2014.

[84] C. Trapnell, 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 abundance estimation from RNA-seq reveals thousands of new transcripts and switching among isoforms,” *Nature Biotechnology*, vol. 28, no. 5, p. 511, 2010.

[85] M.-A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, *et al.*, "A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis," *Briefings in Bioinformatics*, vol. 14, no. 6, pp. 671–683, 2013.

[86] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," *Genome Biology*, vol. 11, no. 3, p. 1, 2010.

[87] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome," *BMC Bioinformatics*, vol. 12, no. 1, p. 1, 2011.

[88] N. Leng, J. A. Dawson, J. A. Thomson, V. Ruotti, A. I. Rissman, B. M. Smits, J. D. Haag, M. N. Gould, R. M. Stewart, and C. Kendziorski, "EBSeq: an empirical bayes hierarchical model for inference in RNA-seq experiments," *Bioinformatics*, vol. 29, no. 8, pp. 1035–1043, 2013.

[89] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[90] H. L. Turner, T. C. Bailey, W. J. Krzanowski, and C. A. Hemingway, "Biclustering models for structured microarray data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 2, no. 4, pp. 316–329, 2005.

[91] A. Tanay, R. Sharan, and R. Shamir, "Discovering statistically significant biclusters in gene expression data," *Bioinformatics*, vol. 18, no. suppl 1, pp. S136–S144, 2002.

[92] S. Kaiser, R. Santamaria, R. Theron, L. Quintales, and F. Leisch, "biclust: Bicluster algorithms," *R package version 0.7*, vol. 2, 2009.

[93] R. Shamir, A. Maron-Katz, A. Tanay, C. Linhart, I. Steinfeld, R. Sharan, Y. Shiloh, and R. Elkon, "EXPANDER–an integrative program suite for microarray data analysis," *BMC Bioinformatics*, vol. 6, no. 1, p. 232, 2005.

[94] K. Glass and M. Girvan, "Annotation enrichment analysis: an alternative method for evaluating the functional properties of gene sets," *arXiv preprint arXiv:1208.4127*, 2012.

[95] G. F. Berriz, O. D. King, B. Bryant, C. Sander, and F. P. Roth, "Characterizing gene sets with funcassociate," *Bioinformatics*, vol. 19, no. 18, pp. 2502–2504, 2003.

[96] E. H. Davidson and D. H. Erwin, "Gene regulatory networks and the evolution of animal body plans," *Science*, vol. 311, no. 5762, pp. 796–800, 2006.

[97] D. H. Erwin and E. H. Davidson, "The evolution of hierarchical gene regulatory networks," *Nature Reviews Genetics*, vol. 10, no. 2, pp. 141–148, 2009.

[98] N. Omranian, J. M. Eloundou-Mbebi, B. Mueller-Roeber, and Z. Nikoloski, "Gene regulatory network inference using fused lasso on multiple data sets," *Scientific Reports*, vol. 6, 2016.

[99] D. Guan, J. Shao, Y. Deng, P. Wang, Z. Zhao, Y. Liang, J. Wang, and B. Yan, "CMGRN: a web server for constructing multi-level gene regulatory networks using ChIP-seq and gene expression data," *Bioinformatics*, p. btt761, 2014.

[100] H. L. Sladitschek and P. A. Neveu, "The bimodally expressed microRNA mir-142 gates exit from pluripotency," *Molecular Systems Biology*, vol. 11, no. 12, p. 850, 2015.

[101] H. Xu, Y.-S. Ang, A. Sevilla, I. R. Lemischka, and A. Ma'ayan, "Construction and validation of a regulatory network for pluripotency and self-renewal of mouse embryonic stem cells," *PLoS Computational Biology*, vol. 10, no. 8, p. e1003777, 2014.

[102] A. K. K. Teo, S. J. Arnold, M. W. Trotter, S. Brown, L. T. Ang, Z. Chng, E. J. Robertson, N. R. Dunn, and L. Vallier, "Pluripotency factors regulate definitive endoderm specification through eomesodermin," *Genes & Development*, vol. 25, no. 3, pp. 238–250, 2011.

[103] S. Muñoz Descalzo, P. Rué, J. Garcia-Ojalvo, and A. M. Arias, "Correlations between the levels of oct4 and nanog as a signature for naive pluripotency in mouse embryonic stem cells," *Stem Cells*, vol. 30, no. 12, pp. 2683–2691, 2012.

[104] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[105] J. M. Lingeman and D. Shasha, *Network Inference in Molecular Biology: A Hands-on Framework*. Springer Science & Business Media, 2012.

[106] A. Irrthum, L. Wehenkel, P. Geurts, *et al.*, "Inferring regulatory networks from expression data using tree-based methods," *PloS One*, vol. 5, no. 9, p. e12776, 2010.

[107] W. A. Whyte, D. A. Orlando, D. Hnisz, B. J. Abraham, C. Y. Lin, M. H. Kagey, P. B. Rahl, T. I. Lee, and R. A. Young, "Master transcription factors and mediator establish super-enhancers at key cell identity genes," *Cell*, vol. 153, no. 2, pp. 307–319, 2013.

[108] F. Petralia, P. Wang, J. Yang, and Z. Tu, "Integrative random forest for gene regulatory network inference," *Bioinformatics*, vol. 31, no. 12, pp. i197–i205, 2015.

[109] R. Bonneau, D. J. Reiss, P. Shannon, M. Facciotti, L. Hood, N. S. Baliga, and V. Thorsson, "The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo," *Genome Biology*, vol. 7, no. 5, p. 1, 2006.

[110] J.-N. Juang, S. J. Shiau, and W. Wu, "A hybrid parameter estimation algorithm for S-system model of gene regulatory networks," *The Journal of the Astronautical Sciences*, vol. 60, no. 3-4, pp. 559–576, 2013.

[111] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[112] L. M. Shaw and B. R. Olsen, "FACIT collagens: diverse molecular bridges in extracellular matrices," *Trends in Biochemical Sciences*, vol. 16, pp. 191–194, 1991.

[113] G. F. Pierce, T. A. Mustoe, B. W. Altrock, T. F. Deuel, and A. Thomason, "Role of platelet-derived growth factor in wound healing," *Journal of Cellular Biochemistry*, vol. 45, no. 4, pp. 319–326, 1991.

[114] S. Ohba, X. He, H. Hojo, and A. P. McMahon, "Distinct transcriptional programs underlie Sox9 regulation of the mammalian chondrocyte," *Cell Reports*, vol. 12, no. 2, pp. 229–243, 2015.

[115] X. He, S. Ohba, H. Hojo, and A. P. McMahon, "Ap-1 family members act with Sox9 to promote chondrocyte hypertrophy," *Development*, pp. dev–134502, 2016.

[116] V. F. Hinman, A. T. Nguyen, R. A. Cameron, and E. H. Davidson, "Developmental gene regulatory network architecture across 500 million years of echinoderm evolution," *Proceedings of the National Academy of Sciences*, vol. 100, no. 23, pp. 13356–13361, 2003.

[117] S. Fisher and T. Franz-Odendaal, "Evolution of the bone gene regulatory network," *Current opinion in genetics & development*, vol. 22, no. 4, pp. 390–397, 2012.

Appendix A

Differential Expression

Table A.1: Differential expression results for the genes most up-regulated in each tissue compared to the other two tissues. Table shows the log2 fold changes, and the genes are sorted based on the minimum log2 fold change. Gene counts for each tissue for each gene are also shown for all three replicates.

Genes most up-reg. (Mature)	logFC (IMM)	logFC.2 (BONE)	MinFC	MAT1	MAT2	MAT3
2200002D01Rik	4.810602	4.325495	4.325495	14	30	169
Abtb1	2.307323	2.758312	2.307323	228	375	1437
AI661453	9.731659	4.970652	4.970652	75	95	169
Apba2	2.772514	4.933579	2.772514	159	106	136
Apod	4.274054	3.931457	3.931457	322	15	66
Arap2	3.509914	2.544457	2.544457	201	250	312
Arsi	2.588097	9.106185	2.588097	1247	1096	3483
Atp6v0a4	3.1244	6.211701	3.1244	75	75	195
Cabp1	4.020776	3.629571	3.629571	15	57	81
Catsper4	7.432629	9.854591	7.432629	32	61	276
Ccdc80	5.824458	5.684647	5.684647	1540	987	4317
Cdh19	5.452143	6.11394	5.452143	118	47	154
Cds1	5.896676	6.0197	5.896676	296	247	955
Col10a1	11.939181	10.279139	10.279139	11410	12681	61834
Comp	2.103194	10.501231	2.103194	10482	6615	25962
Corin	6.377141	4.145676	4.145676	227	63	254
Cpa6	2.50223	5.649675	2.50223	303	179	632
Cttnbp2	2.36133	3.008105	2.36133	363	204	213
Cyp11a1	7.433346	9.15161	7.433346	6	26	195
Dach1	3.456439	4.079639	3.456439	206	107	154
Ddn	6.099884	6.647466	6.099884	96	54	147
Dkk2	5.313394	3.635593	3.635593	1208	133	114
Dusp5	4.505456	2.622123	2.622123	198	444	507
Eps8l2	2.773986	8.710346	2.773986	181	211	1235
Fbln5	7.485709	3.688678	3.688678	591	173	162
Fcer2a	3.02806	6.071101	3.02806	84	45	77
Gcnt2	5.115782	3.634691	3.634691	357	141	367
Gm15712	2.210871	4.338054	2.210871	130	119	261
Gm27249	7.861018	7.86304	7.861018	15	14	66
Hhip	4.563887	4.216572	4.216572	1476	593	29
Hoxa11	5.850657	9.840089	5.850657	353	11	0
Ihh	8.42466	5.636227	5.636227	2030	940	3303
Isg20	5.472439	8.376665	5.472439	15	17	103
Itga1	2.534262	4.371699	2.534262	365	97	456
Itga7	6.422603	6.859952	6.422603	165	229	702
Itgb8	3.270052	3.406617	3.270052	146	118	162
Kirrel3	3.674392	2.740369	2.740369	69	148	349
Klhl31	5.004721	3.320505	3.320505	106	191	184
Lemd1	6.390953	7.092974	6.390953	36	28	110
Lipg	5.799712	5.297901	5.297901	100	65	110

Lypd6	4.83222	4.829945	4.829945	313	49	26
Mbp	3.727502	3.015208	3.015208	282	46	217
Mcoln3	7.906811	5.478464	5.478464	84	20	209
Nfasc	4.339463	10.013455	4.339463	54	89	268
Nhej1	2.580367	2.866777	2.580367	195	205	621
Nim1k	2.307411	5.40575	2.307411	914	602	724
Nt5dc1	4.575578	3.411777	3.411777	437	316	1345
Parm1	5.082532	2.677354	2.677354	1164	411	603
Pde11a	5.848737	6.284956	5.848737	149	73	228
Prkg2	2.115171	3.531984	2.115171	1872	1197	2182
Prom1	4.241879	4.582153	4.241879	752	226	577
Prss50	9.18596	6.790912	6.790912	24	40	169
Pth1r	3.707333	3.359919	3.359919	16229	34093	92554
Rapgef3	2.424302	2.248307	2.248307	120	226	261
Rasgrf2	5.045018	3.404607	3.404607	139	85	129
Rbms3	2.348274	2.144482	2.144482	971	774	621
Rgs7bp	3.013869	2.952551	2.952551	250	207	375
RP24-222G3.1	4.275858	8.023105	4.275858	103	108	132
RP24-475O6.1	5.918157	3.898769	3.898769	21	57	125
Rpl39l	2.714879	2.858102	2.714879	283	181	55
Rspo3	3.66622	5.536452	3.66622	2179	580	382
Serinc5	4.309275	3.44221	3.44221	4940	1206	3461
Sidt2	2.183537	2.022157	2.022157	1710	1219	2557
Slc17a9	2.762031	3.11076	2.762031	382	1235	5780
Slc35g1	2.995354	4.103005	2.995354	185	66	478
Slc43a2	4.415567	2.987216	2.987216	97	86	246
Slco2b1	9.611415	3.167916	3.167916	44	53	217
Stmn2	8.643946	3.812662	3.812662	772	79	18
Stra6	5.582603	3.876393	3.876393	52	24	33
Syna	9.80765	6.559881	6.559881	43	72	242
Thrb	4.601468	4.164985	4.164985	268	133	268
Tmie	2.374371	3.537102	2.374371	132	185	364
Tnmd	10.822619	5.357013	5.357013	527	22	169
Ttll3	3.098853	4.059393	3.098853	544	687	1569
Wnt11	3.069925	4.804596	3.069925	278	56	125
Wnt5b	3.636137	3.751825	3.636137	427	370	955
Zfp185	3.866001	3.874776	3.866001	376	17	169
Znhit6	2.329638	2.533332	2.329638	598	754	3259
Genes most up-reg. (Immature)	logFC (Mature)	logFC.1 (Bone)	MinFC	IMA1	IMA2	IMA3
C4b	9.249958	9.754976	9.249958	252	272	508
Car9	7.37604	8.014918	7.37604	180	858	54
Trhr2	7.147005	9.51244	7.147005	43	10	185
5730596B20Rik	6.166898	8.51908	6.166898	58	25	38
1700049E15Rik	6.952792	5.77329	5.77329	55	126	32
Trabd2b	5.57442	8.84187	5.57442	1018	1216	1606
4933400C23Rik	5.807299	5.541725	5.541725	28	28	39
Fmod	5.507848	7.248587	5.507848	3168	15227	751
Plekha4	4.921619	7.574858	4.921619	93	73	236
Xlr3c	4.910256	8.079711	4.910256	22	6	60
Gm17225	4.905818	8.075264	4.905818	56	29	5
Lin7a	4.773595	5.010597	4.773595	221	924	114
RP23-198G19.1	4.644177	6.483578	4.644177	383	351	803
Sfrp5	4.564137	9.568965	4.564137	139	422	379

Serpina3n	4.472844	10.781695	4.472844	901	640	581
Snph	4.458046	5.71735	4.458046	72	54	75
Gm27202	4.283741	9.502154	4.283741	77	51	110
Hist1h1e	4.21974	5.114965	4.21974	150	126	99
Ucma	4.182678	8.812778	4.182678	189	293	60
Gm13111	4.13892	9.263066	4.13892	67	56	79
Smoc1	4.016777	7.231947	4.016777	291	994	908
Scn9a	3.957163	7.702003	3.957163	106	23	289
RP23-448H3.2	3.949846	4.836995	3.949846	17369	23260	11316
Gm26945	3.947416	10.849907	3.947416	189	170	249
Gm14776	5.085592	3.903751	3.903751	141	195	195
Edn2	5.777074	3.90223	3.90223	25	44	39
Lrrc75b	3.883743	5.447106	3.883743	225	147	984
Trank1	3.84283	5.596119	3.84283	18	78	42
Fam198a	3.810586	3.853473	3.810586	388	134	348
Gm16152	3.680431	5.791181	3.680431	116	112	326
4930545L23Rik	4.194338	3.670655	3.670655	113	102	9
Chdh	4.428763	3.668352	3.668352	40	35	120
Gdf5	4.576696	3.663428	3.663428	93	411	87
Gm25224	3.658465	6.995592	3.658465	73	64	17
Sapcd2	3.48673	4.272382	3.48673	283	543	482
Rbpjl	3.480056	9.808766	3.480056	13	8	17
Fbxo2	3.474793	5.591904	3.474793	82	107	77
Mybl1	3.449473	5.444433	3.449473	243	168	163
Pthlh	3.400384	7.428276	3.400384	332	462	76
Vwa1	3.343207	5.533101	3.343207	200	1176	67
Hoxd4	3.321618	7.799469	3.321618	197	374	102
Chst3	3.295191	4.328497	3.295191	84	287	49
Mfsd2a	3.23883	5.624131	3.23883	57	150	25
Flrt1	3.197749	5.260034	3.197749	175	98	127
2600014E21Rik	3.183158	8.853394	3.183158	268	260	27
Gm16150	3.179214	4.457532	3.179214	88	83	123
Hist1h2ap	3.168437	3.492985	3.168437	896	1096	1346
Gm16326	4.556545	3.159546	3.159546	19	21	57
Fam19a2	3.147345	4.846471	3.147345	101	339	74
Adhfe1	3.136526	3.727971	3.136526	98	47	61
Ephx1	3.414469	3.118784	3.118784	19	97	30
Aim1	4.114407	3.116736	3.116736	55	172	46
Il1rapl1	3.109017	3.7959	3.109017	73	165	134
Col19a1	3.098795	8.033677	3.098795	30	22	34
Unc80	3.085252	7.021364	3.085252	167	121	114
Hist1h1d	3.077059	3.912552	3.077059	23	41	32
Cbr2	3.040547	7.64866	3.040547	150	73	188
1700006J14Rik	3.016552	3.241811	3.016552	121	93	20
Hist1h2ao	3.016257	4.788247	3.016257	144	112	297
Clmn	3.013547	3.99686	3.013547	1088	1529	1037
Prph	4.038076	3.011818	3.011818	18	46	26
Gm16183	3.004755	5.306423	3.004755	205	64	598
Osmr	2.981151	5.183422	2.981151	443	473	100
BC006965	2.939796	8.337577	2.939796	1608	470	6280
Mak	2.896029	6.342479	2.896029	47	73	47
Inhba	2.895669	3.71397	2.895669	92	714	88
Syne4	2.890169	3.699997	2.890169	337	419	546
Rdh12	3.257836	2.874022	2.874022	98	86	34

Casc5	2.81594	3.563772	2.81594	492	499	450
Rgma	2.787285	4.593662	2.787285	151	121	141
Hist1h2ae	3.230079	2.772671	2.772671	168	152	46
Rap1gap	2.7693	4.033521	2.7693	271	106	131
Lrig3	2.748966	4.242207	2.748966	988	1440	1797
Cfap44	3.993681	2.740413	2.740413	68	163	117
RP23-204I16.3	2.714684	6.706832	2.714684	84	32	178
Tbx5	2.70007	10.711909	2.70007	753	1194	1509
Meg3	2.6913	7.929196	2.6913	59621	61080	122676
Ppp1r9a	2.687894	4.375313	2.687894	4834	3596	4133
BC039771	2.623559	2.957934	2.623559	240	267	252
Fxyd6	2.620406	4.260963	2.620406	744	1861	199
Rab36	3.284063	2.60221	2.60221	47	24	36
Matn4	2.569946	7.79374	2.569946	4685	2079	14224
Mtap7d3	2.559913	6.685392	2.559913	1897	1144	2320
Cpm	2.532306	5.345226	2.532306	1127	2339	221
Fan1	2.531	2.99462	2.531	130	101	112
Hoxd9	2.523183	10.549144	2.523183	396	1218	202
6430550D23Rik	2.516636	3.023248	2.516636	76	48	152
Ndufa4l2	2.515878	2.769244	2.515878	257	741	328
B4galnt4	2.489832	4.128476	2.489832	156	185	195
Iqgap3	2.476717	3.374537	2.476717	331	658	328
Mirg	2.475712	7.173399	2.475712	3692	3702	5337
Sox11	2.46664	3.117145	2.46664	5568	10757	5154
H1fx	2.45962	2.758729	2.45962	196	222	417
Gm26603	2.740352	2.450303	2.450303	531	524	400
Rin3	2.44015	2.645907	2.44015	609	561	1273
Fam19a5	2.43538	2.438948	2.43538	124	225	125
Psrc1	2.434505	2.879373	2.434505	200	210	264
Chadl	2.433461	6.255045	2.433461	384	355	1258
Nckap5	2.425943	4.648875	2.425943	410	294	1731
Sox8	2.421676	5.808876	2.421676	587	468	706
Dnm1	2.687658	2.420465	2.420465	410	842	2091
Cdca2	2.42033	3.39143	2.42033	605	566	1182
Prkcz	2.419146	2.872238	2.419146	217	325	677
P4ha3	2.417101	3.263347	2.417101	818	277	1124
Prdm16	2.412959	3.8947	2.412959	335	338	424
Aspm	2.41162	2.766846	2.41162	370	477	178
Mxd3	2.386008	3.645875	2.386008	173	216	322
Arsj	2.363737	6.417907	2.363737	383	185	182
Plcb1	2.362116	2.982614	2.362116	753	1154	200
Zgrf1	2.361901	3.616637	2.361901	611	444	879
C530008M17Rik	2.329135	3.908319	2.329135	398	557	291
Cntn2	2.30023	3.253513	2.30023	190	174	111
Usp51	2.279624	2.417398	2.279624	167	91	202
Enkd1	2.252749	3.467269	2.252749	282	258	731
Nfix	2.245716	2.253558	2.245716	1059	1354	3041
Gpc6	2.201415	2.407498	2.201415	5055	8699	3027
Gabre	2.201119	2.545379	2.201119	147	207	378
Aff2	2.196276	4.154827	2.196276	181	218	340
Arhgef39	2.192644	3.153019	2.192644	310	300	162
Ikzf4	2.180516	2.318342	2.180516	217	235	270
Itpr3	2.431129	2.180217	2.180217	182	219	361
RP23-23C9.1	2.466135	2.16361	2.16361	77	70	52

Ninj1	2.347599	2.157327	2.157327	492	939	377
Map1a	2.142191	6.21035	2.142191	561	452	1714
Sox5	2.132289	5.648823	2.132289	4120	3008	4891
Rian	2.117474	5.605464	2.117474	182	91	310
Limch1	3.424258	2.113269	2.113269	1102	623	1168
Cep135	2.101704	2.901145	2.101704	358	325	551
RP24-338G10.1	2.091957	2.194718	2.091957	221	179	496
Rad51ap1	2.089535	3.529236	2.089535	3799	4413	2856
Trerf1	2.088411	4.665902	2.088411	566	581	1259
Wdr90	2.08618	3.698569	2.08618	473	440	621
Kif22	2.051434	2.404828	2.051434	339	402	643
Fam53b	2.045293	2.486744	2.045293	237	283	258
Lphn3	2.042898	4.320772	2.042898	462	353	1062
Mroh2a	2.03119	2.856586	2.03119	273	369	803
Dlk1	2.010515	4.470405	2.010515	4485	5108	3010
Brca2	2.005431	2.645235	2.005431	506	366	564
Tube1	2.005067	2.400683	2.005067	193	140	221
Genes most up-reg. (Bone)	logFC.1 (IMM)	logFC.2 (Mature)	MinFC	BON1	BON2	BON3
Lhx8	9.136745	9.071637	9.071637	503	583	568
AI606473	9.311965	9.048097	9.048097	56	94	63
Lppr5	7.723725	7.460771	7.460771	20	23	27
Gpr50	9.755557	7.380689	7.380689	87	110	94
Dlx1	7.114804	7.074487	7.074487	31	214	167
Dlx2	10.488019	6.834576	6.834576	49	171	264
Pax3	6.448011	6.592669	6.448011	1475	29	69
Pitx1	7.663759	5.89066	5.89066	276	231	235
BC064078	5.884778	8.370188	5.884778	61	30	42
Tmem132d	8.860258	5.705026	5.705026	50	62	44
Tnfaip8l3	5.684269	5.148943	5.148943	23	19	19
Cd1d2	4.993531	5.484816	4.993531	37	19	33
Crym	12.475711	4.973768	4.973768	587	716	614
Msx1	4.967471	4.856061	4.856061	1232	1175	911
Lhx6	4.70797	5.003821	4.70797	96	108	92
Hist2h3c2	5.897839	4.641575	4.641575	177	125	130
Ovol2	8.662749	4.610689	4.610689	27	42	66
Chgb	7.869674	4.593058	4.593058	134	238	128
Gal	11.428436	4.38991	4.38991	329	322	277
Clec2g	11.014425	4.359401	4.359401	254	204	238
Bcl2a1a	7.53391	4.354274	4.354274	13	25	24
Grm4	10.491861	4.341922	4.341922	238	136	112
Syt6	10.862196	4.331178	4.331178	240	214	172
Arl4d	6.406811	4.316271	4.316271	394	369	333
Ccdc121	5.430921	4.27946	4.27946	52	38	38
Madcam1	8.294017	4.266542	4.266542	10	51	44
Mmp8	5.648194	4.215696	4.215696	20	54	27
5031410I06Rik	5.270554	4.215152	4.215152	13	35	30
Car1	7.359612	4.180348	4.180348	27	15	14
Ramp1	7.053807	4.154912	4.154912	45	45	48
Drd1a	5.511976	4.048314	4.048314	32	41	30
Aifm3	4.213354	4.012438	4.012438	18	67	44
Ranbp3l	7.280341	4.010069	4.010069	1424	2188	1593
Gm16332	6.602178	3.979735	3.979735	52	31	40
Fetub	10.201013	3.956426	3.956426	107	178	111

Cdh12	4.498081	3.941303	3.941303	9	34	23
Mepe	8.423785	3.903966	3.903966	43	38	35
Gprin3	6.122325	3.74407	3.74407	147	173	109
Lingo3	4.349777	3.706382	3.706382	57	40	37
Fhod3	5.604989	3.669986	3.669986	311	558	350
Calcr	8.71084	3.666315	3.666315	18	63	60
Pcbd1	6.097716	3.660692	3.660692	330	329	256
Wif1	3.79022	3.604383	3.604383	3097	3464	3472
Bhlha15	4.253196	3.575559	3.575559	83	100	78
2310030G06Rik	6.584264	3.550023	3.550023	80	104	102
Clec4a2	7.652997	3.537115	3.537115	58	190	174
Insc	11.092047	3.52579	3.52579	1020	847	769
Sall1	5.580406	3.491765	3.491765	34	56	55
Rab38	4.17576	3.440497	3.440497	39	13	36
Cmbl	6.604614	3.429892	3.429892	184	179	171
Kcnj3	10.411635	3.380355	3.380355	140	157	161
Cdh23	4.270056	3.363996	3.363996	183	96	87
Srgn	7.244619	3.346306	3.346306	221	277	290
Fat3	5.825276	3.335312	3.335312	1528	1257	1339
Car3	10.000694	3.315939	3.315939	11527	8698	7192
Cd59a	4.176947	3.309338	3.309338	407	419	333
Gstm6	5.085284	3.280339	3.280339	108	69	88
Slc2a12	5.805997	3.249088	3.249088	321	286	324
Cd1d1	8.359185	3.245766	3.245766	905	832	962
Col1a2	7.363067	3.164504	3.164504	648095	727263	598100
Plekha2	4.999655	3.115288	3.115288	171	161	200
Ctsh	4.964961	3.113365	3.113365	771	916	638
Ccl9	4.109489	3.076699	3.076699	319	1051	857
Mob3b	6.138554	3.066977	3.066977	409	558	496
Foxf1	4.240281	2.971339	2.971339	63	39	52
2010300C02Rik	4.704031	2.942642	2.942642	214	208	154
Prex1	3.759317	2.898032	2.898032	3530	3493	3353
Satb2	9.145641	2.881265	2.881265	3161	2658	2501
Olfml3	5.910038	2.86078	2.86078	6343	6368	5897
Gpr133	9.851301	2.828861	2.828861	1605	1418	1692
Ibsp	10.804709	2.825338	2.825338	290255	269321	234039
Dner	9.627539	2.82529	2.82529	709	119	182
Fyn	4.836725	2.811189	2.811189	2378	4191	3765
RP23-388I22.1	2.799335	3.715813	2.799335	311	298	276
Scn3a	10.0203	2.795545	2.795545	1085	575	607
Smad6	2.950231	2.784969	2.784969	685	658	655
Tdrp	9.43823	2.779462	2.779462	303	270	270
Dcn	6.528736	2.766027	2.766027	3066	4783	4304
Dkk1	9.309895	2.753702	2.753702	2032	1620	2540
Sparc	3.298517	2.751291	2.751291	117578	136556	116792
Ncf1	7.888247	2.720026	2.720026	695	802	622
Kazald1	2.711625	3.189502	2.711625	5946	5801	4817
Pard6g	3.231024	2.700203	2.700203	2140	2380	2666
Dapk2	8.328763	2.672389	2.672389	756	921	770
Ifitm5	7.713749	2.66909	2.66909	5328	5934	5751
Aldh1b1	7.433072	2.645313	2.645313	306	263	169
Serpinf1	3.797382	2.609135	2.609135	854	1189	1008
Prcp	3.700088	2.608368	2.608368	1995	1130	997
Cd109	4.285435	2.598995	2.598995	1412	2924	2377

Shb	3.031154	2.595782	2.595782	247	277	248
Ccdc149	4.174977	2.559771	2.559771	160	143	134
Rassf4	3.136354	2.529738	2.529738	1053	1047	1042
Hrc	6.388662	2.526006	2.526006	1833	1790	1386
Magi2	4.157514	2.503774	2.503774	578	526	456
Hpcal1	3.393296	2.497773	2.497773	785	491	588
Kctd12b	7.226238	2.495244	2.495244	697	655	552
Ddx59	3.435503	2.478726	2.478726	412	416	368
Tmem119	6.840728	2.439901	2.439901	6488	6845	6968
Bmp4	4.79948	2.38646	2.38646	506	333	293
Ptprr	5.227636	2.373628	2.373628	185	224	188
Ttc7	2.348966	2.700116	2.348966	605	681	827
Phex	9.119497	2.347742	2.347742	5037	4701	3690
Cgref1	3.532644	2.32064	2.32064	5082	7422	5323
Ell2	4.420378	2.317539	2.317539	2048	2737	2523
Frmd4b	3.487581	2.309628	2.309628	1199	1018	943
Dlx3	9.169601	2.272572	2.272572	1848	1855	1703
Ust	3.1538	2.263212	2.263212	828	777	793
Gm15417	2.243336	2.701654	2.243336	89	75	61
Pdgfrl	3.555791	2.238034	2.238034	1237	1834	1098
Hist1h1c	2.228665	2.479309	2.228665	1219	1422	1313
Smim14	3.540073	2.220217	2.220217	4141	4548	4152
Fam109b	2.27056	2.205267	2.205267	744	534	622
Ano1	5.236643	2.15554	2.15554	1373	1819	1591
2810025M15Rik	3.215811	2.145351	2.145351	1373	1262	1261
Sema3b	3.614571	2.132613	2.132613	1268	1614	1298
Stk17b	3.705852	2.130371	2.130371	1047	945	1082
Mylk	2.12619	2.165814	2.12619	1121	685	1170
Cd63	2.676787	2.126146	2.126146	33402	22627	24819
BC027582	4.189821	2.111899	2.111899	160	142	160
Sh3bgrl2	4.195357	2.109408	2.109408	565	613	498
Ankrd6	3.574807	2.097861	2.097861	776	743	869
Galm	2.083556	2.308364	2.083556	331	304	313
Slc7a2	2.661316	2.082962	2.082962	1387	1811	2053
Inpp4a	2.673503	2.051824	2.051824	697	672	680
Cadm1	4.06709	2.038146	2.038146	5659	5047	5394
Fam46a	2.369924	2.020256	2.020256	12272	9854	9493
Pls3	4.50806	2.017227	2.017227	3651	4598	5573
Fras1	4.727238	2.009784	2.009784	1099	1061	1142
Sh3pxd2b	4.424895	2.006615	2.006615	3237	3047	3045

Appendix B

Gene Ontology Enrichment Analysis

Table B.1: Gene Ontology results for genes present in FABIA biclusters. N is the number of genes with the associated GO term in a bicluster while X is the total number of genes in the background set that are associated with the GO term.

N	X	p-value	P_adj	attrib ID	attrib name
6	6	3.5879704167483001E-7	1E-3	GO:0005833	hemoglobin complex
6	6	3.5879704167483001E-7	1E-3	GO:0090193	positive regulation of glomerulus development
7	8	2.2327690834081999E-7	<0.001	GO:0090192	regulation of glomerulus development
8	14	4.77256147161806E-6	0.02	GO:0060351	cartilage development involved in endochondral bone morphogenesis
8	14	4.77256147161806E-6	0.02	GO:0071622	regulation of granulocyte chemotaxis
11	21	2.35595435838962E-7	<0.001	GO:0090184	positive regulation of kidney development
11	22	4.3529425554940897E-7	2E-3	GO:0005201	extracellular matrix structural constituent
16	33	1.7340661223660601E-9	<0.001	GO:0031225	anchored component of membrane
10	21	2.62306671942031E-6	6.0E-3	GO:0001968	fibronectin binding
14	31	5.7654916834889098E-8	<0.001	GO:0050840	extracellular matrix binding
12	27	6.4706181619242598E-7	2E-3	GO:0048706	embryonic skeletal system development
13	30	3.1474432892403899E-7	<0.001	GO:0004930	G-protein coupled receptor activity
10	24	1.1533808418117801E-5	3.5E-2	GO:0030858	positive regulation of epithelial cell differentiation

13	32	7.78067056032933E-7	2E-3	GO:0002687	positive regulation of leukocyte migration
10	25	1.77757359173968E-5	4.7E-2	GO:0035137	hindlimb morphogenesis
13	33	1.18524393038482E-6	2E-3	GO:0031214	biomineral tissue development
17	44	3.7462488736548198E-8	<0.001	GO:0050900	leukocyte migration
51	135	2.2957966272059699E-21	<0.001	GO:0005578	proteinaceous extracellular matrix
23	60	1.74891005222263E-10	<0.001	GO:0051216	cartilage development
11	29	1.22854581467478E-5	3.8E-2	GO:0090183	regulation of kidney development
59	162	1.35130950603293E-23	<0.001	GO:0031012	extracellular matrix
13	35	2.60235311068536E-6	5.0E-3	GO:0005518	collagen binding
24	69	7.3450057453294495E-10	<0.001	GO:0009897	external side of plasma membrane
29	85	2.1194797756309901E-11	<0.001	GO:0001501	skeletal system development
14	41	3.4344891183182701E-6	8.0E-3	GO:0005581	collagen trimer
14	41	3.4344891183182701E-6	8.0E-3	GO:0030500	regulation of bone mineralization
16	47	7.2056049050454499E-7	2E-3	GO:0050921	positive regulation of chemotaxis
20	59	3.2038768068600297E-8	<0.001	GO:0030326	embryonic limb morphogenesis
20	59	3.2038768068600297E-8	<0.001	GO:0035113	embryonic appendage morphogenesis
131	423	4.3109749969765002E-43	<0.001	GO:0005576	extracellular region
14	43	6.5082582472223003E-6	2.1E-2	GO:0070167	regulation of biomineral tissue development
14	44	8.8111631760993594E-6	3.2E-2	GO:0002685	regulation of leukocyte migration
19	61	3.2484767129804101E-7	<0.001	GO:0008201	heparin binding
14	45	1.18067970808607E-5	3.7E-2	GO:0005261	cation channel activity
16	53	4.3885048707458198E-6	1.8E-2	GO:0048520	positive regulation of behavior
24	80	2.09951814383663E-8	<0.001	GO:0001503	ossification

30	104	1.0318648253535001E-9	<0.001	GO:0030198	extracellular matrix organization
30	104	1.0318648253535001E-9	<0.001	GO:0043062	extracellular structure organization
21	73	3.53964994121251E-7	1E-3	GO:0035107	appendage morphogenesis
21	73	3.53964994121251E-7	1E-3	GO:0035108	limb morphogenesis
16	59	2.0013644891332301E-5	4.9E-2	GO:0045778	positive regulation of ossification
17	63	1.18996775816344E-5	3.7E-2	GO:0050731	positive regulation of peptidyl-tyrosine phosphorylation
18	67	7.0775972883762301E-6	2.1E-2	GO:0010811	positive regulation of cell-substrate adhesion
24	90	2.5117119768673201E-7	<0.001	GO:0006935	chemotaxis
20	75	2.5054901416722E-6	5.0E-3	GO:0019838	growth factor binding
24	92	3.9206102571434899E-7	2E-3	GO:0042330	taxis
20	77	3.90443286163468E-6	9.0E-3	GO:0044420	extracellular matrix component
20	78	4.8411182914754799E-6	0.02	GO:0005539	glycosaminoglycan binding
30	119	3.2649974110267698E-8	<0.001	GO:0030278	regulation of ossification
28	113	1.3921588684794499E-7	<0.001	GO:0007186	G-protein coupled receptor signaling pathway
25	102	8.1057553084648204E-7	2E-3	GO:0098552	side of membrane
88	379	2.7566695942264198E-19	<0.001	GO:0005615	extracellular space
52	224	8.5508887718875607E-12	<0.001	GO:0005509	calcium ion binding
24	102	2.9319960013772302E-6	7.0E-3	GO:0010810	regulation of cell-substrate adhesion
33	141	4.7898474205072297E-8	<0.001	GO:0004888	transmembrane signaling receptor activity
23	98	4.9262353237078297E-6	0.02	GO:0010632	regulation of epithelial cell migration

25	107	2.0983869853906501E-6	3.0E-3	GO:1901681	sulfur compound binding
24	103	3.5230385454192999E-6	8.0E-3	GO:1901342	regulation of vasculature development
22	97	1.41589662353361E-5	0.04	GO:0001763	morphogenesis of a branching structure
50	225	1.17336323651596E-10	<0.001	GO:0009986	cell surface
39	178	2.06037378564837E-8	<0.001	GO:0038023	signaling receptor activity
45	207	2.1529714974322501E-9	<0.001	GO:0030335	positive regulation of cell migration
24	110	1.1785391692014299E-5	3.6E-2	GO:0030336	negative regulation of cell migration
47	218	1.2525925793602099E-9	<0.001	GO:0040017	positive regulation of locomotion
45	210	3.4820502943171001E-9	<0.001	GO:2000147	positive regulation of cell motility
32	149	6.2158656397240396E-7	2E-3	GO:1903034	regulation of response to wounding
45	215	7.5625168653703296E-9	<0.001	GO:0051272	positive regulation of cellular component movement
24	114	2.22204521078299E-5	0.05	GO:2000146	negative regulation of cell motility
36	176	4.4276581877016699E-7	2E-3	GO:0031226	intrinsic component of plasma membrane
27	132	1.21424719792932E-5	3.7E-2	GO:0090287	regulation of cellular response to growth factor stimulus
73	369	2.4026686373510698E-12	<0.001	GO:0007155	cell adhesion
73	370	2.7594225501547202E-12	<0.001	GO:0022610	biological adhesion
28	139	1.1394645048131599E-5	3.5E-2	GO:0001525	angiogenesis
27	134	1.6201751110553799E-5	4.2E-2	GO:0002521	leukocyte differentiation
68	348	2.5690810149361599E-11	<0.001	GO:0030334	regulation of cell migration
38	192	5.1169269546806001E-7	2E-3	GO:0016337	single organismal cell-cell adhesion
43	220	1.31643145784269E-7	<0.001	GO:0004872	receptor activity

35	180	2.21239386069635E-6	4.0E-3	GO:0002009	morphogenesis of an epithelium
42	217	2.4479838940406002E-7	<0.001	GO:0098609	cell-cell adhesion
54	284	8.6737138171379996E-9	<0.001	GO:0009888	tissue development
68	363	1.81628206128766E-10	<0.001	GO:2000145	regulation of cell motility
78	422	1.4968795464761001E-11	<0.001	GO:0007275	multicellular organismal development
38	202	1.90683069495833E-6	3.0E-3	GO:0043269	regulation of ion transport
40	219	2.2290845680791602E-6	4.0E-3	GO:0048729	tissue morphogenesis
50	277	1.7468167543148699E-7	<0.001	GO:0009887	organ morphogenesis
33	182	2.0046224245379201E-5	4.9E-2	GO:0010721	negative regulation of cell development
38	211	5.6478636866460197E-6	0.02	GO:0098602	single organism cell adhesion
68	386	2.7673116354531601E-9	<0.001	GO:0051270	regulation of cellular component movement
70	398	1.7103270037821999E-9	<0.001	GO:0040012	regulation of locomotion
81	469	2.0356370248101E-10	<0.001	GO:0045597	positive regulation of cell differentiation
104	624	3.82294555661716E-12	<0.001	GO:0048513	organ development
53	311	5.0201339470734897E-7	2E-3	GO:0016477	cell migration
137	849	1.02681139658503E-14	<0.001	GO:2000026	regulation of multicellular organismal development
57	339	2.9292882963249899E-7	<0.001	GO:0048731	system development
99	607	4.9406140125825799E-11	<0.001	GO:0051094	positive regulation of developmental process
46	276	5.3622746657438103E-6	0.02	GO:0030155	regulation of cell adhesion
60	363	2.6084530774391699E-7	<0.001	GO:0003008	system process
181	1186	5.2757987970103801E-17	<0.001	GO:0051239	regulation of multicellular organismal process
55	334	9.4607517169920205E-7	2E-3	GO:0048870	cell motility
62	378	2.1669469423054601E-7	<0.001	GO:0040011	locomotion

41	248	2.1190235060926699E-5	0.05	GO:0022891	substrate-specific transmembrane transporter activity
44	268	1.2719409848960899E-5	3.8E-2	GO:0022857	transmembrane transporter activity
55	340	1.67944037993881E-6	3.0E-3	GO:0045596	negative regulation of cell differentiation
104	663	1.5986855395387201E-10	<0.001	GO:0051240	positive regulation of multicellular organismal process
49	306	8.3362324199080006E-6	2.7E-2	GO:0022892	substrate-specific transporter activity
78	495	3.2491208085246497E-8	<0.001	GO:0051241	negative regulation of multicellular organismal process
124	812	1.3056943189054199E-11	<0.001	GO:0045595	regulation of cell differentiation
115	753	8.2862733266854805E-11	<0.001	GO:0009653	anatomical structure morphogenesis
186	1278	2.5313466974863399E-15	<0.001	GO:0032501	multicellular organismal process
185	1271	3.0671443286448801E-15	<0.001	GO:0044707	single-multicellular organism process
69	449	5.6159199638347598E-7	2E-3	GO:0051093	negative regulation of developmental process
69	453	7.85210333196869E-7	2E-3	GO:0002682	regulation of immune system process
162	1130	1.0812178585085701E-12	<0.001	GO:0050793	regulation of developmental process
73	490	8.8006547814048801E-7	2E-3	GO:0022603	regulation of anatomical structure morphogenesis
74	504	1.2997384598505499E-6	2E-3	GO:0060284	regulation of cell development
125	891	3.0445818161631501E-9	<0.001	GO:0030154	cell differentiation

88	645	2.9713311152811499E-6	7.0E-3	GO:0044459	plasma membrane part
171	1310	6.4579882330220096E-10	<0.001	GO:0005886	plasma membrane
168	1308	3.3641739239291601E-9	<0.001	GO:0048856	anatomical structure development
160	1246	9.1688720551134707E-9	<0.001	GO:0048869	cellular developmental process
264	2177	6.7922345161723396E-12	<0.001	GO:0032502	developmental process
95	721	5.32287189955161E-6	0.02	GO:0042127	regulation of cell proliferation
252	2082	3.5325358310107399E-11	<0.001	GO:0044767	single-organism developmental process
204	1664	2.2798827897708201E-9	<0.001	GO:0031224	intrinsic component of membrane
221	1825	1.0959036746228901E-9	<0.001	GO:0044421	extracellular region part
188	1629	1.1841806460117201E-6	2E-3	GO:0016021	integral component of membrane
6	6	6.4899999999999995E-7	1E-3	GO:0005833	hemoglobin complex
6	6	6.4899999999999995E-7	1E-3	GO:0072124	regulation of glomerular mesangial cell proliferation
5	5	6.9999999999999999E-6	1.8E-2	GO:0003071	renal system process involved in regulation of systemic arterial blood pressure
7	8	4.4299999999999998E-7	<0.001	GO:0090192	regulation of glomerulus development
6	8	1.5400000000000002E-5	3.3E-2	GO:0098801	regulation of renal system process
6	8	1.5400000000000002E-5	3.3E-2	GO:1901722	regulation of cell proliferation involved in kidney development
8	12	1.9599999999999999E-6	3.0E-3	GO:0050919	negative chemotaxis
14	22	5.3700000000000001E-10	<0.001	GO:0008038	neuron recognition

7	11	1.42E-5	0.03	GO:0043395	heparan sulfate proteoglycan binding
7	11	1.42E-5	0.03	GO:0072215	regulation of metanephros development
8	13	4.6800000000000001E-6	9.0E-3	GO:0071772	response to BMP
8	13	4.6800000000000001E-6	9.0E-3	GO:0071773	cellular response to BMP stimulus
11	18	7.5600000000000002E-8	<0.001	GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules
8	14	1.0000000000000001E-5	2.4E-2	GO:0048070	regulation of developmental pigmentation
9	16	3.2200000000000001E-6	4.0E-3	GO:0007413	axonal fasciculation
8	15	1.9700000000000001E-5	3.7E-2	GO:0048521	negative regulation of behavior
9	17	6.2700000000000001E-6	1.4E-2	GO:0043394	proteoglycan binding
11	21	6.4300000000000003E-7	1E-3	GO:2000351	regulation of endothelial cell apoptotic process
10	20	3.6799999999999999E-6	4.0E-3	GO:0003014	renal system process
10	20	3.6799999999999999E-6	4.0E-3	GO:1904036	negative regulation of epithelial cell apoptotic process
9	19	2.0100000000000001E-5	3.9E-2	GO:0045992	negative regulation of embryonic development
16	35	2.1299999999999999E-8	<0.001	GO:0072562	blood microparticle
10	22	1.08E-5	2.4E-2	GO:0030501	positive regulation of bone mineralization
10	22	1.08E-5	2.4E-2	GO:0070169	positive regulation of biomineral tissue development
13	29	6.0999999999999998E-7	1E-3	GO:0090183	regulation of kidney development

11	25	5.7300000000000002E-6	1.3E-2	GO:0034754	cellular hormone metabolic process
14	33	5.1500000000000005E-7	1E-3	GO:0031214	biomineral tissue development
56	135	1.2899999999999999E-23	<0.001	GO:0005578	proteinaceous extracellular matrix
15	36	2.67E-7	<0.001	GO:0010595	positive regulation of endothelial cell migration
12	29	4.6800000000000001E-6	9.0E-3	GO:0098742	cell-cell adhesion via plasma-membrane adhesion molecules
19	46	8.02E-9	<0.001	GO:0045669	positive regulation of osteoblast differentiation
14	34	8.0100000000000004E-7	1E-3	GO:0030193	regulation of blood coagulation
14	34	8.0100000000000004E-7	1E-3	GO:1900046	regulation of hemostasis
16	39	1.3799999999999999E-7	<0.001	GO:1904035	regulation of epithelial cell apoptotic process
18	44	2.3899999999999999E-8	<0.001	GO:0008083	growth factor activity
11	27	1.4100000000000001E-5	2.9E-2	GO:0014068	positive regulation of phosphatidylinositol 3-kinase signaling
12	30	7.1400000000000002E-6	1.8E-2	GO:0004930	G-protein coupled receptor activity
17	43	1.0700000000000001E-7	<0.001	GO:0070167	regulation of biomineral tissue development
11	28	2.12E-5	0.04	GO:0001944	vasculature development
11	28	2.12E-5	0.04	GO:0050715	positive regulation of cytokine secretion
62	162	8.9099999999999994E-24	<0.001	GO:0031012	extracellular matrix

16	41	3.15E-7	<0.001	GO:0030500	regulation of bone mineralization
14	36	1.8300000000000001E-6	3.0E-3	GO:0050818	regulation of coagulation
17	45	2.35E-7	<0.001	GO:0008037	cell recognition
15	40	1.35E-6	2E-3	GO:0042445	hormone metabolic process
13	35	7.7999999999999999E-6	1.9E-2	GO:0005518	collagen binding
15	41	1.95E-6	3.0E-3	GO:0005581	collagen trimer
15	41	1.95E-6	3.0E-3	GO:0048592	eye morphogenesis
15	41	1.95E-6	3.0E-3	GO:0050707	regulation of cytokine secretion
12	33	2.2500000000000001E-5	0.04	GO:0030509	BMP signaling pathway
21	59	3.0600000000000003E-8	<0.001	GO:0045778	positive regulation of ossification
19	54	1.72E-7	<0.001	GO:1904018	positive regulation of vasculature development
14	40	7.8299999999999996E-6	0.02	GO:0007160	cell-matrix adhesion
14	40	7.8299999999999996E-6	0.02	GO:0014066	regulation of phosphatidylinositol 3-kinase signaling
15	43	3.89E-6	4.0E-3	GO:0001570	vasculogenesis
15	43	3.89E-6	4.0E-3	GO:0090596	sensory organ morphogenesis
24	69	5.3700000000000003E-9	<0.001	GO:0009897	external side of plasma membrane
13	38	2.1999999999999999E-5	0.04	GO:0014910	regulation of smooth muscle cell migration
21	62	8.2700000000000006E-8	<0.001	GO:0010594	regulation of endothelial cell migration
25	74	5.14E-9	<0.001	GO:0007411	axon guidance
132	423	6.5999999999999997E-39	<0.001	GO:0005576	extracellular region
25	75	7.0399999999999997E-9	<0.001	GO:0097485	neuron projection guidance
16	48	3.6899999999999998E-6	4.0E-3	GO:0045766	positive regulation of angiogenesis
23	70	3.84E-8	<0.001	GO:0001667	ameboidal-type cell migration

20	61	3.0899999999999997E-7	<0.001	GO:0060560	developmental growth involved in morphogenesis
17	52	2.4899999999999999E-6	4.0E-3	GO:0001936	regulation of endothelial cell proliferation
45	141	3.8399999999999999E-14	<0.001	GO:0004888	transmembrane signaling receptor activity
18	56	1.6700000000000001E-6	3.0E-3	GO:0061041	regulation of wound healing
17	53	3.3500000000000001E-6	4.0E-3	GO:0050673	epithelial cell proliferation
14	44	2.72E-5	4.7E-2	GO:0019199	transmembrane receptor protein kinase activity
14	44	2.72E-5	4.7E-2	GO:0050772	positive regulation of axonogenesis
19	60	1.11E-6	2E-3	GO:0010634	positive regulation of epithelial cell migration
26	83	1.5399999999999999E-8	<0.001	GO:0048754	branching morphogenesis of an epithelial tube
25	80	3.0799999999999998E-8	<0.001	GO:0001503	ossification
29	93	2.57E-9	<0.001	GO:0061138	morphogenesis of a branching epithelium
28	90	5.1300000000000003E-9	<0.001	GO:0006935	chemotaxis
30	97	1.6999999999999999E-9	<0.001	GO:0001763	morphogenesis of a branching structure
26	84	2.0400000000000001E-8	<0.001	GO:0007178	transmembrane receptor protein serine/threonine kinase signaling pathway
24	78	8.1199999999999999E-8	<0.001	GO:0005539	glycosaminoglycan binding
16	52	1.19E-5	2.9E-2	GO:0048514	blood vessel morphogenesis
27	88	1.35E-8	<0.001	GO:0001568	blood vessel development
23	75	1.6199999999999999E-7	<0.001	GO:0048562	embryonic organ morphogenesis
28	92	8.8800000000000008E-9	<0.001	GO:0042330	taxis

23	76	2.1199999999999999E-7	<0.001	GO:0051924	regulation of calcium ion transport
16	53	1.5500000000000001E-5	3.3E-2	GO:0048520	positive regulation of behavior
108	379	6.1699999999999998E-28	<0.001	GO:0005615	extracellular space
52	178	2.2499999999999999E-14	<0.001	GO:0038023	signaling receptor activity
33	113	1.37E-9	<0.001	GO:0007186	G-protein coupled receptor signaling pathway
26	89	7.6500000000000003E-8	<0.001	GO:0050770	regulation of axonogenesis
21	72	1.3999999999999999E-6	2E-3	GO:0010817	regulation of hormone levels
17	59	1.6799999999999998E-5	3.4E-2	GO:0030326	embryonic limb morphogenesis
17	59	1.6799999999999998E-5	3.4E-2	GO:0035113	embryonic appendage morphogenesis
18	63	1.0900000000000001E-5	2.5E-2	GO:0050920	regulation of chemotaxis
19	67	7.0099999999999998E-6	1.8E-2	GO:0031589	cell-substrate adhesion
29	103	3.3899999999999999E-8	<0.001	GO:1901342	regulation of vasculature development
25	89	3.1399999999999998E-7	<0.001	GO:2000027	regulation of organ morphogenesis
29	104	4.29E-8	<0.001	GO:0030198	extracellular matrix organization
29	104	4.29E-8	<0.001	GO:0043062	extracellular structure organization
17	61	2.72E-5	4.6E-2	GO:0008201	heparin binding
22	79	1.88E-6	3.0E-3	GO:0045667	regulation of osteoblast differentiation
61	225	5.5199999999999998E-15	<0.001	GO:0009986	cell surface
27	98	1.6400000000000001E-7	<0.001	GO:0010632	regulation of epithelial cell migration
20	73	7.1999999999999997E-6	1.8E-2	GO:0035107	appendage morphogenesis
20	73	7.1999999999999997E-6	1.8E-2	GO:0035108	limb morphogenesis

18	66	2.1999999999999999E-5	0.04	GO:0045995	regulation of embryonic development
37	138	1.9500000000000001E-9	<0.001	GO:0035239	tube morphogenesis
22	82	3.7000000000000002E-6	4.0E-3	GO:0050795	regulation of behavior
23	86	2.3700000000000002E-6	3.0E-3	GO:0050839	cell adhesion molecule binding
20	75	1.1199999999999999E-5	2.7E-2	GO:0019838	growth factor binding
21	79	7.1899999999999998E-6	1.8E-2	GO:0001649	osteoblast differentiation
57	220	3.67E-13	<0.001	GO:0004872	receptor activity
31	119	8.28E-8	<0.001	GO:0030278	regulation of ossification
54	210	2.1699999999999998E-12	<0.001	GO:2000147	positive regulation of cell motility
25	96	1.4899999999999999E-6	2E-3	GO:0035295	tube development
20	77	1.7200000000000001E-5	3.4E-2	GO:0044420	extracellular matrix component
53	207	4.1899999999999997E-12	<0.001	GO:0030335	positive regulation of cell migration
22	85	7.0199999999999997E-6	1.8E-2	GO:0001501	skeletal system development
46	180	1.2199999999999999E-10	<0.001	GO:0002009	morphogenesis of an epithelium
55	218	3.09E-12	<0.001	GO:0040017	positive regulation of locomotion
91	369	7.0099999999999997E-19	<0.001	GO:0007155	cell adhesion
91	370	8.5100000000000004E-19	<0.001	GO:0022610	biological adhesion
26	102	1.44E-6	2E-3	GO:0098552	side of membrane
55	219	3.7600000000000001E-12	<0.001	GO:0048729	tissue morphogenesis
54	215	5.93E-12	<0.001	GO:0051272	positive regulation of cellular component movement
21	83	1.6500000000000001E-5	3.4E-2	GO:0050679	positive regulation of epithelial cell proliferation
27	107	1.1200000000000001E-6	2E-3	GO:1901681	sulfur compound binding

54	217	8.7600000000000006E-12	<0.001	GO:0098609	cell-cell adhesion
23	93	9.9399999999999997E-6	2.2E-2	GO:0045765	regulation of angiogenesis
25	104	7.1099999999999997E-6	1.8E-2	GO:0070372	regulation of ERK1 and ERK2 cascade
53	224	1.0700000000000001E-10	<0.001	GO:0005509	calcium ion binding
59	253	1.7100000000000001E-11	<0.001	GO:0007167	enzyme linked receptor protein signaling pathway
80	348	7.8299999999999998E-15	<0.001	GO:0030334	regulation of cell migration
45	192	3.8799999999999998E-9	<0.001	GO:0016337	single organismal cell-cell adhesion
63	277	1.0899999999999999E-11	<0.001	GO:0009887	organ morphogenesis
89	398	1.2800000000000001E-15	<0.001	GO:0040012	regulation of locomotion
32	139	1.0699999999999999E-6	2E-3	GO:0001525	angiogenesis
81	363	3.0799999999999999E-14	<0.001	GO:2000145	regulation of cell motility
25	109	1.7099999999999999E-5	3.4E-2	GO:0010770	positive regulation of cell morphogenesis involved in differentiation
30	131	2.61E-6	4.0E-3	GO:0030855	epithelial cell differentiation
32	140	1.2699999999999999E-6	2E-3	GO:0040013	negative regulation of locomotion
47	211	1E-8	<0.001	GO:0098602	single organism cell adhesion
68	311	1.1000000000000001E-11	<0.001	GO:0016477	cell migration
91	422	6.3900000000000001E-15	<0.001	GO:0007275	multicellular organismal development
33	149	1.84E-6	3.0E-3	GO:1903034	regulation of response to wounding
62	284	9.9799999999999994E-11	<0.001	GO:0009888	tissue development
44	200	4.3999999999999997E-8	<0.001	GO:0010769	regulation of cell morphogenesis involved in differentiation

83	386	1.3E-13	<0.001	GO:0051270	regulation of cellular component movement
38	175	5.2300000000000001E-7	1E-3	GO:0050678	regulation of epithelial cell proliferation
27	124	2.1800000000000001E-5	0.04	GO:0051271	negative regulation of cellular component movement
34	158	2.57E-6	4.0E-3	GO:0005887	integral component of plasma membrane
28	130	1.91E-5	3.5E-2	GO:0010959	regulation of metal ion transport
101	490	4.2100000000000002E-15	<0.001	GO:0022603	regulation of anatomical structure morphogenesis
70	334	4.2500000000000002E-11	<0.001	GO:0048870	cell motility
28	132	2.5700000000000001E-5	4.4E-2	GO:0090287	regulation of cellular response to growth factor stimulus
78	378	6.9399999999999999E-12	<0.001	GO:0040011	locomotion
37	176	1.6899999999999999E-6	3.0E-3	GO:0031226	intrinsic component of plasma membrane
57	276	5.2199999999999998E-9	<0.001	GO:0030155	regulation of cell adhesion
96	478	1.1700000000000001E-13	<0.001	GO:0006928	movement of cell or subcellular component
120	607	2.2200000000000001E-16	<0.001	GO:0051094	positive regulation of developmental process
59	290	5.04E-9	<0.001	GO:0004871	signal transducer activity
50	245	6.9499999999999994E-8	<0.001	GO:0051962	positive regulation of nervous system development
32	157	1.6500000000000001E-5	3.4E-2	GO:0045785	positive regulation of cell adhesion
34	169	1.2099999999999999E-5	2.9E-2	GO:0007399	nervous system development
159	849	2.7900000000000002E-19	<0.001	GO:2000026	regulation of multicellular organismal development

67	339	1.57E-9	<0.001	GO:0048731	system development
121	633	2.3499999999999999E-15	<0.001	GO:0007166	cell surface receptor signaling pathway
91	469	4.2800000000000003E-12	<0.001	GO:0045597	positive regulation of cell differentiation
126	663	9.0000000000000003E-16	<0.001	GO:0051240	positive regulation of multicellular organismal process
34	171	1.5800000000000001E-5	3.4E-2	GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway
36	184	1.2999999999999999E-5	2.9E-2	GO:0045666	positive regulation of neuron differentiation
64	332	1.04E-8	<0.001	GO:0060089	molecular transducer activity
209	1186	3.09E-22	<0.001	GO:0051239	regulation of multicellular organismal process
44	228	2.1299999999999999E-6	3.0E-3	GO:0050769	positive regulation of neurogenesis
138	753	7.04E-16	<0.001	GO:0009653	anatomical structure morphogenesis
35	181	2.2200000000000001E-5	0.04	GO:0005911	cell-cell junction
39	202	7.9500000000000001E-6	0.02	GO:0043269	regulation of ion transport
56	293	1.2100000000000001E-7	<0.001	GO:0010720	positive regulation of cell development
57	303	1.6400000000000001E-7	<0.001	GO:0022604	regulation of cell morphogenesis
194	1130	3.6300000000000001E-19	<0.001	GO:0050793	regulation of developmental process
45	242	4.6099999999999999E-6	8.0E-3	GO:0048598	embryonic morphogenesis
60	325	1.5300000000000001E-7	<0.001	GO:0008285	negative regulation of cell proliferation
214	1271	3.1299999999999998E-20	<0.001	GO:0044707	single-multicellular organism process

215	1278	2.7400000000000001E-20	<0.001	GO:0032501	multicellular organismal process
142	812	1.42E-14	<0.001	GO:0045595	regulation of cell differentiation
111	624	5.6199999999999999E-12	<0.001	GO:0048513	organ development
60	329	2.3799999999999999E-7	<0.001	GO:0045664	regulation of neuron differentiation
46	253	6.7599999999999997E-6	1.4E-2	GO:0010975	regulation of neuron projection development
75	424	2.6799999999999998E-8	<0.001	GO:0051960	regulation of nervous system development
48	271	8.9500000000000007E-6	2.1E-2	GO:0003006	developmental process involved in reproduction
123	721	6.7500000000000001E-12	<0.001	GO:0042127	regulation of cell proliferation
212	1310	6.7099999999999997E-18	<0.001	GO:0005886	plasma membrane
86	495	5.7800000000000003E-9	<0.001	GO:0051241	negative regulation of multicellular organismal process
70	401	1.37E-7	<0.001	GO:0008284	positive regulation of cell proliferation
63	363	7.3300000000000001E-7	1E-3	GO:0003008	system process
102	601	7.3199999999999995E-10	<0.001	GO:0005102	receptor binding
67	389	4.3000000000000001E-7	<0.001	GO:0050767	regulation of neurogenesis
86	504	1.39E-8	<0.001	GO:0060284	regulation of cell development
108	645	4.8899999999999997E-10	<0.001	GO:0044459	plasma membrane part
207	1308	2.55E-16	<0.001	GO:0048856	anatomical structure development
80	472	6.2800000000000006E-8	<0.001	GO:0010562	positive regulation of phosphorus metabolic process
80	472	6.2800000000000006E-8	<0.001	GO:0045937	positive regulation of phosphate metabolic process

75	443	1.7499999999999999E-7	<0.001	GO:0048646	anatomical structure formation involved in morphogenesis
58	340	3.5899999999999999E-6	4.0E-3	GO:0045596	negative regulation of cell differentiation
54	318	8.9700000000000005E-6	2.1E-2	GO:0031344	regulation of cell projection organization
60	355	3.3100000000000001E-6	4.0E-3	GO:0032101	regulation of response to external stimulus
76	453	2.11E-7	<0.001	GO:0002682	regulation of immune system process
49	290	2.6699999999999998E-5	4.5E-2	GO:0048468	cell development
74	449	6.2900000000000003E-7	1E-3	GO:0051093	negative regulation of developmental process
270	1825	1.15E-17	<0.001	GO:0044421	extracellular region part
141	891	4.8800000000000002E-11	<0.001	GO:0030154	cell differentiation
56	347	2.8399999999999999E-5	4.7E-2	GO:0043068	positive regulation of programmed cell death
188	1246	1.1E-12	<0.001	GO:0048869	cellular developmental process
71	445	3.5999999999999998E-6	4.0E-3	GO:0098589	membrane region
206	1381	1.9300000000000001E-13	<0.001	GO:0007165	signal transduction
65	407	9.1300000000000007E-6	2.1E-2	GO:0001934	positive regulation of protein phosphorylation
306	2177	5.3899999999999998E-17	<0.001	GO:0032502	developmental process
64	402	1.19E-5	2.9E-2	GO:0030030	cell projection organization
68	428	6.7599999999999997E-6	1.4E-2	GO:0042327	positive regulation of phosphorylation
289	2082	4.3699999999999996E-15	<0.001	GO:0044767	single-organism developmental process
165	1128	4.9299999999999995E-10	<0.001	GO:0032879	regulation of localization

116	785	1.74E-7	<0.001	GO:0048584	positive regulation of response to stimulus
91	615	4.0600000000000001E-6	5.0E-3	GO:0009967	positive regulation of signal transduction
80	540	1.5400000000000002E-5	3.3E-2	GO:0070887	cellular response to chemical stimulus
230	1664	1.8999999999999999E-11	<0.001	GO:0031224	intrinsic component of membrane
99	682	3.5099999999999999E-6	4.0E-3	GO:0023056	positive regulation of signaling
86	594	1.7900000000000001E-5	3.5E-2	GO:0030054	cell junction
222	1629	2.0399999999999999E-10	<0.001	GO:0016021	integral component of membrane
184	1338	6.3799999999999999E-9	<0.001	GO:0023051	regulation of signaling
99	698	9.7100000000000002E-6	2.2E-2	GO:0010647	positive regulation of cell communication
164	1199	7.98E-8	<0.001	GO:0009966	regulation of signal transduction
94	672	2.9099999999999999E-5	4.7E-2	GO:0048585	negative regulation of response to stimulus
209	1578	1.09E-8	<0.001	GO:0048583	regulation of response to stimulus
184	1380	6.9899999999999997E-8	<0.001	GO:0010646	regulation of cell communication
107	783	2.19E-5	0.04	GO:0051174	regulation of phosphorus metabolic process
106	779	2.8799999999999999E-5	4.7E-2	GO:0019220	regulation of phosphate metabolic process
284	2270	3.2700000000000001E-9	<0.001	GO:0044425	membrane part
148	1123	3.7000000000000002E-6	4.0E-3	GO:0065008	regulation of biological quality
296	2388	3.3799999999999999E-9	<0.001	GO:0048522	positive regulation of cellular process
316	2616	1.35E-8	<0.001	GO:0048518	positive regulation of biological process

211	1721	4.16E-6	5.0E-3	GO:0031988	membrane-bounded vesicle
513	4706	5.6799999999999999E-8	<0.001	GO:0065007	biological regulation
227	1884	5.6999999999999996E-6	1.3E-2	GO:0031982	vesicle
494	4552	3.58E-7	<0.001	GO:0050789	regulation of biological process
264	2251	6.2700000000000001E-6	1.3E-2	GO:0048519	negative regulation of biological process
473	4341	5.9500000000000002E-7	1E-3	GO:0050794	regulation of cellular process
248	2128	2.2900000000000001E-5	0.04	GO:0048523	negative regulation of cellular process
446	4108	3.63E-6	4.0E-3	GO:0005515	protein binding
530	5038	7.7100000000000007E-6	1.9E-2	GO:0044699	single-organism process
5	7	4.69142170614867E-7	2E-3	GO:0003009	skeletal muscle contraction
4	6	1.13484012416261E-5	1.9E-2	GO:0010919	regulation of inositol phosphate biosynthetic process
8	25	4.1201915895595998E-7	2E-3	GO:0006941	striated muscle contraction
9	37	1.02944750126797E-6	2E-3	GO:0022900	electron transport chain
26	135	1.9657624069355098E-14	<0.001	GO:0005578	proteinaceous extracellular matrix
27	162	2.4446271205530899E-13	<0.001	GO:0031012	extracellular matrix
10	61	1.17170901212455E-5	1.9E-2	GO:0008201	heparin binding
12	85	7.8190713793110795E-6	8.0E-3	GO:0001501	skeletal system development
11	78	1.89666081314572E-5	0.03	GO:0005539	glycosaminoglycan binding
54	423	2.4993065086753099E-20	<0.001	GO:0005576	extracellular region
34	379	8.3209272376592193E-9	<0.001	GO:0005615	extracellular space
31	363	1.23880253390561E-7	1E-3	GO:0003008	system process
24	284	4.2906296063349796E-6	5.0E-3	GO:0009888	tissue development
31	422	3.2641730427692599E-6	5.0E-3	GO:0007275	multicellular organismal development
74	1278	6.9187352222183001E-9	<0.001	GO:0032501	multicellular organismal process

73	1271	1.2949234596072499E-8	<0.001	GO:0044707	single-multicellular organism process
16	135	2.1707391859897501E-8	<0.001	GO:0005578	proteinaceous extracellular matrix
18	162	7.7580819920006808E-9	<0.001	GO:0031012	extracellular matrix
33	423	4.9683535800453898E-11	<0.001	GO:0005576	extracellular region
18	277	2.1841155519159701E-5	3.3E-2	GO:0009887	organ morphogenesis
35	624	7.8721408172260394E-8	<0.001	GO:0048513	organ development
6	6	4.2696212005413998E-10	<0.001	GO:0005833	hemoglobin complex
4	4	5.7589967481140203E-7	1E-3	GO:0031838	haptoglobin-hemoglobin complex
4	5	2.8166523046157098E-6	1.2E-2	GO:0030825	positive regulation of cGMP metabolic process
4	6	8.2652613645796707E-6	2.4E-2	GO:0019825	oxygen binding
11	35	1.3784435788083099E-9	<0.001	GO:0072562	blood microparticle
7	24	2.6678800903408201E-6	9.0E-3	GO:0055008	cardiac muscle tissue morphogenesis
7	25	3.6183358864521802E-6	1.4E-2	GO:0006941	striated muscle contraction
9	33	1.7998755857775201E-7	<0.001	GO:0031214	biomineral tissue development
8	30	1.0729523558841999E-6	5.0E-3	GO:0060415	muscle tissue morphogenesis
8	31	1.41175994688806E-6	6.0E-3	GO:0050840	extracellular matrix binding
7	29	1.0683093490522701E-5	2.9E-2	GO:1902930	regulation of alcohol biosynthetic process
10	49	7.0848483836871199E-7	1E-3	GO:0006936	muscle contraction
13	66	2.3065524115908601E-8	<0.001	GO:0003012	muscle system process
24	135	2.2657022113402799E-13	<0.001	GO:0005578	proteinaceous extracellular matrix
9	52	1.08356105859414E-5	2.9E-2	GO:0048514	blood vessel morphogenesis
26	162	2.6557523844903202E-13	<0.001	GO:0031012	extracellular matrix

17	104	3.1159503426712301E-9	<0.001	GO:0030198	extracellular matrix organization
17	104	3.1159503426712301E-9	<0.001	GO:0043062	extracellular structure organization
59	423	3.5118516382437701E-26	<0.001	GO:0005576	extracellular region
15	97	5.8545233404851597E-8	<0.001	GO:0001763	morphogenesis of a branching structure
14	93	2.3101454673117499E-7	<0.001	GO:0061138	morphogenesis of a branching epithelium
12	80	1.78374075793079E-6	7.0E-3	GO:0001503	ossification
43	379	1.16790296976662E-15	<0.001	GO:0005615	extracellular space
25	219	1.52313300195994E-9	<0.001	GO:0048729	tissue morphogenesis
29	284	9.8801921614063904E-10	<0.001	GO:0009888	tissue development
18	180	2.3363883670357698E-6	9.0E-3	GO:0002009	morphogenesis of an epithelium
38	422	8.3831069985444898E-11	<0.001	GO:0007275	multicellular organismal development
21	224	9.5773383273316003E-7	4.0E-3	GO:0005509	calcium ion binding
18	196	7.8351232586422001E-6	2.1E-2	GO:0007389	pattern specification process
30	363	6.9885917786189801E-8	<0.001	GO:0003008	system process
19	225	1.4692909506585101E-5	3.3E-2	GO:0009986	cell surface
20	242	1.1994125843516999E-5	3.1E-2	GO:0048598	embryonic morphogenesis
85	1271	6.1626609906293902E-16	<0.001	GO:0044707	single-multicellular organism process
22	277	8.2355556050856005E-6	2.1E-2	GO:0009887	organ morphogenesis
85	1278	8.6109456999171402E-16	<0.001	GO:0032501	multicellular organismal process
21	271	1.93512443454791E-5	4.2E-2	GO:0003006	developmental process involved in reproduction
26	369	1.07445293819627E-5	2.9E-2	GO:0007155	cell adhesion
26	370	1.12736718339205E-5	2.9E-2	GO:0022610	biological adhesion
50	753	3.6248440059090802E-9	<0.001	GO:0009653	anatomical structure morphogenesis

41	633	2.33795832965117E-7	<0.001	GO:0007166	cell surface receptor signaling pathway
32	490	4.8638755667330899E-6	1.6E-2	GO:0022603	regulation of anatomical structure morphogenesis
39	624	1.1883162478938699E-6	6.0E-3	GO:0048513	organ development
36	607	1.05583387628627E-5	2.6E-2	GO:0051094	positive regulation of developmental process
71	1308	6.5828610286661696E-9	<0.001	GO:0048856	anatomical structure development
106	2177	9.6779568692902494E-11	<0.001	GO:0032502	developmental process
63	1186	1.4342743777632399E-7	<0.001	GO:0051239	regulation of multicellular organismal process
100	2082	1.10950940201248E-9	<0.001	GO:0044767	single-organism developmental process
46	849	6.1535790751483699E-6	1.9E-2	GO:2000026	regulation of multicellular organismal development
58	1130	1.63958943701026E-6	6.0E-3	GO:0050793	regulation of developmental process
87	1825	3.9086324583793002E-8	<0.001	GO:0044421	extracellular region part
67	1381	1.5274223757993801E-6	6.0E-3	GO:0007165	signal transduction
61	1246	3.9571887771131096E-6	1.5E-2	GO:0048869	cellular developmental process
61	1310	1.9603235582128199E-5	4.3E-2	GO:0005886	plasma membrane
172	5038	1.6879121263806799E-5	3.6E-2	GO:0044699	single-organism process
5	23	4.9745376934006201E-6	1.2E-2	GO:0042562	hormone binding
5	28	1.38862542500497E-5	2.3E-2	GO:0050715	positive regulation of cytokine secretion
5	33	3.2050548484478999E-5	0.05	GO:0031214	biomineral tissue development
7	53	2.0345766863567202E-6	5.0E-3	GO:0050673	epithelial cell proliferation
8	70	1.13693401348119E-6	2E-3	GO:0007267	cell-cell signaling

9	81	3.0041699300035198E-7	<0.001	GO:0044700	single organism signaling
9	82	3.3429194861118402E-7	<0.001	GO:0023052	signaling
8	88	6.5383891616087396E-6	1.8E-2	GO:0050714	positive regulation of protein secretion
12	135	3.6213547326273002E-8	<0.001	GO:0005578	proteinaceous extracellular matrix
13	162	3.2261875621453698E-8	<0.001	GO:0031012	extracellular matrix
29	423	1.5069389485097999E-15	<0.001	GO:0005576	extracellular region
8	102	1.95630768602668E-5	3.1E-2	GO:0010810	regulation of cell-substrate adhesion
10	131	2.17879990417622E-6	5.0E-3	GO:0030855	epithelial cell differentiation
11	152	1.1176808295683201E-6	2E-3	GO:0050708	regulation of protein secretion
11	157	1.54082368252879E-6	3.0E-3	GO:0045785	positive regulation of cell adhesion
9	141	3.06226303888987E-5	4.8E-2	GO:0004888	transmembrane signaling receptor activity
11	178	5.2505061088622402E-6	1.2E-2	GO:0038023	signaling receptor activity
12	225	8.7494228194370304E-6	2.1E-2	GO:0009986	cell surface
14	277	2.7671275564476198E-6	6.0E-3	GO:0009887	organ morphogenesis
18	379	2.2382476326744001E-7	<0.001	GO:0005615	extracellular space
13	284	1.8774360373409099E-5	3.1E-2	GO:0009888	tissue development
15	339	6.1000040317212299E-6	1.7E-2	GO:0048731	system development
28	753	1.1642168786244801E-8	<0.001	GO:0009653	anatomical structure morphogenesis
22	663	4.0670727714736701E-6	7.0E-3	GO:0051240	positive regulation of multicellular organismal process
22	672	5.0636702248074002E-6	1.2E-2	GO:0048585	negative regulation of response to stimulus
20	607	1.3437935700659501E-5	2.3E-2	GO:0051094	positive regulation of developmental process

51	2177	1.73711264822752E-8	<0.001	GO:0032502	developmental process
25	891	1.5535640447557098E-5	0.03	GO:0030154	cell differentiation
31	1186	4.6699759883394602E-6	7.0E-3	GO:0051239	regulation of multicellular organismal process
33	1308	4.4320725455098598E-6	7.0E-3	GO:0048856	anatomical structure development
32	1271	6.8882026593448999E-6	1.9E-2	GO:0044707	single-multicellular organism process
32	1278	7.7412473531242206E-6	2.1E-2	GO:0032501	multicellular organismal process
31	1246	1.2975467263716299E-5	2.3E-2	GO:0048869	cellular developmental process
46	2082	9.4015002598449898E-7	1E-3	GO:0044767	single-organism developmental process
36	1578	1.40580725229098E-5	2.3E-2	GO:0048583	regulation of response to stimulus

Table B.2: Gene Ontology results for genes present in Plaid biclusters

N	X	p-value	P_adj	attrib ID	attrib name
15	18	1.32373623125089E-5	2.2E-2	GO:0015988	energy coupled proton transmembrane transport, against electrochemical gradient
15	18	1.32373623125089E-5	2.2E-2	GO:0015991	ATP hydrolysis coupled proton transport
14	17	3.4174022449091097E-5	4.9E-2	GO:0042743	hydrogen peroxide metabolic process
20	28	2.8336208197659299E-5	4.3E-2	GO:0016504	peptidase activator activity
24	34	6.1148714894831102E-6	1.1E-2	GO:1902600	hydrogen ion transmembrane transport
23	34	2.8717314646217301E-5	4.3E-2	GO:0030193	regulation of blood coagulation

23	34	2.8717314646217301E-5	4.3E-2	GO:1900046	regulation of hemostasis
30	45	2.76428956831763E-6	3.0E-3	GO:0019003	GDP binding
29	45	1.11460639671914E-5	1.7E-2	GO:0006818	hydrogen transport
29	45	1.11460639671914E-5	1.7E-2	GO:0015992	proton transport
35	55	2.1348354636556E-6	2E-3	GO:0015078	hydrogen ion transmembrane transporter activity
48	78	1.28938065803611E-7	<0.001	GO:0098800	inner mitochondrial membrane protein complex
36	61	1.8373080376874198E-5	3.3E-2	GO:0061134	peptidase regulator activity
36	62	3.01971438818212E-5	4.6E-2	GO:0010594	regulation of endothelial cell migration
40	69	1.16637683833662E-5	1.8E-2	GO:0009897	external side of plasma membrane
55	98	1.18184095341188E-6	2E-3	GO:0010632	regulation of epithelial cell migration
51	92	4.6735541671739796E-6	8.0E-3	GO:0015077	monovalent inorganic cation transmembrane transporter activity
46	83	1.36580454243054E-5	2.5E-2	GO:0008064	regulation of actin polymerization or depolymerization
46	83	1.36580454243054E-5	2.5E-2	GO:0030832	regulation of actin filament length
57	105	3.2071739366463902E-6	5.0E-3	GO:0098798	mitochondrial protein complex
63	117	1.43495121972022E-6	2E-3	GO:0044455	mitochondrial membrane part
70	132	8.1200775357918205E-7	<0.001	GO:0043209	myelin sheath
53	101	2.5378189508670001E-5	4.1E-2	GO:0007162	negative regulation of cell adhesion
63	121	6.3383919330692801E-6	1.1E-2	GO:0044391	ribosomal subunit
53	102	3.6256347802155599E-5	0.05	GO:0098552	side of membrane
99	194	5.8594911117478399E-8	<0.001	GO:0000323	lytic vacuole
99	194	5.8594911117478399E-8	<0.001	GO:0005764	lysosome

72	141	3.5786909701652E-6	5.0E-3	GO:0022890	inorganic cation transmembrane transporter activity
85	169	1.10572630343214E-6	<0.001	GO:0008324	cation transmembrane transporter activity
63	126	3.33351935090832E-5	4.8E-2	GO:0032535	regulation of cellular component size
85	171	2.0595086570090898E-6	2E-3	GO:0032970	regulation of actin filament-based process
111	226	1.30928486559545E-7	<0.001	GO:0005773	vacuole
710	1537	1.4736681496055E-34	<0.001	GO:0043230	extracellular organelle
710	1537	1.4736681496055E-34	<0.001	GO:1903561	extracellular vesicle
708	1533	2.08254310182311E-34	<0.001	GO:0065010	extracellular membrane-bounded organelle
708	1533	2.08254310182311E-34	<0.001	GO:0070062	extracellular exosome
91	187	2.8863958076295201E-6	5.0E-3	GO:0007264	small GTPase mediated signal transduction
100	207	1.44188234123197E-6	2E-3	GO:0030335	positive regulation of cell migration
821	1825	1.4245802710451599E-35	<0.001	GO:0044421	extracellular region part
777	1721	5.5356669608676697E-34	<0.001	GO:0031988	membrane-bounded vesicle
101	210	1.65287269400123E-6	2E-3	GO:2000147	positive regulation of cell motility
77	160	2.6967290849820602E-5	4.1E-2	GO:0032956	regulation of actin cytoskeleton organization
184	386	2.2069265984688799E-10	<0.001	GO:0051270	regulation of cellular component movement
173	363	7.9690413834354295E-10	<0.001	GO:2000145	regulation of cell motility
104	218	1.87014329539657E-6	2E-3	GO:0040017	positive regulation of locomotion
165	348	3.1889988017941999E-9	<0.001	GO:0030334	regulation of cell migration
835	1884	2.3516004352205102E-33	<0.001	GO:0031982	vesicle

187	398	6.9202112931759804E-10	<0.001	GO:0040012	regulation of locomotion
147	312	3.9042401971170998E-8	<0.001	GO:0005912	adherens junction
102	217	5.3743770888666702E-6	0.01	GO:0098609	cell-cell adhesion
101	215	6.16434450561136E-6	1.1E-2	GO:0051272	positive regulation of cellular component movement
85	181	3.3331382211468397E-5	4.8E-2	GO:0005911	cell-cell junction
148	317	7.2367329399200505E-8	<0.001	GO:0070161	anchoring junction
90	192	2.16186585900075E-5	3.5E-2	GO:0016337	single organismal cell-cell adhesion
127	273	7.8313454932435203E-7	<0.001	GO:0005743	mitochondrial inner membrane
131	282	5.7935545194756695E-7	<0.001	GO:0005925	focal adhesion
131	284	9.3637145201601097E-7	<0.001	GO:0005924	cell-substrate adherens junction
92	199	3.4362398209543899E-5	5.0E-2	GO:0005525	GTP binding
161	351	8.3873338609037807E-8	<0.001	GO:0031966	mitochondrial membrane
131	285	1.1850555314874601E-6	2E-3	GO:0030055	cell-substrate junction
109	237	8.9632284013221393E-6	1.5E-2	GO:0006812	cation transport
133	290	1.1431085229435899E-6	1E-3	GO:0019866	organelle inner membrane
173	379	4.1748047087099401E-8	<0.001	GO:0005615	extracellular space
110	242	1.5619139640524201E-5	2.6E-2	GO:0019725	cellular homeostasis
125	276	5.2310312484664401E-6	0.01	GO:0030155	regulation of cell adhesion
192	428	3.7735789792038701E-8	<0.001	GO:0098796	membrane protein complex
177	396	1.7770631807072501E-7	<0.001	GO:0005768	endosome
112	251	3.5613640400051103E-5	4.9E-2	GO:2001233	regulation of apoptotic signaling pathway
200	453	8.6414804294883602E-8	<0.001	GO:0002682	regulation of immune system process
210	480	9.6090625643374899E-8	<0.001	GO:0044429	mitochondrial part
130	297	2.6988580529988501E-5	4.1E-2	GO:0016023	cytoplasmic membrane-bounded vesicle
257	594	1.12500786697113E-8	<0.001	GO:0030054	cell junction

184	423	9.8210445361880801E-7	<0.001	GO:0005576	extracellular region
158	369	1.6721624752456401E-5	2.7E-2	GO:0007155	cell adhesion
255	601	1.04579396313669E-7	<0.001	GO:0005102	receptor binding
158	370	1.9956537751939701E-5	3.4E-2	GO:0022610	biological adhesion
911	2270	1.0996142110218999E-18	<0.001	GO:0044425	membrane part
171	404	1.6860100324754798E-5	2.7E-2	GO:1902533	positive regulation of intracellular signal transduction
205	489	5.2809827087913696E-6	0.01	GO:0031410	cytoplasmic vesicle
195	467	1.2140946692834701E-5	1.8E-2	GO:0016192	vesicle-mediated transport
528	1310	1.20132582108096E-10	<0.001	GO:0005886	plasma membrane
1415	3753	1.73307852914801E-18	<0.001	GO:0044444	cytoplasmic part
438	1083	4.0087255702780299E-9	<0.001	GO:0005739	mitochondrion
1406	3735	5.6987822727163698E-18	<0.001	GO:0016020	membrane
251	615	5.4028954700448998E-6	0.01	GO:0009967	positive regulation of signal transduction
263	645	3.4626844087082899E-6	5.0E-3	GO:0044459	plasma membrane part
277	682	2.7307019802821E-6	3.0E-3	GO:0023056	positive regulation of signaling
317	785	9.1143553349305905E-7	<0.001	GO:0048584	positive regulation of response to stimulus
657	1664	2.9433251516634397E-11	<0.001	GO:0031224	intrinsic component of membrane
642	1629	8.0932384314835596E-11	<0.001	GO:0016021	integral component of membrane
449	1128	2.8223093608457999E-8	<0.001	GO:0032879	regulation of localization
280	698	7.8268348981181199E-6	1.3E-2	GO:0010647	positive regulation of cell communication
375	943	5.4426784849398196E-7	<0.001	GO:0031090	organelle membrane
288	722	9.7822417076479903E-6	1.5E-2	GO:1902531	regulation of intracellular signal transduction
542	1381	8.1916062617619302E-9	<0.001	GO:0007165	signal transduction

462	1199	1.7352464375871299E-6	2E-3	GO:0009966	regulation of signal transduction
414	1074	6.3579294403810503E-6	1.1E-2	GO:0044765	single-organism transport
457	1190	3.0768837406052E-6	5.0E-3	GO:1902578	single-organism localization
512	1338	1.11927596059748E-6	1E-3	GO:0023051	regulation of signaling
603	1586	2.6306585559344101E-7	<0.001	GO:0006810	transport
453	1186	6.7455583076402501E-6	1.2E-2	GO:0051239	regulation of multicellular organismal process
594	1578	1.61548913341542E-6	2E-3	GO:0048583	regulation of response to stimulus
426	1123	2.90335153938905E-5	4.3E-2	GO:0065008	regulation of biological quality
519	1380	1.0039300784261399E-5	1.6E-2	GO:0010646	regulation of cell communication
624	1674	3.7288856409545701E-6	5.0E-3	GO:0051234	establishment of localization
684	1869	1.97953542223332E-5	3.4E-2	GO:0051179	localization
9	13	5.7299654383343997E-6	0.02	GO:0070577	lysine-acetylated histone binding
10	20	8.4669836566384405E-6	1.6E-2	GO:0003014	renal system process
18	52	1.86544996826691E-6	3.0E-3	GO:0001936	regulation of endothelial cell proliferation
24	86	3.2087375311298901E-6	6.0E-3	GO:0050839	cell adhesion molecule binding
26	103	9.5572461894383898E-6	2.1E-2	GO:1901342	regulation of vasculature development
33	135	1.3638991061002901E-6	2E-3	GO:0005578	proteinaceous extracellular matrix
38	162	6.7165841553267E-7	<0.001	GO:0031012	extracellular matrix
57	311	7.5814129638389197E-6	1.5E-2	GO:0016477	cell migration
60	334	8.39096234562273E-6	1.6E-2	GO:0048870	cell motility
84	490	9.7163802154988691E-7	1E-3	GO:0022603	regulation of anatomical structure morphogenesis
68	401	1.58572554684693E-5	3.3E-2	GO:0008284	positive regulation of cell proliferation

78	478	1.5859507448228801E-5	3.3E-2	GO:0006928	movement of cell or subcellular component
112	721	2.4089271563461402E-6	4.0E-3	GO:0042127	regulation of cell proliferation
101	663	1.8433266311205599E-5	3.5E-2	GO:0051240	positive regulation of multicellular organismal process
133	889	2.0153431272333799E-6	3.0E-3	GO:0042221	response to chemical
121	812	7.59118924114161E-6	1.5E-2	GO:0045595	regulation of cell differentiation
125	849	9.6463141056353604E-6	2.1E-2	GO:2000026	regulation of multicellular organismal development
160	1130	4.8232293176944698E-6	1.1E-2	GO:0050793	regulation of developmental process
6	33	3.7659508188293199E-6	6.0E-3	GO:0031214	biomineral tissue development
6	43	1.8623267633530901E-5	2.9E-2	GO:0070167	regulation of biomineral tissue development

Table B.3: Gene Ontology results for genes present in SAMBA biclusters

N	X	p-value	P_adj	attrib ID	attrib name
3	3	6.7242998927886702E-7	<0.001	GO:0005594	collagen type IX
3	4	2.6724838285934402E-6	<0.001	GO:0005593	FACIT collagen
3	5	6.6383615780194199E-6	3.0E-3	GO:0030934	anchoring collagen
4	10	1.15843286945534E-6	<0.001	GO:0001502	cartilage condensation
4	13	3.8652502656835102E-6	<0.001	GO:0007338	single fertilization
4	13	3.8652502656835102E-6	<0.001	GO:0009954	proximal/distal pattern formation
4	19	2.01209171507659E-5	2.1E-2	GO:0035136	forelimb morphogenesis
14	129	5.0438816122287201E-12	<0.001	GO:0005578	proteinaceous extracellular matrix
5	45	4.4840710216580199E-5	4.5E-2	GO:0005581	collagen
7	70	2.5393064828789199E-6	<0.001	GO:0009952	anterior/posterior pattern specification
7	70	2.5393064828789199E-6	<0.001	GO:0035107	appendage morphogenesis

7	70	2.5393064828789199E-6	<0.001	GO:0035108	limb morphogenesis
14	156	6.8153585611196698E-11	<0.001	GO:0031012	extracellular matrix
8	85	7.4702121287196396E-7	<0.001	GO:0001501	skeletal system development
7	78	5.2867336713415004E-6	2E-3	GO:0061448	connective tissue development
8	100	2.5952039821145798E-6	<0.001	GO:0044420	extracellular matrix part
8	117	8.4231791647478101E-6	4.0E-3	GO:0003002	regionalization
22	419	6.3456310474107002E-12	<0.001	GO:0005576	extracellular region
9	180	2.8303743755832499E-5	2.9E-2	GO:0007389	pattern specification process
12	264	3.18079672752497E-6	<0.001	GO:0009888	tissue development
18	681	2.13676527274008E-5	2.1E-2	GO:0009653	anatomical structure morphogenesis
21	903	2.6634733783446199E-5	2.9E-2	GO:0050793	regulation of developmental process
7	18	2.1476094727665501E-7	1E-3	GO:0001968	fibronectin binding
6	21	1.2946937821925199E-5	2.2E-2	GO:0005201	extracellular matrix structural constituent
8	28	4.3929069402221502E-7	1E-3	GO:0030193	regulation of blood coagulation
8	28	4.3929069402221502E-7	1E-3	GO:1900046	regulation of hemostasis
8	30	7.8991383901089397E-7	2E-3	GO:0050818	regulation of coagulation
7	31	1.3193597140846101E-5	2.3E-2	GO:0031214	biomineral tissue development
8	40	8.2452037757126004E-6	1.7E-2	GO:0004222	metalloendopeptidase activity
8	40	8.2452037757126004E-6	1.7E-2	GO:0004930	G-protein coupled receptor activity
8	40	8.2452037757126004E-6	1.7E-2	GO:0050900	leukocyte migration
11	56	1.95386553898746E-7	1E-3	GO:0051216	cartilage development
8	42	1.2085441137275601E-5	2.1E-2	GO:0061041	regulation of wound healing
8	43	1.4510308541343901E-5	2.5E-2	GO:0014706	striated muscle tissue development

8	44	1.7331763917378299E-5	3.4E-2	GO:0008083	growth factor activity
9	55	1.25988601871862E-5	2.1E-2	GO:0001894	tissue homeostasis
12	79	1.0135086931399E-6	2E-3	GO:0005539	glycosaminoglycan binding
13	94	1.08054534735346E-6	2E-3	GO:0030198	extracellular matrix organization
13	94	1.08054534735346E-6	2E-3	GO:0043062	extracellular structure organization
17	135	9.4541102958638404E-8	<0.001	GO:0032844	regulation of homeostatic process
15	119	5.4262435116954098E-7	2E-3	GO:0016337	cell-cell adhesion
18	145	4.9013726830135801E-8	<0.001	GO:0004888	transmembrane signaling receptor activity
23	209	6.67289010910742E-9	<0.001	GO:0051240	positive regulation of multicellular organismal process
17	156	7.7686259502437504E-7	2E-3	GO:0043269	regulation of ion transport
13	120	1.7006065703730101E-5	3.2E-2	GO:0030855	epithelial cell differentiation
19	181	3.0812444769391599E-7	1E-3	GO:0038023	signaling receptor activity
32	327	1.29521023459293E-10	<0.001	GO:0005615	extracellular space
18	182	1.5342483975595201E-6	3.0E-3	GO:0009986	cell surface
22	235	2.6568263275891399E-7	1E-3	GO:0004872	receptor activity
37	425	1.23012468734406E-10	<0.001	GO:0007275	multicellular organismal development
24	293	8.8137008735018296E-7	2E-3	GO:0007155	cell adhesion
24	293	8.8137008735018296E-7	2E-3	GO:0022610	biological adhesion
24	307	2.0284707701572798E-6	3.0E-3	GO:0003008	system process
77	1160	3.4677998641169598E-15	<0.001	GO:0044707	single-multicellular organism process
77	1170	5.5889224631680098E-15	<0.001	GO:0032501	multicellular organismal process
37	510	1.9252484575454499E-8	<0.001	GO:0048513	organ development

76	1179	2.8041357960450999E-14	<0.001	GO:0048856	anatomical structure development
30	412	4.5425741071850101E-7	1E-3	GO:0051094	positive regulation of developmental process
46	678	2.33673618506596E-9	<0.001	GO:2000026	regulation of multicellular organismal development
61	959	4.1233346287006103E-11	<0.001	GO:0051239	regulation of multicellular organismal process
26	384	1.0543778273163499E-5	0.02	GO:0022603	regulation of anatomical structure morphogenesis
68	1213	6.4148987609711403E-10	<0.001	GO:0005886	plasma membrane
73	1326	2.6425542286391201E-10	<0.001	GO:0044421	extracellular region part
47	838	5.5272331065761402E-7	2E-3	GO:0030154	cell differentiation
62	1148	2.0425189390607301E-8	<0.001	GO:0048869	cellular developmental process
97	1995	8.9421028787807098E-11	<0.001	GO:0032502	developmental process
37	656	9.5493325557638202E-6	1.9E-2	GO:0045595	regulation of cell differentiation
92	1900	4.8776232082356105E-10	<0.001	GO:0044767	single-organism developmental process
76	1690	8.1809274391339595E-7	2E-3	GO:0016021	integral component of membrane
166	4851	7.5870970957759897E-7	2E-3	GO:0044699	single-organism process
9	90	2.6095611104555499E-5	4.8E-2	GO:0060249	anatomical structure homeostasis
15	235	1.49552945362614E-5	2.7E-2	GO:0005509	calcium ion binding
6	6	5.1525771335380004E-10	<0.001	GO:0005833	hemoglobin complex
4	4	6.5246180787634404E-7	1E-3	GO:0031838	haptoglobin-hemoglobin complex
4	6	9.3506809547410406E-6	1.9E-2	GO:0019825	oxygen binding
5	12	1.23327513575794E-5	0.02	GO:0042481	regulation of odontogenesis

5	13	1.9574092493763898E-5	2.7E-2	GO:0002673	regulation of acute inflammatory response
6	20	1.4235890707758199E-5	2.5E-2	GO:0001958	endochondral ossification
6	20	1.4235890707758199E-5	2.5E-2	GO:0036075	replacement ossification
7	27	7.8177973219071094E-6	1.6E-2	GO:0014068	positive regulation of phosphatidylinositol 3-kinase signaling
8	35	4.8177706637263797E-6	1.5E-2	GO:0072562	blood microparticle
9	47	5.8351864821531702E-6	1.5E-2	GO:0060560	developmental growth involved in morphogenesis
9	48	7.0043949536272601E-6	1.5E-2	GO:0010634	positive regulation of epithelial cell migration
14	78	3.4206640600406898E-8	<0.001	GO:0001503	ossification
9	52	1.3903546707598601E-5	2.3E-2	GO:0010594	regulation of endothelial cell migration
10	62	8.8777646637907097E-6	1.6E-2	GO:0009897	external side of plasma membrane
11	74	7.1392045823786001E-6	1.5E-2	GO:0010632	regulation of epithelial cell migration
16	112	1.02144336347113E-7	<0.001	GO:0007186	G-protein coupled receptor signaling pathway
18	194	1.04256301848236E-5	0.02	GO:0048729	tissue morphogenesis
30	372	2.3975053213472402E-7	<0.001	GO:0009605	response to external stimulus
47	689	1.39377935547036E-8	<0.001	GO:0007166	cell surface receptor signaling pathway
40	633	1.4468106600721699E-6	2E-3	GO:0042127	regulation of cell proliferation
45	828	1.66721281603162E-5	2.5E-2	GO:0032879	regulation of localization
65	1342	7.5511931389333203E-6	1.6E-2	GO:0007165	signal transduction
104	2386	6.34557268892082E-7	1E-3	GO:0050896	response to stimulus
12	86	3.4484298389747099E-6	8.0E-3	GO:0006935	chemotaxis
12	87	3.9042783378155103E-6	8.0E-3	GO:0042330	taxis

12	87	3.9042783378155103E-6	8.0E-3	GO:1901342	regulation of vasculature development
16	165	1.0981528744944701E-5	1.8E-2	GO:0040017	positive regulation of locomotion
24	332	1.27698723525655E-5	1.9E-2	GO:0040011	locomotion
34	578	1.7530025641233999E-5	2.6E-2	GO:0044459	plasma membrane part
92	2241	8.0897328772972403E-6	1.6E-2	GO:0044425	membrane part
146	4019	2.5154927157424001E-6	4.0E-3	GO:0044763	single-organism cellular process
8	56	1.28600730011157E-5	3.9E-2	GO:0043270	positive regulation of ion transport
12	92	2.3286175626045E-7	<0.001	GO:0034762	regulation of transmembrane transport
11	86	9.2054096415440104E-7	3.0E-3	GO:0034765	regulation of ion transmembrane transport
9	100	5.6271381796036003E-6	7.0E-3	GO:0015672	monovalent inorganic cation transport
8	89	1.92279006335895E-5	2.3E-2	GO:0015077	monovalent inorganic cation transmembrane transporter activity
12	191	6.6680806332757703E-6	0.01	GO:0007610	behavior
15	330	2.33011197117117E-5	3.4E-2	GO:0043565	sequence-specific DNA binding
17	379	7.5400392528364301E-6	1.1E-2	GO:0003700	sequence-specific DNA binding transcription factor activity
17	380	7.8078278140831392E-6	1.2E-2	GO:0001071	nucleic acid binding transcription factor activity
5	26	1.90069714410369E-5	3.5E-2	GO:0048706	embryonic skeletal system development
8	74	5.13898170976169E-6	9.0E-3	GO:0007267	cell-cell signaling
8	82	1.1134832356554401E-5	2.5E-2	GO:0023052	signaling
8	82	1.1134832356554401E-5	2.5E-2	GO:0044700	single organism signaling
10	137	1.18668378716142E-5	3.0E-2	GO:0007154	cell communication

Appendix C

cMonkey2 Results Tables

Table C.1: This presents a tabular view of the information in Figure 9.1. The number of rows and number of columns in each bicluster are shown as well as the names of the rows (genes) and columns (condition). k is the bicluster number, while the residue is a representation of the similarity in gene expression contained in each bicluster. The genes that were predicted by Inferelator as potential regulators are shown, as well as transcription factors within each bicluster.

nrows	ncols	rows	cols	k	resid	outside	weight	TFs
35	8	Smdt1, Rab27b, Manbal, Ube2m, Cmbl, Actb, Tufm, Arid3a, Sh3pxd2a, 2010300C02Rik, Cd59a, Smad6, Syt6, Fam109b, Grb2, Mpdu1, Pitx1, Gaa, Fetub, Abhd17a, Prex1, Rab3d, Fam46a, Frmd4b, Rnaset2b, Sema4d, Rasgrp2, Ptprs, Mrps23, Glrx5, Cd68, Clta, 2810025M15Rik, Arid1b, Sec31a	MAT2, MAT3, MAT1, BON3, BON2, BON1, IMA1, IMA3	38	0.1756	Nfat5, Sox9, Foxn3	(-0.51384, -0.26325, 0.18629)	Arid3a, Smad6, Pitx1, Arid1b
24	5	Sec11c, Zfp827, Cacna1c, Ovca2, Mkl2, Gpr153, Sh3rf1, Fam222b, Ercc6, Sp7, Plekhm3, Mapk12, Rassf3, Msi1, Slc48a1, RP23-380F8.2, Med7, AI480526, Snx15, Tbx3, Cnn2, Btd, Lppr4, Dlx5	MAT2, BON1, IMA1, MAT3, IMA2	250	0.1758	Aff1, Runx1t1, Erf, Aff3, Tle4, Meis2, Smarcd1, Runx2, Atf1, Smad9, Bbx	(0.20778, -0.11947, -0.11498, -0.10411, -0.10078, -0.084329, 0.078725, 0.054123, 0.049653, 0.043559, -0.028171)	Mkl2, Sp7, Tbx3, Dlx5

25	5	Maea, Coa4, Dmp1, Vps37a, Rc3h1, Rnf185, Ptgr1, Kit, RP23-346I1.4, Atp6v1b2, Slc25a11, Dner, Rps17, RP24-546N2.4, Gdi2, Fam102a, Zfp560, Mrpl22, Enpp6, Fam214a, Gadd45b, Atp5h, D1Ertd622e, Myo10, Gars	BON1, IMA1, MAT3, IMA3, MAT1	290	0.1337	Aff1, Aes, Runx2, Bbx, Mef2c, Irx3, Meis1, Mxd1, Sin3b, Sox8, Sox5	(0.15331, 0.14317, 0.13503, -0.11415, 0.10777, 0.09347, -0.088752, 0.050765, 0.039792, -0.036668, -0.029867)	0
29	5	Akr1b10, Zfp280b, Gatad2a, Mrpl36, BC052040, Cdc73, Tpd52, Fap, Pcsk6, Slc31a2, Stk24, Yipf6, Isca2, Rnf19a, Sf3b5, Tmx1, Sec23ip, 2700029M09Rik, Bola3, Zbtb34, Repin1, Srprb, Kirrel, Gadd45b, Egr1, Creb3, Kcnma1, Gbe1, Abhd4	BON3, BON1, MAT1, MAT3, IMA2	297	0.1215	Cebpb, Aff3, Hoxa5, Plagl1, Hoxa9, Sox9, Pcbd1	(0.093631, -0.088493, -0.079995, -0.058402, -0.057215, -0.044339, 0.022497)	Egr1, Creb3
30	6	Ufl1, Ndufa3, Plaur, Psat1, Wif1, Gprc5c, Mrpl17, 2010300C02Rik, Tdg, Fgf13, Myo10, Slc37a2, Aes, Scamp3, Gnb2l1, Wbp1l, Mcoln1, Aamp, Vma21, Sema4d, Pcyt1a, Cebpb, Ddx59, Sike1, Ramp1, Trp53bp2, Zdhhc9, Atp5h, Pcnx, 2810025M15Rik	MAT3, MAT1, BON3, BON2, BON1, IMA2	332	0.1495	Lhx8, Pitx1, Foxa3, Aff3, Arhgef12, Arid3a, Hoxa9, Msx1, Sox9, Smad6, Ostf1	(0.12165, 0.10557, -0.095115, -0.087054, 0.070036, 0.061415, -0.054262, 0.053118, -0.052594, 0.050472, 0.036797)	Tdg, Aes, Cebpb

33	5	Igf1, Runx2, Fam129a, Arap3, Tfdp2, Sepp1, Tgfb3, Zfp282, Ech1, Zfp532, 1700037H04Rik, Prkar2b, Cgnl1, Rars2, Dtna, Arhgap28, Dach2, Akip1, Pacs2, Lacc1, Samd14, Vamp5, Lrrfip2, Sec62, Rbp1, Adamtsl1, Ahcy, Celf2, Efna1, Hs3st3b1, Atp9a, Gas2, Bcar3	BON3, BON1, MAT2, MAT3, IMA3	343	0.1834	0	0	Runx2, Tfdp2
26	9	Stil, Nuf2, Esco2, Rock1, Loxl3, Cyr61, Taf5l, Maml2, Scube3, Cenpk, Cytl1, Kcna6, Spsb4, Fgfrl1, Fam57a, Vrk1, Cpm, Cyp26b1, Mbnl1, Sox5, Papss2, Anln, Enpp2, Clcn5, Zfp131, Peg3	MAT2, MAT3, MAT1, BON3, BON2, BON1, IMA1, IMA3, IMA2	345	0.1841	Creb3l1, Hoxa5, Ezh2, Zbtb16, Hoxd9, Msx1, Sox9	(-0.23248, 0.1899, 0.17186, -0.13872, 0.10672, -0.097227, 0.06576)	Sox5
20	6	Fxyd6, Abhd11, Ddhd2, Nin, 2700097O09Rik, Dok1, Sfrp5, Hk2, Tecpr2, 1110051M20Rik, Sulf2, Zfp318, Ssx2ip, Hoxd4, Ercc6l2, Rab23, Zfp367, Mcmbp, Nacc2, Zfp101	MAT3, MAT1, BON3, BON2, BON1, IMA1	549	0.1521	Runx2, Creb3, Hif1a, Arid5b, Smc1a, Aff3, Dlx3, Aff1, Mbd2, Hoxd8, Aes, Sox8, Taf12	(-0.17095, -0.15654, 0.14414, -0.090853, 0.064771, 0.059541, -0.058945, -0.055944, -0.050454, 0.047497, -0.042976, 0.03527, -0.024977)	Hoxd4

26	8	Ncoa2, Gnas, Msh6, Pde4dip, 1700021K19Rik, Igf2os, Gm24187, Sel1l, 3110079O15Rik, Fam73b, Spsb1, Art3, Fam57a, Cep41, Map2, Ap4e1, Sox9, Fam101a, Gm24270, Csgalnact1, Hoxa5, Hoxc6, Itih5l-ps, Fancb, Gm23935, Dact3	MAT3, MAT1, BON3, BON2, BON1, IMA1, IMA3, IMA2	592	0.1616	Hoxa9, Foxa3, Hoxd9, Id3, Cebpb, Hoxa10	(0.24734, 0.23389, 0.18733, -0.12054, -0.0989, 0.063471)	Ncoa2, Sox9, Hoxa5, Hoxc6
37	6	Colgalt2, Gnas, Islr, Mcoln2, Pde3a, Art3, Il16, Hoxa10, Runx3, Zbtb20, Hoxc6, Tmbim1, Clcc1, Comp, Lrp8, Pds5b, Car12, Wscd2, Myh14, Plod2, Plekhb1, Sox9, Fbln2, Gfpt2, Prelp, Hoxa9, Csgalnact1, Matn1, Matn3, Acan, Col9a2, Col9a1, Itih5l-ps, C1qtnf3, Diap3, Trim47, Ncmap	MAT3, BON3, BON2, BON1, IMA1, IMA3	690	0.1185	Foxn3, Aff3, Foxa3, Aes, Msx1, Zfand3, Hoxd9, Pitx1, Lhx8, Ikzf2	(-0.16009, 0.13877, 0.11165, -0.10233, -0.090299, -0.085989, 0.082594, -0.077011, -0.06043, -0.039318)	Hoxa10, Runx3, Hoxc6, Sox9, Hoxa9
20	7	Gm15654, Map1a, Gm26870, Gm10800, Foxp4, Gm10801, Slc6a8, Mfn1, Cep192, Gm21738, Fam73b, Gm10719, Gm10722, Gm10718, Gm10717, Rrp12, Trpv4, Taf5l, Trip10, Gm10720	MAT2, MAT3, MAT1, BON3, BON2, BON1, IMA2	770	0.1611	Tuba1a, Tdg, Sox9, Mlx, Aes, Kat2a	(-0.32683, -0.28026, 0.19034, -0.17917, -0.092253, 0.073017)	Foxp4

31	4	Dgcr8, Fnip2, Prdm11, AU019823, Slc16a3, Prkdc, Rps19-ps3, Nol8, Cep95, Tex9, Sfi1, Nfkbiz, Rrp12, Ivns1abp, 4930523C07Rik, Heatr3, G2e3, Tacc3, Erf, Stk4, Slc1a5, Hhat, Trim24, Carf, Birc5, Ccdc25, Map7d2, Fance, Ift172, Rbp4, Uhrf2	BON2, MAT2, MAT3, MAT1	780	8.58E-2	Runx2, Arid4a, Ets2, Erg, Gsc, Arhgef12, Kat2a, Hmgn3, Taf1a, Arid1a, Ewsr1	(-0.17486, 0.12582, -0.12206, 0.0913, -0.079742, -0.077608, 0.074218, -0.063882, 0.05556, -0.049714, 0.020405)	Erf, Trim24
27	4	Lgmn, Npnt, Arhgap12, Vim, Intu, Ttc17, Dnaja3, Mxd1, Cd200, Guca1a, Slc10a3, Antxr2, Tspan11, Chn2, Strn, Cdk2ap2, Runx2, Pex7, Nol11, Kitl, Zfp36l1, Msrb2, Trp53inp2, Slc25a20, Ptprm, Dctn6, Mrpl46	BON2, MAT2, MAT3, MAT1	820	7.4E-2	0	0	Mxd1, Runx2