Computational methods for analysing the regulation of genetic systems across species

Djordje Djordjevic

A thesis in fulfillment of the requirements for the degree of

Doctor of Philosophy

UNSW
AUSTRALIA

St Vincent's Clinical School

Faculty of Medicine

The University of New South Wales

August 2017

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family name: **Djordjevic**

First name: **Djordje** Other name/s:

Abreviation for degree as given in the University calendar: **PhD**

School: **St Vincent's Clinical School** Faculty: **Faculty of Medicine**

Title: Computational methods for analysing the regulation of genetic systems across species

Abstract 350 words maximum

Gene regulation in humans and other mammalian organisms is complex with many different layers. Despite huge advances in biomedical research, our ability to build a comprehensive and predictive model of gene regulation is still limited. There are increasing amounts of public genome-wide data available, containing a wealth of untapped knowledge about different aspects of gene regulation from across the tree of life. The main challenge and bottleneck is integrating and analysing this data to extract insights about gene regulation.

In this thesis, I discuss the challenges pertaining to integrating data of different types, and from different species, to better understand gene regulation. Furthermore, I present a series of novel concepts, bioinformatics methods, software tools and case studies to address and overcome some of these issues.

The key contributions of this thesis are:

1. Investigation and method development for constructing and analysing mammalian gene regulatory networks
2. Approaches and tools for large scale systems integration of public data and knowledge
3. Investigation of issues and new methods for cross-species gene set analysis.

FOR OFFICE USE ONLY Date of completion of requirements for Award

Originality Statement

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Djordje Djordjevic
August 10, 2017

Copyright Statement

Djordje Djordjevic
August 10, 2017

Authenticity Statement

Djordje Djordjevic
August 10, 2017

This thesis is dedicated to my ancestors.

Especially the last two.

Abstract

Gene regulation in humans and other mammalian organisms is complex with many different layers. Despite huge advances in biomedical research, our ability to build a comprehensive and predictive model of gene regulation is still limited. There are increasing amounts of public genome-wide data available, containing a wealth of untapped knowledge about different aspects of gene regulation from across the tree of life. The main challenge and bottleneck is integrating and analysing this data to extract insights about gene regulation.

In this thesis, I discuss the challenges pertaining to integrating data of different types, and from different species, to better understand gene regulation. Furthermore, I present a series of novel concepts, bioinformatics methods, software tools and case studies to address and overcome some of these issues.

The key contributions of this thesis are:

1. Investigation and method development for constructing and analysing mammalian gene regulatory networks
2. Approaches and tools for large scale systems integration of public data and knowledge
3. Investigation of issues and new methods for cross-species gene set analysis.

Preface

All research conducted during my PhD candidature involved collaboration with my supervisors and external parties. Most of the results presented in this thesis have been published, or submitted for publication, in international peer–reviewed journals. Only materials in which I have made the most significant contribution are included in this thesis.

List of publications

1. Djordje Djordjevic, Kenro Kusumi, Joshua W. K. Ho (2016), **XGSA: a statistical method for cross-species gene set analysis**, *Bioinformatics*, 32 (17): i620-i628

2. Xin Wang, Helen McCormick, Djordje Djordjevic, Eleni Giannoulatou, Catherine M Suter, Joshua WK Ho (2016), **Epigenomic analysis of chromatin organization and DNA methylation**, *Computational Biology Bioinformatics Gene Regulation*, (Ed. Wong KC), CRC Press, 181-211

3. Kyung-Ah Sohn, Joshua W. K. Ho, Djordje Djordjevic, Hyun-hwan Jeong, Peter J. Park and Ju Han Kim (2015), **hiHMM: Bayesian non-parametric joint inference of chromatin state maps**, *Bioinformatics*, 31(13), 2066-2074

4. Djordje Djordjevic, Vinita Deshpande, Tomasz Szczesnik, Andrian Yang, David T. Humphreys, Eleni Giannoulatou, Joshua W. K. Ho (2015), **Decoding the complex genetic causes of heart diseases using systems biology**, *Biophysical Reviews*, 7, 141-159

5. Djordje Djordjevic, Andrian Yang, Armella Zadoorian, Kevin Rungrugeecharoen, Joshua W. K. Ho, (2014), **How difficult is inference of mammalian causal gene regulatory networks?**, *PLOS ONE*, 9(11), e111661

List of submitted publications

1. Djordje Djordjevic, Yun Xin Chen, Shu Lun Shannon Kwan, Raymond, Ling, Gordon Qian, Chelsea Woo, Samuel Ellis and Joshua W. K. Ho (2017) **GEOracle:**

classification based on free text annotation in Gene Expression Omnibus, submitted to *biorXiv*, doi: https://doi.org/10.1101/150896

2. Joanna Palade*, Djordje Djordjevic*, Elizabeth D. Hutchins, Rajani M. George, John A. Cornelius, Alan Rawls, Joshua W.K. Ho, Kenro Kusumi, and Jeanne Wilson-Rawls (2017), **Identification of satellite cells from anole lizard skeletal muscle and demonstration of increased musculoskeletal potential**, in revision at *Developmental Biology*, * co-first authors

3. Patricia Murphy, Md. Humayun Kabir, Tarini Srivastava, Michele Mason, Andrian Yang, Djordje Djordjevic, Murray Killingsworth, Joshua Ho, David Harman, Michael O'Connor (2017), **Light-focusing human micro-lenses derived from zebrafish-like lens cell masses model lens development and drug-induced cataract in vitro**, in revision at *Development*

4. Atul Kakrana*, Andrian Yang*, Deepti Anand, Djordje Djordjevic, Deepti Ramachandruni, Abhyudai Singh, Hongzhan Huang, Joshua W. K. Ho, Salil A. Lachke (2017), **iSyTE 2.0: a database for expression-based gene discovery in the eye**, submitted to *Nucleic Acids Research*, * co-first authors

List of oral presentations at international conferences

1. Djordje Djordjevic, Kenro Kusumi, Joshua W. K. Ho (2016), **XGSA: a statistical method for cross-species gene set analysis**, presented at *European Conference on Computational Biology 2016*

List of poster presentations at international conferences

1. Djordje Djordjevic, Kenro Kusumi, Joshua W. K. Ho (2016), **XGSA: a statistical method for cross-species gene set analysis**, presented at *European Conference on Computational Biology 2016* and *International Conference on Systems Biology 2016*

2. Djordje Djordjevic, Andrian Yang, Shu Lun Shannon Kwan, Joshua W. K. Ho (2015), **Harnessing a large collection of gene perturbation data to discover mammalian causal gene regulatory networks**, presented at *Intelligent Systems for Molecular Biology / European Conference on Computational Biology 2015*

3. Djordje Djordjevic, Andrian Yang, AmirHossein Kamali, Joshua W. K. Ho (2014), **Harnessing sparse gene perturbation data to discover causal gene regulatory networks**, presented at *International Conference on Systems Biology 2014*

Acknowledgements

I would first like to sincerely thank Joshua. You were the best supervisor I could have asked for and taught me so much. I appreciated your guidance more than you could know. I hope I was a good first PhD student for you.

I would like to thank Diane and Vesna, who believed in me from the start and gave me this wonderful opportunity. Thank you for the warm support and encouragement over the past 4 years.

I would like to thank the entire Ho lab, particularly Dave for all the help and knowledge in the early years, Eleni for being my first real bioinformatics colleague, and Tomasz for drinking with me more than any one else. Also the rest of the staff at the Chang and Garvan that gave me assistance over the years, particularly Bob for running such a great institute and having me on the team, Cath and Sally for always sharing jokes with me, Ash, Mark and Mirana for their bioinformatics guidance, and the IT guys and level 5 social club who kept me laughing every day.

I would like to thank my friends for their alternating support and jeers that kept me motivated. They know the real reasons I embarked on this journey. Δ

I would like to thank my big bro Alex and Jelena for their undying support and making this all happen, Mel for secretly being proud of me the whole time, and my inspirations, Tetka Radmila, Zoki, Chandra and Miljan.

I would like to thank my beloved partner Shann, who kept me alive, happy, healthy, clean and positive over this journey, and for years before it. Thank you for putting up with my insanity and anti–routine lifestyle. $< 3/2$

I could never thank my mum Nada enough, from the bottom of my heart, for giving me life, and then every opportunity I could ever hope for. For getting me through all of my ridiculous adventures, even as I occasionaly put that life at some risk. Volim te mama. Hvala puno.

Finally, I would like to thank my father Miodrag, who started this journey with me 23 years ago. Thank you for giving my life direction.

Abbreviations

AF Atrial fibrillation

AUC Area under the curve

AUROC Area under the receiver operator characteristics curve

BH Benjamini-Hochberg

CHD Congenital heart disease

CNV Copy number variation

DNA Deoxyribonucleic acid

DCM Dilated cardiomyopathy

DE Differential expression

GEO The Gene Expression Omnibus

GO The Gene Ontology

GRN Gene regulatory network

GSE Gene expression experiment

GSM Gene expression sample

GWAS Genome-wide association study

HCM Hypertrophic cardiomyopathy

NGS Next generation sequencing

OMIM Online Mendellian Inheritance in Man

PPI Protein-protein interaction

RNA Ribonucleic acid

RTP Regulator/target pair

SVM Support vector machine

SVM-RFE Support vector machine - Recursive feature extraction

TF Transcription factor

TOF Tetralogy of Fallot

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The key motivation for this study stems from the need for statistically sound and computationally efficient bioinformatics tools to integrate the growing number of published data sets and gain insights into gene regulation. This thesis focuses on developing and applying methods that aid the study of gene regulation in the context of biomedical research. Specifically, this thesis aims to:

1. Improve our ability to infer and analyse reliable causal gene regulatory networks in mammals
2. Develop tools to mine and integrate public gene expression data and functional biological information
3. Investigate and resolve statistical issues encountered during cross-species analyses

The core themes addressed by this thesis include systems biology, integrative bioinformatics analyses, gene regulation and cross-species analyses. These core themes are introduced in this chapter.

1.2 The rise of bioinformatics

1.2.1 From molecular biology to systems medicine

The molecular approach to genetics is built on the premise that physiological and cellular phenotype can be explained by the action of one or more genes. This rationale leads to a strong focus on associating genes with a particular phenotype (e.g., hypertension), biological process (e.g., ageing), or disease (e.g., cardiomyopathy). Commonly, once a candidate gene is identified, it is characterised (e.g., sequenced) and perturbed (e.g., through gene knock-out) to understand its association with the phenotype of interest. Additional experiments might be performed to assess specific functional properties of the gene and the results are then interpreted in the context of existing knowledge about the biological pathways and their relationship with the phenotype of interest.

This idea of associating a disease phenotype to one or several disease genes has dominated human genetics for many years. Linus Pauling and colleagues published their seminal paper, Sickle Cell Anemia, a Molecular Disease, over half a century ago (Pauling and Itano, 1949), in which they identified biophysical differences in haemoglobin molecules between healthy individuals and those suffering from sickle cell anaemia. This raised the concept that human disease can be explained by the biophysical properties and action of individual macromolecules (Strasser, 1999). As a result of this discovery, the focus of human disease studies shifted towards molecular alterations, an approach that has had an enormous impact on virtually all specialities in medicine (Trent, 2012).

Such a gene-centric approach has its merits since it allows us to focus our efforts on the detailed molecular mechanisms involving a small number of genes that may have a direct functional impact on a biological process or disease. However, this gene-centric view of biological processes has its limitations and is potentially misleading, especially in studying complex human diseases where both genes and environmental factors play important roles. A newspaper headline such as scientists found heart disease gene is potentially misleading, since it implies that heart disease (which itself is a heterogeneous class of diseases) can

be understood by the actions of a gene or a small number of genes alone. One practical limitation of focusing on a small number of genes is the restricted range of questions one can ask about the cellular or organ-level systems properties – such as feedback control, feedforward amplification, redundancy, robustness, hierarchy, and self-organisation. The advances in genome-wide omic technologies and computational modelling have enabled us to interrogate a large portion of the genome, transcriptome, and proteome in a much more exploratory manner. This has fuelled many interesting biological questions that we could not previously attempt to answer by reductionist approaches, leading to the emergence of systems biology (Ehrenberg, 2003; Ideker *et al.*, 2001; Westerhoff and Palsson, 2004).

In a systems biology approach, the first objective is to define the system being studied, followed by characterisation of the components of the system. The next step is to examine how these components interact, before inferring the emergent properties of the complex network. This can uncover hidden relationships and establish global principles (MacLellan *et al.*, 2012).

1.2.2 Bioinformatics

As Moore's law has continued unabated, the power of computing has continued to rise and the associated costs to fall. At the same time, the cost of modern biological experiments like DNA sequencing has fallen dramatically. We can currently sequence a human genome for $1000 and in the near future this is expected to fall to $100 (Wire, 2017). This extremely powerful technology will soon become a part of routine clinical use. This trend is repeating throughout many modern research technologies probing the world at finer resolutions, and together they are rapidly transforming cutting edge biomedical research and clinical practice into very specialised fields of big data analytics. Here is where the interdisciplinary field of bioinformatics lies, at the intersection of computer science, biology and statistics.

There are many biological problems to which the power of computers can be applied. Aside from repeating simple operations millions of times a second, computers allow advanced

statistical concepts, models and simulations to be brought to bear on biological questions. The classic bioinformatics application is sequence alignment of biomolecules, which has been an active field of research since 1970 (Needleman and Wunsch, 1970). This ability to automatically compare a sequence of amino-acids or nucloetides to vast databases has facilitated countless breakthroughs, and two bioinformatics algorithms for sequence alignment, CLUSTAL W (Thompson *et al.*, 1994) and BLAST (Altschul *et al.*, 1990), are the 10th and 12th most cited scientific papers of all time (Van Noorden *et al.*, 2014).

Combined with the explosion in DNA sequencing technology since the first human genomes were assembled in 2003, sequence alignment has facilitated the genomics revolution. Modern DNA sequencing is predominantly performed by high throughput next generation systems like those from Illumina, that produce gigabytes of short sequence data in a single experiment. By aligning these millions of sequences to a reference genome, we can seek out genetic variants and structural changes that might be linked to a disease or phenotype of interest.

The technological paradigm of DNA sequencing followed by alignment to a reference genome has been harnessed to measure many other things that are informationally or spatially linked to the genome, most of which give insight into the regulation of gene expression. The most direct example is measuring gene expression activity by sequencing cDNA libraries reversed transcribed from RNA, aligning the data back to the genome and then quantifying how much RNA was produced from each gene. Other methods measure how 'open' or active the genome is at different places, or the folded super-structure of the chromosome, linking distal regulatory elements to their target genes. Crosslinking DNA bound proteins to the genome and selecting only the protein of interest, allows us to pinpoint exactly where transcription factors bind the genome. Modifications to this approach allow us to describe the chemical state of the chromatin, an important regulatory layer. Bisulfite treatment modifies the DNA of methylated cytosines, and when combined with sequencing is a powerful method for measuring DNA methylation, a potent regulator of gene expression.

While very informative on their own, these assays become extremely powerful tools for

biological discovery when performed under multiple experimental conditions, where the results can be contrasted to reveal insights into system dynamics. A classic example is differential expression analysis of gene expression data. This common bioinformatics analysis often produces a list of genes that are regulated by the experimental variables, allowing the elucidation of downstream pathways.

These concepts are based on genomic technologies, but there are other 'omics' as well. Proteomics is the large scale study of proteins, as metabolomics is the study of metabolites. Apart from the 'omics' there are many more areas of bioinformatics, including modelling, simulations, image analysis and more. Additionally, a large portion of biomedical research takes place in non-human model organisms. All of these different sources of data present new challenges and require bioinformatics innovations to process, analyse, integrate and interpret the data.

1.3 A systems perspective on gene regulation

It is remarkable that all cells in a multi-cellular organism contain the same genome yet they express different sets of genes, and respond differently to the same genetic or signalling perturbation, in a highly regulated manner. Indeed it is this dynamic system of genetic regulation that leads to the complex phenotypes, plants, animals and ecological environments in the world today. This is very evident in the field of developmental biology for example, which is essentially the study of growth, differentiation, patterning and regeneration of cells, tissues and organs (Davidson, 2006).

There are many mechanisms through which cells interpret the same genome differently. Transcription of DNA to RNA is controlled by a variety of dynamic biological processes, including changes in the conformation of the DNA that expose some sections while hiding others, and brings distal elements close together such that they interact. Chemical modifications to the DNA molecule such as methylation, or to the histone proteins that bind the DNA also have an effect on transcription, as do the the myriad of ways in which

a combination of transcription factors (TFs) bind to the genome at TF-specific sequence motifs in non-coding DNA regulatory elements.

Additional regulation occurs before messenger RNA is translated into proteins, through alternative splicing, RNA modification and degradation. Finally the translated proteins undergo post-translational modifications such as phosphorylation, form into complexes and undergo conformational changes, are transported or shuttled around the cell and sometimes degraded. The proteins form the bulk of the cellular machinery as well as the complex signalling pathways that feed environmental information back into the regulatory systems, allowing the cell to adapt to internal or external stimuli.

1.3.1 Epigenomics

The word 'epigenetics' can be literally interpreted as that which is "on top of" the genome. This term is used to describe the physical changes that occur around the DNA molecule that can be passed on to the next generation of cells or organisms. These changes are what defines cellular identity and determines cell-type specific phenotype and function. The study of these changes using genome wide techniques is called 'epigenomics'.

Eukaryotic genomes are packaged into chromatin through interactions with histone and non-histone chromosomal proteins along the entire length of DNA. The structure of chromatin is highly dynamic and is regulated through covalent modifications of histones and other forms of chromatin remodelling during development and disease pathogenesis (Chang *et al.*, 2004; Han *et al.*, 2011), thus conferring an additional level of control of gene expression. Chromatin accessibility dictates how available the DNA is to be bound by transcription factors that can promote or inhibit transcription.

Epigenetic mechanisms are heritable and affected by nutritional and environmental factors (Udali *et al.*, 2013). It is the ability of epigenetic mechanisms to be stably transmitted across cell divisions and generations of offspring that creates the diversity of multi-cellular complex organisms. However, epigenetic inheritance also introduces potential points of

system dysregulation, leading to disease.

Histone modifications involve the addition of chemical groups (e.g., methyl, acetyl, phosphate) to the N-terminus of histone proteins. These modifications, individually or in combination, often correlate with specific genomic features such as promoters, enhancers, transcribed regions, Polycomb associated domains, and heterochromatin. For example, active regions such as promoters are typically marked by trimethylation of lysine 4 of histone H3 (H3K4me3) and acetylation of lysine 27 of histone H3 (H3K27ac), while repressed regions such as hetero-chromatin are marked by H3K9me3. Histone modifications are catalysed by histone modifying enzymes including methyltransferases and acetyltransferases, while other groups of proteins read and remove these modifications.

An application of collecting histone modification data is genome-wide chromatin state annotation, which can be applied to discover candidate regulatory elements such as enhancers, especially those associated with a disease. The concept of chromatin states arises from the observation that of the possible combinations of histone modifications, only a subset is confidently observed within a cell. Each unique combination of histone modifications becomes a chromatin state, which can then be used to functionally annotate the entire genome. Computational tools have been developed for this task, such as ChromHMM (Ernst and Kellis, 2012) and Spectacle (Song and Chen, 2014), which are based on Hidden Markov Models.

1.3.2 Causal gene regulatory networks

It is increasingly clear that there is a need to fully unravel cell type-specific gene regulatory networks (GRNs) in order to understand the complex mechanisms underlying many developmental processes (Davidson, 2010; Ho, 2012; Levine and Davidson, 2005). To achieve better understanding of complex genetic causes in organ development and diseases, we need to take a *systems approach* that interrogates causal genetic regulatory relationships (*e.g.*, conditional knockout of *Pax9* reduces the expression of *Msx1* (Kitano, 2002)). Therefore in this thesis we will focus on the inference of *causal GRNs* in which each node represents

a gene, and each edge represents a causal regulatory relationship between two genes.

The identification of causal gene regulatory relationships has a long history in the study of mammalian organ development, despite being primarily driven by hypothesis-based candidate gene investigations. Over the last half a century, developmental biologists have often used low throughput *in vivo* techniques such as *in situ* hybridisation and immunohistochemistry to accurately detect spatio-temporal changes in gene expression in response to targeted gene knock-out, knock-down or over-expression experiments. By summarising the results of many of these experiments, we can incrementally infer reliable causal GRNs. A classic example of this is Eric Davidson's work on constructing and analysing GRNs in sea urchin and other animals (Davidson, 2010; Levine and Davidson, 2005).

With the increasingly widespread availability of genome-wide expression profiling technology, such as microarray and next-generation sequencing, we are now able to measure the expression levels of almost all the genes in the genome simultaneously. This gave rise to the tantalising prospect that we could infer GRNs from many expression profiles, by measuring the correlation between gene activity (Bansal *et al.*, 2007; Friedman *et al.*, 2000). Previous studies have shown that although network inference is partially achievable in prokaryotic organisms, inference in eukaryotic organisms still remains a major challenge (Marbach *et al.*, 2010, 2012).

This thesis aims to improve our ability to infer and analyse GRNs in the mammalian context. First it will investigate whether inference of causal GRNs based on the correlation of expression data is achievable in a mammalian organ, by comparing a large collection of 'gold standard' perturbation data to GRNs inferred by applying a variety of popular approaches to appropriate gene expression profiles. Additionally, it will develop and apply a method for analysing directed GRNs to identify key regulators of a set of target genes. Furthermore, it will present a tool to quickly extract many perturbation data from GEO, and construct large reliable causal GRNs.

1.4 Integrative bioinformatics analyses

Most complex phenotypes occur due to an interaction between multiple biological pathways. There is no single data type that can fully represent all aspects of these pathways. Thus one of the ongoing challenges in bioinformatics is to intelligently integrate an increasing volume and variety of data to discover more complex biological insights.

The raw data from published studies are often stored in databases, many of which are publicly available. Examples of these resources include the Gene Expression Omnibus (GEO – central repository for gene expression data, primarily from microarray studies, Barrett *et al.* (2013)) and Sequence Read Archive (SRA – central repository of DNA sequencing data, Leinonen *et al.* (2011)). Furthermore, the insights from studies are stored in curated knowledge-bases, including Online Mendelian Inheritance in Man (OMIM – links genes and variants to disease phenotypes, Hamosh (2004)), Protein Data Bank (PDB – protein sequence and structure data base, Berman *et al.* (2000)) and Gene Ontology (GO – associations between genes and biological processes, cellular compartments and molecular function, Ashburner *et al.* (2000)). Additionally, large scale projects compile huge amounts of related data into one place, like the ENCODE projects and Human Epigenomics Roadmap (Bernstein *et al.*, 2010; Dunham *et al.*, 2012; Yue *et al.*, 2014). These publicly available databases and resources are a key strength of modern bioinformatics research and allow exploratory analyses to reveal huge amounts of information very quickly.

Often gene set analysis (GSA) is the first step in an exploratory analysis of a genome-wide data set. The ability to represent information from many varied biological assays as sets of genes makes GSA a popular integrative analysis framework. Typically, a set of genes (i.e. differentially expressed genes) is first identified by applying some statistical test on the raw or processed data, then this set of genes is compared against a database of gene sets, such as those derived from GO annotations or known molecular or signalling pathways. This generates rapid insights into which previously discovered biological signals are represented within a data set. Many statistical methods have been adopted or developed to perform GSA, including the Fisher's exact test and its variants (Huang *et al.*, 2009; Rivals *et al.*,

2007).

1.4.1 Integrative analysis for disease gene variant prioritisation

When searching for the genetic cause of a disease, a whole-exome sequencing experiment may return hundreds of predicted deleterious variants, and thus there is an urgent need to intelligently prioritise these variants for functional investigation in the laboratory. In the past, gene prioritisation was done manually and relied mainly on individual, domain-specific expertise. This approach is clearly hard to scale up to the genome-wide level. Most current variant prioritisation pipelines will compare the results to a disease database such as OMIM, and try to focus the search at the gene level. Unfortunately, this heavily biases the results towards what is already known about the disease, and many cases remain unsolved, with no obvious candidates. Systems biology-based variant prioritisation holds a lot of promise to improve this situation. By harnessing genome-scale biological knowledge, we can perform unbiased analyses and potentially identify new mechanisms underlying disease. Emerging forms of systems-level prioritisation are based on a combination of network analysis, gene or protein expression data, and literature curated databases.

In network models of cellular systems the nodes in the network most often represent genes or proteins and the edges between the nodes represent a relationship between the genes. Networks can be built from a variety of biological information, such as protein-protein interactions (PPI) (Chatr-aryamontri *et al.*, 2012; Franceschini *et al.*, 2012), gene expression and other functional relationships, and metabolic pathway information (Kanehisa *et al.*, 2013; Kelder *et al.*, 2011). The majority of network based approaches for gene prioritisation stem from the guilt-by-association principle. This involves ranking candidate genes by their closeness to a set of seed genes, commonly disease causing genes or disease modules, using measures such as shortest path (Krauthammer *et al.*, 2004), random-walk and diffusion kernel (Kohler *et al.*, 2008). Several tool kits are freely available to perform these topological analyses, including the igraph package (Csardi and Nepusz, 2006) in R and NetworkPrioritiser (Kacprowski *et al.*, 2013) in Cytoscape.

Expression-based prioritisation approaches generally rank genes according to differential gene expression or protein abundance between diseased and non-diseased tissues. Apart from the standard differential expression (DE) approach, several alternative expression-based methods exist. Differential variability analysis captures disease-induced changes in regulatory control that might be missed by DE analysis (Ho *et al.*, 2008), and may yield important information about the connectivity or other characteristics of the underlying gene regulatory network (GRN) (Mar *et al.*, 2011; Padovan-Merhar and Raj, 2013). The exact nature of the relationship between gene expression variability and the dynamics of the underlying GRN is still being actively investigated, but we expect up-and-coming single-cell genome wide technologies can further clarify the relationships. Another approach is to subtract away the baseline gene expression in various tissues in the embryos and identify those genes that are differentially expressed in a specific tissue compared to the whole embryo. This approach has been shown to be very useful in prioritising cataract genes using embryonic ocular lens gene expression data (Lachke *et al.*, 2012).

This thesis aims to further develop and apply tools for integrative analyses. Specifically it aims to investigate the utility of propagating functional and gene expression data through a PPI network scaffold in order to improve prediction of disease genes, and to build a tool for rapid large scale extraction and analysis of gene expression data from GEO.

1.5 Cross-species analyses

An incredible diversity of life has come to exist on this planet. Each unique species is just one piece of the biological puzzle that makes up the tree of life, the branches of which contain secrets and adaptations not found anywhere else.

The majority of biomedical research is conducted in *Homo sapiens* (human) and a handful of model organisms, namely, *Mus musculus* (mouse), *Danio rerio* (zebrafish), *Saccharomyces cerevisiae* (yeast), *Rattus norvegicus* (rat), *Drosophila melanogaster* (fly), *Caenorhabditis elegens* (worm), *Escherria coli* (bacteria) and a few others. Different organ-

isms have evolved strikingly different characteristics and solutions to biological challenges, and so studying a range of organisms is of course necessary to understand the variety of biological processes. Working with a particular species brings unique opportunities and challenges and requires specialised expertise, such that each species become an isolated field of biological research. Yet much of biology is conserved and consistent across multiple branches of the tree of life.

There is a huge untapped potential to share information learned from one species to many other species. This is particularly true for transferring knowledge from the well studied model organisms to lesser studied non-model organisms, which also have their own unique biology we can learn from. By investigating and characterising the added variance that cross-species analyses provide, we can better understand which genomic elements drive conserved or unique phenotypes in a species.

1.5.1 Evolution and phenotypic variation

The theory of evolution describes how unique species emerge over time via incremental genetic changes that result in heritable physical characteristics. The classic example of evolution are Darwin's finches, 14 closely related bird species with varying beak morphologies that are suited to the unique food sources on each of the Galapagos islands where the birds live. The phenotypic changes in Darwin's finches are driven by several key genes that regulate beak formation during development (Abzhanov *et al.*, 2006).

While much of evolution is caused by slow genetic drift, there are times when more drastic changes take place. This can be caused by a sudden environmental pressure such as a disruption to a food source, which gives some individuals a fitness advantage. Alternatively drastic genomic changes can occur in a single individual, including whole genome duplication or more localised chromosomal duplications and deletions, which suddenly provide the foundation for evolutionary experimentation to take place. Comparative genomics has allowed us to identify that these events occurred repeatedly throughout evolution, for example vertebrates have undergone two rounds of whole genome duplication (Cañestro

et al., 2013), the ray finned fishes underwent a third round (Meyer and Van de Peer, 2005), as have some African clawed frogs only relatively recently (Uno *et al.*, 2013).

Epigenomic evolution and diversity is less well studied and understood, although it has been acknowledged for many years (King and Wilson, 1975). We know that there is a high conservation of transcription factor binding motifs and histone modifications amongst vertebrates (Schmidt *et al.*, 2010), but what about other epigenomic mechanisms in more evolutionarily distant species? Unanswered questions include, to what extent is epigenomic machinery consistent across the tree of life, how did it evolve and how does it contribute to phenotypic variation? We often don't know how the non-coding genomic variations that drive phenotypic evolution act through regulatory mechanisms. In the case of Darwin's finches it was found that the timing and distribution of certain key transcriptional regulators during development determines beak morphology, as opposed to changes to the protein coding gene sequence (Abzhanov *et al.*, 2006). However the molecular mechanisms that determine those changes in timing remains unclear.

1.5.2 Complex homology

As evolutionary distance increases and more genome altering events occur, our ability to confidently say that genes in different species are functionally equivalent diminishes. This is made particularly challenging when a species retains more than one copy of an ancestral gene after a duplication event, resulting in a complex homology relationship.

Complex homology is a major hurdle to integrating data across multiple species and affects large numbers of genes (almost one quarter of the homology relationships between zebrafish and human are complex (Djordjevic *et al.*, 2016)). In order to lend support to functional equivalence, different measures have been proposed, including protein sequence identity, protein domain equivalence and syntenic block identification, which measures the conservation of larger regions of the chromosome.

Even with such measurements we can often not trivially determine if genes are function-

ally equivalent when multiple homologues exist. A group of genes may all have very similar sequence identity to several homologues in another species, and the epigenomes may determine which are functionally equivalent. Alternatively the homologues may have sub-functionalised and hence share the function of the ancestral gene as well as some new functions. In the case of a gene which has been deleted, dysregulated or mutated, other genes or pathways may have adapted to fill the gap.

The prevailing techniques used today for dealing with complex homology usually involve removing it in some way. This can result in the discard of a large fraction of the homology relationships between two species, especially if they are evolutionarily distant. This becomes particularly problematic if the genes and pathways under investigation are those with complex homology.

This thesis aims to directly address statistical issues in cross-species analyses in order to integrate data and knowledge from across the tree of life. Specifically, this thesis investigates the complex homology issue in cross-species gene set analysis and develops and applies a novel solution that utilises the complete complex homology structure. Furthermore a published method for cross-species joint chromatin state inference is empirically evaluated and applied to determine whether epigenetic mechanisms are conserved between very evolutionarily distant species.

1.6 Structure of this thesis

In this chapter I have introduced the bioinformatics concepts required for analysing the regulation of genetic systems across species. In chapter two I will review in depth the current approaches for applying some of these systems biology concepts to studying one family of disease, heart diseases. This includes the construction of the largest mouse heart specific GRN based on perturbation data, and discussion of many methodologies applied and built on throughout this thesis, particularly in chapters three, five and six which all analyse heart data with implications for heart disease studies. Chapter two is a condensed

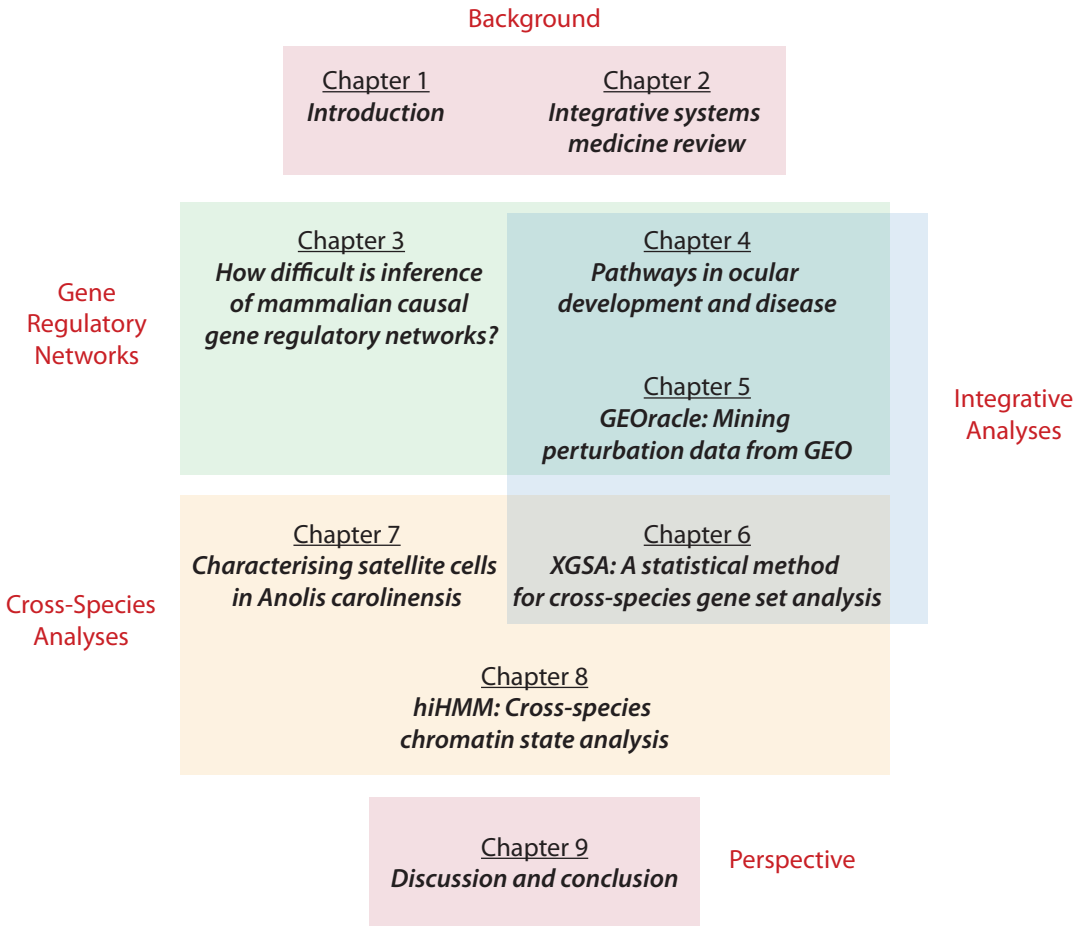

Figure 1.1: **Outline of thesis sturcture showing major themes.**

version of a review article published in Djordjevic *et al.* (2015).

Chapters three to five continue the theme of mammalian causal GRN construction and analysis. Chapter three is an investigation into the feasibility of causal GRN construction in mammals using popular correlation based methods, which has been published in Djordjevic *et al.* (2014). In the context of ocular diseases, chapter four develops a GRN analysis algorithm and applies integrative analyses to prioritising disease genes.

Chapter five introduces a bioinformatics solution to quickly extract large amounts of perturbation data from the GEO database, allowing the rapid construction of large gene regulatory networks, as well as many gene sets.

Chapters six to eight look at issues and applications of direct cross-species analyses. In chapter six I investigate the biases caused by complex homology in cross-species gene set analysis, and propose and validate a statistical solution. I further apply the solution to gain insights from several case studies in widely divergent species. Most of chapter six has been published in Djordjevic *et al.* (2016). In chapter 7 I conduct a cross-species analysis to identify an unknown cell type in *Anolis carolinensis* and investigate the properties that make it unique from the equivalent cell type in mouse.

Chapter 8 describes the empirical evaluation of a new method for cross-species joint chromatin state inference and it's application to investigate the conservation of epigenetic patterns between evolutionarily distant species. Chapter 8 contains the majority of my contributions (around 25%) to a published study Sohn *et al.* (2015).

The final chapter offers my perspective on the outstanding challenges and promise of future research in integrative cross-species bioinformatics analyses of gene regulation.

Chapter 2

Decoding the complex genetic causes of heart diseases using systems biology

2.1 Introduction

The pace of disease gene discovery is still much slower than expected, even with the use of cost-effective DNA sequencing and genotyping technologies. It is increasingly clear that many inherited heart diseases have a more complex polygenic aetiology than previously thought. Understanding the role of gene–gene interactions, epigenetics, and non-coding regulatory regions is becoming increasingly critical in predicting the functional consequences of genetic mutations identified by genome-wide association studies and whole genome or exome sequencing. A systems biology approach is now being widely employed to systematically discover genes that are involved in heart diseases in humans or relevant animal models through bioinformatics. The overarching premise is that the integration of high-quality causal gene regulatory networks (GRNs), genomics, epigenomics, transcriptomics and other genome-wide data will greatly accelerate the discovery of the complex

genetic causes of congenital and complex heart diseases. This review introduces what is known about the genetics of the most prevalent heart diseases, and summarises how state-of-the-art genomic and bioinformatics techniques are being used to characterise the regulatory systems of the heart, accelerating the pace of discovery of heart disease genes.

Many forms of congenital and acquired cardiovascular diseases show familial aggregation, indicating that genetic factors play a role in disease aetiology. Advances in genomic technology have accelerated the transition from single gene assessment in rare Mendelian disorders, such as different types of congenital heart disease (CHD), to studies of more common complex disorders such as cardiomyopathies, ischemic heart disease and atrial fibrillation (AF) (Gelb and Chung, 2014). In the latter group of disorders, many different additive genetic and environmental risk factors, each of relatively small effect size, are hypothesised to be contributing to the disease risk. In recent years, there has been a significant advancement in understanding the genetic contribution to different types of cardiac disease.

CHD is a large collection of structural and functional deficits that arise during cardiac embryogenesis, and exhibit large genetic heterogeneity (Fahed *et al.*, 2013; Gelb and Chung, 2014). CHDs are the most common form of birth defect, affecting up to 68 in 1,000 live-born babies (Blue *et al.*, 2012). Familial CHD can be caused by single-gene mutations, but population prevalence of CHD can also indicate multi-factorial aetiology following the common disease common variant hypothesis. Currently, many known disease causing genes affect developmental signalling pathways involved in cardiogenesis. Examples of such genes include *NKX2-5, NKX2-6, GATA4, GATA5, GATA6, IRX4, TBX20, ZIC3, NOTCH1, NOTCH2* and *JAG1* (Fahed *et al.*, 2013; Hershberger *et al.*, 2013). Mutations in these genes can repress or enhance gene transcription, or affect protein structure and function, which can lead to disruption of developmental signalling pathways. In addition, de novo mutations in histone modifying genes were found to contribute to approximately 10% of CHD cases (Zaidi *et al.*, 2013), indicative of a strong epigenetic component. Following multiple gene discoveries based on highly penetrant gene mutations with Mendelian segregation in CHD, current methods that utilise state-of-the-art genomic techniques aim

to give further insight into the genetic complexity of CHD. On the other hand, a different set of genes are involved in other inherited cardiovascular pathologies such as hypertrophic or dilated cardiomyopathy (HCM and DCM respectively), long-QT syndrome (LQTS), or Marfan syndrome. These mainly arise from mutations in sarcomeric proteins or ion channel components with specific functions in cardiac biology such as TTN, MYH11, MYH6, MYH7, SCN5A and KCNH2 among others (Fahed *et al.*, 2013; Hershberger *et al.*, 2013).

Traditional gene discovery methods such as linkage and candidate gene studies were successful in identifying causal genes in cardiac diseases such as HCM (Kimura *et al.*, 1997), idiopathic DCM (Krajinovic *et al.*, 1995; Messina *et al.*, 1997), AF (Chen *et al.*, 2003) and LQTS (Keating *et al.*, 1991), among others. For more complex cardiac diseases, a hypothesis-free genome-wide approach was needed to assess their polygenic cause. The advent of Genome-Wide Association Studies (GWAS) facilitated the high-throughput screening of single nucleotide polymorphisms (SNPs) for the identification of marker alleles or genotypes that are more frequent in diseased individuals compared to healthy control individuals.

In the past 8 years, GWAS have identified multiple genetic variants associated with many complex diseases including cardiac diseases (Cappola *et al.*, 2010; Gudbjartsson *et al.*, 2009). Although our understanding of the genetic architecture of many complex diseases has been improved, the findings so far confer small increments in risk, explaining a small proportion of familial clustering (Manolio *et al.*, 2009). This has led to scepticism regarding the potential clinical applicability of these findings, while the question of what may contribute to the missing heritability remains open. The current hypothesis is that this may include contributions of rare inherited or de novo variants and epistatic effects, among others (Eichler *et al.*, 2010; Manolio *et al.*, 2009).

To explore the impact of GWAS on identifying genetic variants that are associated with heart diseases or traits, we searched the GWAS catalogue provided by the National Human Genome Research Institute of the USA (`http://www.genome.gov/gwastudies/`). As of 6 August 2014, this catalogue contains results from 1,960 publications, and 14,001 reported SNP associations. Using the keywords heart OR cardiac OR cardio OR fibrillation OR

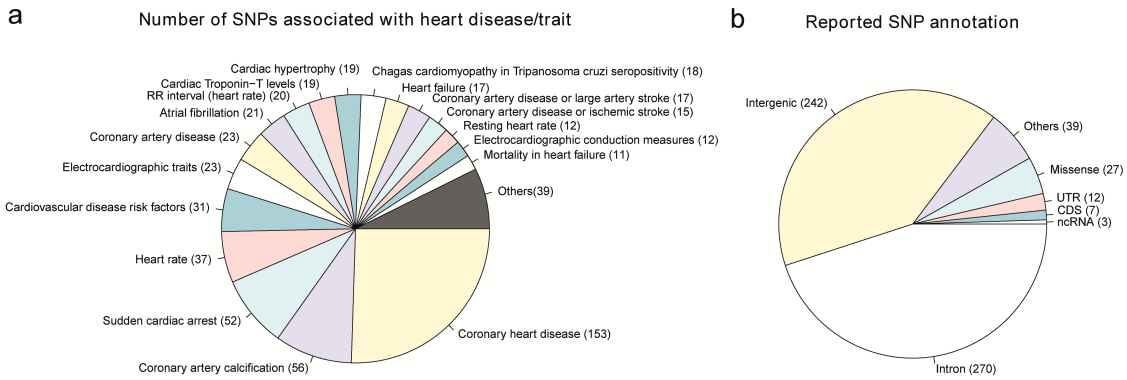

Figure 2.1: **Summary of the 600 SNPs that have been reported to be associated with a heart disease or trait.** These SNPs are associated with about 300 genes. Data were downloaded from the NHGRI GWAS catalogue

coronary OR arrhythmia, we found 75 publications satisfying this criterion, resulting in 600 reported SNP associations, which map to about 300 genes. These publications were published from 2007 to 2013. As shown in Fig. 2.1, many SNPs have been reported to be associated with coronary diseases and heart rate/arrhythmia, but relatively little is known about cardiomyopathy and other heart muscle conditions (Fig. 2.1a). More importantly, the majority of the SNPs are located in intronic or intergenic regions (Fig. 2.1b). Currently, many of these SNPs are assigned to their closest genes, yet the mechanistic relationships between the genetic variant and the aetiology are in most cases unknown. Clearly, a lot more work needs to be performed to further dissect the genetic functions of these SNPs.

The advent of next-generation sequencing (NGS) has facilitated the identification of causal genetic variants given its potential to discover the entire spectrum of sequence variation within a sample. It has revolutionised genomic and genetic analyses and is now becoming the primary discovery tool in human genetics (Cirulli and Goldstein, 2010). Sequencing individuals within families, apart from identifying genetic causes of Mendelian disorders, can improve power to detect rare variants that are associated with common diseases, since predisposing variants will be present at a much higher frequency in affected relatives of a proband (Roach *et al.*, 2010). In addition, family-based studies can allow the detection of de novo mutations, parent-of-origin effects as well as genegene interactions when affected

relatives share two nearby epistatic loci in linkage disequilibrium.

The discovery of extensive copy number variation in the genomes of normal and diseased individuals provided new hypotheses to account for the phenotypic variability among inherited disorders and new leads for the detection of the molecular basis of common complex disorders (Beckmann *et al.*, 2007). Moreover, since most of the single nucleotide variants associated with diseases confer relatively small increments in risk, copy number variation was thought to account for some of the remaining missing heritability (Manolio *et al.*, 2009). To address this, the Wellcome Trust Case Control Consortium undertook a large genome-wide study of association between common copy number variants (CNVs) and eight common human diseases (Wellcome Trust Case Control Consortium *et al.*, 2010). CNVs were found to be well tagged by SNPs, while the CNV association analysis was shown to be susceptible to a range of artefacts, which can lead to false positive associations. As a conclusion, common CNVs, which are typed using existing technologies, were found to be unlikely to contribute greatly to the genetic basis of common human diseases. On the other hand, rare CNVs, detected by either array or sequencing technologies, have often been found to contribute to the risk of rare or common cardiac diseases such as CHD (Fakhro *et al.*, 2011; Soemedi *et al.*, 2012), specifically Tetralogy of Fallot (TOF) (Greenway *et al.*, 2009) and early-onset myocardial infarction (Myocardial Infarction Genetics Consortium *et al.*, 2009).

Remarkable progress has been made in understanding the genetic basis of many cardiac diseases, including rare CHD as well as more common complex cardiac diseases. However, advances in technology have now made it possible to interrogate the genetic causes of these diseases at a genome-wide level using the concepts of systems biology.

2.2 Systems biology of heart development and disease

The vertebral heart is a complex adaptive biological system, with attributes and organisational features spanning multiple orders of magnitude of space, time and energy re-

quirements. It undoubtedly has many characteristics of complex adaptive systems, including adaptive capacity, inter-component communication, homeostasis, spatiotemporal self-organisation and specialisation, emergent properties and cascading failures.

At the cellular level, cardiomyocytes are studded with heart specific voltage-gated and mechano-sensitive ion channels resulting in a unique electrophysiological profile. Cardiomyocytes are packed full of long chains of myofibrils which provide the fundamental contractile forces needed for the heart to function. These tubular cardiomyocytes are connected to each other via intercalated disks - porous regions containing electrical and mechanical connections that enable the individual cells to communicate. Bundles of aligned cardiomyocytes constitute the extensive muscular mass of the heart, with altered 3D organisation and gene expression signatures depending on location and function (Barth, 2005).

The electrical signal that sparks the heartbeat originates at the sinoatrial node, a group of specialised cardiomyocytes. The impulse propagates through the right and left atria causing them to contract and leading to activation of the atrioventricular node. The impulse then spreads via specialised cardiomyocytes called the Purkinje fibres, which act as electrical superhighways, coordinating muscular contraction of the ventricles, the main blood pumping chambers. It is this highly organised biomechanical and electrically coupled propagation system that generates the coordinated contraction of the billions of cardiomyocytes that constitute the four human heart chambers robustly over a lifetime.

The vertebral heart is known to be adaptive in response to external stimuli during all stages of life. During heart development, the formation of a complex multi-chambered 3D organ from a collection of multi-potent cells is tightly regulated, but is also easily influenced by external factors, as is evident from the vast literature on cardiac developmental perturbations from multiple species. In the reverse role, the maternal human heart has been observed to undergo eccentric hypertrophy, changes in sphericity and decreased left ventricular strain during pregnancy (Savu *et al.*, 2012). Extensive mechanical and electrical remodelling of the heart is also observed in many types of adult onset heart disease, and indeed reverse remodelling with treatment of disease has been observed (De Jong

et al., 2011; Glukhov *et al.*, 2012; Merlo *et al.*, 2011).

In terms of system failures, the healthy heart is in general fairly robust to physiological levels of strain in the short term. However, many pathological environmental factors and diseases will eventually lead to a cascading set of failures, ultimately resulting in a collapse of the system. This often results in a class of clinical phenotypes known as heart failure (HF), characterised by very poor functional output of the heart, shortness of breath and an inability to adapt to exercise. As such, HF is not necessarily the result of an underlying molecular problem, but an inability to maintain homeostasis of the hearts emergent properties. While risk factors including age, previous myocardial infarcts, smoking, exercise and diet play a strong role in determining the occurrence and severity of HF, clear genetic links between familial occurrence of cardiomyopathies, AF and HF are emerging (Hershberger *et al.*, 2013; MacRae, 2010).

Perhaps the strongest example of this is sudden cardiac death, which is believed to be caused when a transient event, like intense exercise, meets an underlying disease substrate that may not have previously displayed symptoms. Mendelian inheritance of some cases of disease aetiologies including HCM, DCM, AF, LQTS and CHD is well known (Hershberger *et al.*, 2013; Kamisago *et al.*, 2000; MacRae, 2010), and several studies have suggested similar inheritance patterns for sudden cardiac death (Deo and Albert, 2012; Myerburg, 2001). In a post-mortem screening of 173 cases of sudden unexplained death, 26% showed novel putatively pathogenic mutations in a targeted screen of 5 LQTS disease genes and RYR2 (Tester *et al.*, 2012). Considering the high hit rate from such a limited screening of only 6 genes, this finding suggests that mutations in cardiac electrical genes are potentially a strong precursor to catastrophic systems failure in the heart.

The final goal of systems biology is often a complete and comprehensive computational model of the complex biology being studied. Such a complete model theoretically allows the functional consequences of any perturbation to the system to be simulated, tracked and characterised. Multi-scale modelling of the heart is an extremely exciting field that has developed immensely over half a century, from simple mathematical models of blood flow volume through the heart, to computationally solving millions of equations simulta-

neously to model the mechanical forces during a ventricular contraction. These nested models include the contribution of ion channels to the action potential within individual cardiomyocytes, the subsequent mechanical forces at the sarcomeric, cellular, tissue and chamber levels, and the propagation of electrical stimuli and contraction throughout the heart (Smaill and Hunter, 2010). While computational cardiac models can be personalised from clinical imaging technologies (Wang *et al.*, 2013b), and have been successfully used to recapitulate the contribution of faulty proteins to heart disease phenotypes (Sadrieh *et al.*, 2014), they are still far from an exhaustive genome-wide model that could be applied as a high-throughput variant prioritisation framework.

With these limitations in mind, our review will focus on systems biology approaches for discovering the genetic causes of CHD and complex heart diseases, from the overwhelming pool of candidate genetic variants identified through whole-genome sequencing and whole-exome sequencing of patients. Although complex heart phenotypes with later onset and milder progression have distinguishably different genetic aetiology compared to rare congenital heart phenotypes, the difficulties of discovering disease-causing variants can be very similar. For example, in family-based NGS studies, large lists of coding variants can be found to be segregating with disease. In such cases, variant prioritisation is necessary to help us identify the disease-causing genes. Conversely, many families will show incomplete penetrance or missing heritability, resulting in a lack of obvious pathogenic mutations after standard filtering. We focus on using systems biology resources and approaches that exist today and are increasingly based on genome-wide omics data.We discuss how these methods can be applied for prioritising variants and elucidating the mechanism by which mutations in both coding and non-coding regions affect the whole system and ultimately lead to disease.

2.3 Network approaches to decode the genetic causes of heart diseases

It is possible that many inherited heart diseases are a result of allelic variants or damaging mutations, the effects of which are propagated through multiple interconnected molecular levels, rather than only affecting a single gene. While genes associated with disease can be identified, the underlying molecular mechanism in cardiac development and pathogenesis remains largely elusive (He *et al.*, 2011a; Lin *et al.*, 2010). Such situations, which entail a large number of molecules spanning across several biological pathways that potentially contribute to pathogenicity, demand a global, system-level view of the functional genetic architecture (Lage *et al.*, 2010) and how its dysregulation leads to disease. Molecular networks are a valuable means of assembling this information, in the form of nodes representing molecules, connected by edges representing interactions, which can be directed or undirected. Networks unify what may appear as disparate biological pathways from single experiments, driving the formulation of novel hypotheses and models to explain the mechanisms of pathogenesis. They can also be applied in the discovery of novel disease network biomarkers (Chen and VanBuren, 2014) and inform new therapeutic treatments, conceptualised by the idea of network medicine (Barabasi *et al.*, 2011).

Many biological networks have been constructed and analysed in the context of heart diseases (Table 2.1). Protein protein interaction (PPI) networks utilise proteomics data to model physical interactions between proteins as undirected edges. Gene co-expression networks are based on the premise that genes that have similar expression profiles across a number of samples are likely to have similar biological functions. These networks are generated from transcriptomic data such as microarrays, where it is expected that co-expressed genes will cluster together. A unifying feature of bioinformatics analysis of these networks is the identification of sub-networks of molecules, commonly referred to as modules. Investigations of various biological networks have demonstrated that hub molecules (i.e., those that have many incoming or outgoing networks connections) are often the most critical for normal cellular function, are closely related to disease, and

perturbing them often results in embryonic lethality or pathogenic phenotypes (Dickerson *et al.*, 2011). Further studies have shown that common disease causing genes often cluster together, representing important protein complexes or biological pathways, where the failure of any single component can lead to a similar disease phenotype (Goh *et al.*, 2007). These findings validate the general approach of disease module identification (Barabasi *et al.*, 2011).

Systems biology approaches have previously been used in the study of cardiac development and diseases, although not as prevalently as one might expect (Sperling, 2011). Berger *et al.* (2010) used PPI data to investigate the functional neighbourhood of 13 known LQTS genes. By building and analysing the PPI network around known disease genes, they were able to rank molecules, diseases and drugs in relation to LQTS. They found that the LQTS neighbourhood represented the convergence of cardiac channelopathies in the diseasome. They also found that the gene targets of drugs designed as QT prolonging and drugs that have an undesired QT-related side effect were enriched in the LQTS neighbourhood. Furthermore, they were able to use the ranked LQTS neighbourhood proteins to classify drugs with adverse QT-related side effects with reasonable accuracy, based on the FDA Adverse Event Reporting System database.

Lage *et al.* (2010) used phenotype annotations from gene knock-out mouse models to annotate 255 genes into 19 spatio-temporal cardiac developmental phenotype modules. Based on PPI data, they identified 49 statistically significant novel heart development candidate genes linked to one or more modules. Using immunohistochemistry, they were able to validate that 11 of 12 tested novel candidates were in fact expressed in the specific heart tissue and at the developmental time predicted by their assigned phenotype module.

Modules form the basic functional unit of most networks; however, the methods used to derive these differ substantially (Table 2.1). These methods generally restrict a gene to be part of one module only, which may not be a true reflection of its activity if it performs multiple functions in multiple pathways. A novel clustering algorithm, recently developed to construct a transcriptional regulatory network of mouse heart development, permits genes to participate in multiple modules (Chen and VanBuren, 2014). Other limitations of

Table 2.1: A summary of published heart disease studies that involve construction and analyses of protein-protein interaction networks and gene co-expression networks

Disease	**Data used to construct network**	**Network Construction Method**	**Network Analyses**	**Reference**
19 heart phenotypes / diseases	PPI, cardiac developmental genes	Phenotype specific networks, literature based module annotation	Complexity measure, temporal phenotype ordering, experimental validation of novel candidate genes	(Lage *et al.*, 2010)
CHD	PPI, gene sets for CHD genetic and environmental risk factors and responses	Networks for separate anatomical structures as described in Lage et al. (2010)	Network permutation tests	(Lage *et al.*, 2012)
DCM	PPI, DCM gene expression profiles, GO annotations	Pearsons correlation between co-expressions and GO annotations to identify DCM and non-DCM modules	Identification of hub genes, topological measures, functional enrichment of modules, performance as disease classifier	(Lin *et al.*, 2010)
AF	Human microarray profiles grouped by disease state	Weighted gene co-expression network analysis and connectivity of to derive modules	Module preservation, PCA of modules to identify eigengenes, Identification of hub genes, functional enrichment using IPA	(Tan *et al.*, 2013)
Cardiac Hypertrophy, Heart Failure	Microarrays of mouse myocardium	Weighted gene co-expression network analysis and topological overlap to derive modules	Identification of hub genes, topology, PCA of modules to identify eigengenes, GO annotation, module preservation	(Dewey *et al.*, 2011)
MI	Microarrays from mice with MI	Coexpression and topological modules	GO annotation, hub gene (Col5a2) as disease classifier	(Azuaje *et al.*, 2013)
LQTS	PPI, disease genes	Immediate neighbours	Ranking genes, diseases and drugs based on LQTS neighbourhood	(Berger *et al.*, 2010)

CHD = Congenital Heart Disease, ASD = Atrial Septal Defect, DORV = Double Outlet Right Ventricle, AV = Atrioventricular, DCM = Dilated Cardiomyopathy, AF = Atrial Fibrillation, MVD = Mitral Valve Disease, CAD = Coronary Artery Disease, MI = Myocardial Infarction, LQTS = Long QT Syndrome, PPI = Protein-Protein Interaction Network, DE = Differentially expressed, GO = Gene Ontology, IPA = Ingenuity Pathway Analysis, IHC = Immunohistochemistry, qRT-PCR = Quantitative Real-Time PCR

network-based approaches include heterogeneity of available data sets and indeed changes in the underlying biology across different tissues and organisms, and the generalisability of animal models to humans (Dewey *et al.*, 2011; Tan *et al.*, 2013). Furthermore, predictions of novel disease genes are made through computational analysis of networks, and thus need to be experimentally validated in animal models.

Whilst networks based on PPI and gene co-expression data are useful for identifying genes associated with diseases (Fig. 2.2c), to understand the underlying mechanism we need to model causality using directed networks such as gene regulatory networks (GRNs). Causal relationships can be robustly inferred through perturbation of molecules, allowing nodes to assume roles of regulator and target depending on the direction of the interaction. This type of data can come from large-scale omic studies, but they could also come from the vast collection of previously published molecular biology data probing the effect of perturbing individual genes and proteins. In fact, one effective way to gain an understanding of a causal GRN is by piecing together many previously experimentally identified molecular interactions. In other words, we want to discover the truth based on many previous discoveries, as epitomised by Sir Isaac Newton's famous quote - "If I have seen further, it is by standing on the shoulders of giants".

2.4 CardiacCode: a manually-curated resource for cardiac-specific GRN analysis

There is a vast amount of high quality genetic or molecular perturbation data in the published literature that largely remains computationally inaccessiblemostly hidden in figures, tables or text in developmental biology papers. We manually collected over 700 pieces of genetic perturbation evidence from 43 published primary research papers on *in vivo* mouse cardiac development, from embryonic day E6.5 – E13.5 heart tissues. We integrated this data into a causal GRN using a statistical model (see Chapter 3). The network consisted of 280 unique edges between 33 regulators and 129 target molecules. The majority (59%) of these edges represent activating regulatory relation-

Gene Prioritisation.png

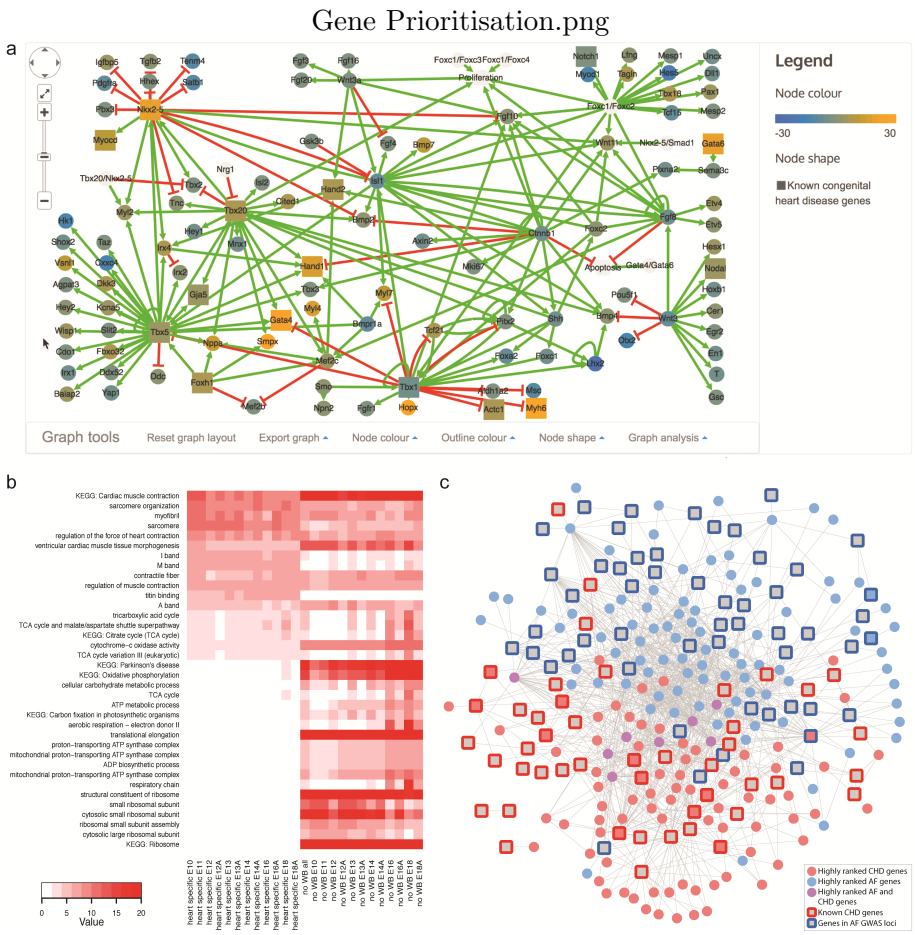

Figure 2.2: **Common approaches for integrating biological information in the study of disease. a** A screenshot of a cardiac-specific GRN from the CardiacCode website. Nodes are coloured based on differential expression of mouse E10.5 heart against whole embryo body (WB). Node shape is related to whether a gene is known to cause congenital heart disease. **b** Gene set enrichment of the 200 most highly specific (or expressed) genes at E10.5 and E11.5 for whole heart and at E12.5E18.5 for ventricles and atrial chambers,with (or without) normalisation against WB. The gene set enrichment results demonstrate that the normalisation of cardiac-specific gene expression patterns from WB can increase the enrichment of cardiac-specific gene sets, and decrease the enrichment of housekeeping gene sets. This suggests that WB normalisation may be a simple method for enriching for potential cardiac disease genes. **c** A PPI network showing the differential clustering of AF GWAS genes and known CHD genes in the human protein interactome. Candidate disease genes were prioritised based on proximity to all known disease genes for each of the two diseases. The sub-network containing the connections between the combined top 100 ranked candidates from each disease is shown

ships, 13% are inhibitory and 28% appear to have no-causal effect. We complemented this with 86 microarray expression profiles from the GEO database. The curated per-

turbation data set, the assembled microarray data, and the inferred cardiac development network can be accessed through our newly developed interactive web resource, CardiacCode (`http://CardiacCode.victorchang.edu.au/`; Fig. 2.2a). The resource is built on an SQL database and interfaces with Javascript and HTML5 through PHP. The network visualisation and analysis tools are supported by the cytoscape.js plugin and Python scripts running on the server. One limitation of the current CardiacCode network representation is that it does not include information from many of the 'no-effect' relationships. It would be useful to differentiate these relationships from those for which we simply have no data, perhaps by using negative distances or a different edge type. This is an important area for future research in GRN representation and analysis.

The CardiacCode network represents the core knowledge about the causal gene regulatory interactions that drive mammalian cardiac development. Key cardiac markers and transcription factors (TFs), including *Nkx2-5, Gata4/6, Tbx1/5/20, Isl1* and *Mef2c*, are the hub genes in the network, potentially reflecting their important roles in the network or the widespread studies of these genes. Interactions between many of these key TFs have been shown to be critical for cardiac development and fetal viability. *Nkx2-5* is a key regulator of heart morphogenesis (Lyons *et al.*, 1995) and differentiation of cardiomyocytes (Tanaka *et al.*, 1999). Gata TFs are required for the formation of the primitive heart tube (Molkentin *et al.*, 1997) and myocardium maturation (Peterkin *et al.*, 2003). Tbx TFs are crucial for morphogenesis of the outflow tract and aorto-pulmonary septum (Xu *et al.*, 2004) and chamber differentiation (Stennard *et al.*, 2005). *Isl1* is required for outflow tract and right ventricle formation (Cai *et al.*, 2003), and *Mef2c* plays a role in heart tube looping (Lin *et al.*, 1997). It is therefore not surprising that the breakdown of any link in this core regulatory network has been shown to cause CHD (McCulley and Black, 2012; Schlesinger *et al.*, 2011).

Many critical signalling pathways are also represented as molecules from the WNT, BMP, SHH, NOTCH, FGF, TGF-beta and other growth factor families that have been linked to CHD, as well as their receptors (Li *et al.*, 2014a). CardiacCode also encodes regulatory knowledge about known CHD genes, including the key TFs and markers mentioned before,

as well as *Myocd, Foxh1, Myh6, Actc1, Nodal* and *Notch1*.

The CardiacCode website has a growing suite of inbuilt analysis options. Apart from node position, three node attributes can be customised using uploaded data sets: node colour, node border colour and node shape. By default, the known CHD genes are coloured red. These colour and shape features allow the user to simultaneously visualise continuous values such as gene expression levels and rankings, or discrete values such as module membership, gene ontology (GO) annotation, disease association and candidate disease gene status. Therefore, CardiacCode can be very useful in facilitating disease gene prioritisation in NGS studies where a large number of candidate genes are found.

2.5 Integrative analysis for heart disease gene prioritisation

A new resource for murine heart development gene expression data from E7.5 all the way to adult tissues has recently been published (Li *et al.*, 2014a), which in combination with other published data will provide a foundation for expression-based prioritisation for human heart disease genes in the future.

More recently, several groups have developed gene prioritisation tools that combine multiple data types and analysis approaches in order to better rank candidate genes. Barriot *et al.* (2010) created an online collaborative wiki-based resource for studying CHD that implements a mixed data gene prioritisation analysis. They implement the ENDEAVOUR algorithm (Tranchevent *et al.*, 2008) and allow the user to specify data sources including GO annotations, PPIs, cisregulatory information, gene expression data sets, sequence information and text-mining data. ENDEAVOUR was applied to prioritise genes located in an 850-kb locus associated with CHD, using gene expression microarrays and gene homology data (Thienpont *et al.*, 2010). Seven training sets of genes representing different cardiac developmental and genetic phenotypes were used as input to the algorithm, which identified *TAB2* as the top-ranking gene from all 105 gene candidates. The role of this gene in heart development was experimentally verified using knock-downs in zebrafish which

led to heart defects, and mutation analysis in 402 patients with CHD revealed a translocation disrupting *TAB2* that co-segregated with familial CHD (Thienpont *et al.*, 2010). ENDEAVOUR was also applied to study two patients with CHD, identifying *CRKL* and *MAPK1* as the most likely causal genes out of those lost in a large deleted region of the genome (Breckpot *et al.*, 2012).

Another tool, Gentrepid, applies two algorithms to prioritise candidate disease genes (Ballouz *et al.*, 2013). Common Pathway Scanning (CPS) utilises PPI and pathway data to determine relationships between candidate and disease genes based on membership in the same protein complex or pathway. Common Module Profiling (CMP) performs sequence comparisons of Pfam domains to assess functional similarity of candidates with disease genes. Gentrepid was applied to GWAS data on hypertension (Ballouz *et al.*, 2013) and coronary artery disease (Ballouz *et al.*, 2014), identifying disease genes that were consistent with those previously reported as well as several novel candidates.

Li *et al.* (2013) built three cardiomyopathy subtype-specific networks [DCM, Arrhythmogenic Right Ventricular Cardiomyopathy (ARVC), HCM] based on the PPI neighbours of known disease proteins from the OMIM database. They used the STRING PPI or functional link database which provides a confidence of interaction score, and combined this with gene ontology similarity between proteins to weight each edge of the network. Based on the weights of a proteins immediate neighbourhood, along with the number of disease genes it neighboured, each protein in the network received a disease relevance score and hence a rank. From the top 50 DCM candidate proteins, 9 have been previously linked to DCM in the literature; 7 additional candidates were directly associated with other cardiomyopathies such as HCM, and 4 additional proteins were related to cardiac arrhythmias, cardiac dysfunction and cardiac cell damage. Similar results were achieved for the HCM and ARVC candidates.

A web resource (`http://www.esat.kuleuven.be/gpp`) developed by Tranchevent *et al.* (2011) contains detailed information about a large selection of gene prioritisation tools to help researchers decide which tool is most appropriate for their analysis. While existing tools are very useful in disease gene discovery, they are predominately gene-centric. Inter-

actions between genes, microRNAs (miRNAs), non-coding RNA (ncRNAs), disruption of transcription factor binding sites and epigenomic modifications have all been found to play a role in heart development and disease, and emerging knowledge about these processes needs to be incorporated into future gene prioritisation efforts (Klattenhoff *et al.*, 2013; Matkovich *et al.*, 2011; Schlesinger *et al.*, 2011; Smemo *et al.*, 2012; Zaidi *et al.*, 2013).

2.6 Epigenomics of heart diseases

Heart development involves intricate programs of gene expression that differ across different cell types and stages. This necessitates looking deeper into chromatin structure and regulatory elements such as enhancers to elucidate the regulatory mechanisms in cardiac development and disease. This urgency is exemplified by the discovery that heart diseases can arise from mutations in non-coding, regulatory regions (Fig. 2.3 Smemo *et al.* (2012)).

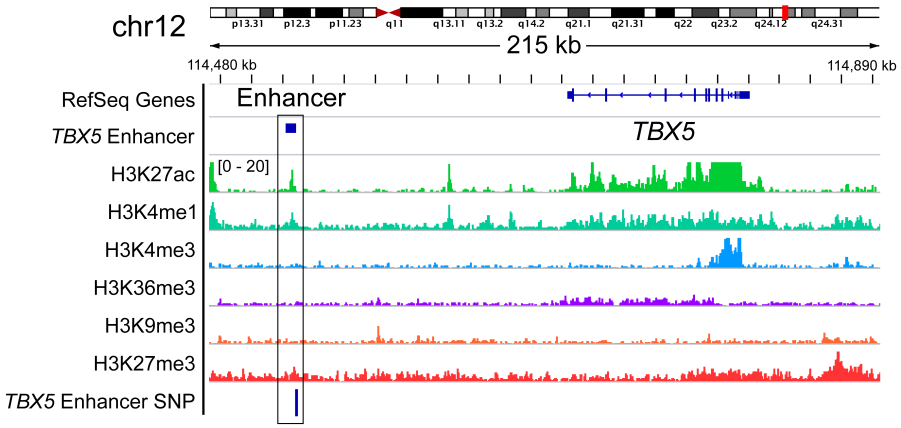

Figure 2.3: **Analysis of human genome-wide chromatin landscape of cardiac cells can be used to identify cardiac-specific enhancers that are associated with congenital heart disease.** ChIP-seq of histone modifications in human left ventricular heart tissue (GEO Accession: GSE16256) shows the presence of enhancer-associated marks H3K27ac and H3K4me1 surrounding a SNP that is associated with a congenital heart disease. This SNP may exert its function through affecting this enhancer, which is approximately 90 kb downstream of the *TBX5* gene (Smemo et al. 2012). This example illustrates how disease-associated enhancers can be identified using ChIP-seq

2. Decoding the complex genetic causes of heart diseases using systems biology

Zaidi *et al.* (2013) performed exome sequencing of children with severe CHD and their parents, identifying de novo missense, frameshift and truncation mutations in genes that deposit, read and remove H3K4 methylation. These mutations were located in the *MLL2* methyltransferase, *KDM5A* and *KDM5B* demethylases, the H2BK120 ubiquitination complex and the *CHD7* helicase. Mutations were also found in *SMAD2*, which encodes a protein that binds chromatin and leads to downstream demethylation of H3K27, facilitating increased activation of normally repressed genes (Zaidi *et al.*, 2013).

Several studies have implicated the Ezh2 subunit of the methyltransferase PRC2, which trimethylates H3K27, in heart development and disease. In mouse E12.5 ventricles, ChIP-qPCR revealed enrichment of Ezh2 and H3K27me3 at the promoters of the transcription factors *Six1, Pax6* and *Isl1* (He *et al.*, 2012). This was concordant with de-repression of these genes when Ezh2 was conditionally knocked-out in mouse embryonic hearts. The inactivation of Ezh2 also resulted in lethal congenital heart defects including myocardial hypoplasia and ventricular septal defect (He *et al.*, 2012). In a similar study, *Ezh2* knock-out in adult mice hearts revealed a decrease in H3K27me3 at the promoter of the transcription factor *Six1*, which was accompanied by an increase in PolII binding (Delgado-Olguin *et al.*, 2012). This de-repression of *Six1* was associated with activation of skeletal muscle genes and hypertrophy of cardiomyocytes. Thus, epigenetic repression by *Ezh2* is established during embryonic development and persists in late development (He *et al.*, 2012) and postnatally to maintain the normal phenotype in adult hearts (Delgado-Olguin *et al.*, 2012).

In mouse postnatal hearts with pathologically induced cardiac hypertrophy, over-expression of the JMJD2A demethylase resulted in increased expression of fetal cardiac genes such as *Myh7* and the disruption of cardiac function (Zhang *et al.*, 2011). The re-occurrence of these results in human patients with cardiac hypertrophy, who also displayed increased levels of JMJD2A, is suggestive of its role in aggravating cardiac hypertrophy in pathological conditions (Zhang *et al.*, 2011).

Histone acetyltransferases (HATs) and deacetylases (HDACs) have been shown to play critical roles in cardiac development and disease. Through the addition of acetyl groups

to histone molecules, HATs mark regions associated with active transcriptional activity. The HAT activity of CBP/ p300 leads to a hypertrophic phenotype when over-expressed in cultured ventricular myocytes taken from neonatal rats (Gusterson *et al.*, 2003). In transgenic adult mice subjected to myocardial infarction, over-expression of a mutant form of p300 lacking the HAT domain prevented extensive left ventricular remodelling, a process leading to heart failure, when compared to over-expression of the normal p300 (Miyamoto *et al.*, 2006). HDACs remove acetyl groups, and are therefore associated with repression of genes. Mice with deletion of Hdac3 specifically in the heart displayed extensive and lethal cardiac hypertrophy (Montgomery *et al.*, 2008). Single deletions of *Hdac5* and *Hdac9* led to an exacerbated hypertrophic response in mouse models, while double deletion of these genes increased susceptibility to lethal ventricular septal defects (Chang *et al.*, 2004).

Chromatin remodelling involves reorganisation of nucleosomes and affects DNA accessibility, particularly the ability of transcriptional activators and repressors, and other proteins such as RNA polymerase II, to bind to DNA during transcription (Han *et al.*, 2011). Deletion of *Brg1*, the major ATPase subunit of the BAF chromatin remodelling complex, in mouse myocardium, reduced the incidence of cardiac hypertrophy, thus making *Brg1* an attractive therapeutic target (Hang *et al.*, 2010). Allelic imbalance between Brg1 and cardiac transcription factors (*Nkx2-5, Tbx20, Tbx5*) also results in heart defects (Takeuchi *et al.*, 2011).

Working in synergy with histone modifications and chromatin remodelling, DNA methylation also acts as a key epigenetic regulator of gene expression during development and disease. During cardiogenesis, differential DNA methylation is believed to regulate the expression of specific cardiac developmental genes (Chamberlain *et al.*, 2014). In particular, the expression of *Has2*, required for proper heart valve formation, is predicted to be controlled by methylation of its enhancer by DNA methyltransferase 3b (Chamberlain *et al.*, 2014). In patients with TOF, DNA methylation profiling revealed greater methylation of the promoters of genes associated with CHD, which corresponded to decreased mRNA expression levels (Sheng *et al.*, 2014). Another key player believed to contribute to the pathogenesis of TOF is *retinoid X receptor a* (*RXRA*), a member of the

retinoic acid signalling pathway (Zhang *et al.*, 2014). Patients with TOF exhibited significantly less *RXRA* mRNA in their right ventricular outflow tract myocardium, which was likely due to an observed increase in methylation of the *RXRA* promoter (Zhang *et al.*, 2014). Similarly, in paediatric patients with CHD, attenuated transcriptional activity was hypothesised to be caused by hypermethylation of the promoter of the *CITED2* gene, which encodes the transcriptional co-activators CBP/p300 (Xu *et al.*, 2014). Regions of methylated DNA are read by proteins such as MeCP2, whose over-expression in mice was associated with hypertrophy of the septum in embryonic hearts, resulting in lethality around E14.5 (Alvarez-Saavedra *et al.*, 2010).

Differential methylation of heart development genes *EGFR* and *GATA4* inmothers can result in CHD in their offspring (Chowdhury *et al.*, 2011). Furthermore, chronic treatment of pregnant rats with hypoxia directly resulted in methylation and subsequent repression of the *PKC-epsilon* gene promoter (Patterson *et al.*, 2010). This was associated with greater susceptibility to cardiac ischemia and injury in adult offspring. Evidence for trans-generational epigenetic inheritance has also been demonstrated in mice, where mutations in the Mtrr (methionine synthase reductase) enzyme in maternal grandparents led to detrimental uterine environmental conditions in their wild-type daughters (Padmanabhan *et al.*, 2013). This in turn resulted in growth deficiencies and congenital defects of various organs, including the heart, in grand-progeny, that was independent of the maternal genotype. Together, this alludes to the idea that heart diseases can occur as a result of both genetic and environmental risk factors that have been inherited from previous generations. Thus, a complete understanding of all potential pathogenic mechanisms is warranted to develop therapeutic strategies and reduce the incidence of heart disease in current and future generations.

We have compiled a summary of published epigenomic data sets related to heart development and disease which is available at `http://cardiaccode.victorchang.edu.au/download.php`.

2.7 Future direction

The advancement of microarray and NGS technologies underlies the widespread adoption of GWAS and whole- genome sequencing for disease gene discovery. These technologies have fundamentally changed the way we do research and have enabled us to address systems-level biological questions at a genome-wide scale. More recently, two innovative technologies are promising to transform the landscape of biomedical research: single cell genomic analysis, and CRISPR-Cas9-based genomic editing.

Single-cell genomic technology was named the Method of the Year in 2013 by the journal Nature Methods. This innovation means that we can now perform genome-wide genomic and transcriptomic sequencing at the level of individual cells, instead of averaging the signals from a heterogeneous cell population. This has facilitated the fine-scale identification of gene expression status during cellular differentiation (Trapnell *et al.*, 2014). Highly parallel single-cell qPCR has found that there is greater variability in gene expression between individual B lymphocyte cells than between different human donors, and that this variation has been largely masked by lower experimental resolution in previous studies (Wills *et al.*, 2013).

Existing microarray and NGS profiling techniques are applicable to millions of cells. Therefore, it is important to recognise the potential effect of averaging molecular profiles from heterogeneous cells, which is potentially a problem in analysing primary heart tissues. It has been suggested that a single-cell approach would be beneficial for the study of heart development and pathology (Sperling, 2011). The human heart consists of a mixture of cell types; the most prominent including fibroblasts, myocytes, endothelial cells, and epicardial cells. These cells interact with one another during development, and each play a key role in cardiac development, repair, and pathogenesis (de la Pompa and Epstein, 2012; Deb, 2014; Runyan and Markwald, 1983). Although the application of this technology to the heart is in its infancy, in an organ where defects in a single cell can trigger a life threatening arrhythmia, the ability to investigate cardiac cells at this resolution might prove to be the driver of the next major breakthrough.

Clustered regularly inter-spaced short palindromic repeats (CRISPR)-associated (Cas) (CRISPR-Cas) is an exciting new technology that exploits a bacterial defence mechanism against viruses and plasmids to perform genome editing with unprecedented precision and efficiency (Ran *et al.*, 2013). The CRISPR-Cas mechanism relies on small RNAs to act as sequence-specific templates and recognition factors to cleave and silence undesired nucleotides. Recently, the CRISPR mechanism has been demonstrated as a viable alternative to siRNA to knock-down expression of almost any gene in many model organisms important in heart disease research, including zebrafish, mouse, goat and human (Wu *et al.*, 2013). Jinek *et al.* (2012) discovered that a single chimeric short-guide RNA (sgRNA) molecule can program the Cas9 protein to cleave double-stranded DNA with single base-pair accuracy. Because it directly modifies the DNA of the cell, the effect of the gene knock-out persists through cell division, unlike siRNA interventions. Since then, CRISPR-Cas9 has also been used to knock-in an inducible protein knock-down, by inserting a Shield1 conditional destabilisation domain into a target protein (Park *et al.*, 2014).

The ability to edit DNA with single base pair accuracy means that CRISPR-Cas9 can be used to systematically investigate functional effects of mutations anywhere in the genome, including non-coding regulatory regions, in a much more efficient workflow than cloning and breeding approaches (Dickel *et al.*, 2014). This should greatly improve the pace of discovery of functionally active cell-type-specific regulatory elements in mammalian systems. By creating libraries of the sgRNAs that direct the Cas9 protein, genome-scale knock-down screenings in human cell lines have systematically perturbed thousands of genes in a robust and lasting way with minimal off-target effects (Shalem *et al.*, 2014; Wang *et al.*, 2014a). The recently demonstrated ability to introduce multiple mutations *in vivo* in a single step will further facilitate research into complex polygenic diseases including heart diseases (Wang *et al.*, 2013a). This possibility opens up new avenues for large-scale genetic perturbation experiments that will greatly enable inference of tissue specific causal GRNs.

Perhaps the most exciting application of CRISPR-Cas9 is its ability to edit germline DNA

and remove disease-causing mutations from future generations (Lokody, 2014; Wu *et al.*, 2013). Wu *et al.* (2013) injected the CRISPR-Cas9-sgRNA complex into the zygote of a cataract mouse model, designing the sgRNA such that the site of the cataract-causing 1-bp deletion in the Crygc gene was precisely cut. 6 of 22 of the resulting pups did not display a cataract phenotype and were healthy. This tantalising result demonstrates the potential to use genome editing as a remediation of congenital diseases, such as CHD.

The highly specific and programmable binding of CRISPR-Cas9 has been further transformed into a general purpose platform for modifying the epigenome of living cells in a highly controlled way (Lopes *et al.*, 2016). This is achieved through two main approaches: editing the genome at non-coding regulatory regions such as enhancers (Canver *et al.*, 2015; Li *et al.*, 2014b); directing a nuclease-deactivated Cas9 protein (dCas9) fused with transcription factors or epigenetic modifiers, to bind specific regulatory regions, thereby interfering with or activating gene transcription (Dominguez *et al.*, 2015; Thakore *et al.*, 2015). The fusion between dCas9 and KRAB is often used for targeted transcriptional repression, as KRAB recruits a heterochromatin-forming complex that causes histone methylation and deacetylation, whereas the VP64 transactivation domain cam be fused to dCas9 for targeted transcriptional activation.

This approach can be extended to target several genes simultaneously by using multiple guide RNAs, facilitating the study of regulatory networks and genetic interactions, as well as genome wide activation screens (Konermann *et al.*, 2014). By incorporating RNA aptamers onto the guide RNAs, multiple transcriptional co-factors can be recruited to the dCas9 complex, allowing increasingly complex experiments in genetic regulation to be designed (Thakore *et al.*, 2015).

The playing field of cardiac disease gene discovery has improved substantially in the last few years thanks to the rapid advancement of experimental genomic technologies. Generation of genomic data is no longer a bottleneck for decoding the genetic cause of heart disease. We can sequence a patients entire genome with roughly $1,000 in 1-2 weeks. The real bottleneck is our ability to unravel how changes in one gene can propagate through the entire system, and how genetic variation in coding and non-coding regions can contribute

to a cardiac phenotype. When this wealth of genomic data is integrated with other omic and clinical information through state-of-the-art analytical and modelling techniques, we can begin to truly decode the complex genetic causes of heart disease. In this article, we have reviewed many recent developments in cardiac systems biology in the context of decoding the genetic causes of heart diseases. We are sure to see new advancements in the coming few years as new technologies and resources are applied to heart disease research.

Chapter 3

How difficult is inference of mammalian causal gene regulatory networks?

3.1 Inferring gene regulatory networks

Large-scale community challenges, such as the Dialogue for Reverse Engineering Assessments and Methods (DREAM), have been conducted to evaluate gene regulatory network (GRN) inference methods using *in silico* simulated data or a number of known GRNs in bacteria or yeast. Many approaches perform better than random when comparing to 'gold standard' perturbation experiments, although distinguishing true from false positives in even the most confident predictions from the best performing algorithms is infeasible given the total search space, usually several orders of magnitude larger(Maathuis *et al.*, 2010). Ongoing evaluations of the DREAM challenge have shown that although network inference is partially achievable in prokaryotic organisms, inference in eukaryotic organisms still remains a major challenge (Marbach *et al.*, 2010, 2012).

Several methods have been designed to infer causal networks based on perturbation data

and have been applied to study mammalian development. Wagner showed that theoretically, if there is no noise and missing value in the data, it is possible to infer a causal GRN of n genes in $O(n^2)$ steps (Wagner, 2001). Nonetheless, real data contains noise and missing values. More sophisticated methods must be used. Nested effects models (Fröhlich *et al.*, 2009b; Markowetz *et al.*, 2005) and methods based on deterministic effects propagation networks (Fröhlich *et al.*, 2009a; Pinna *et al.*, 2010) are effective at reconstructing the causal network between genes for which systematic (genome-wide) perturbation experiments exist. These algorithms are *not* the main focus of this study as these data sets are not yet widely available in mammalian contexts.

Recently other types of data have also been used to infer tissue-specific GRNs. Connecting DNA binding transcription factors (TF) to their consensus binding sequence motifs throughout the genome can provide a putative regulatory scaffold, but many predicted binding sites are not consistently bound across dynamic cellular contexts. Integrating experimental data on chromatin accessibility or regulatory activity can drastically reduce false positive rates of predicted binding sites (Marbach *et al.*, 2016; Pique-Regi *et al.*, 2011). However such prediction based networks remain incomplete largely due to gaps in our understanding of dynamic TF binding, including variable binding affinity, TFs having multiple non-canonical binding sequences (Wong *et al.*, 2011), incomplete databases and the complexity introduced by transcriptional co-factors. Even if this information was known, there are multiple classes of molecular interactions that can causally modify gene regulation independent of TF binding, including RNA binding, modification or degradation, and protein interaction, modification, complex formation and transport. These types of effects would be missed in TF based networks.

With these limitations in mind, the appeal of reverse-engineering networks from measurements of the causal end-points of gene regulation becomes evident. Transcriptome wide measurements provide a high confidence representation of the output of the cells regulatory network, and can be generated relatively cheaply and simply. This is likely what drives the ongoing desire to use gene-expression based methods in experimentally challenging mammalian contexts, where these approaches are least likely to work (Mar-

bach *et al.*, 2012). Although correlation between two genes does not imply causation, the converse is commonly implicitly assumed by many GRN inference algorithms — that causal gene regulation leads to *observable* gene co-expression. This assumption implies that if one can properly remove non-causal edges from a network constructed on measures of gene co-expression, the remaining edges are likely causal (Barzel and Barabási, 2013; Wang *et al.*, 2014b). In other words, many people attempted to infer a GRN from gene co-expression data alone without explicitly making use of gene perturbation experimental data (Glass *et al.*, 2013). Even though the developers of these methods were likely aware of the underlying assumptions and limitation on interpreting a GRN inferred from gene co-expression data, it is quite possible that the end-users might treat each edge in the inferred GRN as having a causal regulatory role.

In this study we mainly focus on assessing the underlying assumption behind the algorithms that make use of gene co-expression data alone. Two such popular expression-based GRN inference algorithms that we will assess are GENIE3 (Huynh-Thu *et al.*, 2010) and ARACNE (Margolin *et al.*, 2006a), the latter being specifically targeted at mammalian systems with over 100 citations in the year 2016 alone. A fundamental question arises, 'Can we reverse engineer mammalian developmental causal GRNs from a collection of gene expression profiles?'. To fully address this question, we will need high quality causal GRNs for comparison, but there is currently no gold standard for mammalian GRNs. Nonetheless, we have observed that there is a vast amount of experimentally validated genetic or molecular perturbation data in the published literature, but these data remain largely computationally inaccessible — mostly buried in figures, tables or text in developmental biology papers. Indeed it is exactly this type of data that is most often used as a gold standard for validation of predicted regulatory relationships and construction of high quality causal GRNs (Buckingham *et al.*, 2005; Maathuis *et al.*, 2010; Marbach *et al.*, 2012; Olsen *et al.*, 2014).

There have been significant efforts in the field of automated mining of biomedical texts and images for knowledge discovery, some of which could be applied to the problem of mining perturbation data. In biomedical texts work has been done linking gene perturbations to

disease phenotypes (Rodriguez-Esteban *et al.*, 2009). The same ideas could be applied to linking regulator genes to target genes, but this approach has not been as widely pursued. The TRRUST database has collected 8,015 perturbation data from text-mining of journal abstracts (Han *et al.*, 2015). Unfortunately TRRUST is not fully automated, requiring manual verification of the predicted relationships, of which 66% were false positives. Furthermore this method has not been applied to full texts where the majority of the results reside, and these data do not come with cellular context information such as the tissue of origin or developmental time.

Several promising studies have focused on analysing biomedical figures which could theoretically be used for the automatic extraction of perturbation data. Of particularly relevance are the analyses of gels (Kuhn *et al.*, 2014) and bar charts (Al-Zaidy and Giles, 2015) the most common formats used to report perturbation data. Unfortunately these approaches are incomplete and not at a functional stage where automated tools exist without significant human intervention.

Considering the limitations of current approaches for automated mining of perturbation data from the literature, targeted reading of the literature in the tissue of interest is still the only way to get reliable tissue-specific 'gold standard' perturbation data.

3.2 Methods

3.2.1 Data summary

In this study, we assembled two manually-curated mouse GRN data sets (embryonic development of tooth and heart), summarising experimental evidence for causal regulation (or lack of causal regulation) between 1,177 regulator-target gene pairs, and a compendium of matching microarray expression profiles, to systematically investigate the difficulties of GRN inference in mammalian cells, especially in the context of organ development. The tooth GRN and microarray data set was downloaded from ToothCODE, and the data were

generated to study epithelial-mesenchymal interactions during early tooth organogenesis (O'Connell *et al.*, 2012). It contains over 1,500 pieces of genetic perturbation evidence from 120 primary research papers, and 105 matching microarray profiles. Using a similar curation approach, we specifically assembled the heart dataset for this study. We manually collected over 700 pieces of genetic perturbation evidence from 43 published primary research papers on *in vivo* mouse cardiac development. We complemented this with 86 microarray expression profiles from the GEO database. The curated perturbation data set representing information for 137 regulators and 371 targets, the assembled microarray data, and the inferred cardiac development network (see below for more details on inference of mode of regulation) can be accessed through our newly developed interactive web resource, CardiacCode (Figure 3.1). It was built on an SQL database and interfacing with javascript and HTML5 through PHP. The network visualisation was supported by the cytoscape.js plugin (Figure 3.1). The tooth and heart GRN and microarray gene expression data sets are available via ToothCode (`http://compbio.med.harvard.edu/ToothCODE/`) and CardiacCode (`http://CardiacCode.victorchang.edu.au/`).

3.2.2 Manual curation of genetic perturbation evidence from the literature

We recorded genetic perturbation experimental evidence from primary research papers. Each piece of evidence consists of 11 crucial pieces of information: regulator gene; target gene; perturbation performed on the regulator (+ or -); effect on the expression of the target gene (up-regulated, no change, down-regulated); species; developmental stage; tissue in which the perturbation was performed; tissue in which the expression of the target gene was measured; measurement technique; type of molecule measured (mRNA or protein); citation. If we were not confident about any of these pieces of information, the evidence was discarded. We further recorded the experimental context and additional information where it was available, including the genotype and phenotype of the perturbed mouse embryo.

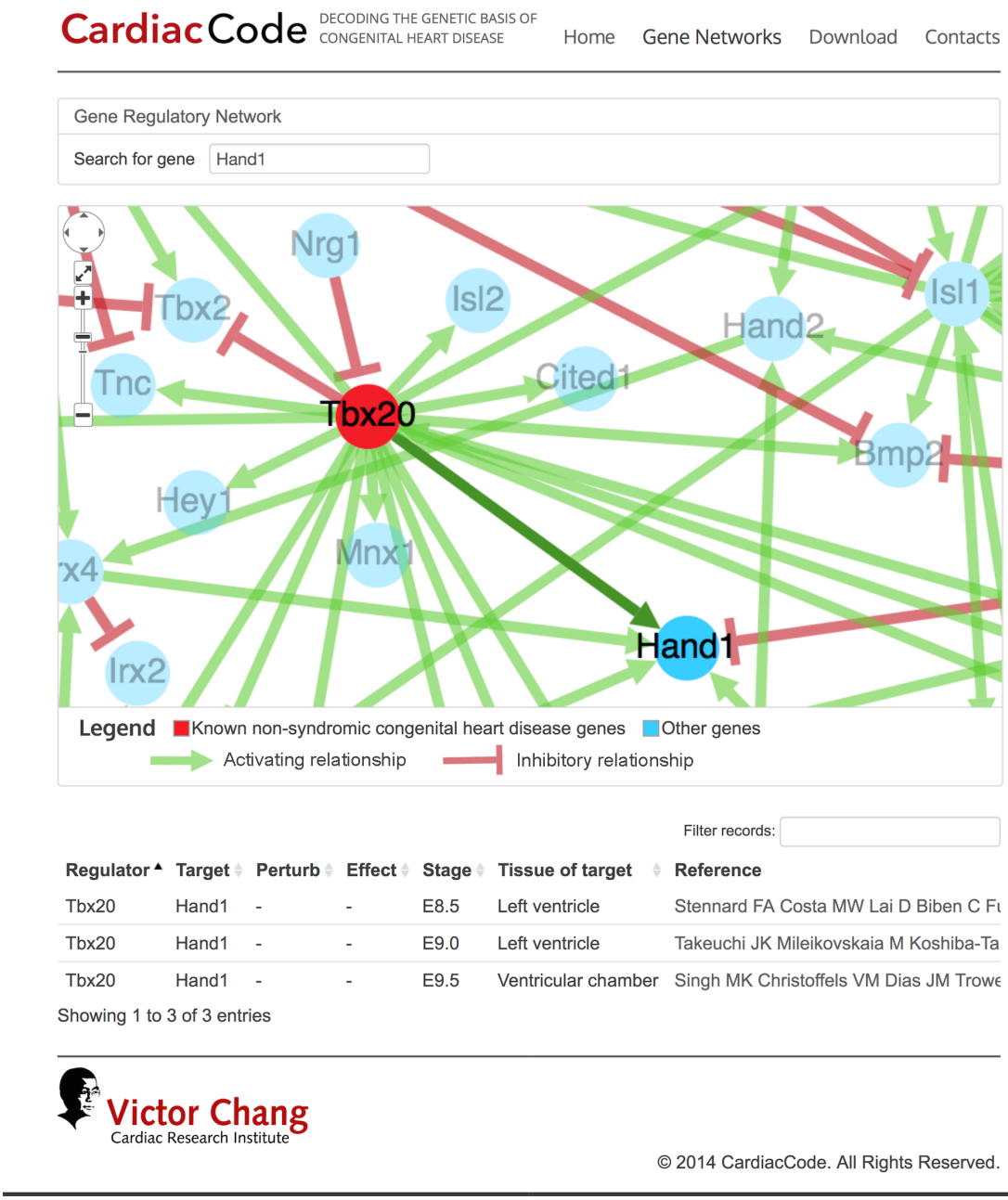

Figure 3.1: **CardiacCode is a public online resource allowing interactive visualisation of the heart GRN, and download of the heart data collected and used in this study.**

3.2.3 Inferring mode of regulation of a regulator-target pair

In this study, we only consider experimental evidence that comes from an *in vivo* embryonic mouse model (*i.e.*, not in cultured cells, or not in adult tissues), and was measured by *in situ* hybridisation, qRT-PCR, or similar well-established expression measurement techniques.

The regulator and target genes in each piece of experimental evidence form a regulator-target-pair (RTP). We define three possible *modes of regulations* for each RTP: activating, no interaction, and inhibiting. An edge is placed between two nodes in a GRN if its mode of regulation of the corresponding RTP is activating or inhibiting. We do not distinguish between direct and indirect interactions in this study, instead focusing solely on observable functional regulatory relationship between a regulator and a target gene. Since each RTP may be supported by multiple pieces of evidence, and they may not always be in total agreement, it is important to infer the mode of regulation of each RTP using a principled means.

First, we removed all RTPs that have opposite regulatory evidence in any tissues or time points — *i.e.*, observing both 'activating' and 'inhibiting'. Afterwards, we used a probabilistic model to integrate the occasionally noisy data $D = \{d_1, d_2, ..., d_k\}$ and estimate the mode of regulation $M = \{act, no, inh\}$ for each RTP. This method was first proposed by (O'Connell *et al.*, 2012). We specified a likelihood model $L(M; D)$ for each RTP,

$$L(M; D) = \prod_{i=1}^{k} P(d_i|M)$$

We then specified the likelihood model of observing each piece of evidence given the mode of regulation, $P(d_i|M = m_j) = p_{ij}$, as a conditional probability matrix that describes the likelihood of observed experimental evidence for $i = \{\text{positive}, \text{no}, \text{negative}\}$ (rows of the matrix) given the true mode of interaction $j = \{\text{activating}, \text{no}, \text{inhibiting}\}$ (columns of the matrix),

$$p_{ij} = \begin{bmatrix} \alpha & \frac{1}{2}(1-\alpha) & (1-\alpha)(1-\beta) \\ (1-\alpha)\beta & \alpha & (1-\alpha)\beta \\ (1-\alpha)(1-\beta) & \frac{1}{2}(1-\alpha) & \alpha \end{bmatrix}$$

where k is the number of pieces of evidence corresponding to a RTP, α represents the probability of a correct experimental observation and β represents the probability of a missing observation due to insensitivity of the detection technology given that the correct experimental observation is not obtained. Here we used $\alpha = 0.9$ and $\beta = 0.9$, but our results are not sensitive to reasonable changes in these parameters (O'Connell *et al.*, 2012).

The inferred mode of regulation is the mode M that maximises the likelihood function $L(D; M)$.

3.2.4 Microarray preprocessing

The tooth (Illumina MouseWG-6 v2.0) microarray gene expression data were downloaded from GEO (GSE32321) (O'Connell *et al.*, 2012). The heart (Affymetrix Mouse Genome 430 2.0) microarray data were assembled from multiple studies from GEO (Table 3.2). The assembled heart microarray profiles were quality checked, RMA normalized and $\log_2$ transformed (Figure 3.3, 3.4). In both data sets, low signal probes were removed (mean probe expression < 7.14 (Illumina) and < 5.6 (Affymetrix) respectively). For differential gene expression analysis, we use the limma package (Smyth, 2005) to determine statistically significantly up- or down-regulated genes (Benjamini-Hochberg adjusted $p-$value < 0.01). For inference of GRNs from microarray data, we use the 5000 most variable probes and all of the probes that matched regulator or target genes in our corresponding literature data sets were retained for further analysis.

3.2.5 Network inference based on gene expression

Correlation

Correlation coefficients (Pearson and Spearman) were calculated on the subset of probes that matched the RTPs in the corresponding dataset. A representative correlation cut-off of 0.5 was used to define co-expression of the two genes represented by the two probes.

Mutual information

We use the *minet* R package (Meyer *et al.*, 2008) to calculate mutual information between the probes for all the RTPs.

Table S1. Summary of cardiac microarray data set.

Time series

	GSE1479							GSE11040	
	E10.5	E11.5	E12.5	E13.5	E14.5	E16.5	E18.5	E12.5	E17.5
Whole heart	3	3							
Both ventricle			3	3	3	3	3		
Atrial chamber			3	3	3	3	3		
Endocardial cushions								2	
Atrioventricular valves									2

Genetic perturbation

	GSE28186 IP3R1/IP3R3		GSE41179 Ilk		GSE9124 Sp3		GSE50426 Fog2		GSE6770 Hdac2		GSE45583 Lsd1	
	WT	Mut	WT	Mut	WT	Mut	WT	Mut	WT	Mut	WT	Mut
Whole heart, E9.25	2	2										
Neural crest, E10.5			2	2								
Whole heart, E12.5					2 (/3)	2 (/3)						
Whole heart, E16.5							3	3				
Ventricle, E17.5									2	2		
Whole heart, E18.5											5	3

Phenotypic difference

	GSE32078	
	WT	Diabetic
Whole heart, E13.5	3	3
Whole heart, E15.5	3	3

Figure 3.2: **Summary of cardiac microarray data set.**

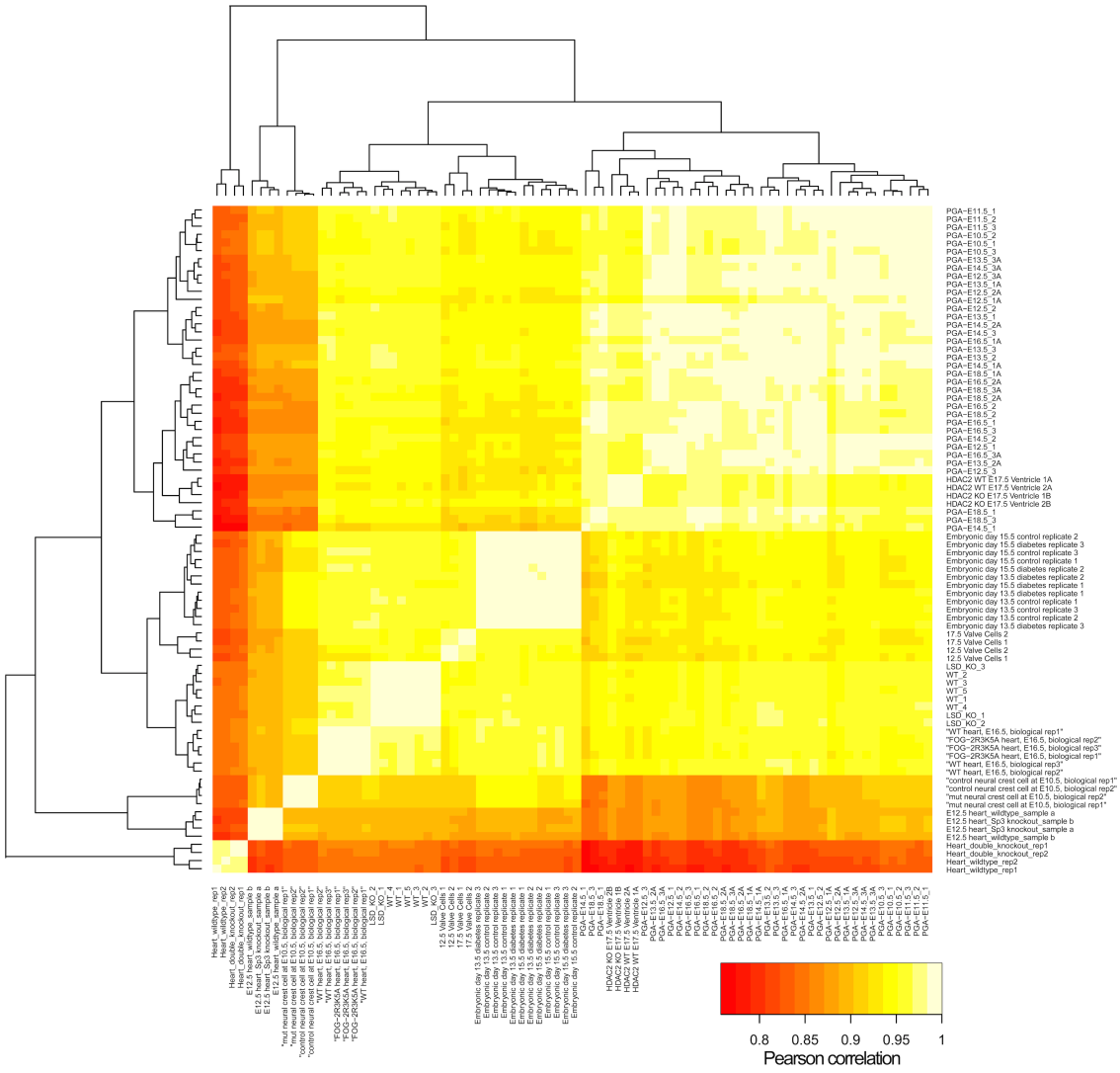

Figure 3.3: **Correlation matrix of cardiac microarray data downloaded from GEO.**

ARACNE

We use the ARACNE algorithm (Margolin *et al.*, 2006a,b) as implemented in *minet*. Default settings were used, including 'eps=0' for ARACNE to avoid prematurely throwing edges away.

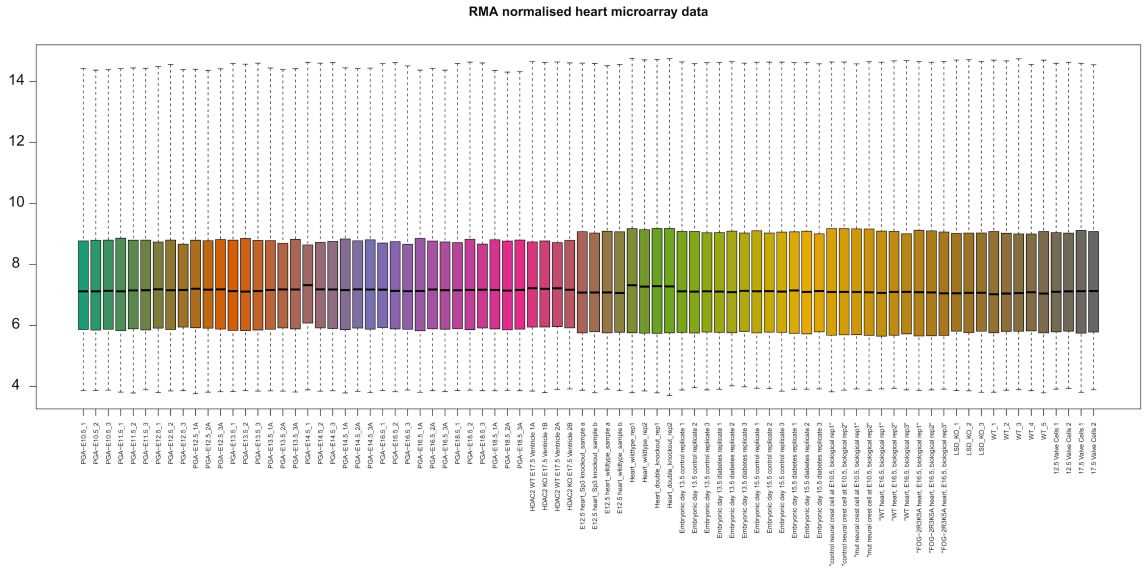

Figure 3.4: **Boxplots showing RMA normalised cardiac microarray data downloaded from GEO.**

GENIE3

The GENIE3 (Huynh-Thu *et al.*, 2010) algorithm was run using the R code provided by the authors. The random forest training step was parallelised using the *foreach* and *doParallel* libraries to improve efficiency on multi-core processors. In an attempt to standardise the network sizes between methods, the number of edges retrieved by the most unrestricted ARACNE adjacency matrix in each analysis was used to determine how many edges to retrieve from the GENIE3 weight matrix.

3.2.6 Network inference based on other molecular networks

Protein-protein interactions

Protein-protein binding data was collected using the 'iRefR' R package (Mora and Donaldson, 2011). Both human and mouse interaction data were used. Human data were converted to mouse gene symbols and combined with mouse data, resulting in a network

of 448147 edges.

Pathway Commons data

Pathway information was downloaded from Pathway Commons (`http://www.pathwaycommons.org/`). Mouse and human specific edges were downloaded in .SIF format, giving 35088 and 392309 unique RTPs respectively.

3.2.7 Calculating sensitivity and specificity of edge inference in GRNs

All of the networks, including those generated from the curated literature data and those inferred from other data sources, were encoded into graph structures using the 'igraph' R package (Csardi and Nepusz, 2006). Overlap of edges between two networks was calculated using the functions in the 'igraph' package. Area under the receiver operator characteristic curve was calculated using the ROCR R package (Sing *et al.*, 2005).

3.3 Results

3.3.1 Causal gene regulation does not necessarily result in observable gene co-expression

The assumption that a causal gene regulatory interaction should lead to an observable correlation of gene expression between the regulator and target is an attractive hypothesis that underlies many GRN reverse engineering approaches. If this assumption is true then we would expect certain trends, including an activating or inhibiting relationship having a positive or negative co-expression, respectively; and gene pairs that have been shown to have no regulatory relationship should have correlation coefficient close to zero. For each literature RTP, we calculated the Pearson and Spearman correlation, as well as the

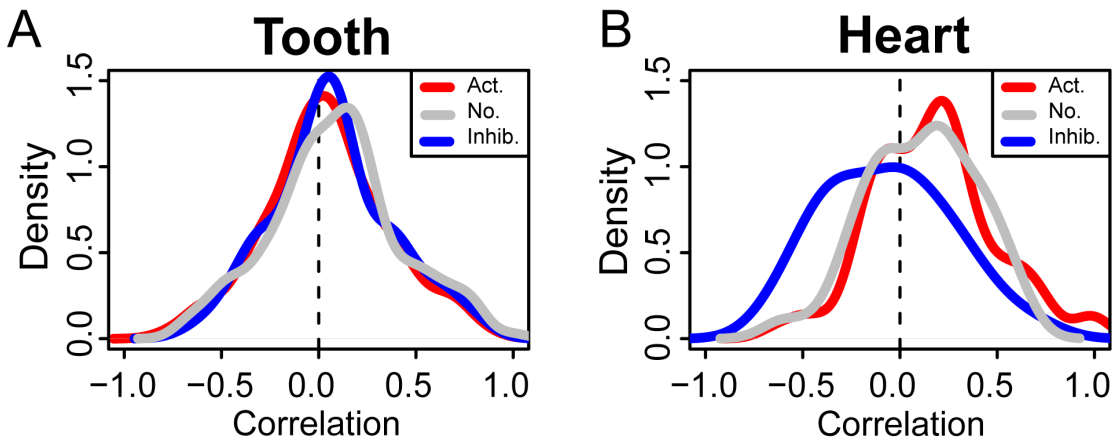

Figure 3.5: **Spearman correlation of different classes of RTP** in the tooth data (A) and the heart data (B). RTP classes are activating (Act.), no effect (No.) and inhibitory (Inhib.).

mutual information between the two genes across all of the matched microarray profiles (Figure 3.5).

In the tooth data, all RTPs, regardless of activating, inhibiting and no effect, have no or very weak Spearman correlation coefficients (Figure 3.5A). Neither do we observe a difference in Pearson correlation (data not shown) or mutual information values (Figure 3.7). This agrees with previous findings based on the *S. cerevisiae* GRN in the DREAM challenge (Marbach *et al.*, 2012). In the heart data, there is a weak shift of the activating and inhibiting RTP towards higher and lower correlation values respectively (Figure 3.5B, 3.6A,). Nonetheless, the gene co-expression patterns of the no-effect RTPs seem to be similar to that of the activating RTPs, suggesting in practice it would have been hard to distinguish true from false positive edges in a GRN if it was constructed based on gene co-expression.

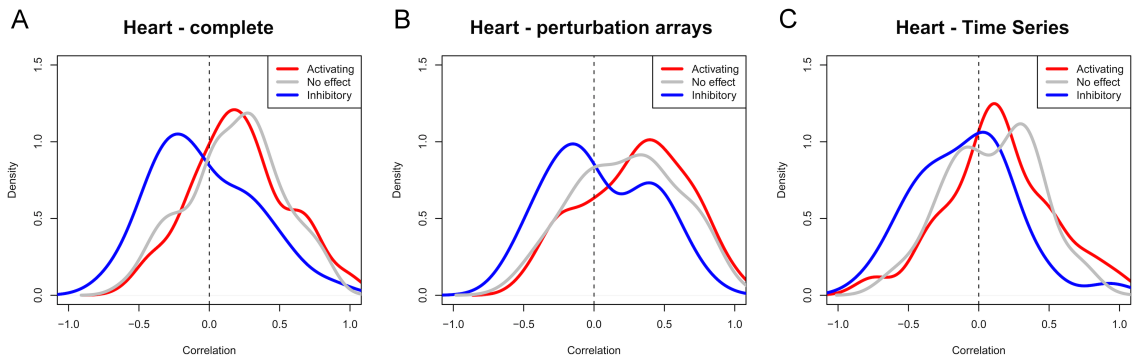

Figure 3.6: **Pearson correlation kernel density plots for each class of RTP in heart**, based on the complete microarray set (A), only the perturbation arrays (B) and only the time series (C). The tooth data showed no shift and was omitted.

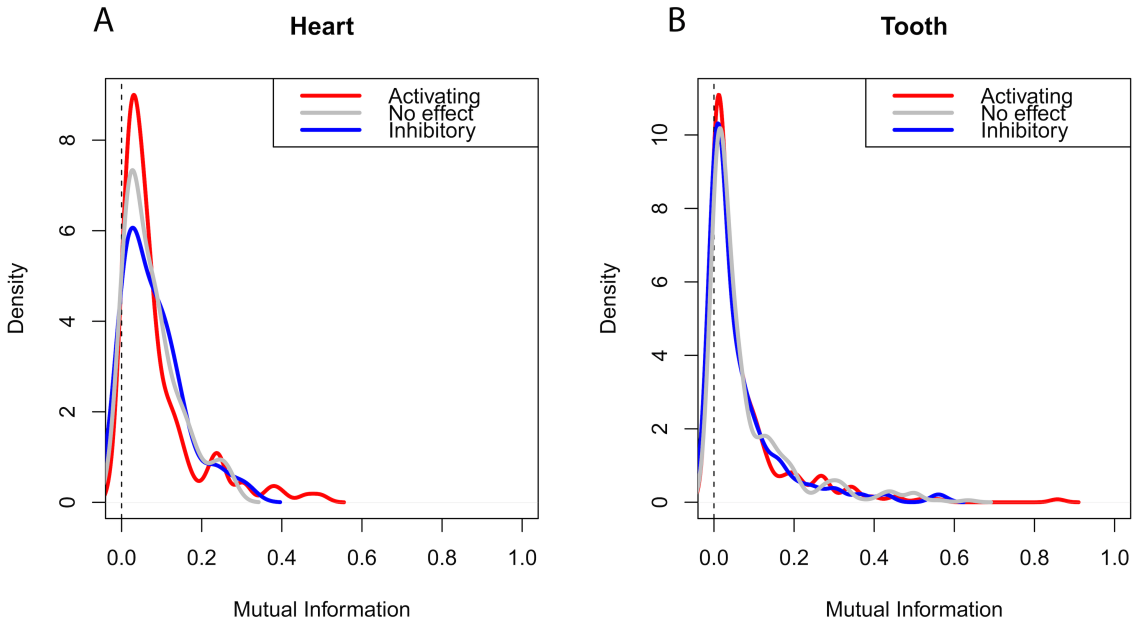

Figure 3.7: **Mutual information kernel density plots for each class of RTP in heart (A) and tooth (B).**

3.3.2 Common expression-based inference methods cannot reliably recover mammalian causal GRNs

In order to investigate the usefulness of current GRN reverse engineering approaches applied to mammalian developmental gene expression data, we ran the GENIE3 and

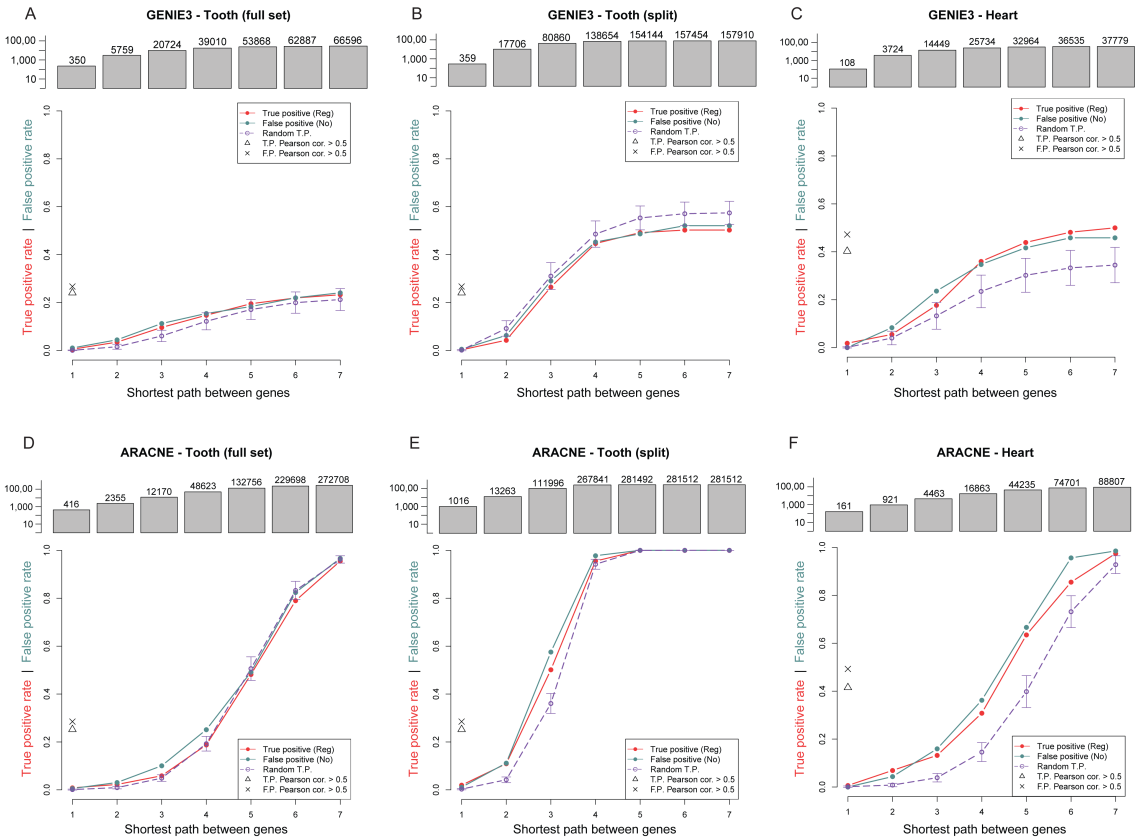

Figure 3.8: **Evaluation of sensitivity (true positive rate) and specificity (1-false positive rate) of edge discovery** by GENIE3 (A-C) and ARANCE (D-F) using the tooth and heart microarray data sets. To account for the possibility that our literature-curated RTP may represent indirect regulatory interactions, we allow matching of a RTP with a linear path of multiple edges (x-axis). The bar chart above each plot shows the size of the network. Dotted lines shows control background of 1,000 node-label-permuted randomised networks

ARACNE algorithms on the tooth and heart microarray data sets and compared the resulting networks to our literature curated GRNs. These two algorithms were chosen because GENIE3 was shown to perform well in the recent DREAM challenge (Marbach *et al.*, 2012), and ARACNE was developed for inferring human GRNs. In general, we found that the algorithms did not offer a tangible improvement over random background in terms of detection sensitivity or specificity (Figure 3.8), *i.e.*, the area under the receiver operator characteristic curve (AUROC) is close to 0.5.

First we observed that at the first-neighbour level, neither algorithm returned more than

one or two true positives on any data set. Because ARACNE and GENIE3 both work by pruning supposedly indirect edges and we do not assume that our RTPs are all direct regulatory relationships, we also considered matching each literature-based RTP with the terminal genes of each 2- to 7-edge path. Although the true positive rate increased as expected, it was accompanied by an almost equivalent increase in the false positive rate. Furthermore, we found that the algorithms trained on the tooth data set did not perform better than the randomly permuted networks of the same structure, and only performed slightly better than random in the heart data set. This slight improvement might be due to the slightly stronger discriminatory gene co-expression signals between activating and inhibitory RTPs (Figure 3.5). Nonetheless, reliable GRN inference in both data sets is virtual impossible in practice based on these two algorithms.

We found that using an absolute Pearson correlation threshold of 0.5 identified 2871 unique RTPs from the heart data and 3528 unique RTPs from the tooth data once self loops have been removed. From these RTPs we could reproduce 24% of our activating and inhibitory edges from the tooth literature (true positives), however 26% of our no-effect edges (false positives) were also identified. In heart we observed a 42% true positive rate, coupled with a 49% false positive rate. The overall result is the same even if we use a different Pearson correlation cut-off, and the overall AUROC is close to 0.5 (Figure 3.14). The size of the inferred networks that must be analysed in order to retrieve the same true positive rate as Pearson correlation was often an order of magnitude larger than the correlation based networks. This indicates that in practice, interpretation of the results of GENIE3 and ARACNE may be more challenging, less beneficial and less intuitive than analysing a Pearson correlation based network, although neither will consistently return more true positives than false positives.

3.3.3 Microarray perturbation results are consistent with the literature-curated RTPs

To examine whether the microarray data actually contain any information for identifying causal regulatory interactions, we investigated whether the set of differentially expressed genes from perturbation experiments can be used to infer causal regulatory relationships. The tooth microarray data set contains 6 perturbation experiments, including transgenic knockdowns of *Msx1* and *Pax9*, and exogenous stimulation of the *BMP*, *Wnt*, *sonic hedgehog* and *FGF* pathways. Based on the pathway information provided by (O'Connell *et al.*, 2012), we identified 39 RTPs from the literature at stage E13 that corresponded to the microarray perturbation experiments. Encouragingly, the observed directions and fold changes of differential expression as determined by the microarray experiments were consistent with the regulatory relationships predicted by the literature (Figure 3.9). We found that using a fairly conservative absolute $\log_2$ fold change cut-off of 1 (*i.e.*, 2-fold change) would result in a edge detection sensitivity (true positive rate) of 30%, increasing up to $> 70\%$ as the cut-off is relaxed. The false positive rate (=1-specificity) is consistently much lower than the true positive rate, suggesting that it is possible to distinguish causal gene regulation from non-regulatory ones with a reasonable sensitivity and specificity. We repeated the analysis considering all developmental stages, which increased the number of RTPs to 144. The trends are still visible although with increased noise (Figure 3.10).

3.3.4 Tissue and temporal specificity is a confounding factor in network reconstruction

We sought to investigate the extent to which different tissues display different genetic responses to the same stimulus. Using the ToothCODE microarray profiles on genetic perturbation experiments, we found that the magnitude of tissue specific responses varies considerably between different perturbations. First we examined epithelial and mesenchymal tissue microarray profiles from *Pax9*$^{-/-}$ and *Msx1*$^{-/-}$ mice (Figure 3.11A,B). We identified hundreds of genes are significantly differentially expression in only one tissue

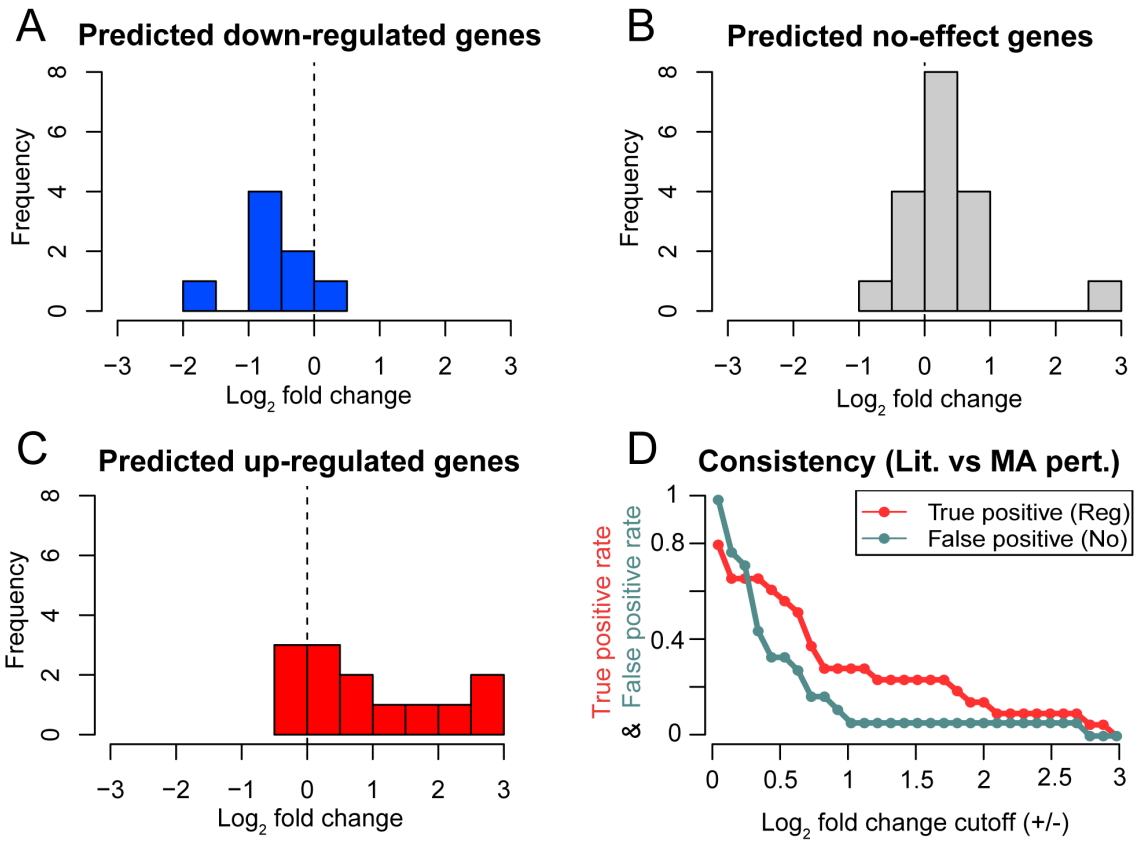

Figure 3.9: **Fold changes ($\log_2$) from tooth microarray perturbation experiments that matched the perturbation evidence in the literature show consistency with expected trends.** RTPs that are inhibiting (A), have no effect (B), or are activating (C) trend to have negative, close to zero and positive fold changes respectively. (D) shows the consistency of the literature based RTP type (Lit.) and microarray data (M.A.) as fold change cut-off varies between 0 and 3 (both up- or down-regulation).

type and not the other, even in the same genetic mouse model (FDR$<$ 0.01). Similarly, we observed that distinct sets of genes are differentially expressed in response to the same signalling pathway stimulation (BMP and Wnt) in dental epithelium versus dental mesenchyme (Figure 3.11C,D; see also Figure 3.12). In addition, we also observed many tissue and/or temporal specific causal gene regulation in our tooth and heart literature data sets (Table 3.13). These results suggest that the causal gene regulatory network structure may be specific to individual cell or tissue types. Therefore, it is important to consider cell-type specificity when constructing GRNs in multicellular organisms (Li and White, 2003;

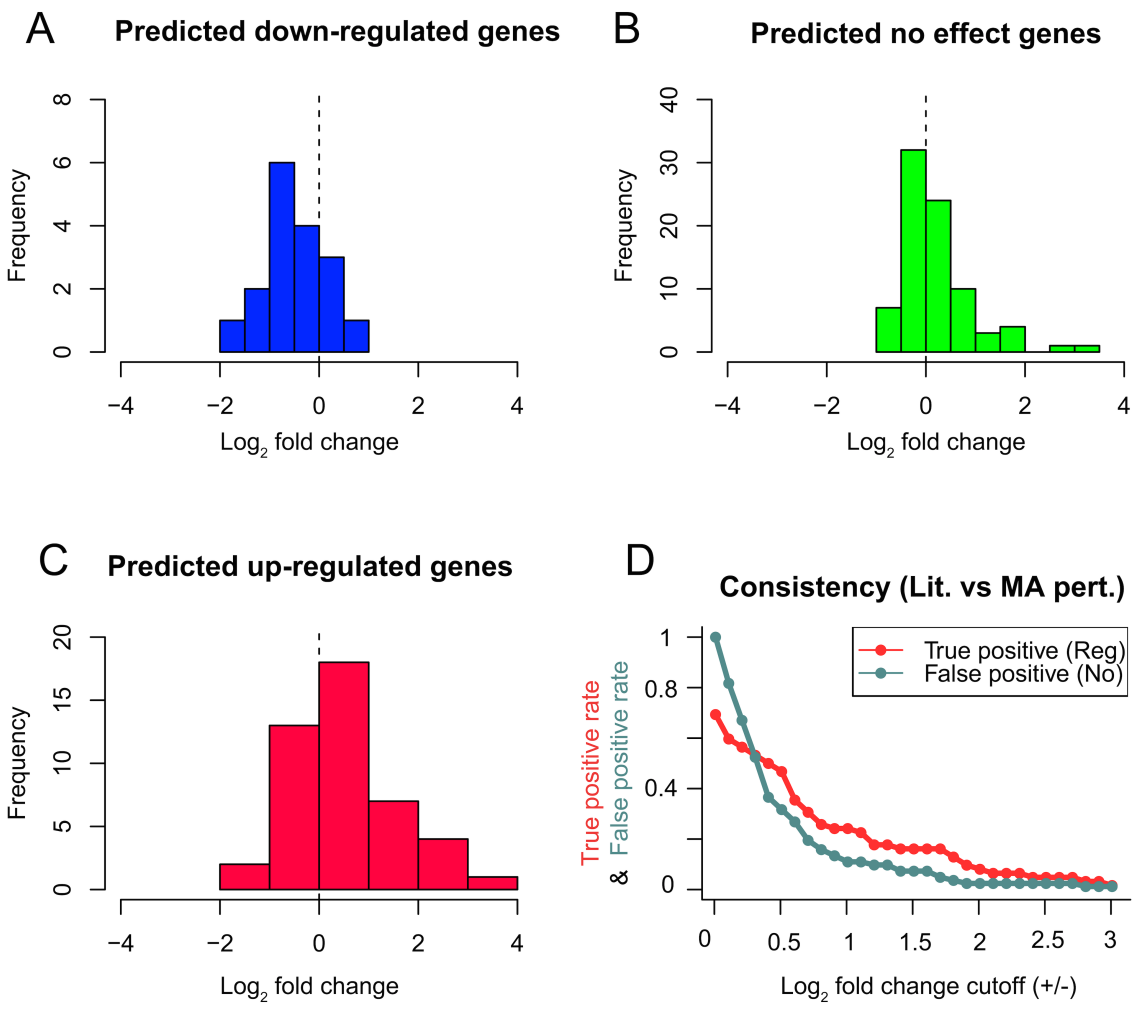

Figure 3.10: **Fold changes ($\log_2$) from tooth microarray perturbation experiments that matched the perturbation evidence in the literature (all stages)** show consistency with expected trends. Regulatory relationships that are inhibitory (A), have no effect (B), or are activating (C) trend to have negative, close to zero and positive fold changes respectively. (D) shows the consistency of the literature based predictions and microarray data as fold change cutoff is increased.

Odom, 2004).

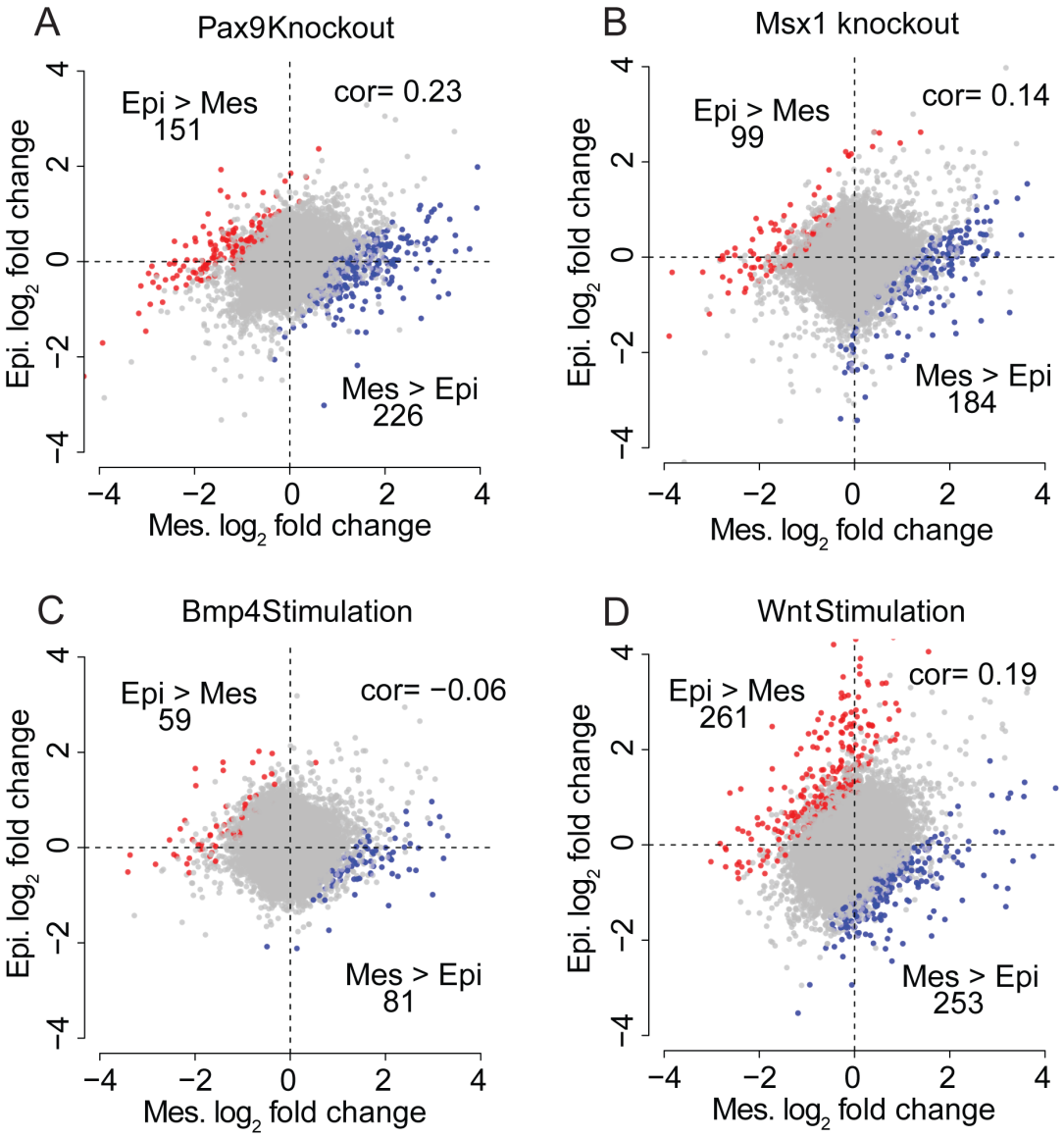

Figure 3.11: **Scatter plots show the extent of tissue-specific differential expression in dental epithelium (y-axis) and dental mesenchyme (x-axis)** as a result of *Pax9* knockout (A), *Msx1* knockout (B), *Bmp4* stimulation (C) and *Wnt* stimulation (D). Coloured points represent probes of differentially responsive genes between the two tissues. Pearson correlation is also shown.

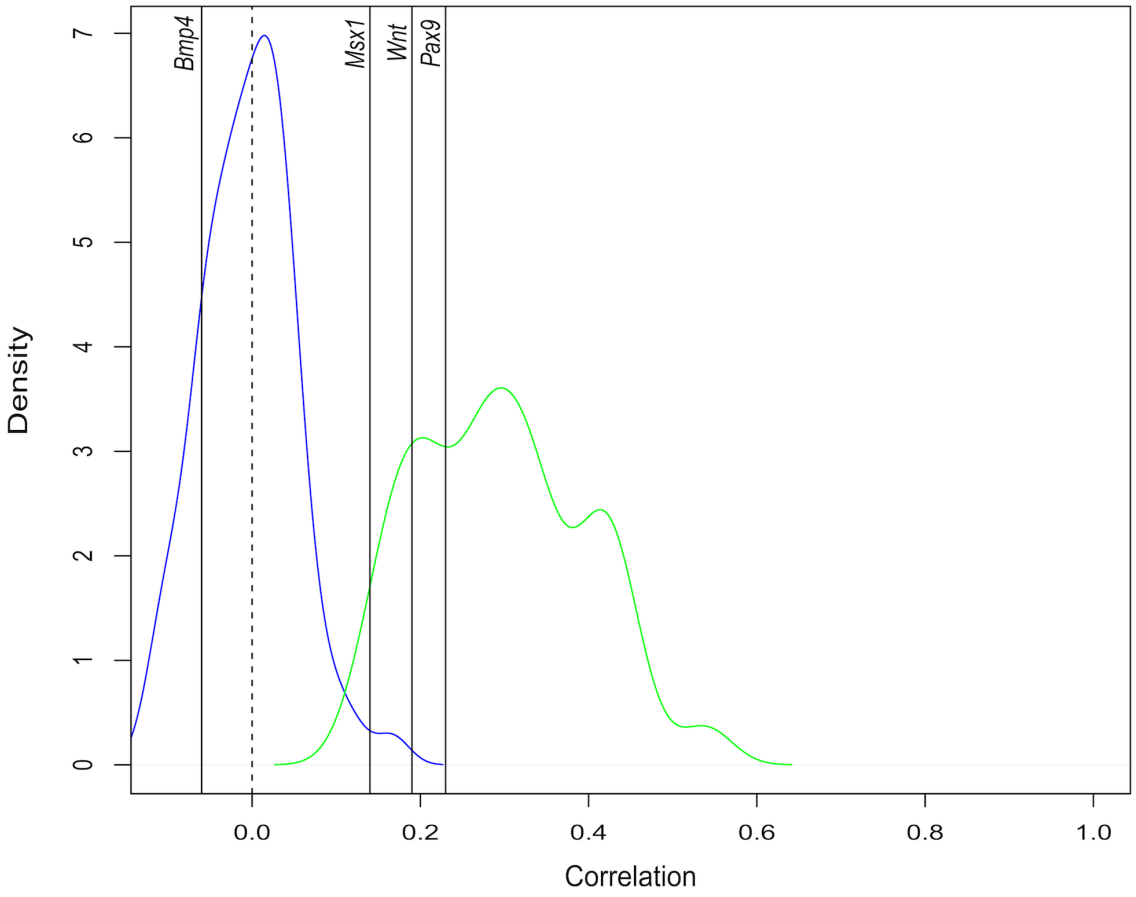

Figure 3.12: **Negative (blue) and positive (green) control distributions for analysing tissue-specific genetic responses to the same perturbation.** Positive control is generated by correlation of fold change of biological replicates. Negative control is correlation of independent experiments.

3.3.5 The value of using perturbation data for GRN inference

It has been commonly believed that it is best to infer GRNs using expression profiles from a broad range of diverse conditions. To achieve such a diversity, we might collect samples from multiple cell types, multiple genetic perturbation, or developmental time series. To investigate the relative value of using perturbation vs. time series data, we split the heart microarray data set into: 1) arrays from wild-type time-series experiments; 2) arrays from perturbation experiments. To see if the correlation shift trends arose we plotted the Pearson correlation for each type of RTP (Figure 3.6). We observed that the

correlation of activating RTPs in the perturbation subset is generally higher than that observed in the time-series subset. The combined set seems to yield the best result.

We have shown that there can be a slight shift in the overall distribution of correlation values between activating and inhibiting causal relationships, but not to the extent where a cutoff can accurately differentiate these two classes from false positives (Figure 3.5, 3.6). How much information can be gained by exploiting perturbation experiments? We calculated the AUROC for Pearson correlation of all our activating or inhibiting RTPs compared to our no-effect RTPs, and similarly for the fold change values observed in matching perturbation microarray experiments (Figure 3.14). We clearly see that fold change from direct perturbation experiments is a much better predictor of causal gene regulation than Pearson correlation, with AUROCs of 0.63 - 0.87 compared to 0.55 based on gene co-expression alone.

Table S2. Summary of tissue and time specific regulatory actions. Full references can be found at ToothCode (http://compbio.med.harvard.edu/ToothCODE/) and CardiacCode (http://cardiacCode.victorchang.edu.au) websites

Heart

Regulator	Target	Tissue	Stage	Mode	Reference
Pax3	Plxna2	Hypaxial muscle	E12.5	act	Brown et al. (2001)
Pax3	Plxna2	Left ventricle	E12.5	inhib	Brown et al. (2001)
Tbx20	Hand2	Right ventricle	E9	act	Takeuchi et al. (2005)
Tbx20	Hand2	Right ventricle, Outflow tract	E9.5	inhib	Singh et al. (2005)
Tcf21	Tcf21	Pharyngeal mesoderm	E9.5	act	Harel and Maezawa (2012)
Tcf21	Tcf21	Pharyngeal mesoderm	E9.75	inhib	Harel and Maezawa (2012)
Wnt3	Lhx1	Mesoderm	E7.5	act	Liu et al. (1999)
Wnt3	Lhx1	Anterior viseral endoderm	E7.5	inhib	Liu et al. (1999)

Tooth

Regulator	Target	Tissue	Stage	Mode	Reference
Bmp4	Dlx2	Epi	E10 and E13	act	Thomas et al. (2000), Liu et al. (2005)
Bmp4	Dlx2	Mes	E11	act	Bei et al. (1998)
Bmp4	Dlx2	Mes	E10	inhib	Thomas et al. (2000), Liu et al. (2005)
Bmp4	Shh	Epi	E11, E12 and E14	act	Fujimotri et al. (2010)
Bmp4	Shh	Epi	E11 and E14	inhib	Zhao et al. (2000)
Ednra	Dlx2	Epi	E10	act	Ruest et al. (2004)
Ednra	Dlx2	Mes	E10	inhib	Ruest et al. (2004)
Fgf8	Dlx2	Mes	E11	act	Thomas et al. (1997), Bei et al. (1998)
Fgf8	Dlx2	Epi	E10	inhib	Thomas et.al (2000)
Sostdc1	Shh	Epi	E14	act	Ohazama et al. (2008)
Sostdc1	Shh	Epi	E13	inhib	Ahn et al. (2010)

Figure 3.13: **Summary of tissue and time specific regulatory actions.** Full references can be found at ToothCode (http://compbio.med.harvard.edu/ToothCODE/) and CardiacCode (http://cardiacCode.victorchang.edu.au) websites.

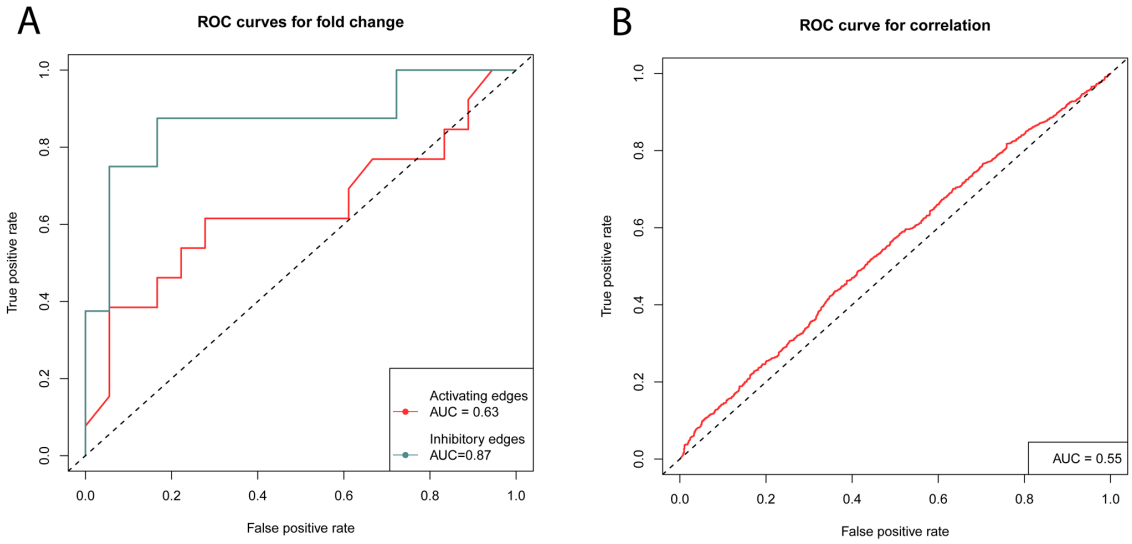

Figure 3.14: **ROC curves showing the ability of perturbation experiments (A) and gene expression correlation (B) to differentiate regulatory from non-regulatory edges.**

3.3.6 GRN inference based on protein interaction network and other molecular pathways

Using co-expression (as determined by Pearson correlation), we could achieve a true positive rate of 25%, but with almost a 30% false positive rate. We found that only 3-6% of the edges in Pathway Commons pathways or protein-protein interaction networks overlap with activating or inhibiting RTPs, however in all cases a similar proportion of false positives was also retrieved (Figure 3.15, Figure 3.16). By explicitly taking into account the perturbation design (as in Figure 3.9), we can significantly increase the true positive rate while keeping the false positive rate low (Figure 3.15, Figure 3.16).

3.4 Discussion

This study aims to evaluate the practical utility of genome-wide expression profiles to infer causal gene regulatory networks in mammalian organ development. In particular,

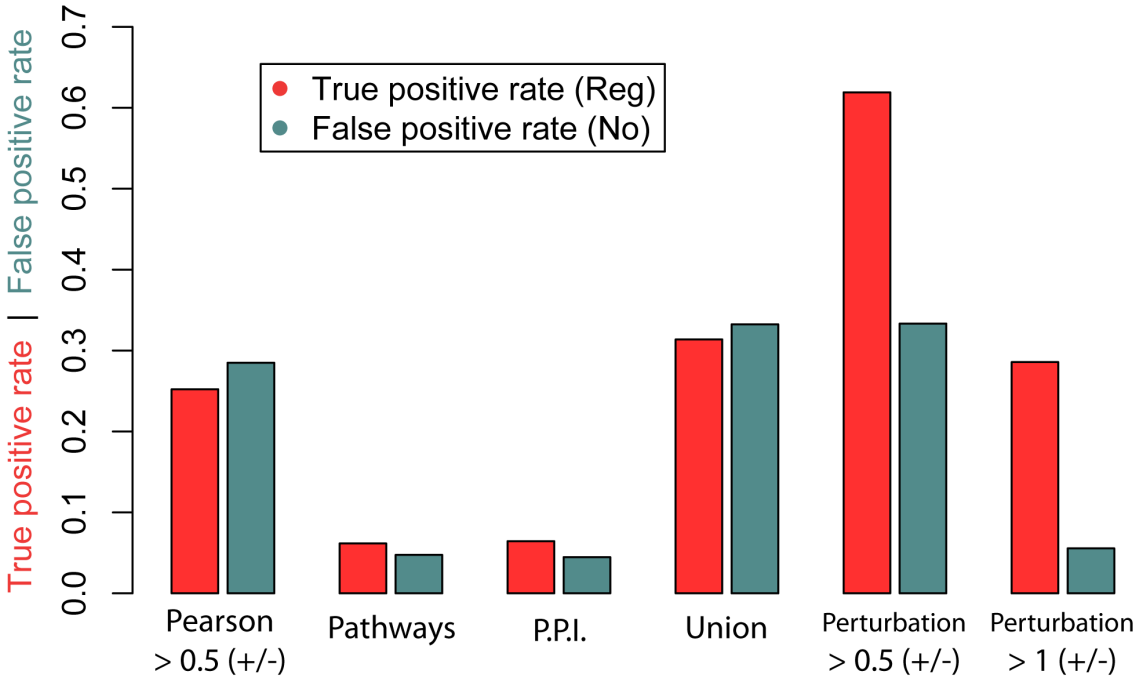

Figure 3.15: **Comparison of the true positive and false positive rates as determined by different network inference approaches on the tooth data set:** Pearson correlation, Pathway Commons database, protein-protein interactions (PPI), the union of the previous three methods and direct effect on genetic perturbation ($\log_2$ fold change cut-off or 0.5 and 1). Note: the TP and FP rates for the first 4 methods were calculated based on the subset of 686 RTPs that were represented in the microarray, PPI and pathway data. The TP and FP rates for perturbation data were based on the subset of 39 RTPs with a regulator matching the pathway being perturbed.

we assessed whether it is possible to observe gene co-expression in experimentally verified causal gene regulatory relationships — a common assumption in most GRN inference algorithms (Bansal *et al.*, 2007; Marbach *et al.*, 2012). One of the major results from the DREAM5 challenge is that many inference methods performed well when analysing *in silico* data sets and prokaryotic (*E. coli*) data sets, but inference of eukaryotic GRNs (in *S. cerevisiae*) is very poor regardless of which method was used (Marbach *et al.*, 2012). Marbach et al. (2012) attributed the reduced inference accuracy to an increased regulatory complexity and prevalence of post-transcriptional regulation in eukaryotes. Ensemble-based approaches that combine multiple inference methods have been shown to slightly improve the inference accuracy (Marbach *et al.*, 2012).

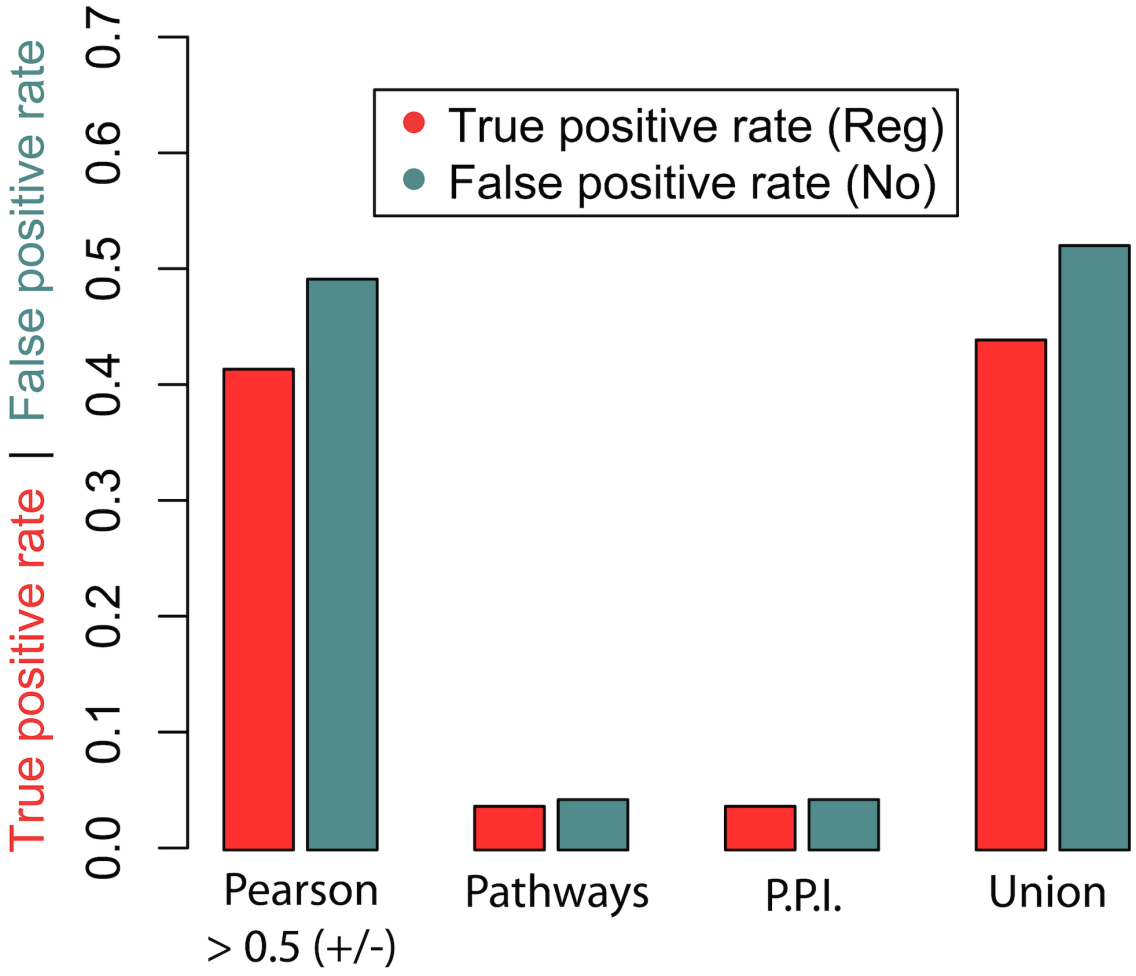

Figure 3.16: **Comparison of the true positive and false positive rates as determined by different network inference approaches on the heart data set:** Pearson correlation threshold on 82 microarray profiles, Pathway Commons database, protein-protein interactions (P.P.I.), and the union of the previous three methods.

We have gathered two well validated literature-curated data sets and matching microarray gene expression data sets to systematically evaluate the challenges of causal GRN inference. Our data sets are unique because they contain thorough annotation of tissue types and embryonic stages, as well as the type of regulation observed (activation, repression and no effect), which importantly allows us to estimate both sensitivity and specificity of inference of GRN edges. To our knowledge this study contains the most extensive evaluation of commonly applied GRN inference paradigms to mammalian embryogenesis and

the first quantification of the difficulty of their application to this context. Our results show that inference of causal GRNs for mammalian developmental systems by considering gene co-expression alone is likely not an effective approach. Nonetheless, perhaps not too surprisingly, it is possible to infer causal regulatory relationships with good sensitivity and specificity if perturbation data are used. This result supports the importance of considering these data when reconstructing causal regulatory networks.

Our study place a strong emphasis on embryonic organ development. From a practical point of view, we chose this emphasis because of the wealth of data we have already collected (*e.g.*, the published ToothCode data), the availability of a large amount of matching published microarray gene expression data from GEO, and the many reported successful applications of GRN to study developmental biology problems, such as Eric Davidsons work (Davidson, 2006, 2010; Levine and Davidson, 2005). In this sense, the process of GRN inference should be easier than other non-developmental GRNs. From a conceptual point of view, the inference of developmental GRN is at least as difficult as, if not more difficult than, the inference of other GRNs since a useful developmental GRN will need to deal with regulatory relationships between multiple cell types, and the regulatory relationship between two genes may change dramatically during successive developmental stages. Therefore, we expect the lessons learned from our study will be informative to the inference of other non-developmental GRNs.

Our tooth and heart microarray data sets each have about 100 microarray samples, containing about 30 conditions. It is conceivable that better performance can be achieved by profiling more samples in additional conditions. Nonetheless, we noticed that it is practically not easy to obtain such data when studying *in vivo* gene expression patterns in embryonic animal models. Embryonic dissection, tissue collection and processing all require time, money and labour.

One potential limitation of our study is the imbalance in classes of our gold-standard perturbation based edges. The negative classes (no regulatory effect) represented 25% and 36% of our heart and tooth data respectively. For a more confident interpretation of our measured true / false positive / negative rates and the AUROC, the negative class

should represent 50% of the total samples, however considering we were not using these data to train any models, only to evaluate the inferred networks, we do not expect the results would have differed significantly. Also due to our manual curation of gold-standard data from the literature it is likely that this data does not represent an unbiased sample of regulatory edges, but those genes that are most interesting to researchers and play some role in the cell type being profiled. The inference algorithms are however agnostic to this type of information, which may reduce the expected overlap of the gold-standard and inferred edges.

Another potential limitation of this study is that our network inference model is based on the conclusions of O'Connell *et al.* (2012), particularly in the robustness of their results to various fixed model parameters. As the data we collected for this study has very similar characteristics as their data, including distribution of effect types and high sparsity for most RTPs, we assume that their chioce of parameters will be suitable for our models. In the future we should collect larger and less sparse data sets to investigate estimating the parameters directly from the data.

We did not extensively test the effect of microarray probe normalisation procedure. We simply follow common practices in microarray analysis, and ask whether this is quantitatively sufficient to recover information for GRN construction. It is conceivable that a more extensive enumeration of microarray processing procedure may yield higher correspondence with the literature network, but considering the high false positive rate observed in our current data set, we do not expect the major results to change.

There are always new methods for correlation-based GRN inference being released with differeing sets of assumptions and strategies, for example by explicitly modelling time delays in time-series data. With this in mind, this study should be repeated in the future to determine if the conclusions still hold.

3.4.1 Lessons for mammalian causal GRN inference

During the course of manually curating the literature data, we observed that there is a vast amount of genetic or molecular perturbation data in the published literature that largely remains computationally inaccessible. Unlike microarray or high throughput sequencing data, most people do not deposit the results of their perturbation results into a centralised database such as EBI ArrayExpress (Rustici *et al.*, 2013) or NCBI GEO (Edgar, 2002). Based on our experience, an undergraduate-level biology student can read 2-3 papers a day, and each paper contains on average 12 useful pieces of perturbation data. In one month, a single person can curate up to 700 pieces of perturbation data. Ultimately we would like to see a similar centralised repository where authors and researchers submit their own spatio-temporally annotated perturbation results at the time of publication, but there is currently no standard for reporting and annotating these data set. Our experience on manual curation has been generally very positive and rewarding.

Considering the amount of gene perturbation data that one can obtain from simply computerising existing records, we believe this suggests that the community of computational systems biologist investigating mammalian disease and development should perhaps re-prioritise their research effort, *e.g.*, instead of focusing on inferring causal GRNs from high throughput genome-wide data sets, committing resources to systematic generation and curation of relevant genetic perturbation data, and developing algorithms to construct cell type and developmental stage specific GRNs from these potentially sparse and noisy perturbation data.

Chapter 4

An integrative systems biology approach to discover cataract disease genes

4.1 Introduction

Cataracts are an opacification of the ocular lens that are the leading cause of blindness, affecting tens of millions of people worldwide. Most cataracts develop through natural ageing, however around 25% of cataracts are inherited and congenital cataracts can also present at birth. While some congenital cataracts present as just one defect in a developmental syndrome, many appear to be spontaneous without any other disease phenotypes. Previous studies have identified multiple genes and loci responsible for congenital cataracts, and these are recorded in a database called CatMap (Shiels *et al.*, 2010).

A computational tool called iSyTE was recently constructed to help prioritise potential cataract causing genes by analysing gene expression during early lens development in the mouse (Lachke *et al.*, 2012). The rationale was that genes that are uniquely expressed in the lens during development are more likely to be congenital cataract causing genes.

In order to discover lens specific genes, Lachke *et al.* (2012) computed differential gene expression between the micro-dissected lenses and the remainder of the embryo with ocular tissues removed, during several stages of mouse development. Using this approach, iSyTE could successfully prioritise many human congenital cataract causing genes within a given genomic interval, but not all of the cataract disease genes were differentially expressed in lens during murine development. There is a wealth of additional biological knowledge that could be applied to this problem, including regulatory relationships between genes during eye development, relevant protein complexes and qualitative gene lists that have been annotated as important to lens development. While it is not clear which of these data sources will prove most useful for the cataract context, this type of integrative systems level analysis has previously been useful for prioritising disease causing genes and pathways in neurodegenerative diseases (Zhang *et al.*, 2013).

In this study we build on the foundation of iSyTE by investigating the utility of integrating additional types of biological data for predicting human congenital cataract genes through systems-biology inspired approaches. We construct a lens development specific GRN, and develop an algorithm to prioritise upstream regulators of sets of phenotype specific genes in the network. We also use machine learning approaches to identify patterns in a combination of gene expression data sets during mouse lens development and after genetic perturbations, protein interaction data and gene ontology information, to improve our ability to prioritise congenital cataract causing genes.

4.2 Method

4.2.1 Integrative cataract gene analysis workflow

This chapter implements the dual methodology approach shown in figure Fig. 4.1.

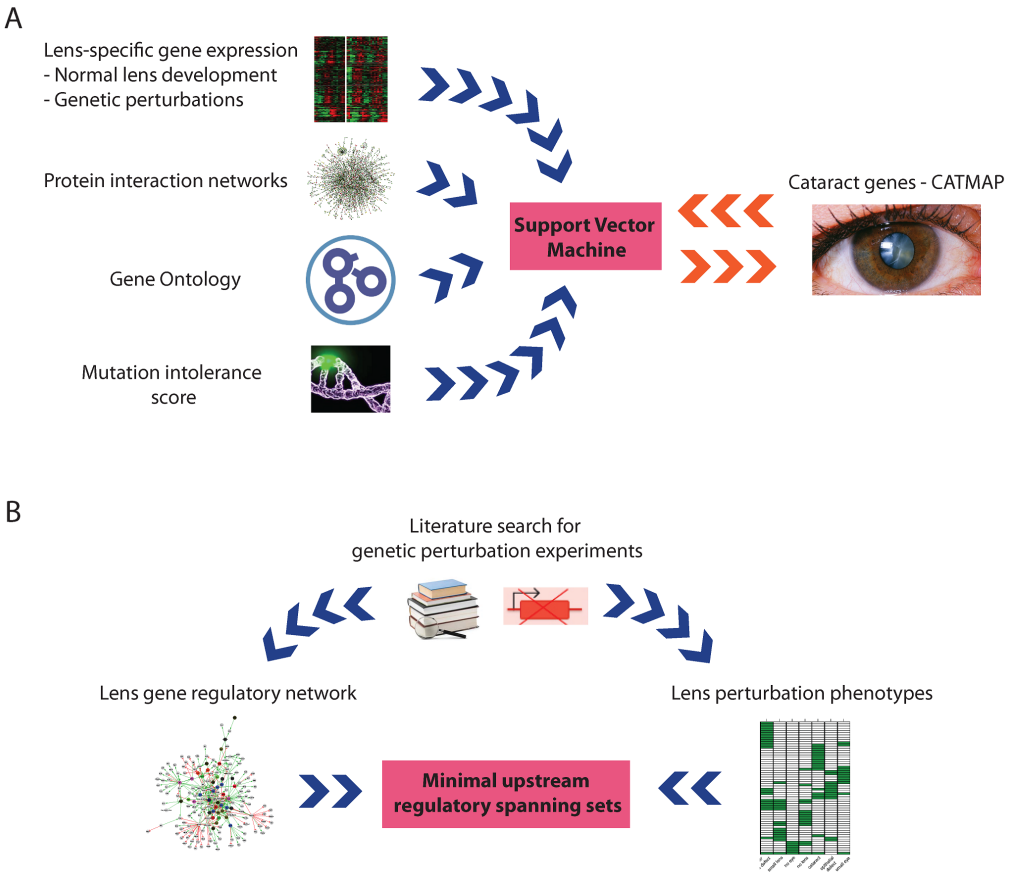

Figure 4.1: **Integrative cataract gene analysis workflow.** A) Integrating mutiple data sources through a machine learning approach to predict cataract disease genes B) Identifying likely regulators of specific cataract phenotypes by constructing and analysing a lens-specific gene regulatory network

4.2.2 Predicting cataract genes through integrative analysis of lens gene expression

Microarray data processing

Microarrays representing the breadth of available experimental data in developing mouse lens were collected. These data contain normal mouse lens development as well as multiple perturbation experiments (e.g. knock down of *Pax6* and *Notch2*) and are derived from 4 different microarray platforms: Affymetrix 430_2; Affymetrix 430A_2; Illumina MouseWG-

6 v1.0; Illumina MouseWG-6 v2.0. 127 arrays passed quality controls including separation of phenotypic classes in PCA and clustering (several other arrays and experiments were omitted at this stage). The remaining 127 arrays were jointly processed from the raw data within each platform.

Within each Affymetrix platform all arrays were jointly normalised using the robust multi-array average (RMA) method from the oligo R package. Within each Illumina platform, RMA equivalent normalisation was performed using the lumi R package, including background subtraction, quantile normalisation and transformation to Log_2 scale.

Differential expression (DE) analysis was conducted using the limma method (Smyth, 2005). The control samples used for the lens developmental time points consisted of whole mouse embryos with the lens removed from E11-13. Log_2 fold change, t-statistics, t-statistic ranks, p-values, average expression, and f-statistics were generated. All of this data is available for download from our website: `http://iSyTE.victorchang.edu.au`.

Probe presence / absence calls were determined by comparison to 351 data points from the published lens developmental literature, primarily measured by immuno-fluorescence and *in-situ* hybridisation techniques. Each gene expression measurement from the literature was paired with the closest matching microarray, and an ROC analysis was performed to identify the optimal cutoff value for each platform.

Data sources

The combined human and mouse protein-protein interaction (PPI) network was constructed using the iRefR (Mora and Donaldson, 2011) and igraph (Csardi and Nepusz, 2006) packages in R. The final PPI was converted to mouse gene symbols using homology data from MGI and the Ubc node was removed, resulting in 170,872 interactions between 14,433 protein nodes.

The gene ontology (GO) annotations for human and mouse were downloaded from `http://geneontology.org/page/download-annotations` on the 6/9/2014 and the 19/9/2014

respectively. 37 GO terms specifically related to lens or cataract were identified as the target set.

Genic variation intolerance scores were downloaded from Petrovski *et al.* (2013). CatMap genes were downloaded from `http://CatMap.wustl.edu/` on 29/8/2014. Three classes of CatMap genes were defined; classic cataract genes as identified by (Lachke *et al.*, 2012); all non-syndromic CatMap genes; and all CatMap genes.

Feature construction

All microarray DE results from lens developmental and perturbation experiments were used as features. The fold changes from each developmental microarray DE analysis were propagated through the PPI network to calculate neighbourhood enrichment of lens specific gene expression. The average lens enrichment fold change value of a protein's first degree interaction partners (excluding itself) was calculated for each developmental time point.

A combined GO score was defined as the sum of the number of target GO terms annotated to each gene in mouse or human. The combined GO score was propagated throughout the PPI network in order to score neighbourhood enrichment of GO terms. The average and sum of the combined GO scores of 1st degree interaction partners (excluding itself) were calculated. The sum was also normalised by the degree, and separately the $\log_2$ of the degree, of interactions of each protein node, producing three features in total.

Machine learning

Support vector machine (SVM) learning was implemented using the e1071 package. Area under the receiver operator characteristics curve (AUC) was calculated based on leave-one-out cross-validation (LOOCV), alleviating the need to separate the training and test sets. As such, the positive class of genes consisted of all those genes in the CatMap group of interest for each analysis. The negative class was twice the size of the positive class,

and was randomly sampled from all genes that were not in the CatMap lists or associated with the terms eye, lens, cataract, or ocular, in the Online Mendelian Inheritance in Man (OMIM) knowledge-base (Hamosh, 2004). AUC distributions were calculated from 100 separate LOOCV runs, each run re-sampling the negative class.

Feature selection

Feature selection was performed using the SVM-recursive feature extraction (RFE) method, provided by: `http://www.uccor.edu.ar/paginas/seminarios/Software/SVM_RFE_R_implementation.pdf`. Feature selection was performed 1000 times using a randomly sampled negative class as described above. Each run produced a list ranking all 55 features. For each feature, the mean rank across 1000 runs was taken as the final value, and the features with the lowest 10 values were selected.

4.2.3 Analysing lens-specific gene regulatory networks

Manual curation of genetic perturbation evidence from the literature

Dr. Deepti Anand at the University of Delaware recorded genetic perturbation experimental evidence from primary research papers. Each piece of evidence consists of 11 crucial pieces of information: regulator gene; target gene; perturbation performed on the regulator (+ or -); effect on the expression of the target gene (up-regulated, no change, down-regulated); species; developmental stage; tissue in which the perturbation was performed; tissue in which the expression of the target gene was measured; measurement technique; type of molecule measured (mRNA or protein); citation. If we were not confident about any of these pieces of information, the evidence was discarded. We further recorded the experimental context and additional information where it was available, including the genotype and phenotype of the perturbed mouse embryo.

Annotation of perturbation evidence onto a developmental ontology

We used the e-Mouse Atlas Project (EMAP) (Richardson *et al.*, 2014) developmental ontology to provide a scaffold for consistent annotation and integration of experimental evidence from a variety of tissue types across developmental stages. We added two additional terms to the ontology, "TS13 lens pre-placodal ectoderm" and "TS14 lens pre-placodal ectoderm". Each piece of collected evidence was assigned the EMAP ontology term corresponding to the tissue and stage of the observed perturbation. Lens related sub-trees of the ontology were then used to select data for further analysis.

The data were also split by developmental time into four different stages of lens development: initiation (318 pieces of evidence, E8-E10.5); primary fiber cell differentiation (PFCD, 434 pieces of evidence, E11-E13.5); secondary fiber cell differentiation (SFCD, 576 pieces of evidence, E14-E19.5); post natal (348 pieces of evidence, P0-P665).

Inferring mode of regulation of a regulator-target pair

See section 3.2.3.

Minimal upstream regulator spanning sets

The spanning set of upstream regulators for a set of seed genes is calculated via a novel algorithm, the Upstream Regulator Spanning Set (URSS) algorithm. This algorithm tackles a similar problem as the Ingenuity Upstream Regulator Analysis (URA) but with different assumptions about the data and hence a different statistical approach (Kramer *et al.*, 2014). The main difference is that URA is designed to work on signed gene expression data such as fold-change, where as URSS is designed to work on an unsigned gene set. The URSS algorithm is a recursive depth first search, exhaustive and exact, and hence pre-filtering the data is essential for reducing the search space. Upstream regulation is converted into a Boolean matrix where rows represent regulators and columns seed genes.

First, seed genes without any regulators are removed. Second, redundant regulators (those whose targets are a strict subset of another regulator's targets) are removed. Third, critical regulators, (ie. those that are the only regulator for a seed gene) are removed and stored, as are all the seed genes which are covered by these critical regulators. Finally the URSS algorithm is called on the resulting matrix, returning all upstream regulator spanning sets for the seed genes.

From this complete collection of sets the minimum sized sets are found, as they contain the regulator genes that regulate the largest number of seed genes, and therefore the most relevent regulators. Several enhancements were made to improve the URSS algorithm with the specific goal of efficiently returning the minimal spanning sets, including: not searching deeper than the minimum depth of a previously found spanning set, and prioritising regulators by their coverage of the seed genes. Thus, the minimal upstream regulator spanning sets (MURSS) are the smallest possible sets of regulators that lie upstream of all the seed genes in the GRN, given a maximum depth to search.

The pseudo code for the enhanced MURSS algorithm is provided below:

```
initialise global variable max.d = 999 #maximum depth
initialise variable cur.d = 0 #current depth
initialise variable rm = pre-processed regulator matrix
MURSS (rm, cur.d)
if cur.d > max.d
   return <reg> , FALSE
end if
for each row of rm [sorted in order of coverage of rm] (reg)
   if rm [ reg , ] is all TRUE
      max.d = cur.d
      return <reg> , TRUE
   else
      return <reg> , MURSS ( rm [ !reg , !(rm [ reg , ]=TRUE) ], cur.d + 1 )
```

```
   end if
end for
```

Where $< reg >$ indicates a marker to keep track of the order of the path through the regulators.

The critical regulators removed in pre-processing are now added to the minimum spanning sets. From these minimum sets an occurrence frequency for each regulator is calculated. This occurrence frequency is multiplied by the number of seed genes that each regulator regulates, giving a weighted score for each regulator.

P-values were calculated using the monte-carlo simulation approach. For each set of seed genes, 1000 random seed gene sets of the same size were generated from the genes in the network, and the weighted scores were computed by MURSS. The reported P-value is the fraction of simulated weighted scores that are greater than the observed weighted scores for each minimum regulator in each seed set.

4.3 Results

4.3.1 Lens specific gene expression is predictive of cataract disease genes

Lens specificity of gene expression is predictive for CatMap genes compared with the negative class (Fig. 4.2A). As the 22 classic non-syndromic (NS) iSyTE genes are almost perfectly classified by lens specific gene expression (AUC 0.95), from here on we discuss the results for the larger group of 52 NS CatMap genes. Although post natal time-points significantly improve performance compared to any other developmental stage ($p = 3.94 \times 10^{-09}$), they do not significantly improve on all stages combined ($p = 0.05745$) (Fig. 4.2B). The improvement using post natal time-points may be a factor of data depth, as early embryo is represented by five time points between, late embryo is represented by two time points, and post natal is represented by 11 time points. Choice of DE statistic also

made little difference (data not shown). For the sake of simplicity, dimension reduction and interpretability, we will only use fold change (FC) in further analyses.

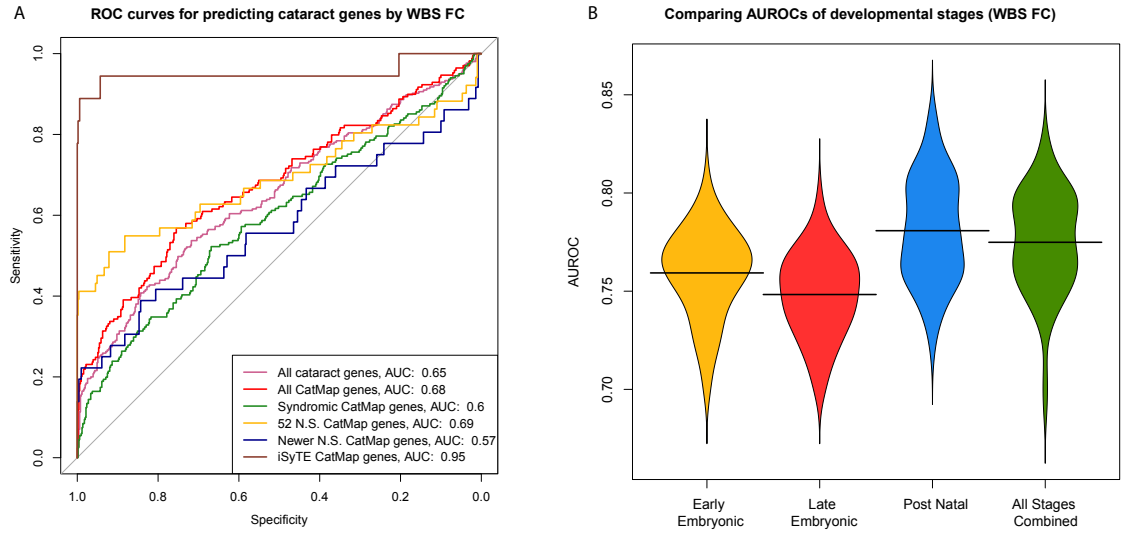

Figure 4.2: **Lens specific gene expression is predictive of cataract disease genes.** A) AUROC analysis of the predictive power of lens specific gene expression on multiple classes of cataract genes. B) Bean plots show the relative predictive power of lens gene expression from different stages of mouse development. WBS = whole embryo background subtracted (lens specific gene expression)

4.3.2 Certain perturbation experiments in mouse lens are predictive of cataract genes

Changes in gene expression due to genetic perturbations (knock down of *CBP/p300*, *Pax6*, *E2f* or *Hsf4*) are predictive of CatMap genes, but are less informative than lens specific gene expression (Fig. 4.3). *CBP/p300* early embryonic knock-out is the most informative individual perturbation experiment. Combining *CBP/p300* with *Tdrd7*, *E2f* and *Hsf4* outperforms the complete set of all perturbations. Note that the *Tdrd7* knock down experiment did not have a high predictive value by itself, but increases prediction performance when combined with the others. This is in stark contrast to the *Pax6* knock down data, which was somewhat predictive on its own but reduced overall performance when combined with the three informative knock down experiments. Using this specific combination

of perturbation data sets is significantly better than using all of them ($p = 3.825 \times 10^{-05}$). However, combining them with lens specific gene expression does not increase prediction performance over just lens specificity alone.

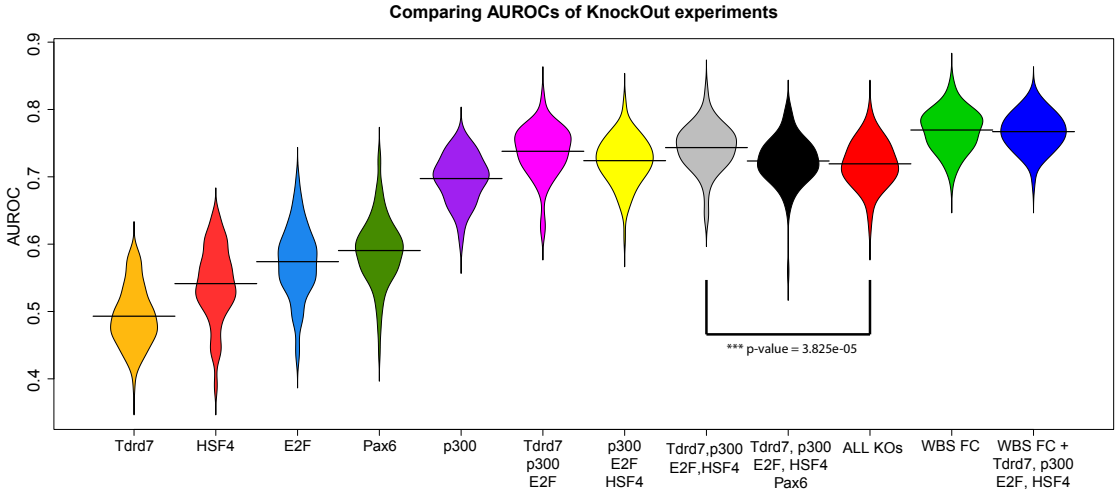

Figure 4.3: **Certain perturbation experiments in mouse lens are predictive of cataract disease genes.** WBS FC = whole embryo background subtracted fold-changes (lens specific gene expression)

4.3.3 Incorporating functional knowledge increases predictive power

We next investigated the classification power of three important sources of functional biological information; genetic variance intolerance scores; PPI based features; and GO based features.

Genetic variance intolerance scores had minimal classification power alone, with an average AUC of 0.54 (Fig. 4.4). PPI based features were slightly more predictive of CatMap genes, with an average AUC above 0.6. GO based features showed a strong predictive power for CatMap genes, with an average AUC of 0.75, comparable to lens specificity and genetic perturbation features. Combining GO based features with gene expression features increased classification performance.

The most informative 10 features for CatMap classification identified by SVM-RFE contain

a mixture of GO term and lens specific expression enrichment of PPI neighbourhood, lens specific gene expression and perturbation experiments. Using the top 10 features for classification produced an average AUC of 0.84, a significant improvement over lens specific expression ($p = 2.33 \times 10^{-63}$) as well as the combination of all features ($p = 1.86 \times 10^{-39}$).

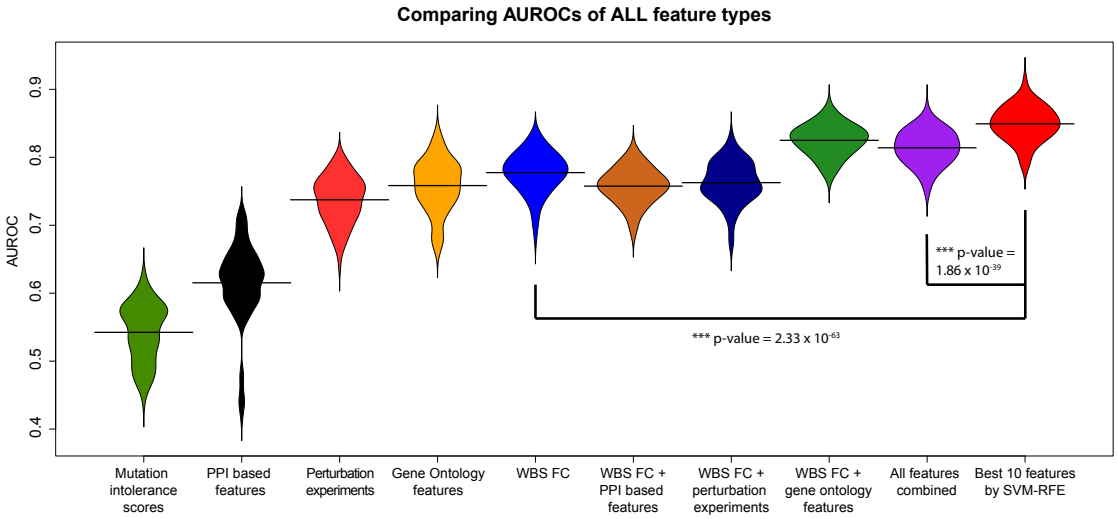

Figure 4.4: **Incorporating functional knowledge increases predictive power.**

4.3.4 Predicting novel cataract genes

SVMs were trained using the top 10 features identified by SVM-RFE to classify each class of CatMap genes, resulting in three SVM models. Predictions were then made for each gene in the genome for which the top 10 features were available, resulting in a predicted cataract score for 8747 genes. In total, 531 genes (6%) received a score greater than zero (indicating a potential cataract gene) from at least one model. Of the classic iSyTE genes, only *Chmp4b* was misclassified by all three models. 27 / 52 NS CatMap genes and 56 / 167 of all CatMap genes were predicted as cataract causing by at least one model.

50 genes received a score above zero from all three models (Fig. 4.5). These were taken as our highest confidence predictions. 16 of the classic iSyTE cataract genes were in this category, along with 8 other CatMap genes and 26 novel predictions. A literature search

showed that five of these high confidence novel predictions (*Six3, Capn3, Sox1, Aqp5* and *Cyp4v3*) have causal links to cataracts in published genetic studies. Investigating the literature for some of the genes predicted by one or two models yielded at least another six verified cataract genes *Wfs1, Gss, Anxa1, Rbp3, Trpm3* and *Birc7*. Our data integration and machine learning approach thus helped us prioritise at least 11 known cataract genes that were not in the CatMap database, with many more potential new leads.

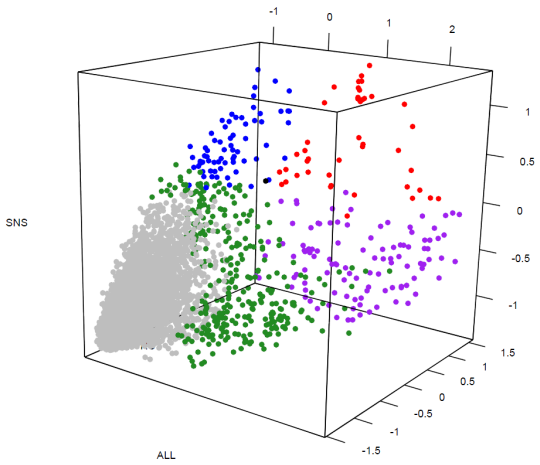

Figure 4.5: **Combining predictions from multiple SVM classifiers generates different levels of cataract prediction confidence.** Each predicted gene appears as a point in this 3D scatter plot. Genes are predicted to be cataract disease genes either by all three models (red), two models (blue and purple), one model (green) or no models (grey)

4.3.5 Lens developmental gene regulatory networks

We generated 5 networks from the collected mouse eye developmental data, split by the time point at which the regulatory relationship was observed: lens induction covered evidence from E8 - E10.5 (Fig. 4.6); primary fiber cell differentiation from E11 - E13.5 (Fig. 4.7); secondary fiber cell differentiation from E14 - E18.5 (Fig. 4.8); postnatal stages (Fig. 4.9); and a complete developmental network containing all the pre-natal evidence (Table 4.1).

Table 4.1: Description of inferred lens GRNs

Stage	Regulators	Targets	Edges	Activating	Inhibitory
Induction	46	73	140	114	26
PFCD	59	109	195	120	75
SFCD	44	188	322	193	129
All Pre-Natal	63	155	409	266	143
Post-Natal	28	167	207	134	73

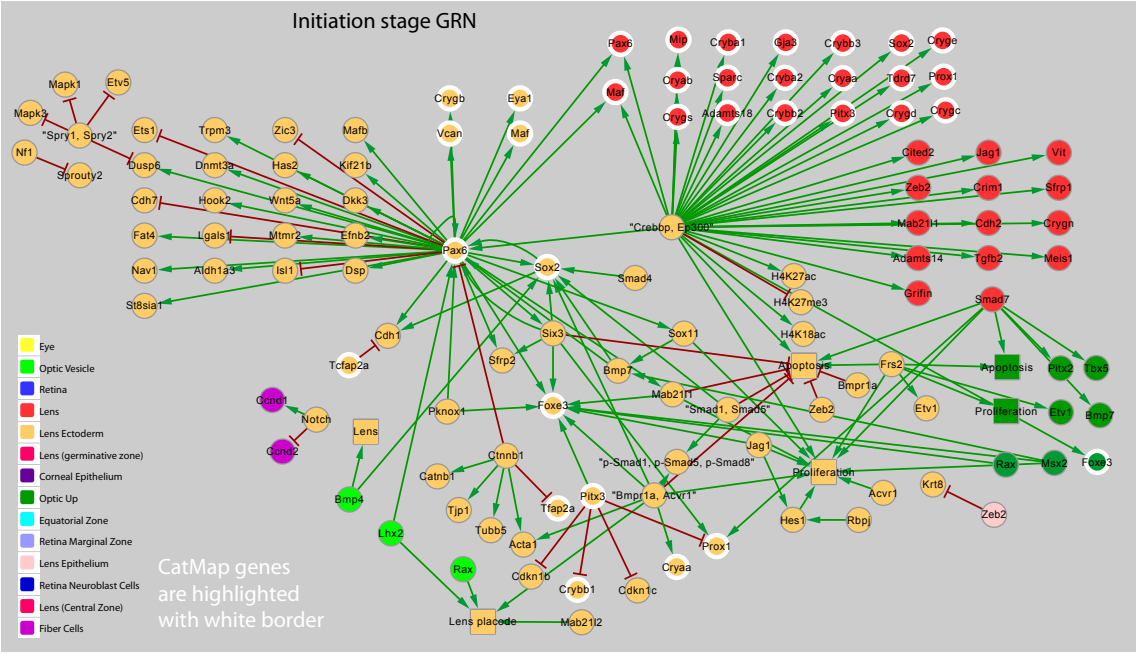

Figure 4.6: **Initiation stages lens gene regulatory network.**

4.3.6 Phenotype Drivers

We performed a network analysis on the complete lens developmental network to identify driving regulators of specific ocular developmental phenotypes. During the literature curation process we recorded the phenotype of mutant animals. We Identified 7 distinct phenotype groupings: no eye; no lens; small eye; small lens; cataract; epithelial defect; fiber cell defect (Fig. 4.10).

We collected the regulatory relationships of the genes in the total lens GRN and submitted them as a regulatory matrix into our MURSS algorithm, where the seed genes were those

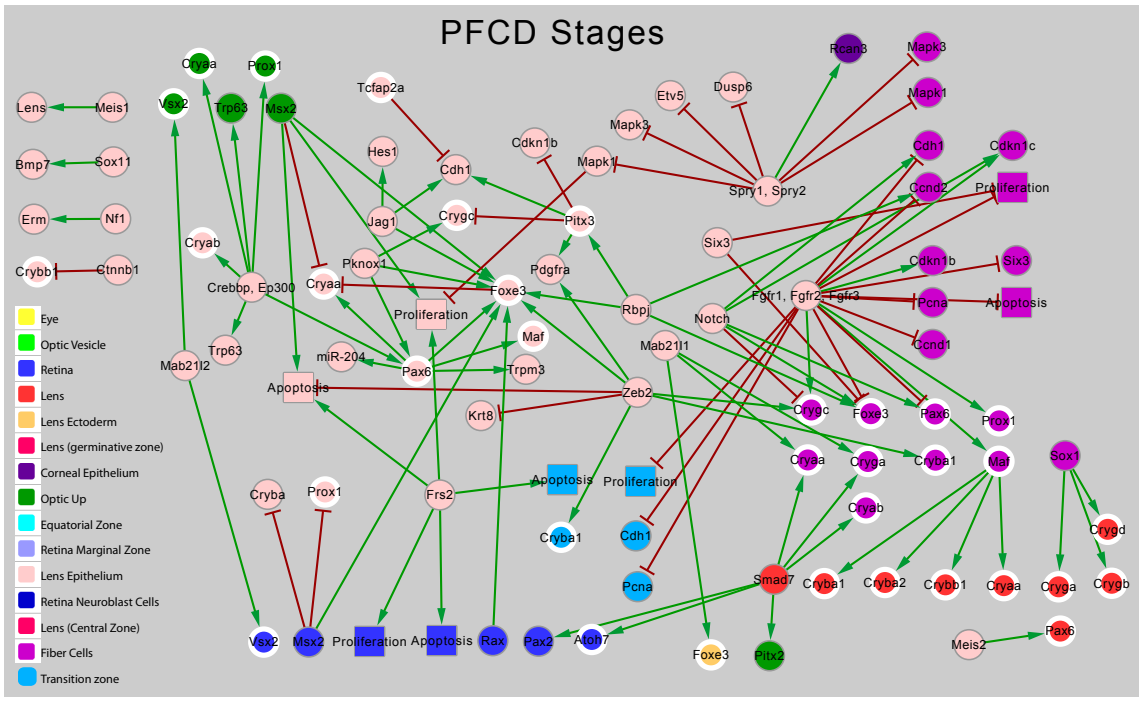

Figure 4.7: **Primary fiber cell differentiation stages lens gene regulatory network.**

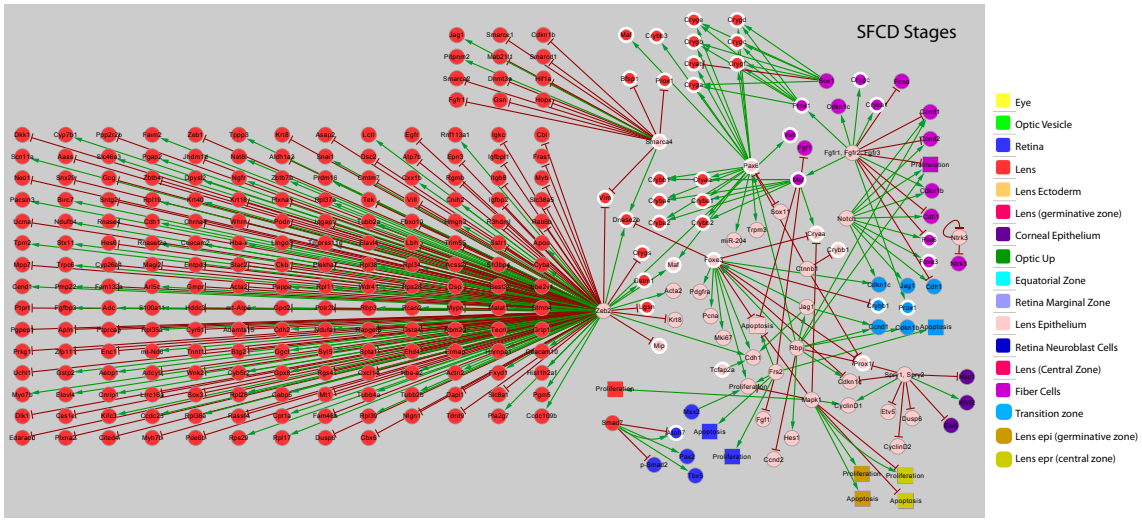

Figure 4.8: **Secondary fiber cell differentiation stages lens gene regulatory network.**

annotated to each phenotype. Calculating P-values for the weighted scores of the minimum

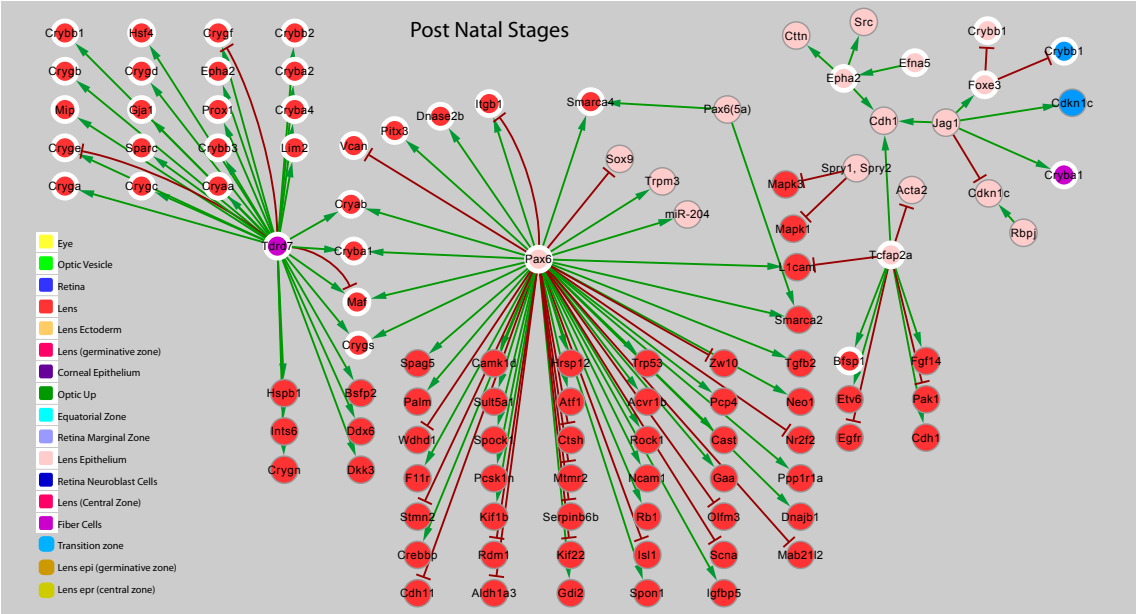

Figure 4.9: **Postnatal stages lens gene regulatory network.**

spanning sets of regulators allowed us to identify those regulators that are most highly significant to a phenotype gene set, relative to the rest of the network (Table 4.2). This analysis revealed that although *Pax6* is an important regulator for all observed phenotypes, this result is expected by chance considering it's prevalence in dominating the inferred network structure. For fiber cell defects, *Notch1* was the most outstanding regulator. For general cataract phenotypes, *Pknox1* was most significant, followed by *Ctnnb1*, double knockout of *Frs2* + *Ptpn11*, and *Notch1*. *Rbpj* stood out as the most important regulator for the no lens phenotype. Several regulators emerged as highly significant to the small lens phenotype, including *Jag1*, *Rbpj* and *Notch 1*, as well as the *Fgf* receptors and *Spry* genes. *Ralbp1* was the most significant result to the small eye phenotype. No regulators emerged as highly significant for the epithelial defect and no eye phenotype genes.

We repeated the MURSS analysis but considered higher levels of upstream regulators as direct regulators of the seed genes, including regulators 2 and 3 edges upstream in the network. In this way we could identify those super-regulators that might be acting through some intermediary direct interactions. *Notch1* remained the most significant regulator for

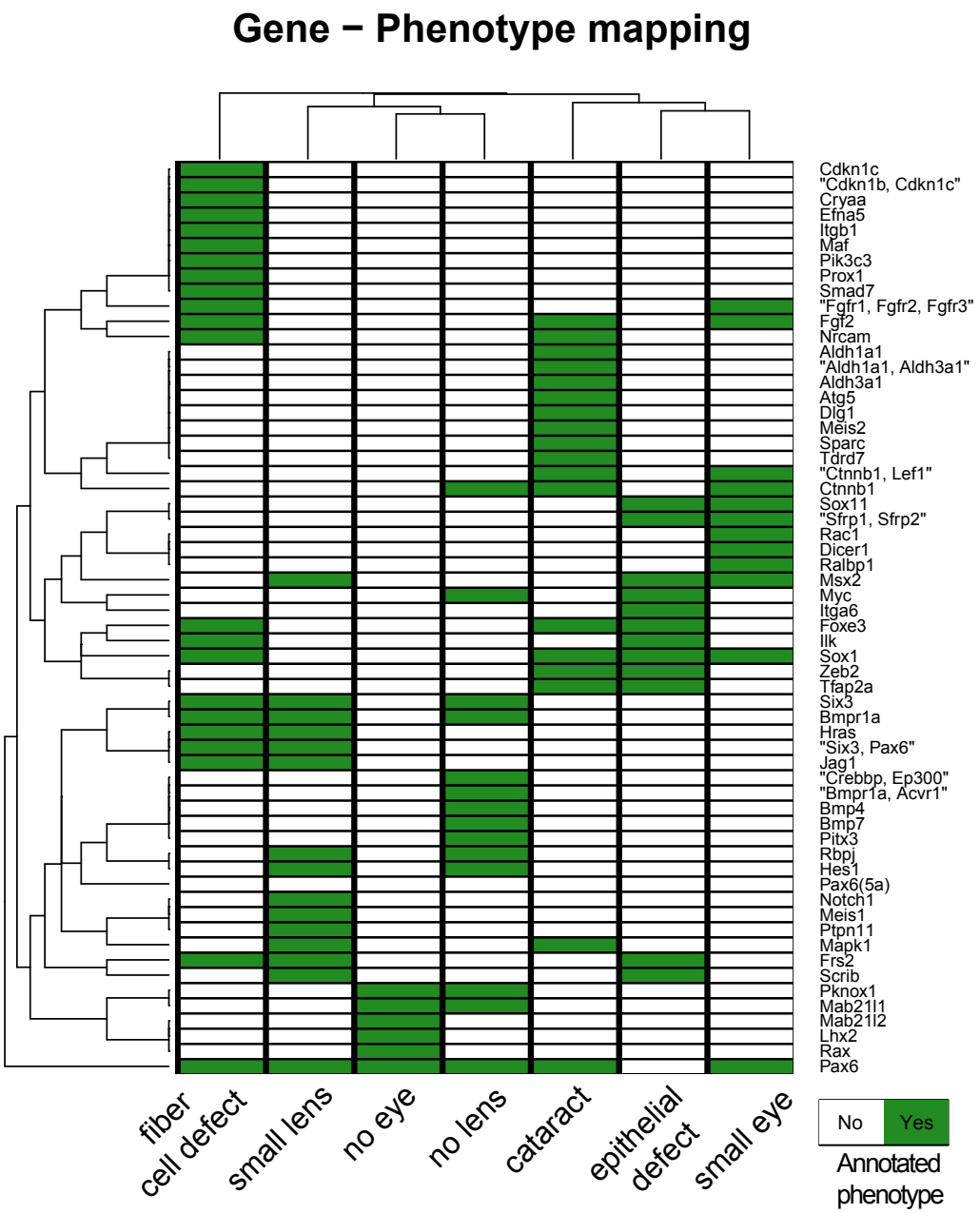

Figure 4.10: **Lens gene - phenotype associations matrix.** Matrix showing which genes are associated with each phenotype based on perturbation studies

fiber cell defects and the small lens phenotype, absorbing the regulatory influences of *Pax6* and *Jag1* respectively. The most significant regulators for cataract phenotypes are lost as

Table 4.2: Significant regulators for lens defect phenotypes based on the MURSS algorithm

Depth	Phenotype	Gene	Score	Pval
1	Cataract	Pknox1	2	0.001
1	Small lens	Jag1	4	0.003
1	Cataract	Ctnnb1	21	0.005
3	Small eye	Ralbp1	6	0.005
2	Fiber	Notch1	8	0.007
3	Fiber	Notch1	8	0.007
1	Cataract	Frs2, Ptpn11	2	0.012
1	Small lens	Rbpj	12	0.012
1	No lens	Rbpj	12	0.013
1	Small lens	Notch1	4	0.013
2	Small lens	Notch1	4	0.013
3	Small lens	Notch1	4	0.013
1	Fiber	Notch1	4	0.014
2	Small eye	Ralbp1	3	0.014
1	Small eye	Ralbp1	2	0.022
1	Cataract	Notch1	2	0.025
1	No lens	Sox11	1	0.028
3	No lens	Rbpj	7	0.029
1	Small lens	Fgfr1, Fgfr2, Fgfr3	4	0.03
2	No lens	Rbpj	6	0.03
2	Fiber	Itgb1	8	0.034
3	Fiber	Itgb1	8	0.034
1	Small lens	Spry1, Spry2	8	0.037
1	No lens	Ctnnb1	6	0.044
1	No lens	Msx2	1	0.044

they or their targets are regulated by *Ctnnb1*. *Rbpj* remains the most important regulator of the no lens phenotype, as *Sox11* and *Msx2* are lost. Interestingly *Pax6* disappeared from most of the lists of minimal regulator sets as longer paths of regulation were considered.

4.4 Discussion and conclusion

In this study we performed a machine learning based integrative analysis using mouse gene expression data and functional information to predict and prioritise congenital cataract causing genes. Our approach was able to prioritise at least 11 known cataract genes that were not in the CatMap database, with many more potential new leads that should

be targets of follow up investigations. Our final models performed reasonably well, and certainly improved over the previous iteration of iSyTE that only used lens specific gene expression information.

We also developed and applied a method for GRN analysis (MURSS) that revealed potential master regulators of disease phenotypes, which could represent therapeutic targets worthy of future investigation. The major limitation of this approach was the incompleteness of the underlying GRN due to the sparse nature of the perturbation data we could collect from the literature.

Chapter 5

GEOracle: classification based on free text annotation in Gene Expression Omnibus

5.1 Issues with automated GEO analyses

The NCBI Gene Expression Omnibus (GEO) is one of the largest public repositories for genome-wide omic data, including mostly transcriptomic data (Barrett *et al.*, 2013). As of March 2017, GEO contains over 79,000 data series (GSE), consisting of over 1.6 million individual gene expression samples (GSM). This database harbours biological insights that are not apparent when studying each data set individually (Rung and Brazma, 2013). Several packages are available to programmatically access GEO data, including GEOquery (Davis and Meltzer, 2007), GEOmetadb (Zhu *et al.*, 2008), compendiumdb (Nandal *et al.*, 2016) and shinyGEO (Dumas *et al.*, 2016), allowing keyword based search and download of GSE and GSM, with few standard analysis options.

One major challenge in effectively reusing public gene expression data is the availability of good quality metadata. The need for standardisation of metadata is the reason for

the development of the Minimum Information About a Microarray Experiment (MIAME) standard (Brazma *et al.*, 2001), and more recently the MINSEQE standards for sequencing data (Rung and Brazma, 2013). While some fields in GEO metadata use controlled vocabularies (*e.g.*, species name, gene symbols), the majority of the metadata appears as free text, describing the context of samples (*e.g.*, tissue type or developmental stage) and the experimental design (*e.g.*, perturbation experiment). Although this free text is often readily interpretable by humans, there is no simple means to process this information from GEO in an automated fashion. Ultimately this imposes a major limitation on effectively re-using the huge amount of public data in GEO (Rung and Brazma, 2013). While we believe it is important to push for the use of standard annotations, we nonetheless wish to reuse the large amount of data that exists in GEO.

A gene expression experiment can typically be classified based on its experimental design (*e.g.*, perturbation, time-series and case-control experiments). In many cases, data sets from perturbation experiments (*e.g.*, gene knock-out, signalling stimulation, or physical stimulation) are valuable because they allow us to identify the set of genes that are causally downstream of the perturbation agent. This has important applications in determining signalling pathway targets and regulatory networks (Djordjevic *et al.*, 2014; Parikh *et al.*, 2010; Schubert *et al.*, 2016; Xiao *et al.*, 2015). There are tens of thousands of perturbation studies in GEO, likely containing millions of experimentally determined perturbation data. Nonetheless, currently there is no simple way to determine whether a GSE contains perturbation data. Furthermore, even when a GSE is known to contain perturbation data, it is not trivial to automatically match the treatment samples with their respective control samples since a single GSE may contains multiple treatment and control groups. Only around 5% of GSM have additional fields of metadata denoting comparison groups, while the rest appear as individual samples containing only free-text metadata within a single 'characteristics' field. As no systematic evaluation has been performed of the challenges presented by unstandardised free-text metadata in GEO, we do not know to what extent each step of this process can be automated or otherwise sped-up using computational approaches.

In light of these challenges and lack of knowledge, we use text mining and machine learning techniques to classify GSE that contain perturbation data, and to identify and match the treatment and control samples in a perturbation data set. We test various approaches on a manually curated set of perturbation experiments and quantify our performance during each step of the process. Text mining of free text metadata has previously been used to identify related experiments through semantic similarity (Galeota and Pelizzola, 2016), and to automatically process large amounts of the GEO database with limited quality control or user oversight (Wang *et al.*, 2016; Zinman *et al.*, 2013). Using our R Shiny tool called GEOracle, we can quickly annotate many perturbation experiments from GEO in a semi-automated fashion with full user control. GEOracle then performs differential expression analysis to identify gene targets of the perturbation agent.

5.2 Implementation

The GEOracle workflow follows the same steps a bioinformatician would employ when analysing perturbation data on GEO (Fig. 5.1).

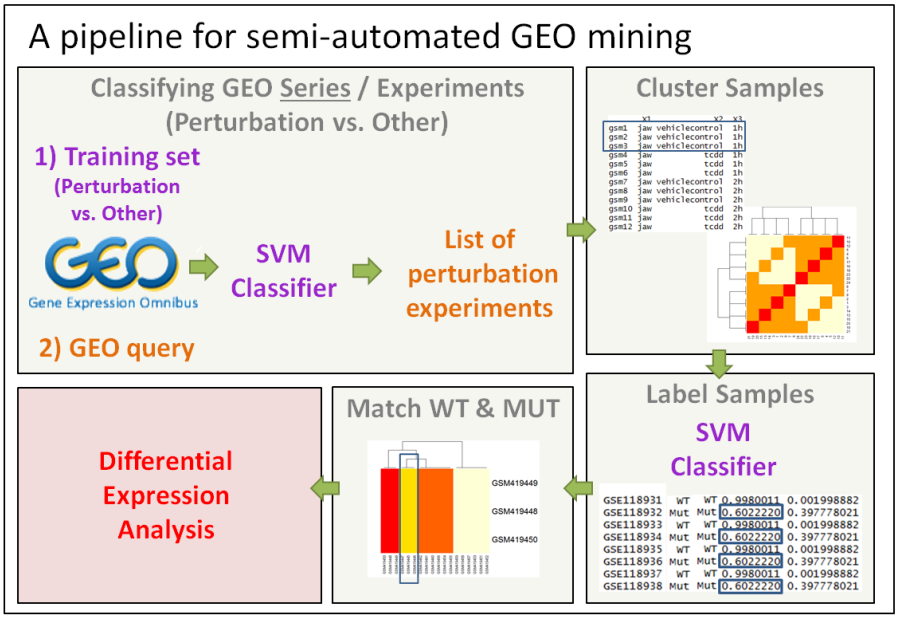

Figure 5.1: **The GEOracle workflow** follows the same steps a bioinformatician would employ when analysing perturbation data on GEO

This begins with identifying whether a GSE is a perturbation experiment. Next comes grouping of replicate samples and identifying the perturbation group relevant for the analysis. Finally the appropriate control group is selected and differential expression analysis is performed. In this section we describe our methodology for performing these steps and evaluate GEOracle's performance on manually curated training and test sets. Given a list of GSE accession numbers, GEOracle begins by extracting their metadata via the R package GEOmetaDB (Zhu *et al.*, 2008).

5.2.1 Classifying perturbation GSE

To build a classifier for identifying perturbation experiments, we manually curated a training set of 277 randomly selected GSE IDs, which we annotated with the experimental design.

Based on 31 manually defined textual features from the free text metadata that can differentiate perturbation experiments (including keywords such as 'knockout', 'KO', 'wildtype', 'WT', 'null', '-/-', 'transgenic', ''TG'), a support vector machine (SVM) classifier was built to predict perturbation GSE. Performance was maximised by the radial basis function kernel (Fig. 5.2). When evaluated on our training set by 100 rounds of 10-fold cross-validation with internal feature selection, our model produced a mean Area Under the Receiver Operating Characteristic curve of 0.89, suggesting high sensitivity and specificity.

5.2.2 Grouping replicate samples

To evaluate our automated grouping of GSM samples and subsequent matching of control and perturbation groups, we manually curated a second set of 73 perturbation GSE. Half of these GSE were chosen from the previous training set (including the particularly difficult GSE) and the other half were randomly selected perturbation GSE. We annotated the 832 constituent GSM samples into 259 groups labelled as 'perturbation' or 'control', and paired the 'perturbation' sample groups with their appropriate 'control' groups.

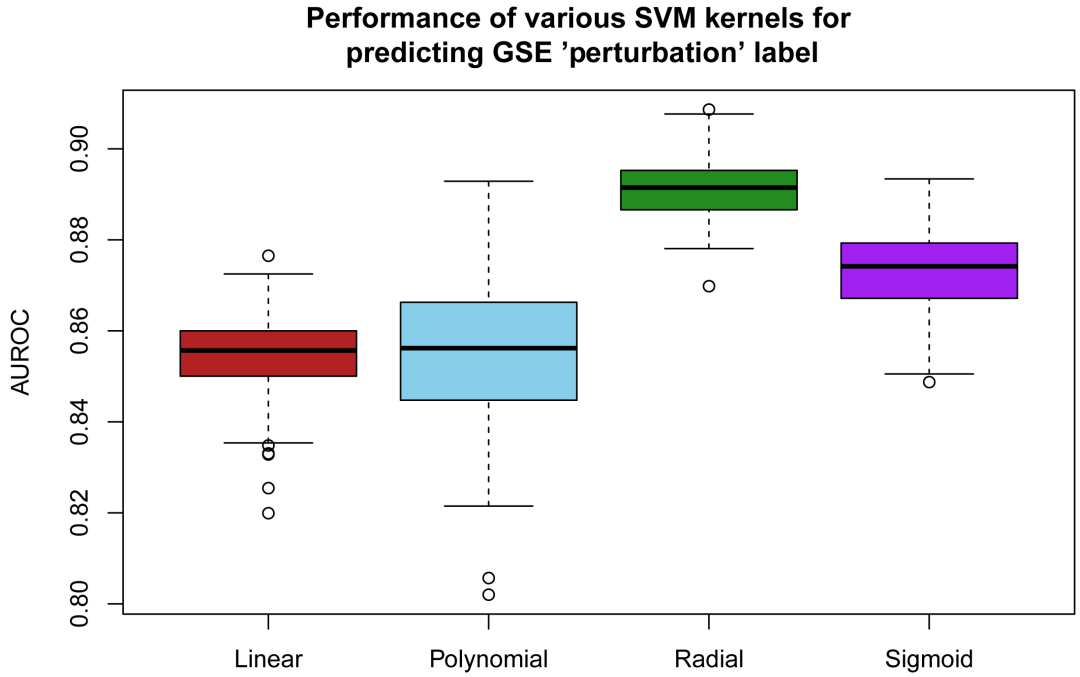

Figure 5.2: **Comparison of the performance of different SVM kernels for predicting GSE 'Perturbation' label** based on the manually curated training set of 277 GSE IDs. Shown are boxplots of the Area Under the Receiver Operating Characteristic (AUROC) curve from 100 repetitions of 10-fold cross-validation.

For each identified perturbation GSE, GEOracle groups replicate samples using the available GSM metadata. Replicates could mean biological or technical replicates that together form a unit of analysis for differential expression. GSM titles are processed via a series of string manipulations to remove replicate identifiers and tokenise the titles. A simple hierarchical clustering approach is used, based on Gower distance between tokenised GSM titles, with the tree cut at height 0, resulting in identical GSM titles being assigned to one cluster. The same approach is applied to GSM characteristics to produce a second clustering of samples. Based on these two sample clusterings, we identify the most valid clustering outcome and assign confidences to the output, removing data-sets with insufficient metadata or invalid clustering results from further analysis (Fig. 5.3).

Our multi stage clustering approach produces a grouping sensitivity of 93.2% at the GSE

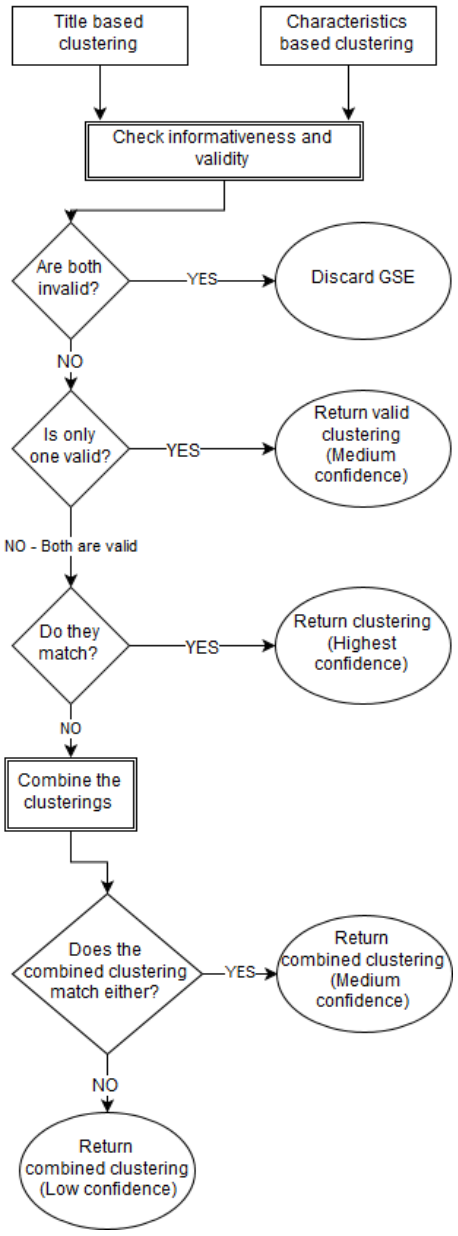

Figure 5.3: **The logic flow for assessing the most valid clustering of GSM samples.** This schematic diagram shows the decision making process during the multi-stage clustering procedure that combines information from the GSM titles and characteristics. Informativeness and validity means that there is more than 1 cluster in the GSE and that there are fewer than N clusters, where N is the number of GSM in the GSE.

level (meaning every sample in a GSE must be correctly grouped for that GSE to be considered a positive result) based on our training set. All incorrectly clustered GSE can be explained by typographical errors and other anomalies in the metadata. This

was an improvement over more naive clustering approaches, based solely on the GSM characteristics, GSM titles, or a simple concatenation of the two, producing sensitivities of 64.4%, 86.3% and 74% respectively (Fig. 5.4). Although samples can often be grouped by either the titles or the characteristics, the process of deciding which information to use is non-trivial. Fig. 5.5 shows a complex example where a simple concatenation of GSM titles with GSM characteristics fails to group samples correctly, while our multi-stage decision process succeeds.

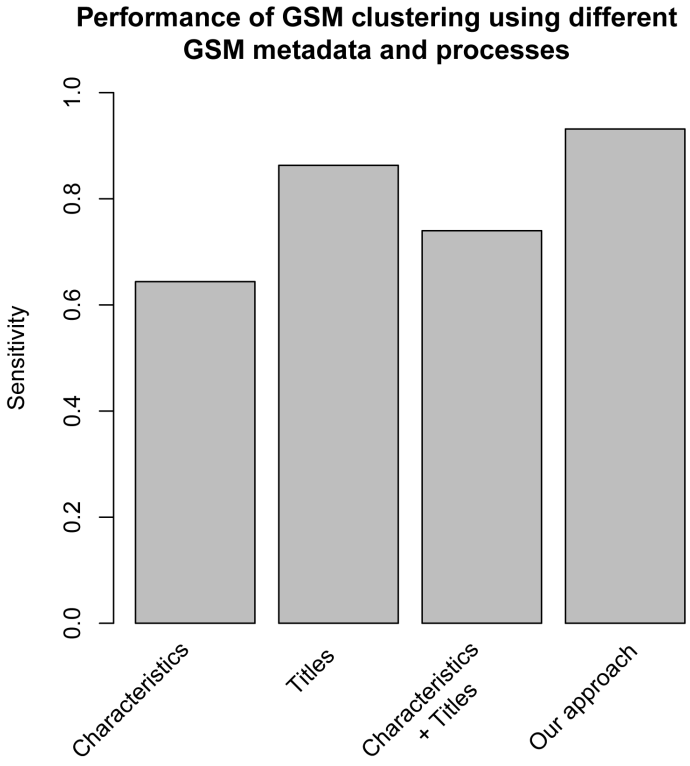

Figure 5.4: **Comparing the performance of clustering using GSM titles and characteristics.** Shown is the relative sensitivity of different clustering methods, using GSM characteristics only, GSM titles only, a simple concatenation of GSM characteristics and titles and our multi-stage clustering approach.

5.2.3 Classifying sample groups

Both the GSM titles and characteristics were analysed for the presence of 33 textual features that represent molecular concepts that can differentiate 'perturbation' from 'control' samples. We trained another SVM classifier to label the groups as 'perturbation' or 'con-

GSM	TITLE	gender	strain	tissue	stage	genotype.variation
GSM1022194	ZJGXZ3_E15-4	male	b6/129sv	embryonic heart	15.5 dpc	zic3 +/y
GSM1022195	ZJGXZ3_E15-6	male	b6/129sv	embryonic heart	15.5 dpc	zic3 +/y
GSM1022196	WT-1	male	b6/129sv	embryonic heart	15.5 dpc	zic3 +/y
GSM1022197	WT-2	male	b6/129sv	embryonic heart	15.5 dpc	zic3 +/y
GSM1022198	WT-5	male	b6/129sv	embryonic heart	15.5 dpc	zic3 +/y
GSM1022199	ZJGXZ3_678-4	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y
GSM1022200	ZJGXZ3_678-5	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y
GSM1022201	ZJGXZ3_628-1	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y
GSM1022202	680-5	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y
GSM1022203	741-3	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y
GSM1022204	741-4	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y
GSM1022205	ZJGXZ3_628-4	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y; sox2-cre
GSM1022206	ZJGXZ3_754-2	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y; sox2-cre
GSM1022207	ZJGXZ3_754-4	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y; sox2-cre
GSM1022208	913-1	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y; sox2-cre
GSM1022209	882-1	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y; sox2-cre
GSM1022210	882-2	male	b6/129sv	embryonic heart	15.5 dpc	zic3 flox/y; sox2-cre
GSM1022211	ZJGXZ3_1191-2	male	b6/129sv	embryonic heart	15.5 dpc	zic3 -/y
GSM1022212	ZJGXZ3_1191-3	male	b6/129sv	embryonic heart	15.5 dpc	zic3 -/y
GSM1022213	ZJGXZ3_1235-8	male	b6/129sv	embryonic heart	15.5 dpc	zic3 -/y
GSM1022214	1322-3	male	b6/129sv	embryonic heart	15.5 dpc	zic3 -/y
GSM1022215	1493-1	male	b6/129sv	embryonic heart	15.5 dpc	zic3 -/y

Figure 5.5: **Sample titles and characteristics from GSE41674.** Title based clustering was not able to correctly cluster this GSE, whereas GEOracle's multistage clustering approach could, by utilising the information in the GSM characteristics.

trol'. We found the linear kernel for the SVM gave the best results (Fig. 5.7). We adjust the predicted labels of some groups when only one label is predicted for all samples in a GSE. A confidence associated with the final outcome of group labelling is determined (Fig. 5.6). We observe a sensitivity of 94.6% for group classification at the GSE level. This is a large improvement over the 73.3% sensitivity produced by the basic approach of choosing the highest scoring label based on the occurrence of the subset of 20 features that unambiguously distinguish between 'perturbation' and 'control' samples.

We examined the features weights from the trained SVM for insights into the relative importance of features (Fig. 5.8). We found that the limited vocabulary used for denoting 'control' samples results in high feature weight for these features, as opposed to low

weights of many different textual features and acronyms that denote different types of 'perturbation' samples. This suggests that searching for the presence of the few 'control' textual features may be able to correctly classify the majority of samples.

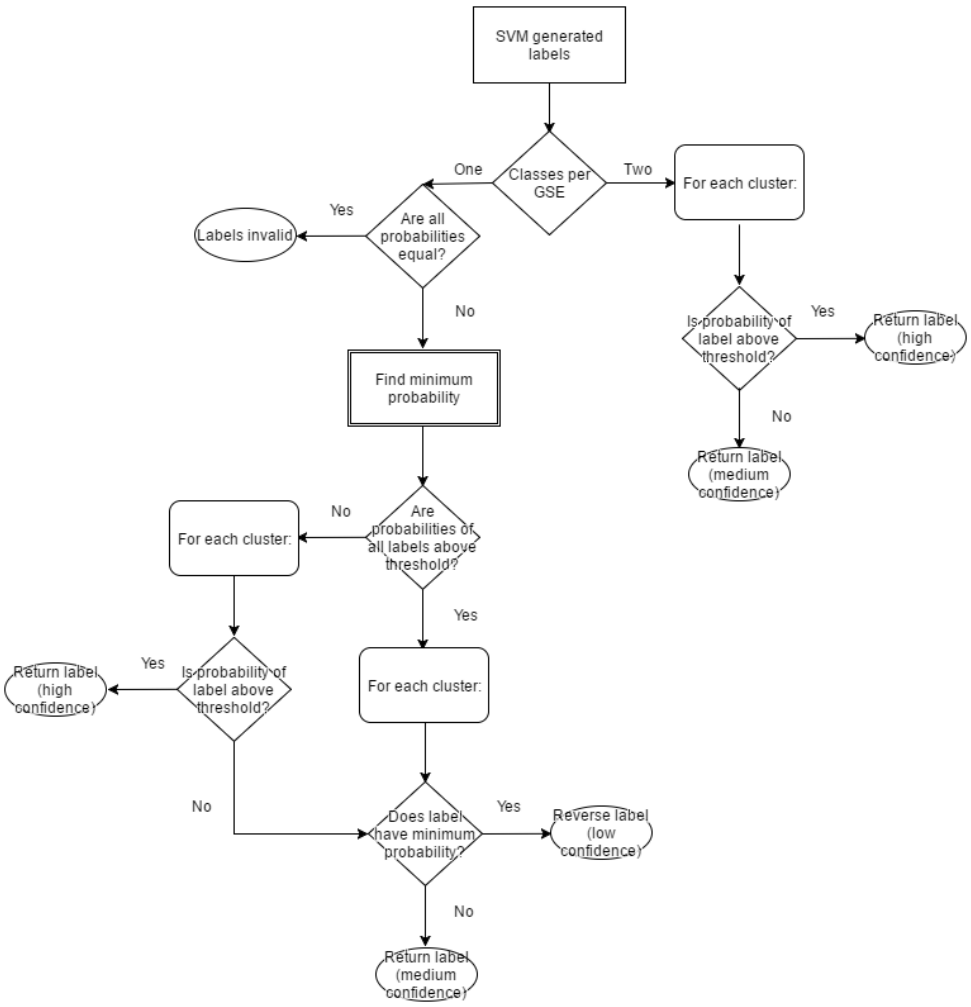

Figure 5.6: **The logic flow for assessing the most valid label for a cluster of GSM.** This schematic diagram shows the decision making process for fixing labels (perturbation or control) predicted by the SVM based on textual features. This process is particularly important when only one cluster label is generated for every cluster in a GSE.

5.2.4 Matching perturbation with control groups

GEOracle matches each predicted 'perturbation' group to the 'control' group with the lowest Gower distance based on the tokenised GSM titles and characteristics, and determines

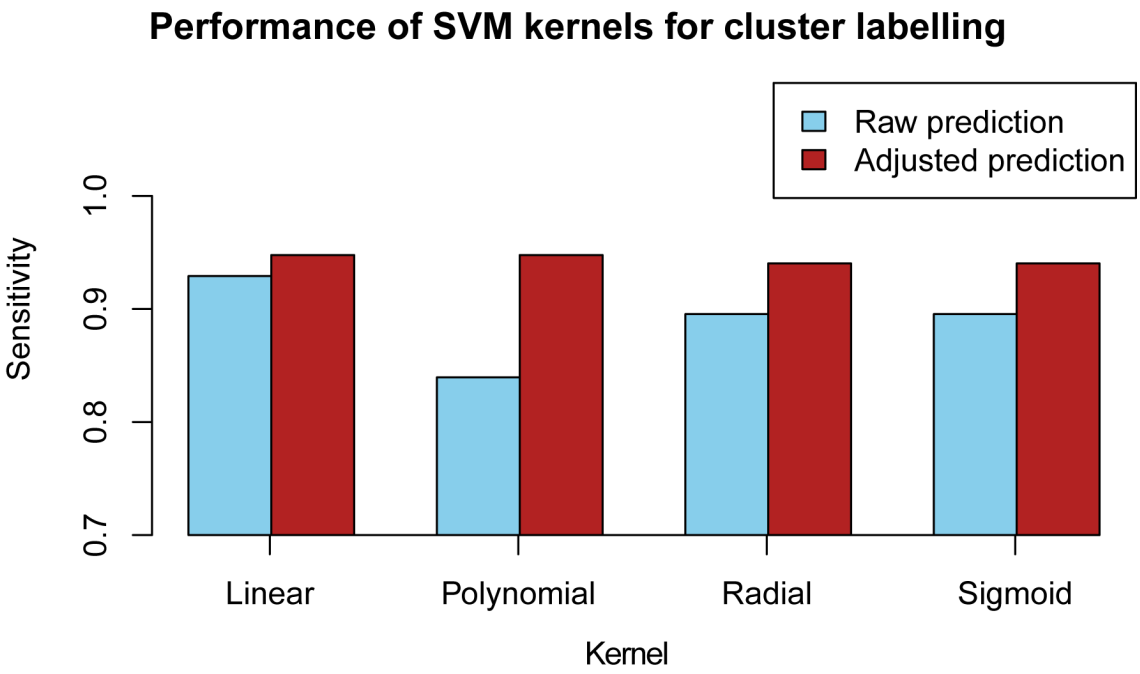

Figure 5.7: **Comparing the performance of different SVM kernels to predict the label of GSM clusters (perturbation vs control).** Sensitivity is calculated as the fraction of GSE for which the GEOracle output perfectly matches the manually annotated set of 73 GSE. Shown is sensitivity calculated on the raw label predictions (blue) and after cluster label adjustment (red).

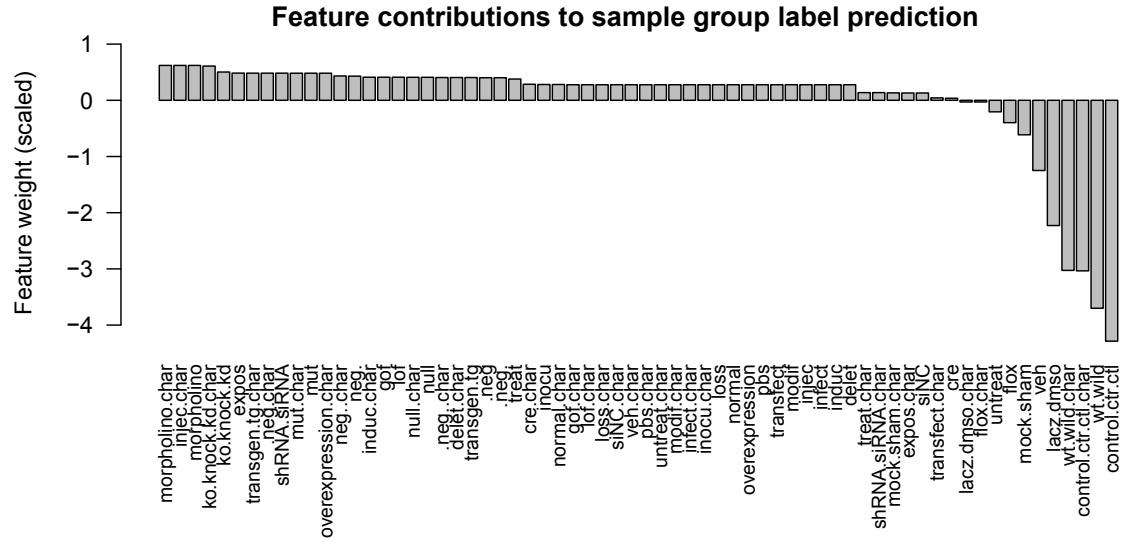

Figure 5.8: **Contribution of textual features to sample group label prediction.** Feature weights were extracted from the SVM model.

the confidence for each of groups (Fig. 5.9). We observe a sensitivity of 83.1% for group matching at the GSE level. Furthermore, we attempt to determine the identity of the perturbation agent and perturbation direction for each group pair by searching for gene names and keywords in the GSM and GSE metadata. The keywords used represent the concepts of addition (i.e. 'overexpress') and removal (i.e. 'knockout') of a perturbation agent. The direction with the most keyword matches becomes the assigned direction.

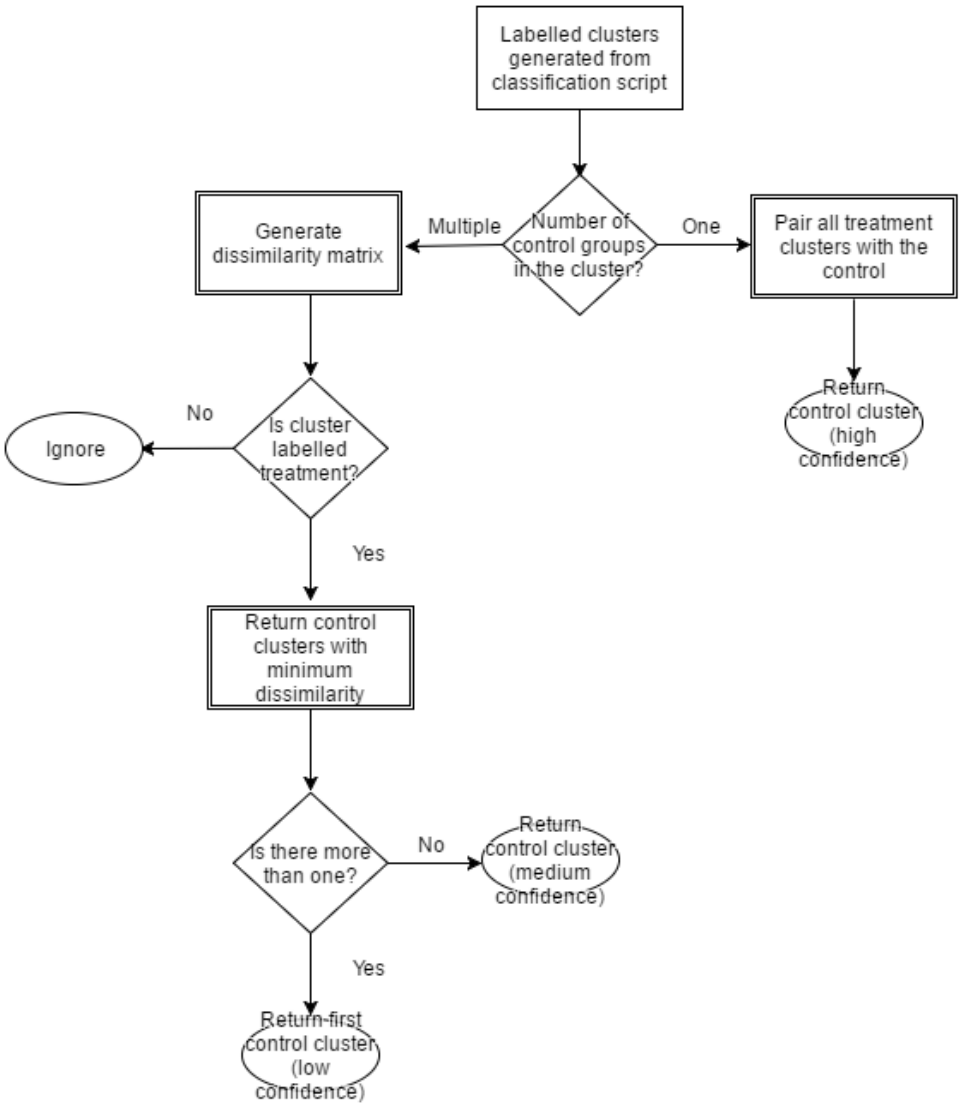

Figure 5.9: **The logic flow for pairing labelled clusters.** This schematic diagram shows the decision making process for matching a perturbation cluster which its closest control cluster. This can be non-trivial when multiple control clusters exist within a GSE. Dissimilarity is Gower distance.

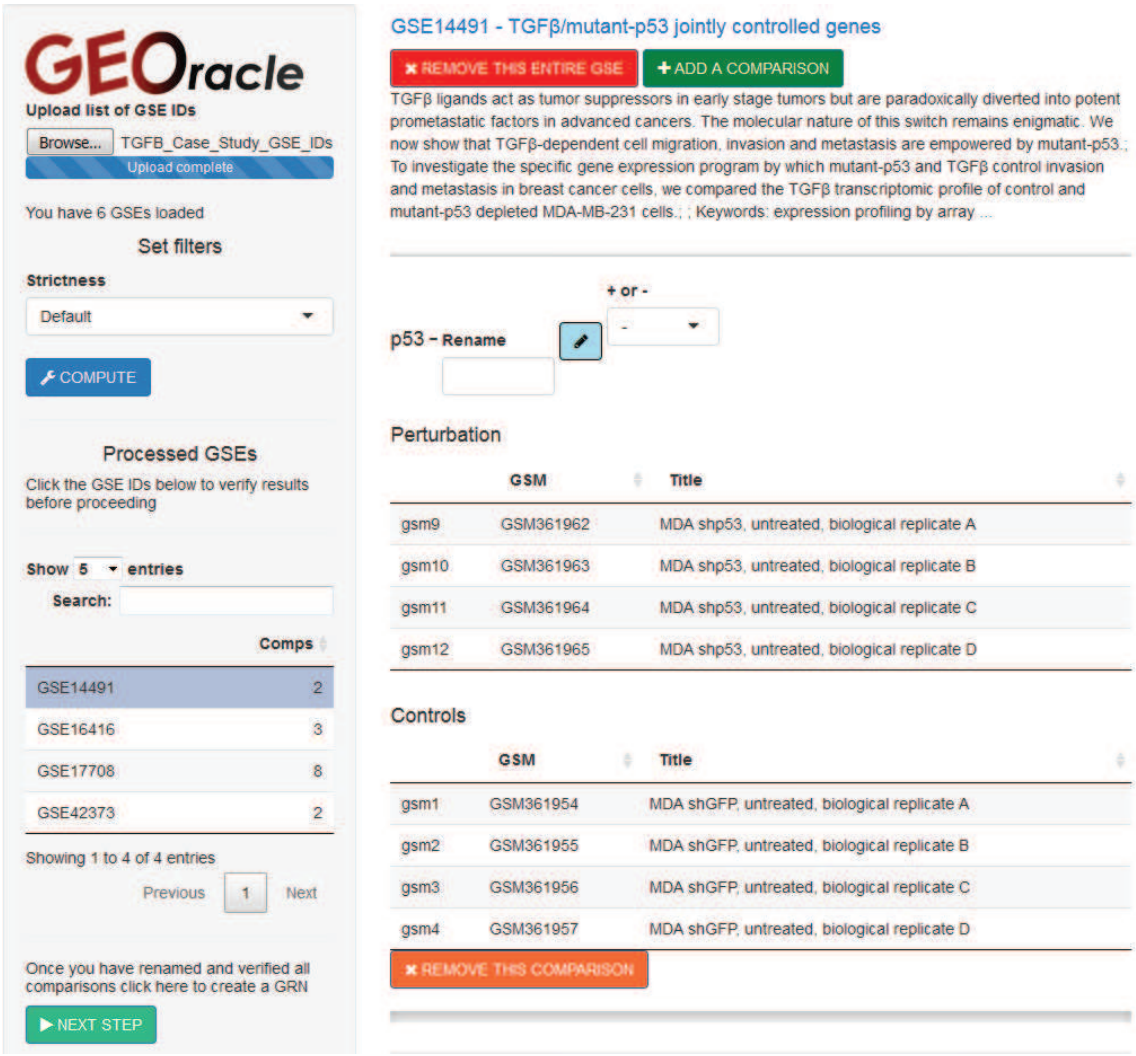

Figure 5.10: **The GEOracle user interface.**

5.2.5 Manual adjustment using the graphical user interface

The GEOracle interface (Fig. 5.10) guides users through the entire process. Importantly the interface allows the user to manually adjust and verify all details of the predicted GSM labels and pairings, and create their own pairings from all GSM within each GSE. This allows the user to be 100% confident in the setup of samples for differential expression analysis.

5.2.6 Differential expression analyses

The paired 'perturbation' and 'control' groups are then used to compute differential gene expression using GEO2R, which implements the limma pipeline (Ritchie *et al.*, 2015). The results can then be downloaded by the user. GEOracle is currently tailored for microarray data analysis as this is the most prevalent data type in GEO, but it can be extended to analyse RNA-seq data or even other functional genomic data sets such as ChIP-seq.

5.3 Case studies

5.3.1 A conserved response to TGFβ stimulation in human cells

We used GEOracle to process six GSE containing TGFβ perturbation experiments and discover the consensus target genes of TGFβ signalling stimulation in human cells. The total time required for classifying the GSE and GSM groups, matching the treatment and control samples, manually verifying the results, downloading the gene expression data from GEO and performing differential expression analysis is less than 12 minutes. This analysis required minimal human intervention and essentially no bioinformatics expertise.

Based on these results we could identify a consensus TGFβ target gene signature in human cells consisting of 82 genes (Fig. 5.11). Many of the observed trainscriptional changes matched the literature about the TGFβ pathway, including increased transcription of *CTGF*, *JUN*, *JUNB* and *WNT5B*, and repression of *TGFBR3*, *FZD7* and *SPRY1*. A GO analysis of the 82 genes from the consensus signature using g:Profiler (Reimand *et al.*, 2007) showed significant enrichment for the term 'response to transforming growth factor beta' (Benjamini-Hochberg (BH) adjusted p-value = 8.93×10^{-08}).

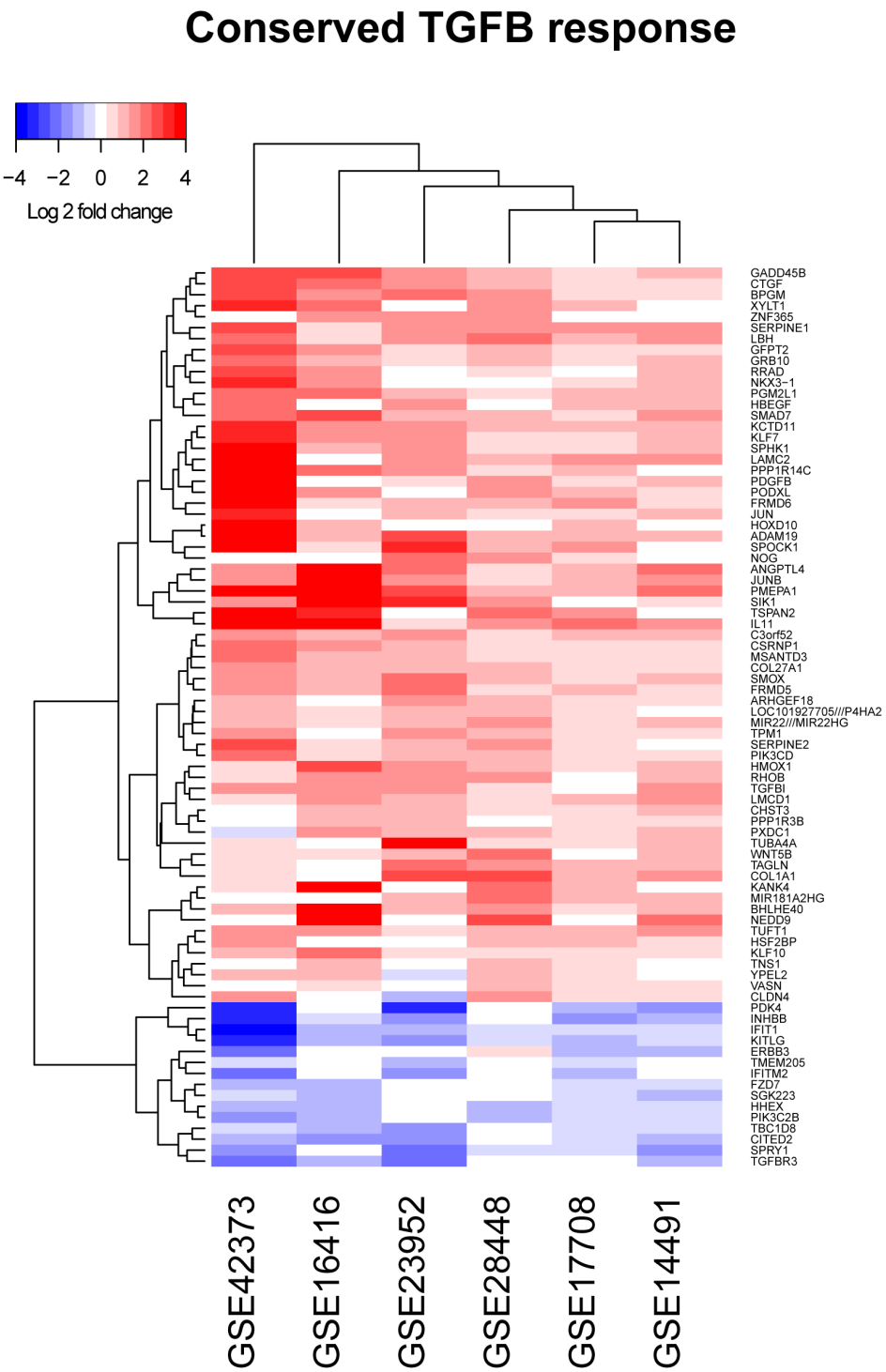

Figure 5.11: **A heat map showing the discovered conserved response to TGFB stimulation in human cells.** This plot is generated in case study 1.

5.3.2 Mouse heart specific perturbation based causal GRN

We further used GEOracle to analyse all perturbation microarray data from mouse cardiac tissues. We searched GEO using the following query: "mus musculus"[Organism] AND ("heart"[MeSH Terms] OR heart[All Fields] OR cardiac[All Fields]) AND ("gse"[Filter] AND "Expression profiling by array"[Filter]). This resulted in 851 GSE.

Processing these 851 GSE though GEOracle, including the most user intensive steps of manually verifying and modifying the predicted GSM sample comparisons and excluding those non-cardiac GSE, required approximately 8 hours of user time, again with essentially no bioinformatics expertise required. 164 relevant GSE were included for further processing. We obtained significantly differentially expressed genes for 87 genetic perturbations (i.e. gene knockdown or over expression) and 10 non-genetic factors (diet, chemicals etc.) using standard thresholds (absolute log 2 fold change > 1 and BH adjusted P value < 0.05). GEOracle automatically outputs significant differentially expressed genes as an edge list for gene regulatory network construction. From the genetic perturbation experiments we constructed a gene regulatory network of 23,347 causal and directed relationships between 9,152 genes (Fig. 5.12). Of these 14,120 were activating relationships and 9,681 were inhibitory. This case study illustrates how we can construct a large organ-specific gene regulatory network from published experimental perturbation data in GEO.

We used the MURSS algorithm to investigate upstream regulators of several published gene sets, including a set of known congenital heart disease (CHD) genes (Blue *et al.*, 2012), and atrial fibrillation genes from GWAS and published familial disease studies (Beck *et al.*, 2014; Tucker and Ellinor, 2014). For CHD, at the first depth level MURSS identified *Tbx20, Tbx1* and *Mesp1* as significant upstream regulators. *Mesp1* is a known CHD gene (Werner *et al.*, 2016). At the second depth level, MURSS identified the *Foxc1/2* double perturbation, as well as *Dhx36* at the second and third depth levels. As a regulator of *Nkx2-5, Dhx36* can exert a lot of influence during heart development and is known to cause heart defects and embryonic lethality when deleted in mice (Nie *et al.*, 2015). In atrial fibrillation, MURSS identified *Tbx3* at depth 2. *Tbx3* is well described as a key

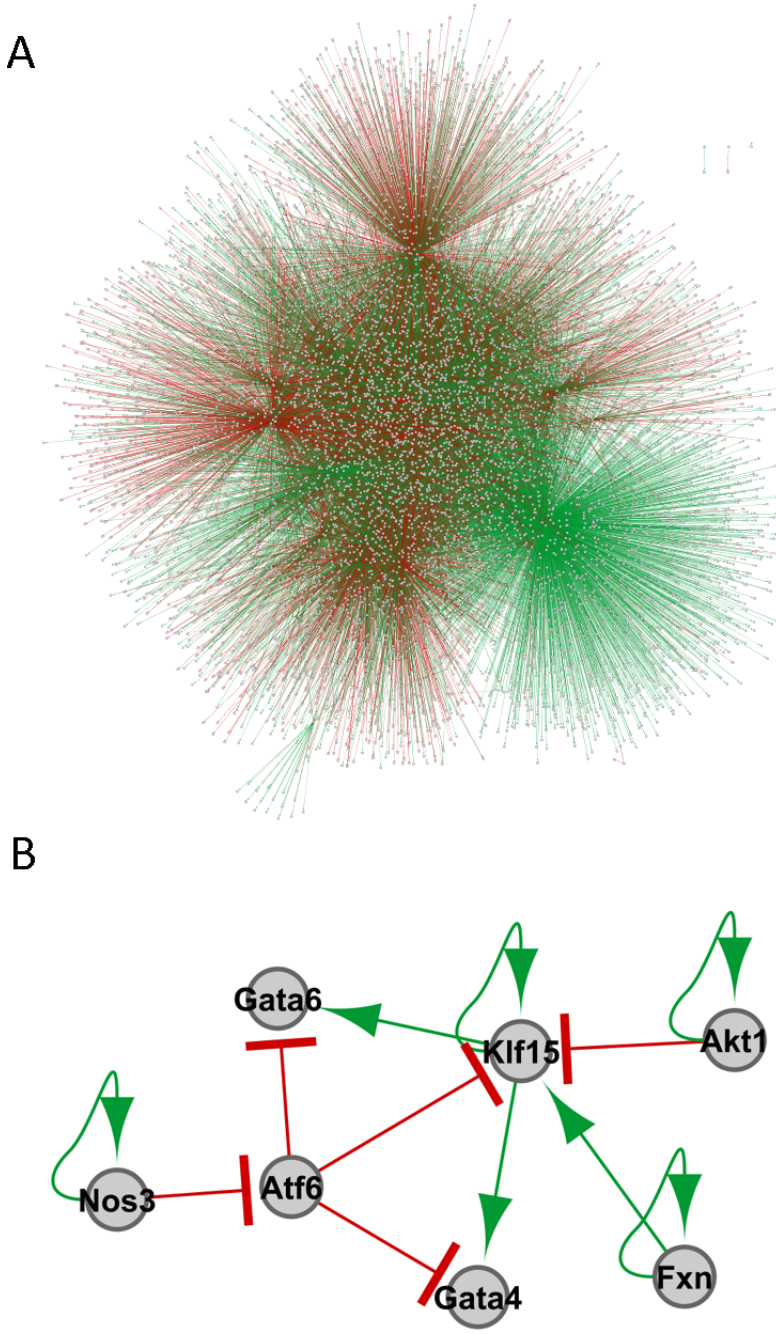

Figure 5.12: **Mouse heart causal gene regulatory network.** A) Overview of the network, showing 23,347 edges between 9152 genes. B) Zoom view, showing the directed and signed (activating vs inhibiting) information contained in a subset of the network.

gene in atrial development and assigns pacemaker function to the heart, making it a clear risk gene for human atrial fibrillation (Hoogaars *et al.*, 2007).

Chapter 6

XGSA: A statistical method for cross-species gene set analysis

Gene set analysis (GSA) is a powerful tool for determining whether an experimentally derived set of genes is statistically significantly enriched for genes in other pre-defined gene sets, such as known pathways, gene ontology terms, or other experimentally derived gene sets.

One central assumption in GSA is that the gene sets being compared are subsets of the same set of genes, in practice meaning from the same species. With an increasing variety of genetic resources available in evolutionarily diverse model and non-model organisms, there is an increasing interest in utilising these cross-species gene set resources in GSA. Cross-species GSA is the process of determining whether a gene set in one species overlaps with a gene set from another species more than would be expected by chance. This is an important problem in the emerging field of comparative transcriptomics, which aim to integrate knowledge on regulation of biological pathways across the tree of life (Roux *et al.*, 2015). For example, consider the regeneration of organs and appendages, an ability present in several diverse vertebrate organisms but apparently missing from humans.

As the consistent gene set universe assumption fails when more than one species is involved,

it becomes increasingly problematic as the number of many-to-many homologues increase between evolutionarily distant species. Several largely *ad hoc* methods have been proposed and used in the literature and have become the standard analysis options. The naïve cross-species mapping approach is to apply an 'at least one homolog' function to map a gene set from one species to another. This approach is the most common method for homology mapping in general and is used by the majority of existing cross-species gene set analysis web based platforms, including g:Profiler, Gene Weaver and GSGator (Baker *et al.*, 2012; Kang *et al.*, 2014; Reimand *et al.*, 2007).

When performing comparative analyses between evolutionarily distant species, many researchers remove the increased complexity from homology assignment by applying the BLAST best reciprocal hits (BRH) approach (Britto *et al.*, 2012; Gohin *et al.*, 2010; Labbé *et al.*, 2012). BRH reduces complexity by restricting homology assignments to at most one per gene, choosing the best hits for each gene (highest sequence similarity or lowest E value) and only assigning homology if the two genes are each others best hits. This implies the assumption that the best hit is the only valuable hit, which is particularly problematic when there are multiple closely scored hits in one gene family. For distantly related organisms a large amount of non 1 - 1 homology information is discarded before any analysis is done, reducing the potential insight that can be gained.

Another alternative approach to reduce complexity is to perform significance testing at the level of gene family, also called an orthologous group (OG) (Kristiansson *et al.*, 2013; Rittschof *et al.*, 2014; Zheng *et al.*, 2011). In this approach, the entire OG is assigned a representative value summarising the constituent genes (often the normalised minimum p-value), discarding homology information after this assignment. Traditional statistical enrichment tests are then applied at the level of the OG. The OG structure between a large selection of species can be retrieved from databases such as eggNOG, OrthoDB and InParanoid (Kriventseva *et al.*, 2015; Powell *et al.*, 2014; Sonnhammer and Ostlund, 2015). One strength of the OG framework is the ability to test gene sets from more than two species simultaneously. Nonetheless, similar to the BRH approach, one major limitation of this approach is the loss of information regarding the signal from individual genes in

the same OG, and that the exact gene responsible for the final result can be unknown, making interpretation and validation challenging.

Another approach is to computationally transfer the functional annotations (based on protein domains for example) to a less studied organism from well studied ones, as facilitated by PANTHER (Mi *et al.*, 2016). This annotation transfer reduces confidence in annotation quality and relies on the assumption that the relationship between the protein domain and functional annotation is known and true, which limits it's utility to molecular function annotation as opposed to more general biological pathways. Several other studies harness the strength of sequence information in microarray probes to transfer information between species (Le *et al.*, 2010; Lu *et al.*, 2009; Xie *et al.*, 2011). While these and other approaches lend more confidence and resolution than simple ID mapping, they do not create a general and principled cross-species gene set analysis framework that specifically addresses complex homology (Lu *et al.*, 2010; Yang *et al.*, 2014).

To our knowledge, there has not been any systematic investigation on the issues of cross-species GSA. An approach that utilises the full and complex homology structure between two species is not available. In this study we discuss the statistical issues associated with cross-species gene set analyses and define an informative homology complexity score. We show that the naïve implementation of homology mapping followed by Fisher's exact test can lead to false positive discovery. To alleviate this bias, we propose a straightforward statistical test, called XGSA, to perform cross-species GSA by considering the complete homology structure between two species. Our simulations show that XGSA can indeed remove the false positive bias, while maintaining good statistical power when analysing gene sets with complex homology structure. We apply XGSA to two real biological applications that involve comparing gene sets from distantly related organisms.

6.1 Methods

6.1.1 Performing cross-species gene set analysis

Problem definition

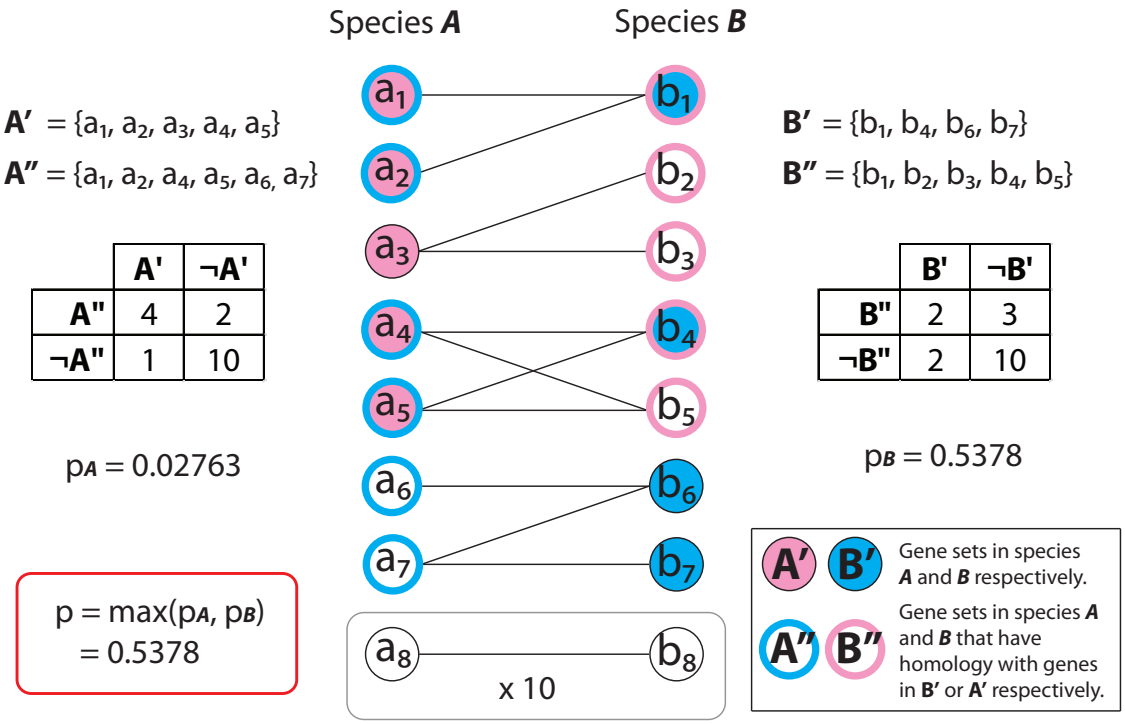

Figure 6.1: A schematic diagram illustrating the XGSA method. Nodes represent genes in species A and B respectively, with edges representing homology, and shading and outlines representing gene set membership. The grey box represents the remainder of the homologous universe of 1 - 1 relationships not assigned to either gene set. The tables are contingency tables describing the observed overlap of the homologous gene sets in the two species. p_A and p_B show the different p-values derived from performing Fisher's exact test in each species. The red box indicates the final value p produced by XGSA.

Let $A = \{a_1, a_2, ..., a_l\}$ and $B = \{b_1, b_2, ..., b_k\}$ denote the set of all homologous genes (the homologous gene universes) in two species A and B, respectively. We further define subsets A' and B' as the *gene set of interest* in species A and B respectively, where $A' \subseteq A$ and $B' \subseteq B$.

Let there be a *homology mapping function*, $m(a, b)$, that describes the sequence homology

relationship between any gene a in species A and any gene b in species B:

$$m(a,b) = \begin{cases} 1, & \text{if } a \text{ and } b \text{ are homologous} \\ 0, & \text{otherwise.} \end{cases}$$

Given gene sets of interest A' and B', we can further define their homologous partners in the other species as $B'' = \{b \in B : m(a,b) = 1, \exists a \in A'\}$ and $A'' = \{a \in A : m(a,b) = 1, \exists b \in B'\}$

The **Cross-species Gene Set Analysis** problem can be defined as a hypothesis test where the null hypothesis H_μ is that the membership of A' and A'' are independent and B' and B'' are independent (Fig. 6.1).

XGSA

We calculate the probability p_A of co-membership of A' and A'' equal to or greater than the observed co-membership if H_μ is true, using the hypergeometric distribution,

$$p_A = \sum_{k=|A' \cap A''|}^{min(|A'|,|A''|)} \frac{\binom{|A'|}{k}\binom{|A_u|-|A'|}{|A''|-k}}{\binom{|A_u|}{|A''|}}$$

where A_u is the gene universe in A that has homology to the gene universe in species B, $A_u = \{a \in A : m(a,b) = 1, \exists b \in B\}$. This is equivalent to an upper tail Fisher's exact test. Similarly, we compute the probability p_B for observing the co-membership of B' and B'' if H_μ is true,

$$p_B = \sum_{k=|B' \cap B''|}^{min(|B'|,|B''|)} \frac{\binom{|B'|}{k}\binom{|B_u|-|B'|}{|B''|-k}}{\binom{|B_u|}{|B''|}}.$$

We calculate a statistic p to estimate the probability of H_μ being true, as the maximum of p_A and p_B,

$$p = max(p_A, p_B).$$

We take the maximum in order to reduce the false positive rate caused by complex homology, as illustrated in (Fig. 6.1).

Naïve approach

We compared the performance of XGSA with other *ad hoc* approaches for cross-species GSA. The naïve approach is equivalent to doing the above test in only one of the species, *e.g.* species A. In this case, the p-value is the same as p_A.

Best reciprocal hits (BRH)

The best reciprocal hits approach only differs from the naïve approach in that it reduces the complexity first. We created a subset of homology mappings for which the retained human and zebrafish homology mappings were each others highest scoring partners, based on sequence similarity percentages from Ensembl.

Orthologous group (OG)

We downloaded OG annotations from OrthoDB with no filtering applied. We mapped genes to OGs and calculated p_A at the OG level.

6.1.2 Automatically identifying homology between species using Ensembl BioMart

Following standard practice (Reimand *et al.*, 2007; Yates *et al.*, 2016), we accessed Ensembl BioMart programmatically through the R package *biomaRt* (Durinck *et al.*, 2009) and retrieved homology mapping between Ensembl gene ids in two species. We turned this mapping into a sparse matrix using the R package *Matrix*.

6.1.3 Homology complexity score

We define a measure of complexity for a gene set in one species with respect to its homology mapping to another species (A and B), as the fraction of genes in the gene set GS_A in

species A which have more than one homologue in species B,

$$Complexity(GS_A, B) = \frac{\left|a \in GS_A : \sum_{b \in B} m(a,b) > 1\right|}{|GS_A|}.$$

6.1.4 Statistical power analysis

For each human gene ontology (GO) term we find all of the zebrafish homologues for that GO term. Intuitively, when gene sets devoid of any homologous genes are tested in a cross-species gene set enrichment test, the p-value for that GO term should be 1 (no match). Alternatively, when the entire set of homologues is tested, the p-value should be close to zero (perfect match). Based on this logic, if we incrementally add homologous genes to the gene set enrichment test, the p-value should decrease. We can then interpret the rate at which several methods reached significance as an indicator of their relative power for that cross species gene set enrichment test.

We start with a zebrafish gene set consisting of the same number of non-homologous genes as there are zebrafish homologues to the chosen human GO gene set. We incrementally substitute each non-homologous gene in the zebrafish set with a homologous gene, and perform enrichment testing after each substitution.

6.1.5 Data preprocessing for the vertebrate regeneration case study

We downloaded four spinal cord regeneration data sets from three species, zebrafish (*Danio rerio*), lizard (*Anolis carolinensis*) and Western clawed frog (*Xenopus tropicalis*). We reprocessed the zebrafish and frog results from the raw microarray data because the lists of DE genes were not available in the original papers. All processing was done in R using the *limma* package and custom scripts unless otherwise noted. Benjamini-Hochberg multiple hypothesis testing correction was applied in each case.

Zebrafish 1 - Hui et al. (2014)

Raw Agilent microarray data were downloaded from GEO (accession GSE39295), corrected for background effects (offset = 16), log-transformed and quantile normalised. Probes with an average expression less than 8 were removed as 'not present' probes after visual inspection of probe intensity distribution. Differential expression at each post injury time-point compared to time zero control was computed to match the published study design. We applied an absolute T-statistic threshold of 7 resulting in 404 significantly differentially expressed genes with Ensembl gene IDs across the 5 time-points.

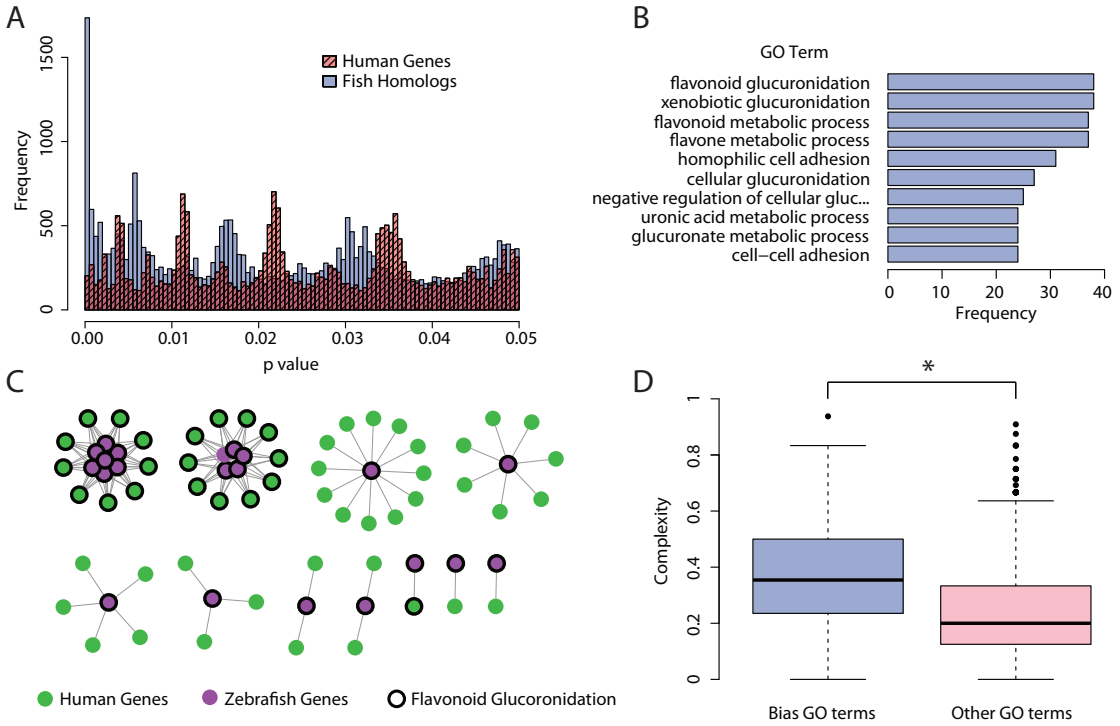

Figure 6.2: Identification of bias in naïve cross-species gene set analysis. A) *P*-values from human GO analysis of 1,000 randomly selected human genes and 1,000 randomly selected zebrafish genes naïvely mapped to their human homologues (Note: the apparent periodicity is an artefact of TopGO). B) Frequency of the 10 most common false-positive GO terms from 100 repeats of the experiment. C) Human and zebrafish homology relationships for genes assigned to the GO term 'flavonoid glucoronidation'. D) Complexity scores of false positive GO terms vs. all other GO terms.

Zebrafish 2 - Guo et al. (2014)

Preprocessing of raw data (accession GSE20460) as above. Differential expression at 4 hours and 12 hours post injury compared to matched sham time-points was computed to match the published study design, although we ommitted the 264 hour time-point due to poor data quality. We applied an absolute T-statistic threshold of 4 resulting in 62 significantly differentially expressed genes across the 2 time-points.

Frog - Love et al. (2011)

Raw Affymetrix CEL files were downloaded from Array Express (accession E-MEXP-2420) corrected for background effects, log-transformed and quantile normalised using the RMA method. Probes with an average expression less than 6 were removed as 'not present' probes after visual inspection of probe intensity distribution. Differential expression followed a time series design with 6 hour post amputation (PA) vs 0 hour control, 24 hour PA vs. 6 hour PA, and 60 hour PA vs. 24 hour PA, to match the published study design. We applied an absolute T-statistic threshold of 4 resulting in 666 significantly differentially expressed genes across the 3 time-points.

Lizard - Hutchins et al. (2014)

We retrieved the differentially expressed gene lists from the supplementary files of the published study.

6.1.6 Data sources for the mouse heart perturbation case study

We constructed 8 sets of genes that have been linked to human heart development and disease in published studies. Two heart failure gene sets were the top 1000 genes correlated with pulmonary arterial pressure and *Natriuretic Peptide B* in human myocardium from patients with failing hearts (Min *et al.*, 2010). Two modules of 174 and 679 atrial

fibrillation associated genes were extracted from another recent study (Tan *et al.*, 2013). Another two atrial fibrillation related gene sets were extracted from a comparison of fibrillating atrium to normal sinus rhythm atrium, consisted of 196 down-regulated genes and 61 up-regulated genes (Barth, 2005). Early heart development genes were determined from the top 1000 up and down regulated genes in iPS induced cardiomyoctyes relative to mature human heart (Babiarz *et al.*, 2012).

6.2 Results

6.2.1 Human and Zebrafish gene sets exhibit a broad range of complex homology

We chose two model organisms with well annotated genomes, *Homo sapiens* (human) and *Danio rerio* (zebrafish), retrieving homology mappings between 15,908 human Ensembl gene IDs and 18,777 zebrafish Ensembl gene IDs. Henceforth the term 'genes' refers to Ensembl gene IDs. 4179 human genes map to more than one zebrafish gene, and 2218 zebrafish genes map to more than one human gene, corresponding to 26.3% and 11.8% of the respective homologous genomes.

We constructed a 'best reciprocal hits' (BRH) subset of homology mappings (see methods), henceforth referred to as the BRH set. The BRH set has 1807 fewer human genes and 4493 fewer zebrafish genes than the complete set, corresponding to a reduction of 11.4% and 23.9% of the respective homologous genomes.

We calculated human-zebrafish complexity scores (see methods) for each gene set in the gene ontology (GO). We observe a wide range of complexity occurs in GO.

6.2.2 Naïve cross-species GSA approach results in a systematic bias

When a random selection of 1,000 human genes is tested against the human GO using the Fisher's exact test as implemented in TopGO (Alexa and Rahnenfuhrer, 2010), there are no significant results passing the significance threshold after multiple testing correction, and a relatively uniform distribution of p-values is observed, as expected (Fig. 6.2A, red bars). The same is true when 1000 zebrafish genes are tested against the zebrafish GO, and these results were consistent for 100 different random selections of genes.

In contrast, when all the human homologues of 1000 randomly selected zebrafish genes were tested against the human GO (using the homologous gene universe as a background), a very strong enrichment of small p-values is observed (Fig. 6.2A, blue bars). An enrichment of small p-values passing multiple hypothesis testing thresholds can be interpreted as evidence for a strong signal in the data. Considering that the original selection of genes was entirely random, this indicates that these significant p-values are false positives.

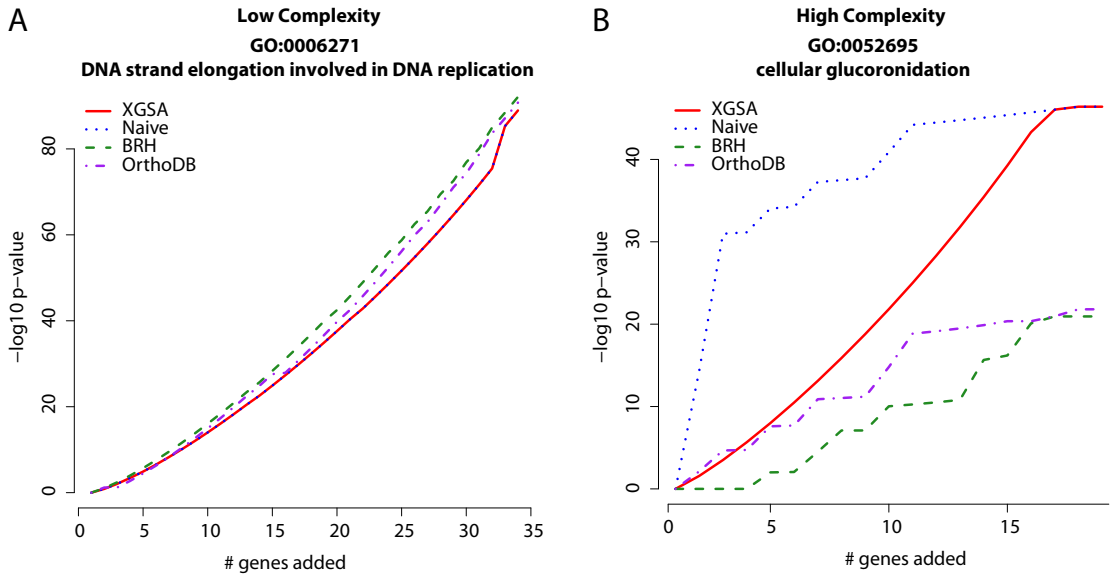

Figure 6.3: Simulations show an increased power of XGSA for high complexity gene sets. A) Low complexity gene sets show comparable performance, with all methods demonstrating a smooth near linear power curve as expected. B) With high complexity gene sets the difference in the methods becomes obvious. OrthoDB represents the orthologous group implementation using the OrthoDB database

Repeating this virtual assay 100 times reveals that several GO terms appear repeatedly in the list of enriched gene sets, with the most recurrent gene set 'flavonoid glucoronidation' enriched in 38% of trials (Fig. 6.2B). This shows a systematic bias leading to false positive results when using Fisher's exact test, introduced by naïve homology mapping from zebrafish to human genes. Importantly this bias is species-dependent; as different pairs of species show different biased GO terms. For example, mapping 1000 random genes from *Xenopus tropicalis* to *Mus musculus* and performing GSA results in a very different set of biased GO terms, including 'sensory perception of chemical stimulus' in 65% of virtual assays. This indicates that the bias may result from the complex homology mapping between two species. When we look at the genetic homology between genes annotated with the GO term 'flavonoid glucoronidation' we see several striking examples of complex homology (Fig. 6.2C). Comparing the complexity scores of the repeated biased GO terms vs. all other GO terms shows that the biased GO terms have a significantly higher gene set complexity on average (two-sided t-test, p-value = 1.156e-07) (Fig. 6.2D). When we use the same sets of randomly selected zebrafish genes but map them to human genes using the BRH homology mapping, the bias disappears (data not shown). Taken together, these findings provide evidence that the cause of the bias is the introduction of complex homology mapping into the testing framework without compensation.

6.2.3 XGSA alleviates the bias in the naïve method

Using a toy example of the cross-species testing problem, we observe that the directionality of the complex homology mapping creates the bias (Fig. 6.1). Our solution involves performing testing in both species / directions, and combining the results. This means that both species act as the host for a Fisher's exact test, with the test set being naïvely mapped from the gene set in the other species. We then return the maximum p-value of the pair of tests. We call this approach XGSA (see methods). Intuitively this means that the gene set overlap must be significant in both species – that is, in both directions of testing (Fig. 6.1). In this way we reduce the effect of complex homology on the resulting p-value. When we applied our method to the same 100 repetitions of 1000 randomly selected

zebrafish genes we saw that the systematic bias disappears – zero out of 100 repetitions had any significantly enriched human GO terms. By accounting for the effects of complex homology in our statistical testing framework we can remove the bias while still utilizing the full complex homology structure.

6.2.4 Simulation studies shows XGSA maintains good statistical power even when analysing gene sets with complex homology

As this problem has not been studied in depth before and no gold standard exists against which we can evaluate our method, we devised a novel testing approach to compare the power of different methods (see methods). Briefly, after choosing a human GO gene set, we incrementally replace zebrafish genes that are not homologous to the GO set with genes that are homologous and calculate significance using different cross-species gene set enrichment testing approaches. Based on the assumption that zero homologous genes should return a p-value of 1 and all homologous genes a p-value close to zero, we can compare the rate at which the p-value decreases as genes are replaced (Fig. 6.3).

We found that for low complexity gene sets, the four methods of naïve mapping, BRH, OG and XGSA perform comparably with no practical difference at commonly used thresholds (Fig. 6.3A). However when testing higher complexity gene sets the power of XGSA becomes clearer (Fig. 6.3B). By retaining the full homology structure XGSA continues to gain power from genes assigned to complex gene families, as opposed to BRH and OG in which the power curve plateaus when complex genes are added. The over-sensitivity of the naïve method to high complexity gene sets can be observed as abrupt rises in the curve above the diagonal. In contrast, XGSA maintains a near linear diagonal power curve as with low complexity gene sets.

We can summarise these curves by measuring the relative area under them (Fig. 6.4). We find that for zero complexity gene sets all methods perform similarly with small differences due to various gene universe sizes - XGSA receives a lower score because it uses the most extensive gene universe. ODB then quickly drops in detection power as complexity

increases in the tested gene sets and plateaus are introduced to the power curve. As gene set complexity increases the advantage of XGSA over both ODB and BRH becomes clear. While it seems that XGSA may be too sensitive as complexity increases, this is because the zero p-value saturates earlier in large and complex gene sets, causing the power curve to change shape and the AUC to increase.

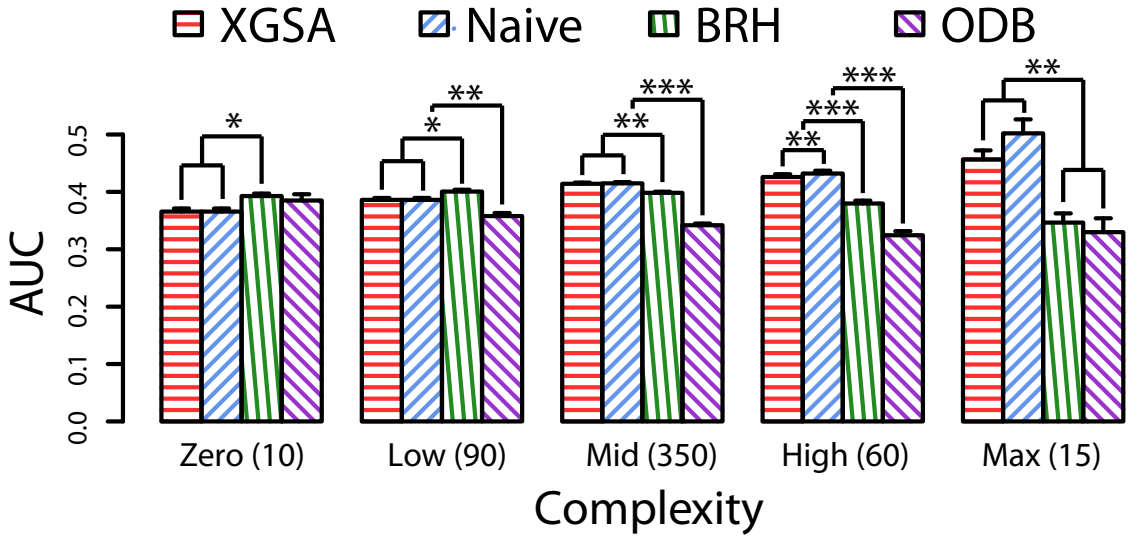

Figure 6.4: Area under the curve of different methods performance during the GO simulation study. Error bars indicate the standard deviation around the mean. In parentheses on the X-axis labels is indicated the number of GO terms in each complexity bracket. The apparent improved sensitivity of the naïve method for maximum complexity gene sets is in fact the detection of false positives as shown in Fig. 6.2 . $* = p < 0.01$, $** = p < 1 \times 10^{-4}$ and $*** = p < 1 \times 10^{-8}$ by two sided t-test

6.2.5 Case study 1: Discovering conserved pathways in social challenge in evolutionarily distant organisms

Ritschoff et al. (2014) studied the transcriptomic changes associated with social challenge in three species: stickleback fish (*Gasterosteus aculeatus*), mouse (*Mus musculus*) and honey bee (*Apis mellifera*). They performed a ranked GSA (Sartor *et al.*, 2009) on their DE genes for each species by assigning GO membership based on protein domain (sequence) information using PANTHER. They also performed a cross-species analysis using the homologous triplet OG approach by harnessing the OrthoDB database, and used the mouse GO as the reference gene sets. We downloaded their lists of DE genes from each

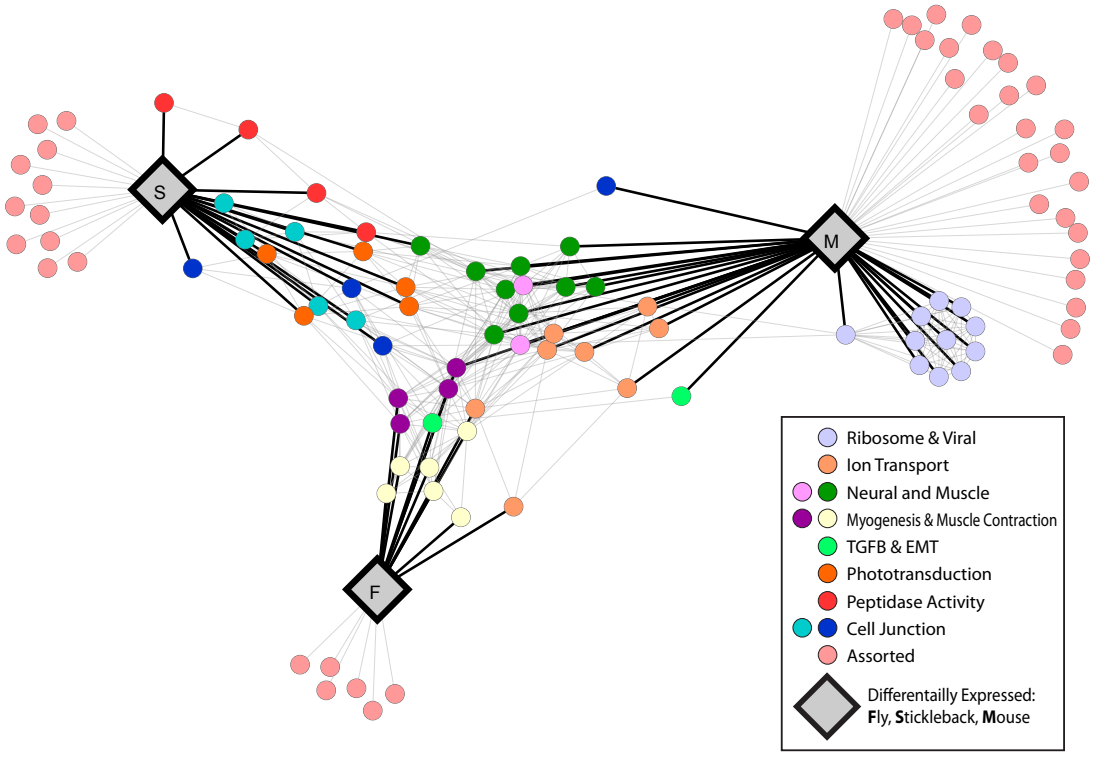

Figure 6.5: Cross-Species gene set analysis of transcriptional response to social challenge. Molecular concept map showing the results of the XGSA pipeline, nodes represent gene sets and edges represent a significant overlap between gene sets. Solid lines represent edges between experimental gene sets and clusters of reference GO / pathway gene sets. Light gray lines represent within cluster similarity of reference gene sets, and overlap between experimental gene sets with outlier reference gene sets. Network is drawn in Cytoscape.

species and sought to recreate their analysis using XGSA. Because the honey bee *Apis mellifera* is not yet included in the Ensembl BioMart homology database, we mapped the 182 honey bee genes to 153 fly (*Drosophila melanogaster*) genes using OrthoDB as suggested by FlyBase (Attrill *et al.*, 2016), and used the fly genes to continue the analysis.

We visualise our results as a molecular concept map (MCM) (Rhodes *et al.*, 2007) – a network diagram where each node represents a gene set and each edge represents a significant overlap between gene sets (Fig. 6.5). Unlike Ritschoff et al., we are directly comparing the experimental gene sets from all three species against the standard mouse GO terms which allows us to interpret them in a single MCM. This approach is different from the approach used by Ritschoff et al. where they used computationally inferred GO membership for each species.

As with the original study we found very little in the way of shared significant gene sets between two or more species when comparing the tests performed for each individual species. However, gene set similarity clustering shows that several gene set categories span multiple species, including ion transport and regulation of neuronal and muscle activity, particularly between fly and mouse. We also see a mouse specific viral and ribosomal response, as well as a fish specific phototransduction response. Furthermore we included KEGG pathways into the MCM analysis, allowing us to identify interesting and relevant pathways for social challenge such as Long Term Depression and Long Term Potentiation, and the only gene set significant in two species (mouse and fly), Dilated Cardiomyopathy.

We found that 39%, 33% and 12% of our significant GO gene sets overlapped with Ritschoff et al. results in mouse, stickleback and honey bee, respectively. Furthermore, 21% of our total significant GO gene sets were significant in the Ritschoff et al. homologous triplet OG analysis, including representative gene sets spanning their major result categories.

Our comparison with the study by Ritschoff et al. raises several issues. As a limitation, our analysis used the DE gene sets as opposed to the ranked list used with GSEA in the original study. As we also do not know the GO terms universe or term – gene assignment used in that study, we cannot declare how closely our results matched theirs. That our results, although retrieving fewer and different GO terms, spanned their species-specific and homologous triplet categories indicates that we recreated many of their key findings. Mapping from honey bee to fly was clearly not ideal and so it is not surprising the fairly low correspondence of significant gene sets for that species. An alternative is to create homology mappings from honey bee to stickleback fish and mouse using BLAST, to enable direct comparisons.

6.2.6 Case study 2: XGSA reveals conserved molecular pathways in vertebrate organ regeneration

Many vertebrates display the ability to regenerate entire appendages, but humans or other common mammalian animal models have very limited capacity to regenerate. With the

availability of whole genome sequences and functional genetic technologies for reptilian and amphibian species with significant regenerative capacity, genome-wide comparative studies of gene expression dynamics during organ regeneration are now possible. Lizards, which are amniote vertebrates like humans, are able to lose and regenerate a functional tail with regrowth and patterning of cartilage, muscle, vasculature, spinal cord, and skin (Hutchins *et al.*, 2014). In addition to the lizard, tadpoles of the African clawed frog, *Xenopus laevis*, are also capable of regenerate their tails and fins, and there are extensive genomic resources available for this model. One important task in regenerative biology is to identify molecular pathways that are conserved in multiple regenerative vertebrates during organ regeneration.

Here we performed a case study to investigate spinal cord regeneration across three species for which transcriptomic profiling of regenerating spinal cord tissues was available; zebrafish (Guo *et al.*, 2011; Hui *et al.*, 2014), lizard (Hutchins *et al.*, 2014) and frog (Love *et al.*, 2011). We sought to explore the biology captured in these data sets by leveraging the extensive gene sets available for human and zebrafish in GO, MSigDB and SPEED (Subramanian *et al.*, 2005). In total our analysis included 97,079 XGSA tests between 2804 gene sets which took 30 minutes on a single core and resulted in 175 significant overlaps. We analysed the results using an MCM (Fig. 6.6).

Zebrafish and tadpole regeneration gene sets show a direct overlap between them as well as many shared enriched gene sets, including *TNFa* and *E2F* signalling, cell cycle, DNA repair and oocyte maturation signals. The lizard gene sets are more isolated from the other two, which is itself not surprising due to their different experimental designs and tissues being profiled. We found that Lizard and tadpole share endothelial to mesenchymal transition and extra cellular matrix assembly related signals. In the base of the lizards regenerating tail we see a very strong enrichment of muscle-related gene sets, likely due to the dominance of this tissue in this regenerative region.

When we compared our gene sets against the MSigDB perturbation gene sets, we find a gene set related to human carious teeth that overlap significantly (adjusted p-value $<$ 0.05) with all three species. The pulpal tissue of human carious teeth has been reported

to be a source of active multipotent mesenchymal stem cells and may represent a tissue with limited regenerative capacity in human (Rajendran *et al.*, 2013). When focusing on the TF targets and immune gene sets, we found the motif for *SRF* (a known regeneration stimulant (Stern *et al.*, 2013)) is enriched in DE genes from both lizard and zebrafish, and that there is a conserved immune response in zebrafish and tadpole.

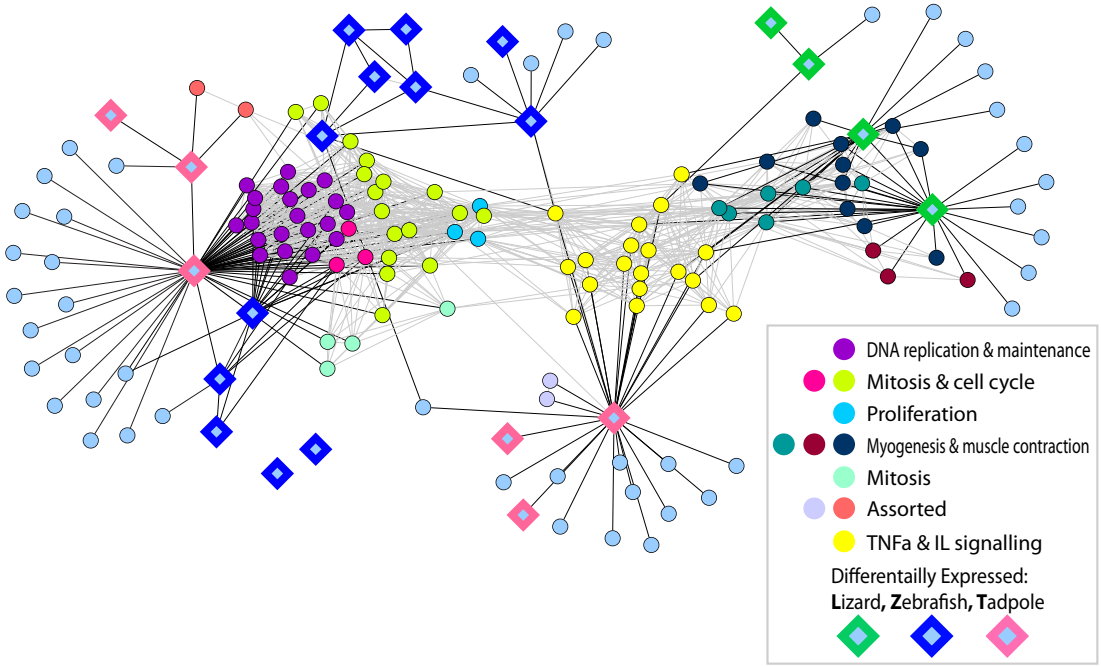

Figure 6.6: Molecular concept map (MCM) showing the overview of the cross-species spinal cord regeneration gene set analysis, nodes represent gene sets and edges respresent a significant overlap between gene sets. Solid lines represent edges between experimental gene sets and reference GO / pathway gene sets. Light gray lines represent similarity between reference gene sets. Network is drawn in Cytoscape.

We further looked at which genes are commonly DE between multiple species. Aurora Kinase A (*AURKA*), which plays a crucial role in spindle assembly, was DE in lizard and zebrafish regeneration experiments, and is known to be required for regeneration in mouse (Pérez de Castro *et al.*, 2013). Furthermore, *AURKB* was DE in the tadpole regeneration experiment, suggesting an evolutionarily conserved role for Aurora kinases in regeneration. Another gene of interest is Keratin 19 (*KRT19*), a marker of hepatic stem cells, endothelial mesenchymal transition and TGFβ signalling. *KRT19* was DE in both lizard and tadpole regeneration experiments, with other keratins being DE in zebrafish. Thirteen more DE genes were conserved between zebrafish and tadpole regeneration, including *PLK1* which is

required for cardiac regeneration in zebrafish (Jopling *et al.*, 2010), *KIF23* which controls G2/M arrest and is also DE in axolotl limb regeneration, *SOCS3* which has been shown to suppress optic nerve regeneration in mice (Smith *et al.*, 2009), and several hepatocyte regeneration markers (*KIF20a*, *MCM4* and *LIG1*).

6.2.7 Case study 3: Mouse heart perturbation target gene sets

We also used XGSA to compare the downstream targets of each mouse genetic perturbation from the previous chapter of this thesis, to human gene sets that are either cardio-developmentally active or implicated in adult heart disease phenotypes (Min *et al.*, 2010). Our results visualized as an MCM, confirm knowledge such as *Gata4* and *Tbx5* being central to early cardiac development, and are also risk genes for atrial fibrillation along with *Tbx3* (Fig. 6.7).

Several interesting new candidates also emerge. *Atp2b4*, a relatively under described gene which codes for a calcium pump, shares many connections with *Tbx5* and *Gata4*, including being implicated in atrial fibrillation and early cardiac development. *Atp2b4* has recently been identified as a mediator of cardiac hypertrophy (Mohamed *et al.*, 2016). Another clear result was the affiliation of *Klf15* with heart failure genes correlated with pulmonary arterial pressure and Natriuretic Peptide B (Fig. 6.8). Indeed *Klf15* deletion in mice has previously been found to lead to heart failure (Haldar *et al.*, 2010). We also found a connection between *Dmpk* and atrial fibrillation, which was also supported by the literature (Berul *et al.*, 1999). A promising novel candidate as an atrial fibrillation gene was Frataxin, implicated in Friedrich's ataxia, a syndrome with well described heart conduction defects.

6.3 Discussion

The main contributions of this work are: (1) formulation of the cross-species gene set analysis problem, (2) investigation of the statistical bias that may arise when comparing

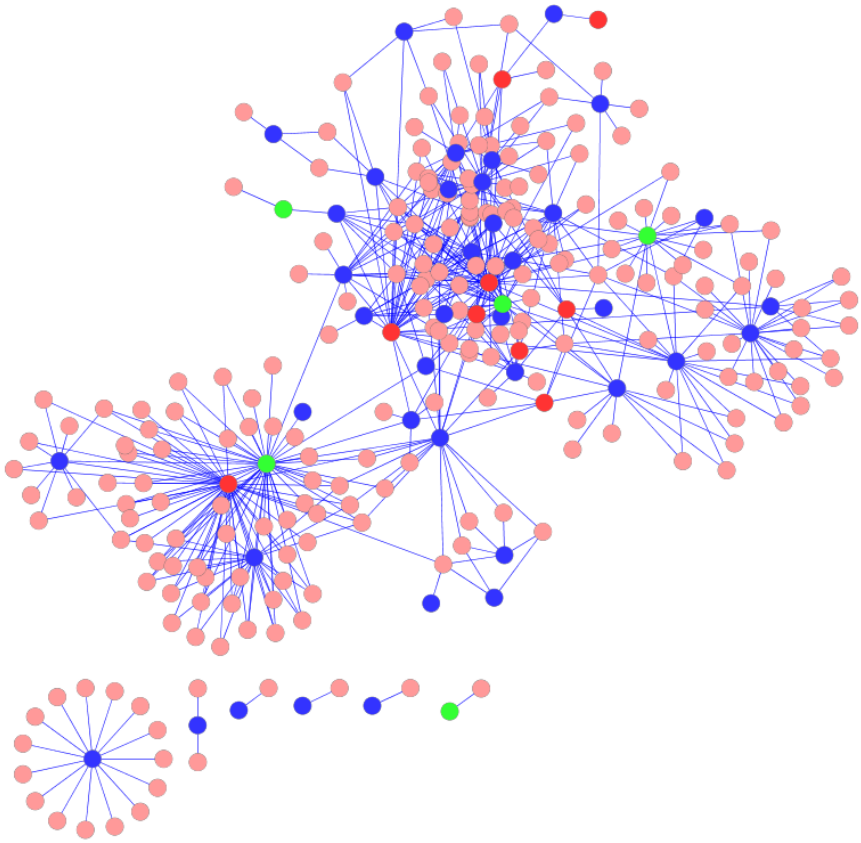

Figure 6.7: **Molecular concept map of gene sets from mouse perturbations (blue), human heart diseases (red) and human gene ontology terms (pink)**

gene sets with complex homology relationships, (3) development of a statistical hypothesis testing approach called XGSA, and (4) demonstration of how XGSA can be used in conjunction with MCM to identify evolutionarily conserved and species-specific molecular pathways using three real datasets.

Effectively, current GSA approaches deal with the complex homology mapping issue by reducing the complexity of the homology mapping (*i.e.,* by removing non one-to-one homologous gene pairs or abstracting the test to a higher level). In contrast, XGSA takes into account the entire homology structure when performing GSA. The benefit of XGSA is increased power to detect enrichment of gene sets with complex homology. XGSA also alleviates the false positive bias introduced by the naïve testing framework by ensuring gene set enrichment is significant in both species, overcoming the main limitation of 'at

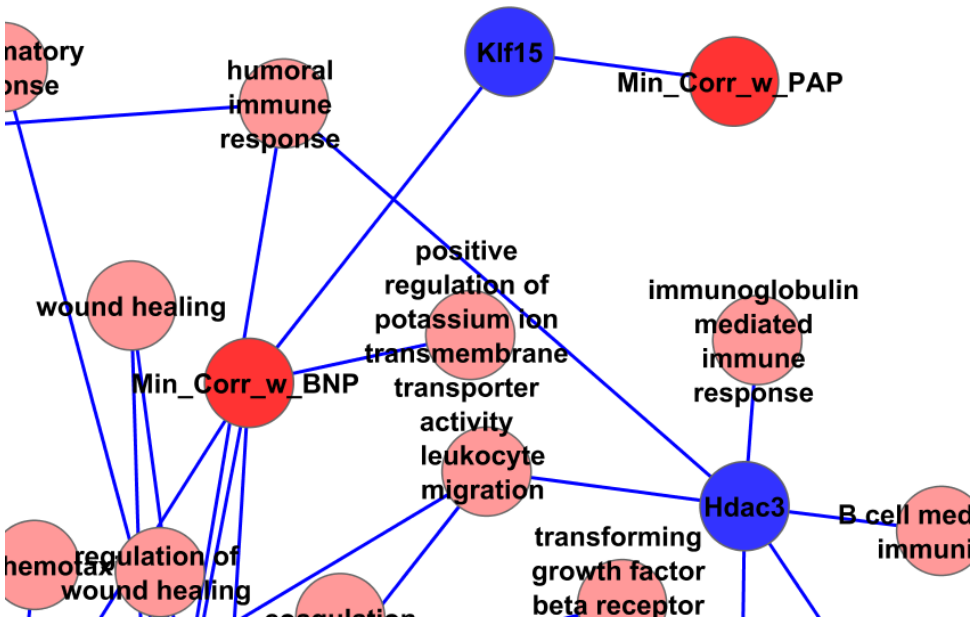

Figure 6.8: **Klf15 targets showing overlap with heart failure gene sets**

least one homolog' mapping. When compared to existing cross-species GSA approaches, XGSA balances both sensitivity and specificity for all gene sets.

Our work does not deal with the issues of comparing ranked lists, like the GSEA method (Subramanian *et al.*, 2005). Also, our method currently treats each homology relationship equally (absent or present), whereas the information about the extent of sequence homology was not used. Nonetheless, based on the formulation of the problem statement, we aim to extend our method to incorporate these features.

We have implemented the source code for XGSA in R. By harnessing the Ensembl BioMart portal our framework utilises the latest homology structure on a growing number of species supported in Ensembl (currently 69). Due to the flexibility and simplicity of our R frame-

work such that users can include custom homology matrices for unsupported species, the potential for XGSA to unlock cross-species gene set analyses is widespread. The typical use case is when investigating gene sets from an organism without a comprehensive gene set database. If the organism is supported by Ensembl the XGSA workflow is trivial, otherwise the user needs to compute homology to a genomic model organism to unleash XGSA. A second use case is when cross-species analysis is central to the biological questions being studied, such as in our case study of spinal cord regeneration. Our third case study shows how XGSA can be used to gain clinically relevant insights from cross-species analyses. The ability to integrate gene sets from different species together into a unified network-based visualisation such as a MCM improves speed and confidence when interpreting insights from traditionally problematic cross-species gene set analyses. This improved workflow is expected to be valuable for researchers in practice (Huang *et al.*, 2009).

Chapter 7

Cross-species identification of satellite cells from anole lizard skeletal muscle

7.1 Introduction

Lizards are evolutionarily the closest vertebrate group to humans with the ability to regenerate an entire appendage (Eckalbar *et al.*, 2012; Koshiba-Takeuchi *et al.*, 2009). Furthermore, many lizard species are able to autotomize, or self-amputate, their tails to avoid predation and then subsequently grow a replacement (Gilbert *et al.*, 2013). The regenerated lizard tail is a structurally complex appendage with skeletal muscle group, tendons, a hylainecartilage endoskeleton, peripheral motor and sensory nerves, vasculature, and skin (Fisher *et al.*, 2012; Hutchins *et al.*, 2014). Mammals have some regenerative capacity of appendages, limited to digit tip formation in neonatal mice and humans under age two (Yu *et al.*, 2010). Neonatal mice can also regenerate limited damage to heart ventricular muscle during the first week of life (Darehzereshki *et al.*, 2015; Porrello *et al.*, 2011).

Tail regeneration in A. carolinensis likely occurs through a stem cell mediated process,

rather than de-differentiation as occurs during epimorphic regeneration in salamanders (Fisher *et al.*, 2012; Hutchins *et al.*, 2014). Regeneration of a complex, multi-tissue structure such as the tail require pools of proliferative stem cells capable of differentiating into different lineages. Regeneration-capable species employ distinct strategies to generate these stem cell populations. In urodele amphibians, de-differentiation of injured tissue results in proliferative, lineage restricted progenitors (Kragl *et al.*, 2009), and trans-differentiation of cells that change their fate can contribute to more than one tissue (Jopling *et al.*, 2010). Resident progenitor cells play a role in the the urodele limb, with PAX7 positive cells activated and migrating to the site of injury following amputation (Sandoval-Guzmán *et al.*, 2014).

Studies of skeletal muscle repair in response to injury in mammals have provided considerable insight into the signalling pathways associated with satellite cell activation, proliferation and differentiation during repair. In response to acute damage, the myofibers are repaired by resident PAX7 positive satellite cells (Lepper *et al.*, 2011; Sambasivan *et al.*, 2011). Mammalian satellite cells are limited in their function to the repair of existing myofibers (Chen and Goldhamer, 2003; Dhawan and Rando, 2005; Relaix and Zammit, 2012; Wang and Rudnicki, 2011). There are cells present in a similar niche on the muscle fibers of anole lizards (Kahn and Simpson, 1974).

Several previous studies have profiled the transcriptomes of satellite cells in mammalian species such as the mouse (Ryall *et al.*, 2015), human (Charville *et al.*, 2015), pig (Jeong *et al.*, 2013), and cow (Lee *et al.*, 2014). However, comparison of gene expression across vertebrate species remains a bioinformatic challenge due to difficulties in identifying orthologous genes and differences in baseline gene expression. A useful framework for comparing transcriptome-wide expression profiles across species is based on testing whether a gene set that is specifically expressed in a species is shared with similar tissues or cell types in other species (Djordjevic *et al.*, 2016).

An earlier transcriptomic analysis of the regenerating *A. carolinensis* tail demonstrated that there were 326 differentially expressed genes along the proximal-distal axis, many of which are involved in the development of the skeletal system and muscle (Hutchins *et al.*,

2014). In this study we aim to determine which mammalian cell type is the closest match to the PAX7 positive cells isolated from the anole lizards. In order to achieve this, we characterised the lizard muscle progenitor cells using XGSA cross-species comparison of their transcriptome with that of mammalian cell types and tissues, including satellite cells (Djordjevic *et al.*, 2016). Our data demonstrated that the transcriptomic profile of the *A. carolinensis* muscle progenitor cells was most similar to mammalian satellite cells. It is likely that changes in the regulation of gene expression underlies the ability to regenerate appendages, thus we also compared expression of musculoskeletal and TGFβ/BMP pathway genes between mouse and lizard satellite cells using this same bioinformatics approach. We found key regulators of myogenesis and chondrogenesis that showed much higher ranked expression in lizard satellite cells, potentially giving insight into the mechanisms that drive myofiber regeneration.

7.2 Method

7.2.1 Bioinformatic Analysis of RNA-Seq Data

RNA-Seq analysis of A. carolinensis satellite cells (3 biological replicates) has been described previously by our group (Hutchins *et al.*, 2014) and the data are deposited in the NIH Sequence Read Archive (SRR1502189, SRR1502190, and SRR1502191; BioProject PRJNA253971). RNA-Seq data of mouse C57Bl/6J satellite cells (Ryall *et al.* (2015); SRA accessions SRR1726676, SRR1726677) were supplemented by our RNA-Seq analysis of mouse CD1 satellite cells isolated as described above using protocols as described in Hutchins *et al.* (2014). Transcript reads were mapped to A. carolinensis (AnoCar2.0) or mouse (GRCm38) Ensembl annotated genomes with HISAT2 v2.0.1 using default parameters (Kim *et al.*, 2015). Gene level read counts were generated using HTSeq v0.6.0 in intersection-nonempty mode (Anders *et al.*, 2015). For lizard and mouse satellite cell transcriptomes, Reads Per Kilobase of transcript per Million mapped reads (RPKM) were generated using edgeR (Robinson *et al.*, 2010). For human satellite cells, Fragments Per Kilobase of transcript per Million mapped reads (FPKMs) were generated from two biolog-

ical replicates analysed by RNA-Seq (Charville *et al.*, 2015) using TopHat (Trapnell *et al.*, 2009) and Cuffnorm (Trapnell *et al.*, 2010). For comparison with reported gene expression profiles for a library of different tissues, ENCODE transcriptome profiles summarized as FPKMs were obtained for human (hg19; 139 tissues) and mouse (mm9; 94 tissues).

7.2.2 Cross-species gene set analysis

For each RNA-Seq experiment containing RPKM or FPKM values, rank products were calculated for each gene using all available replicates (Breitling *et al.*, 2004). We identified the 1,500 most highly expressed genes in each cell type or tissue type. We then remove genes that are deemed highly expressed in more than 10% of the cell types, as these genes are likely ubiquitously expressed genes. The remaining genes are considered specifically expressed in each cell type or tissue type. We call this collection of gene sets the cell-type specific gene sets. Using the human and mouse ENCODE data collections, we generated a compendium of cell-type specific gene sets in humans and mice. Using a similar procedure, we used our lizard RNA-Seq data to generate a highly expressed gene set. We compared the lizard gene set against the human and mouse cell-type compendium using XGSA. A p-value is generated to represent whether the lizard gene set has significant overlap with a mouse or human cell-type-specific gene set.

To confirm that our gene set analysis is robust against our choice of parameters, we repeated the same analysis with a range of parameters (number of top highly expressed genes: 500, 1,000, 1,500, 2,000; % of cells containing marker gene: 1%, 2%, 3%, 5%, 10%, 15%, and 20%).

7.3 Results

Analysis of genes expressed in lizard satellite cells was carried out using RNA-Seq transcriptomic analysis (Hutchins *et al.*, 2014). Based on our previous data, we wanted to determine the similarity of the lizard PAX7 positive cells to the satellite cell population in

the mouse and human. We carried out this comparison using XGSA, a statistical method for cross-species gene set analysis (Djordjevic *et al.*, 2016). We compared highly expressed genes in the lizard satellite cell transcriptome to a compendium of cell-type-specific gene sets including 94 tissues from the mouse ENCODE project (Yue *et al.*, 2014). This comparison demonstrated that out of the tissues examined, the lizard PAX7 positive cell has a significant similarity with the mouse satellite cell based on expression of cell-type-specific marker genes ($p-value < 1.9 \times 10^{-26}$). It has a much more significant overlap with mouse satellite cell markers than markers of any other mouse cell types studied here (Fig. 7.1A), and this result is robust against analysis parameters (Fig. 7.2). Similarly, we compared lizard satellite cells with 139 tissues from the human ENCODE project and identified the greatest similarity with activated and quiescent human satellite cells (Fig. 7.1B, 7.2).

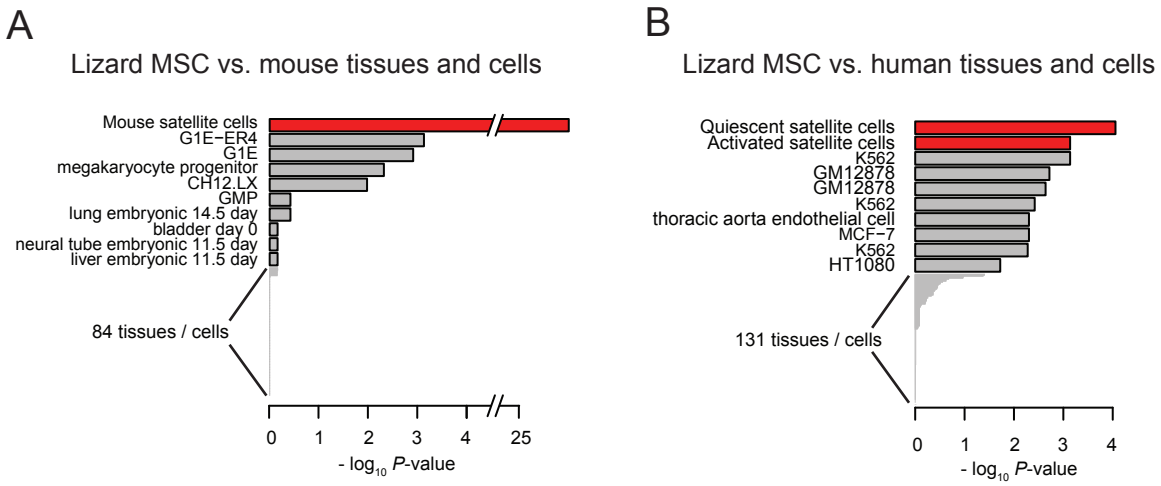

Figure 7.1: **XGSA analyses comparing the transcriptome from lizard satellite cells to multiple tissues from the mouse and human ENCODE projects.** XGSA comparison of marker genes with 94 mouse (A), and 139 human (B) tissues reveals that the lizard satellite cell transcriptome (Hutchins et al., 2014) displayed highest similarity with mouse and human satellite cells. Depicted are the top 10 most similar for each species comparison.

Bone morphogenic protein (BMP) and transforming growth factor β (TGFβ) signaling pathways have important regulatory roles both in embryonic myogenesis and postnatal muscle regeneration (George *et al.*, 2015; Lee *et al.*, 2012; McFarlane *et al.*, 2011; Sartori *et al.*, 2013). We examined the differences in the expression of the Tgfβ/Bmp pathway genes by comparing the relative rankings of 105 Tgfβ/Bmp pathway genes (KEGG

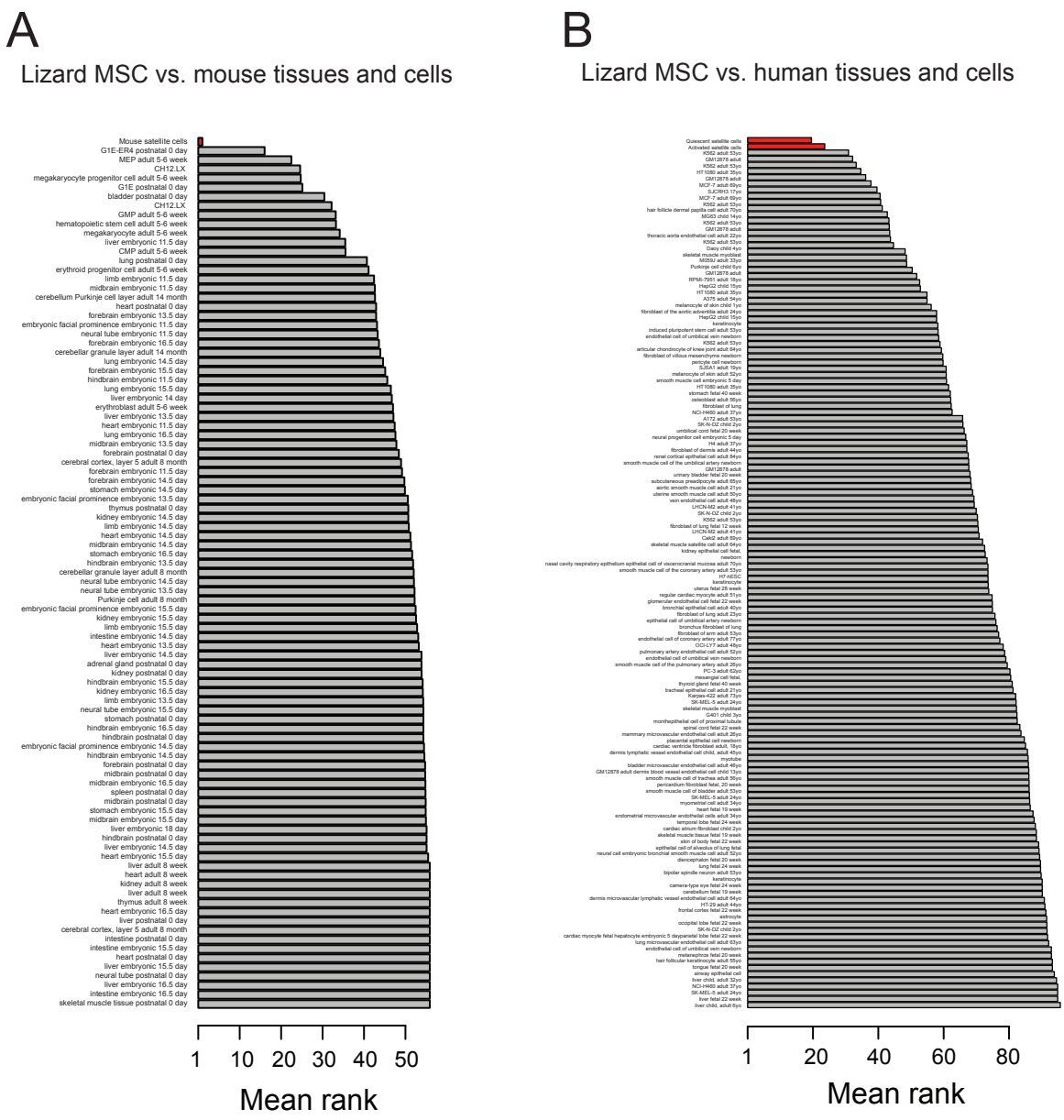

Figure 7.2: **Average XGSA rank for all tissues from the mouse and human ENCODE projects compared to anole satellite cells across a large range of parameters.** Over a large range of parameters (number of top highly expressed genes: 500, 1,000, 1,500, 2,000; % of cells containing marker gene: 1%, 2%, 3%, 5%, 10%, 15%, and 20%), XGSA comparison against 94 mouse and 139 human tissues reveals that the lizard satellite cell transcriptome (Hutchins et al., 2014) consistently displayed highest similarity with mouse and human satellite cells.

category mmu04350 and GO term GO:0030509) from lizard and mouse satellite cells (Fig. 7.3A). Genes with differential expression, higher rankings in lizard, included *Bmp2, Bmp5, Bmp7, Dcn, Fst, Id4, Inhba, Pitx2, Smad9, Tgif2, Zfyve16, Msx1, Msx2, Rnf165, Grem2*, and *Sostdc1*. Included in this list were genes with important roles in myogenesis, such as follistatin (*Fst*) and *Pitx2* that induce proliferation of satellite cells in regenerating muscle (Lee and McPherron, 2001; Lozano-Velasco *et al.*, 2011; Ono *et al.*, 2011; Winbanks *et al.*, 2013). *Msx1* also was among the highly ranked genes in lizard satellite cells, this gene regulates the cellularisation of myofibers, *i.e.*, the conversion of multi-nucleated skeletal muscle fibers into mononuclear cells, during amphibian limb regeneration. *Msx1* also regulates cell cycle re-entry of the resulting myoblasts. Ectopic expression of either *Msx1* or *Msx2* in mouse myofibers induced cellularisation and inhibited myoblast differentiation (Echeverri and Tanaka, 2002; Kumar *et al.*, 2004; Odelberg *et al.*, 2000; Yilmaz *et al.*, 2015).

Our data demonstrated that the expression of *Bmp2, Bmp5*, and *Bmp7*, genes encoding proteins that inhibit muscle differentiation and regulate chondrogenesis and osteogenesis, ranked considerably higher in satellite cells of the lizard compared to the mouse (Fig. 7.3A;(King *et al.*, 1994; Ono *et al.*, 2011; Shen *et al.*, 2009; Snelling *et al.*, 2010). While mouse and human satellite cells and myoblasts typically only differentiate into skeletal muscle, high levels of these BMP ligands can induce these cells to undergo osteogenesis and chondrogenesis in vitro (Asakura *et al.*, 2001; Katagiri *et al.*, 1994; Shea *et al.*, 2003; Wada *et al.*, 2002). Decorin (*Dcn*) encodes an extracellular matrix protein that decreases the fibrotic response during muscle regeneration (Fukushima *et al.*, 2001; Li *et al.*, 2008; McCroskery, 2005). Similarly, *Tgfb2* and *Fzd1*, implicated in fibrosis in dystrophic and aged skeletal muscle, are ranked much lower in the lizard myoprogenitors (Biressi *et al.*, 2014; Brack *et al.*, 2007). Other genes with lower expression levels in satellite cells of lizard, as compared to mouse, included *Bmp6, Bmpr1b, Bmper, Fzd1, Lef1*, and *Sox11*.

We observed that the transcriptomic profile of lizard satellite cells was marked by the high expression of genes involved in not only myogenesis, but also chondrogenesis and osteogenesis (Fig. 7.3B). Mammalian satellite cells are limited to differentiating into skeletal muscle

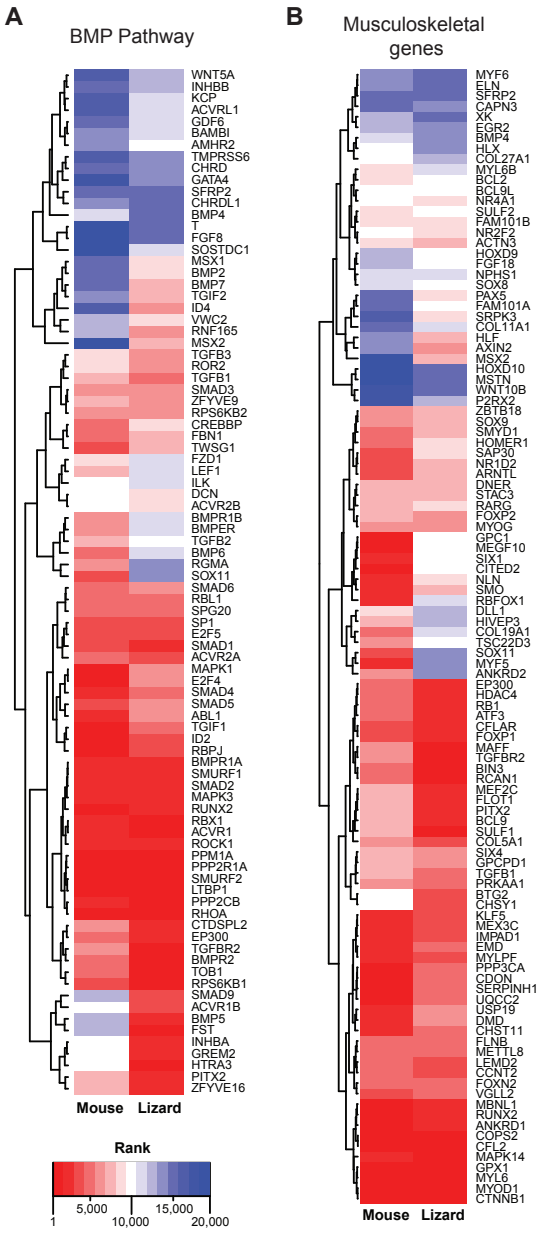

Figure 7.3: **Gene rank comparison of mouse and lizard satellite cell transcriptomes.** Heatmaps of genes involved in the TGFβ/BMP signaling pathway (panel A) and musculoskeletal development (panel B) show differential rankings between mouse and lizard satellite cells. Genes with the highest rank (highest level of expression) are closer to 1, and those with the lowest rank (lowest expression in that species) are closer to 20,000.

in vivo (Starkey *et al.*, 2011) however, murine satellite cells can be induced to become other musculoskeletal cell types or adipocytes *in vitro* when cultured in the presence of high concentrations of morphogens (Asakura *et al.*, 2001; Haas and Tuan, 1999; Ozeki *et al.*, 2007;

Takács *et al.*, 2013; Usas and Huard, 2007; Wada *et al.*, 2002). Focusing on genes in the gene ontology (GO) terms skeletal muscle and cartilage development and differentiation, we found that expression of several key genes involved in satellite cell activation and proliferation, and muscle differentiation ranked higher in lizards compared to mouse. This set included *Mef2c, Pitx2, Sox8, Srpk3, Hdac4, Axin2, Bcl9*, and *Sulf1* (Fig. 7.3B; Brack *et al.* (2008, 2009); Edmondson and Olson (1989); Gill *et al.* (2010); Knopp *et al.* (2013); Martínez-Fernandez *et al.* (2006); Molkentin *et al.* (1995); Raines *et al.* (2015)). *Nr4a1* whose expression was also ranked higher in lizard plays a key role in tissue repair and represses TGFβ target genes to limit fibrosis (Palumbo-Zerr *et al.*, 2015). Another receptor ranked higher in lizards was *Nr2f2*, a negative regulator of differentiation of many stem cells, including those in the skeletal muscle, cartilage, and bone marrow lineages (Gao *et al.*, 2017; Zhu *et al.*, 2016), and is up-regulated in activated satellite cells (Gill *et al.*, 2010). Genes that demonstrated a lower level of expression in lizard satellite cells included genes with known regulatory roles in muscle development and satellite cell function such as, *Dll1, Six1, Ankrd2, Cited2, Gpc1, Fzd1, Megf10*, and *Smo* (Fig. 7.3B). Among these *Megf10, Ankrd2, Fzd1*, and *Smo* induce differentiation of satellite cells (Brack *et al.*, 2008; Holterman *et al.*, 2007; Kemp *et al.*, 2000; Zhao and Hoffman, 2004). Together these data suggest that these myoprogenitors express the myogenic transcriptional program but are not differentiating, consistent with their status as proliferating single cells.

One of the most striking observations from our analyses were the very high expression of *Bmp2, Bmp5, Bmp7*, and many other cartilage-promoting genes in the lizard satellite cells, which would be predicted to result in greater potential of these cells to differentiate along the cartilage lineage.

7.4 Discussion

Transcriptomic and histological analysis of regeneration in the A. carolinensis tail points to a stem cell-mediated process (Fisher *et al.*, 2012; Hutchins *et al.*, 2014). Previous work has demonstrated that lizard skeletal muscle contains PAX7 positive progenitor cells and

7. Cross-species identification of satellite cells from anole lizard skeletal muscle

RNA-Seq analysis showed that the regenerating tail expressed marker genes of activated satellite cells and myoblasts (Hutchins *et al.*, 2014). Using the cross-species analytical tool XGSA, we carried out comparative transcriptomic analysis of proliferating satellite cells from the lizard with mouse and human tissues from the ENCODE project. As might be expected, we found the closest match with satellite cells (Fig. 7.1). However, the expression level of genes associated with chondrogenesis and osteogenesis were higher in lizard satellite cells compared with their mouse counterparts (Fig. 7.3).

BMP signalling induces different responses depending on ligand concentrations and exposure times. BMPs and their inhibitors define where and when muscle formation occurs (Hirsinger *et al.*, 1997; Re'em-Kalma *et al.*, 1995; Reshef *et al.*, 1998). During skeletal muscle regeneration, low concentrations of BMP2, 4, and 7 maintain proliferation of satellite cells and myoblasts (Amthor *et al.*, 1998; Friedrichs *et al.*, 2011; Ono *et al.*, 2011; Ozeki *et al.*, 2007; Sartori *et al.*, 2013; Wang *et al.*, 2010). Whereas, high levels of BMP2, 4, 5, or 7 inhibit myogenesis, induce chondrogenesis, and ultimately, osteogenesis of satellite cells, C2C12 myoblasts, C3H10T1/2 MSCs, and other MSCs, with continued exposure in culture (Bandyopadhyay *et al.*, 2006; Friedrichs *et al.*, 2011; Katagiri *et al.*, 1994; Knippenberg *et al.*, 2006; Liao *et al.*, 2014; Ozeki *et al.*, 2007; Schmitt *et al.*, 2003; Shea *et al.*, 2003; Takács *et al.*, 2013; Zhou *et al.*, 2016). Proliferating lizard satellite cells expressed *Bmp2, Bmp5* and *Bmp7* at high levels, unlike their murine counterparts (Fig. 7.3).

The expression of *Msx1* and *Msx2* was ranked higher in lizard compared to mouse satellite cells. These transcriptional repressors play important roles in inhibition of chondrogenic cell differentiation in mammals. (Ishii, 2005; Odelberg *et al.*, 2000; Satokata *et al.*, 2000). In amphibians, these proteins are important for de-differentiation of muscle during limb regeneration and are expressed in the blastema (Echeverri and Tanaka, 2002). The expression of these two genes in proliferating lizard satellite cells may indicate that they are necessary for maintenance of proliferation and plasticity.

Taken together, these data suggest that lizard satellite cells have increased musculoskeletal potential governed by changes in the regulation of gene expression. Further inquiry into these changes will not only shed light on the mechanisms of lizard tail regeneration, but

also provide a means for approaching improved tissue engineering of mammalian muscle and cartilage from satellite cells. Understanding the regeneration strategy employed by reptiles could provide important genetic algorithms that are exploitable in mammals.

Chapter 8

hiHMM: Bayesian non-parametric joint inference of chromatin state maps

8.1 Introduction

Readout of genetic information in eukaryotic genomes is mediated by the dynamic chromatin environment, which regulates DNA accessibility for the gene expression machinery through chromatin compaction, associated histone modifications and incorporation of histone variants. Chromatin immunoprecipitation experiments followed by genome-wide microarray (ChIP-chip) or sequencing (ChIP-seq) have revealed that distinct genomic regulatory regions are associated with different covalent modifications of histone proteins across various organisms (Kharchenko *et al.*, 2011; Liu *et al.*, 2011; Mikkelsen *et al.*, 2007; Park, 2009; Roudier *et al.*, 2011). For example, H3K4me3 (trimethylation of histone H3 at residue lysine 4) marks active promoters, H3K4me1 marks enhancers, H3K36me3 marks transcribed gene bodies, H3K27me3 marks Polycomb-repressed regions, and H3K9me3 marks heterochromatin. Although there are theoretically up to 2^n possible combinations

of n histone modifications at any given locus in the genome, in practice we only observe a small number of distinct dominant combinations, thus giving rise to the concept of chromatin states (Ernst and Kellis, 2010; Heintzman *et al.*, 2009; Hon *et al.*, 2008; Kharchenko *et al.*, 2011; Liu *et al.*, 2011; Mikkelsen *et al.*, 2007; Roudier *et al.*, 2011), in which each state consists of a combination of histone modifications.

A key idea underlying chromatin state analysis is to computationally identify the number and composition of chromatin states in the genome based on multiple genome-wide profiles of histone modifications, and to annotate the genome with these chromatin states. These states were found to be strongly correlated with various functional genomic features such as promoters, actively transcribing gene bodies, enhancers and heterochromatins. Although many chromatin states are common across different cell types or organisms, there are indeed clear examples of cell-type specific chromatin states consisting of unique co-occurrence of histone modifications. The H3K4me3/H3K27me3 bivalent promoter state that is prevalent in embryonic stem cells but mostly absent from terminally differentiated cells is such an example (Bernstein *et al.*, 2006). Investigating co-occurrence of multiple histone marks facilitates the differentiation of more subtle features in chromatin state, such as identifying tissue specific strong and weak enhancer regions (Ernst *et al.*, 2011) and changes in co-occurrence patterns between evolutionarily distant species (Ho *et al.*, 2014). Therefore a chromatin state map is a powerful means to infer potential genome function in a systematic and automated fashion. In conjunction with transcriptomic, DNase I and transcription factor binding data, chromatin state analysis was used to infer putative biochemical functions to a large fraction of the non-coding genomic regions (Dunham *et al.*, 2012).

Various machine learning algorithms, such as ChromHMM (Ernst and Kellis, 2012), Segway (Hoffman *et al.*, 2012), TreeHMM (Biesinger *et al.*, 2013), and tiered HMM (Larson *et al.*, 2013), have been developed to generate such maps to facilitate cell type-specific genome annotations in a systematic and automated fashion. All of them are based on probabilistic graphical models such as the hidden Markov model (HMM) and dynamic Bayesian network. One essential task for these algorithms is to learn the prominent com-

bination of histone modifications. Similar to any clustering problem, it is often difficult to identify a reasonable number of combinations that can adequately capture the major variation in the data. One possibility is to estimate the adequate number of states by performing exploratory analysis such as the Principal Component Analysis (PCA) (Julienne *et al.*, 2013). Another common approach is to run the HMM learning multiple times with varying state numbers, and identify the best fitting model using measures such as the Bayesian Information Criterion (BIC). The inferred states do not necessarily have a one-to-one correspondence with distinct functional regions in the chromatin, but they do give a very good data-driven description of the chromatin that can act as a starting point for further bioinformatics and experimental analysis (Baker, 2011). Therefore it is still of great interest to develop principled methods for identifying chromatin states within and across multiple genomes.

The cross-species chromatin state comparison problem was motivated by a recent model organism encyclopedia of DNA elements (modENDCODE) project that aims to systematically compare chromatin organisation in *Homo sapiens* (human), *Drosophila melanogaster* (fly) and *Caenorhabditis elegans* (worm) (Ho *et al.*, 2014). A naïve approach to this problem would be to compute the state map for each organism separately and then try to compare them afterwards. However, this causes significant problems for interpretation because what was defined as an enhancer state in one organism is likely not identical with that from another organism. In the other extreme, we could simply concatenate the three genomes into one and infer states, but then the inferred result would be highly biased by the species with the largest genome size or by other species-specific biases in the ChIP-seq signals. Similar problems exist when comparing multiple developmental stages or cell types in the same organism. In essence, we require a method that allows the information of the state definition to be shared across multiple genomes while retaining the ability for each genome to have its own chromatin state definition.

In the context of that project, a novel Bayesian non-parametric method was developed, called hierarchically linked infinite hidden Markov model (hiHMM), to infer chromatin state maps across multiple genomes simultaneously. The application of hiHMM in the

human/fly/worm cross-species comparison setting indicates that the chromatin state segmentations in individual organisms generated by hiHMM are highly comparable to the maps generated by ChromHMM (Ernst and Kellis, 2010, 2012) and Segway (Hoffman *et al.*, 2012) - two widely used chromatin state segmentation algorithms (Ho *et al.*, 2014). Furthermore, hiHMM is designed to address species-specific confounding factors such as variations in ChIP signal strength, genome size and co-occurrence patterns. In this chapter we will demonstrate the utility of this method using real data sets.

8.2 Methods

8.2.1 hiHMM

To address the problem of inferring consistent chromatin state definition across multiple related genomes, we employ an infiniteHMM (iHMM) (Beal *et al.*, 2002), a non-parametric extension of a finite state hidden Markov model, as a base model and extend it to model data from multiple conditions. For ease of conceptualisation, we consider the problem of chromatin state segmentation on multi-species histone modification data, in which case multiple conditions correspond to multiple species. The same statistical model can be used to describe data from different types of conditions such as multiple developmental stages or cell types.

To obtain a consistent state definition for principled comparison of chromatin states between multiple species, we propose to model the multi-species data by using an iHMM as a base model for each species data and then by coupling species-specific iHMM parameters together through the use of hyper-parameters, so that state definitions can be shared across species. We denote the proposed model as hiHMM. As the majority of histone modification co-occurrence patterns are conserved in animals, the probabilistic basis of hiHMM allows for largely consistent chromatin state inference across samples, while still allowing for minor species-specific differences. For more details about the statistical foundations of hiHMM, see (Sohn *et al.*, 2015).

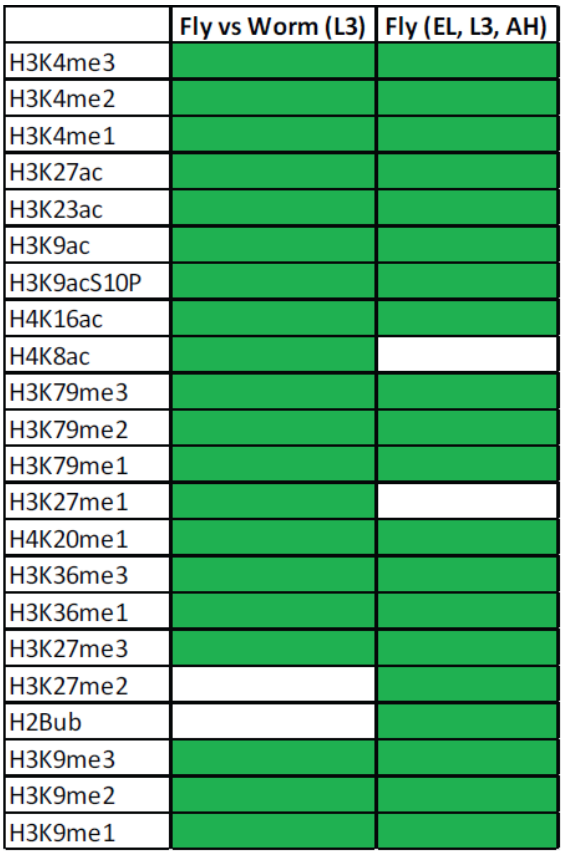

Figure 8.1: **Common histone modification marks profiled between samples.** Different stages of organism development were compared in this study: late embryo (EL); third instar larvae (L3); adult head (AH).

8.2.2 Running hiHMM on fly and worm ChIP-seq data

The fly (genome assembly version dm3) and worm (genome assembly version WS220) ChIP-seq and RNA-seq data were generated by modENCODE consortia. All Input-normalised ChIP-seq signal tracks were downloaded from the ENCODE-X interactive faceted browser: `http://encode-x.med.harvard.edu/data_sets/chromatin/`. The original fly and worm ChIP-seq data were in 10-bp resolution. All tracks were re-binned to 100 bp resolution by taking the mean of 10 consecutive bins. Data from multiple histone modifications were concatenated as columns into a tab-delimited format. Bins that overlapped unmappable regions were removed (Mappability regions were downloaded from `https://www.encodeproject.org/comparative/chromatin/#mappability`).

hiHMM was run in Matlab with default parameters: 200 burn-in iterations which means the first 200 samples are discarded during iterations for posterior inference, and then 10 consecutive posterior samples are collected to produce the final Maximum-A-Posterior output. For each comparison all available histone modification profiles produced by ChIP-seq experiments that are common across the targeted species and cell-types were used (Fig. 8.1). Chromatin states were trained on representative fly chromosomes 2L, 2LHet, X and XHet and worm chromosomes II, III and X, as per the modENCODE study (Ho *et al.*, 2014). Our prior experience suggests that training with all or only this representative subset of chromosomes in these organisms make very little difference in terms of the resulting chromatin state definition. Nonetheless, the hiHMM program is scalable to analyse all the chromosomes — which would be useful for exploring any previously uncharacterised chromatin landscapes.

Emission matrices from hiHMM output were examined and states were named based on chromatin state definitions in previous studies as well as overlap with expressed or unexpressed genes (Ernst *et al.*, 2011; Kharchenko *et al.*, 2011) (Fig. 8.2, Fig. 8.3). A custom R script is used to rename the states and re-introduce unmappable regions as State 0.

8.2.3 Chromatin state statistics

Genomic coverage was calculated as the percentage of the mappable genome that is occupied by each state, at the bin level. Expression odds ratio was calculated as the ratio of the number of expressed versus silent genes that overlapped with each chromatin state, divided by the genome-wide ratio of the number of expressed versus silent genes. A gene was considered expressed if its mRNA expression levels were greater than 1 RPKM. Gene body overlap was calculated as the percentage of bins annotated to each chromatin state that occur between the transcription start site (TSS) and transcription end site (TES) of an annotated gene.

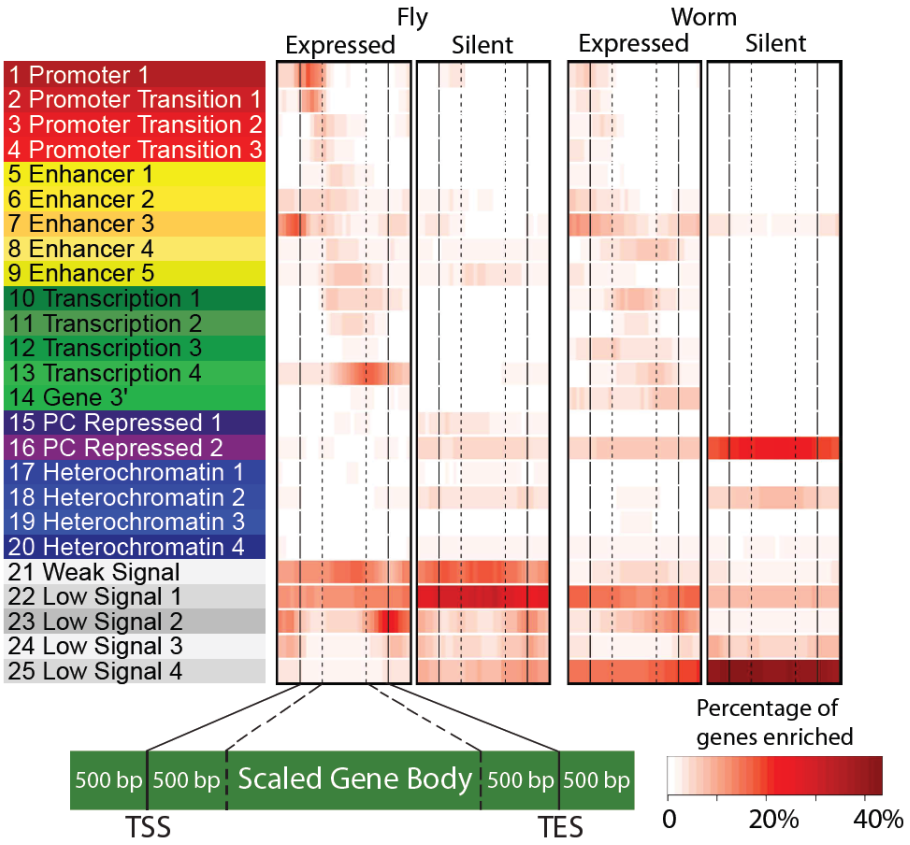

Figure 8.2: **Chromatin State Meta Gene Enrichment Profiles - Fly vs Worm** Heatmaps showing the spatial enrichment of each chromatin state in relation to the average 'meta gene.

8.2.4 Meta-gene chromatin state enrichment profile

A meta-gene matrix was constructed from all annotated protein coding genes that were at least 1300bp in length and do not overlap another gene within 500bp of its TSS or TES. Protein-coding gene annotation was downloaded from `https://www.encodeproject.org/comparative/transcriptome/`. We further excluded genes that occurred within 1000bp of a chromosome start or end. The meta-gene matrix contains the chromatin state annotations of each 'representative' gene extending to 500bp upstream of the TSS and 500bp downstream of the TES. Enrichment profiles are presented as heatmaps where the colour indicates the percentage of genes that have been annotated with that particular chromatin state at that relative genomic position. Meta-gene profiles of expressed and silent genes were computed separately.

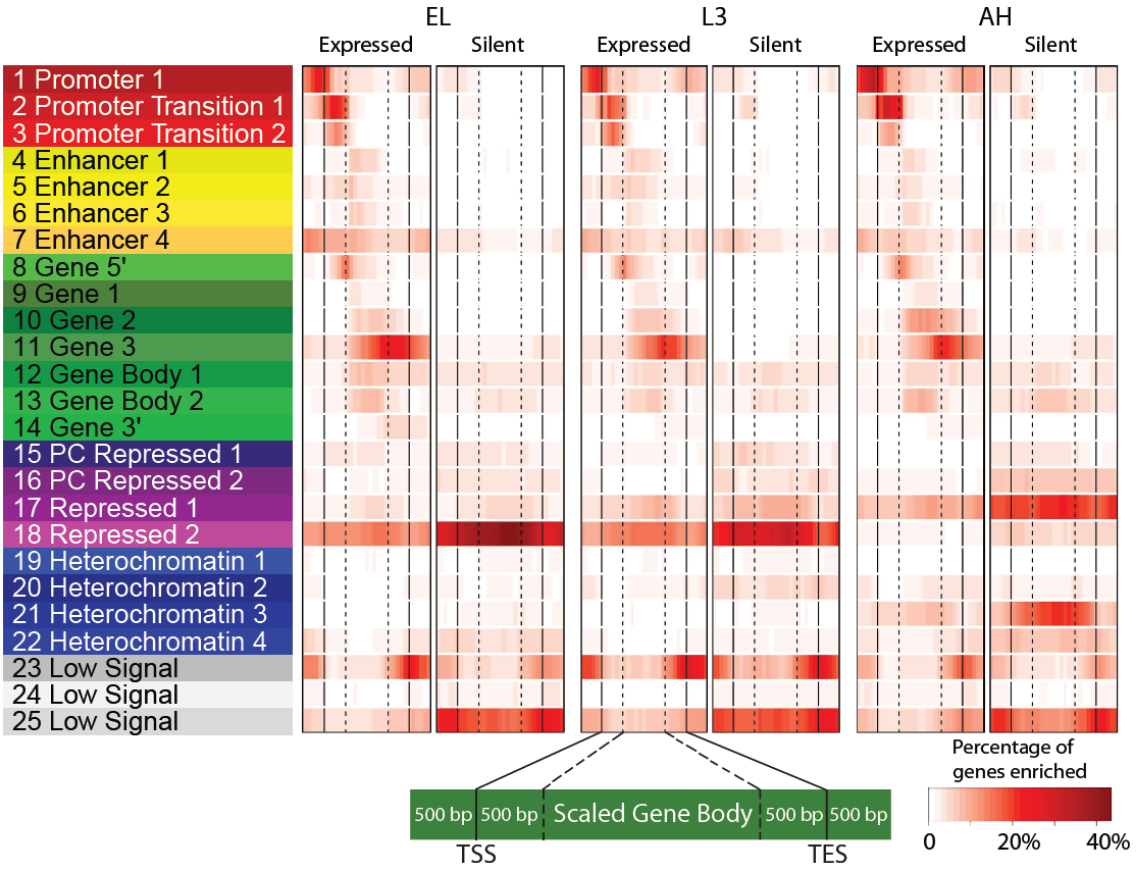

Figure 8.3: **Chromatin State Meta Gene Enrichment Profiles - Fly 3 Stages - Model 2** Heatmaps showing the spatial enrichment of each chromatin state in relation to the average 'meta gene.

8.2.5 Inter-sample chromatin state co-occurrence

Fold change was calculated as the observed number of bins that transitioned between any two chromatin state annotations divided by the expected number. The expected number was the mean number of bin transitions between those two states in 1000 Monte Carlo simulations with a randomised chromatin state assignment, preserving the relative genomic coverage of each state. Fold change was truncated between 1 and 5 for simplified visualisation and interpretation.

8.2.6 Co-occurrence matrices

Co-occurrence of genomic chromatin state annotation between experiments was calculated as the number of bins that were annotated as a particular chromatin state combination in two experiments divided by the total number of bins annotated to those states in the respective experiments. This gives a value between 0 and 1 which is presented in a heatmap.

8.2.7 Gene ontology enrichment of target genes in a region

Official gene symbols for all genes that overlapped with the selected regions were submitted to the DAVID bioinformatics tool (Huang *et al.*, 2009). The Benjamini and Hochberg adjusted P-value (Benjamini and Hochberg, 1995) of the 10 most significant gene ontology (GO) biological process results are presented for each analysis.

8.3 Result

8.3.1 Case study 1: hiHMM identifies species-specific chromatin states in fly and worm

The analogous developmental stage of stage 3 larva (L3) in fly and worm was selected for this cross-species chromatin state analysis. hiHMM was run using 25 starting states with Models 1 and 2 on the combined data, and 30 starting states using Model 2 to capture more species specific states.

Chromatin state analysis of fly L3 vs. worm L3 shows both shared and unique patterns of chromatin mark co-occurrence between the two species (Fig. 8.4A). We found 25 states that grouped into 6 categories: promoters, enhancers, gene body, heterochromatin, repressed and low signal. Most states are conserved and have similar compositions, but with some clear differences. Fly promoter states (red) show a distinct lack of H3K23ac when

compared with worm (green highlight in Fig. 8.4A). Conversely, worm promoter states lack H3K79me1 when compared with fly (blue highlight in Fig. 8.4A). Genic and transcription states (green) in fly show enrichment of H4K16ac, H3K79me1 and H4K20me1, all of which are largely absent in the same states in worm (orange highlight in Fig. 8.4A). H4K8ac on the other hand, is enriched in these states in worm but completely absent in fly (yellow highlight in Fig. 8.4A). Further differences are visible in the repressed (purple) and heterochromatin (blue) states. In fly there is a clear differentiation of repressive histone modifications between the two state classes whereas the marks consistently co-occur in worm (purple highlight in Fig. 8.4A).

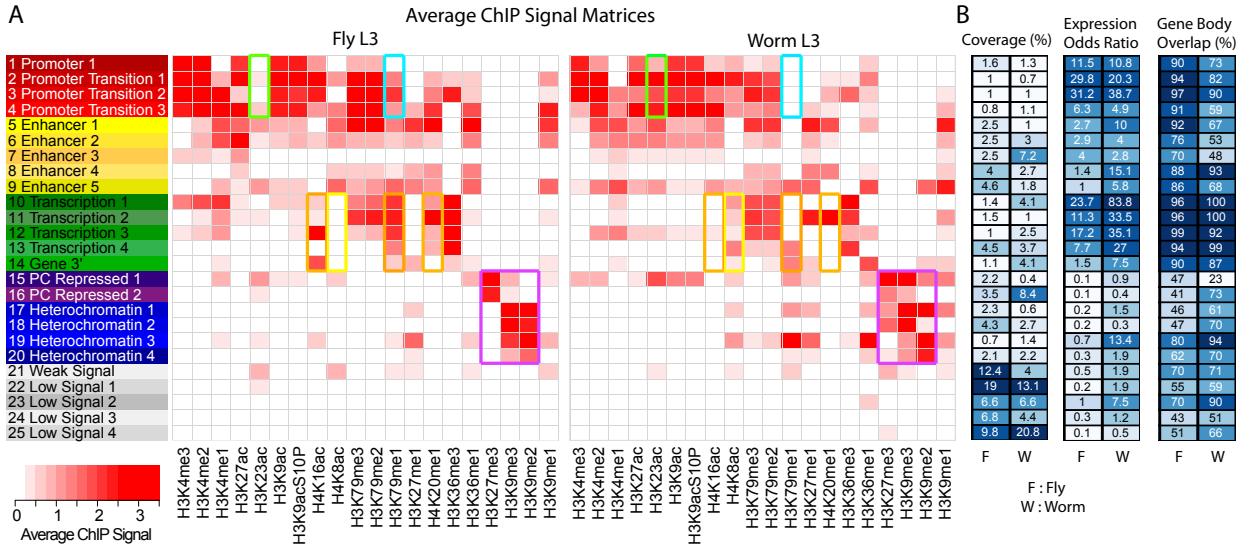

Figure 8.4: **Cross-species chromatin state analysis. A)** ChIP signal matrix showing the average observed histone modification profiles for each of 25 states inferred by the hiHMM algorithm (Model 1) in fly and worm. Species specific differences are highlighted. **B)** Percentage of genome covered by the state (Coverage), relative enrichment of expressed genes per state (Expression Odds Ratio) and the percentage of state annotations that occur between the TSS and TES of annotated genes (Gene Body Overlap).

In fly all low signal (grey) states are associated with an expression odds ratio of less than or equal to one, indicating a low overlap between the state and expressed genes (Fig. 8.4B). In worm however, most low signal states have an expression odds ratio above one and low signal state 2 shows largely increased odds for overlapping with expressed genes. We also see reduced gene body overlap in worm promoter and enhancer states when compared to fly, also visible in the meta-gene enrichment profiles (Fig. 8.2).

8.3.2 Case study 2: hiHMM identifies developmental stage specific loci in fly

Three fly developmental stages were chosen for chromatin state comparison: Late embryo (EL), third instar larvae (L3) and adult head (AH). hiHMM was run on the combined data sets using 25 starting states with Models 1 and 2.

Jointly learned average ChIP signal matrices for three developmental stages in fly show that the majority of observed histone modification combinations and their genomic occurrence remains stable during development (Fig. 8.6). H3K79me1, however, shows a marked reduction in enrichment in active states in AH compared with EL and L3 stages (Fig. 8.7). While this difference is interesting it is secondary to our ensuing analysis of differential state co-occurrence during development.

Using these chromatin state maps we implemented an unbiased approach for identifying developmentally regulated genes from chromatin state co-occurrence between two developmental stages — EL and AH — in fly (Fig. 8.5). 1659 genes from regions that transitioned from an active (promoter, enhancer, gene) state in EL to an inactive (repressed, heterochromatin, low signal) state in AH (Fig. 8.5C, top right) were strongly enriched for multiple developmental GO terms, including 'Respiratory system development' (P-value 1.66×10^{-8}) (Fig. 8.5E). Similarly, 1889 genes that changed from inactive in EL to active in AH (Fig. 8.5C, bottom left) were strongly enriched for terms expected from a fully developed organism, including 'Transmission of nerve impulse' (P-value 1.48×10^{-7}). These transitions in developmental regulation are clearly visible in a genome browser when the chromatin state tracks are visualised (Fig. 8.5D).

8.4 Discussion

The main difficulty of joint analysis of related data-sets is how to obtain a consistent state definition, or the mapping between a set of states defined in one species and the one in

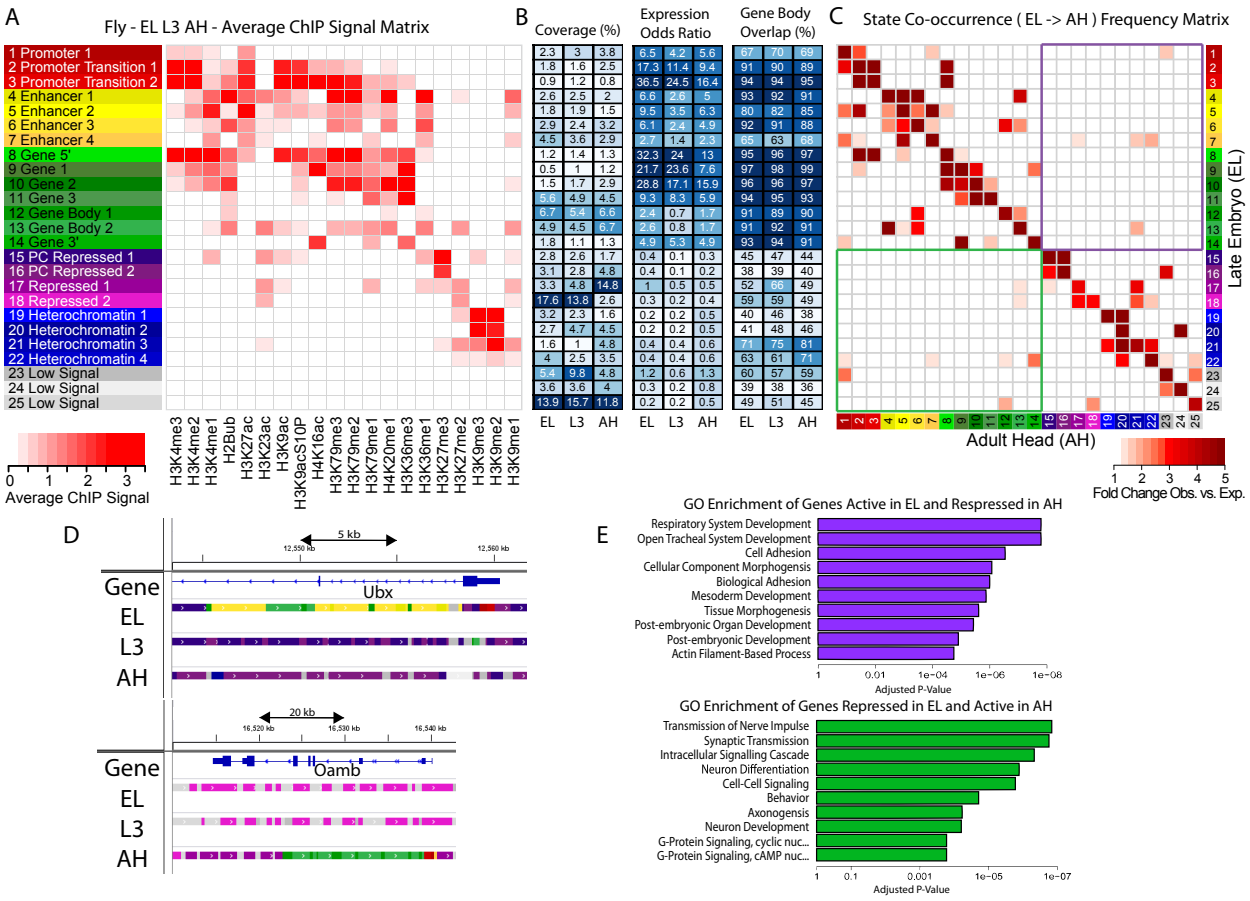

Figure 8.5: **Chromatin state characterisation and analysis across developmental stages in fly. A)** ChIP signal matrix showing the average observed histone modification profiles for each of 25 states jointly inferred by the hiHMM algorithm (Model 2) for three stages of fly development: late embryo (EL), stage 3 larvae (L3) and adult head (AH). **B)** Percentage of genome covered by the state (Coverage), relative enrichment of expressed genes per state (Expression Odds Ratio) and the percentage of state annotations that occur between the TSS and TES of annotated genes (Gene Body Overlap). **C)** Chromatin state co-occurrence between two developmental stages in fly (EL and AH). Shown is the observed vs. expected fold change of the co-occurrence of each state in EL and each state in AH. Based on this analysis we selected significantly over represented co-occurrence regions to investigate and characterise the genes involved through gene set enrichment analyses. **D)** Genome browser views of representative genes Ubx and Oamb with three stage chromatin states. These genes were identified through chromatin state co-occurrence analysis as having different chromatin states in EL and AH. **E)** The top 10 GO biological processes enriched in the genes that are within regions of the genome that changed from an active state in EL to a repressive state it AH (top panel) or vice versa (bottom panel).

another species. If one were to apply iHMM separately on each of the data-sets, we face the problem of mapping state definitions between species. The proposed hiHMM solves

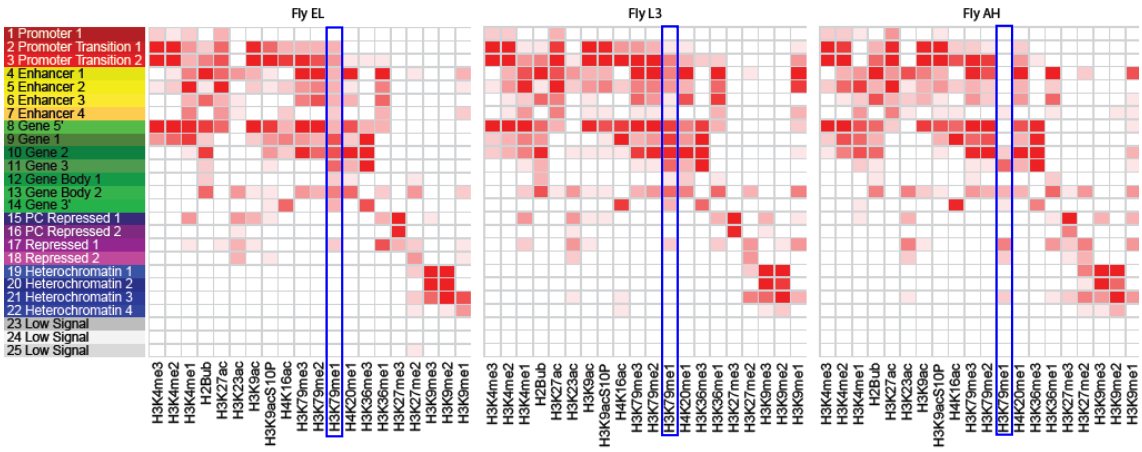

Figure 8.6: **Chromatin state characterisation across developmental stages in fly. A)** Jointly inferred chromatin states for each developmental time point show that H3K79me1 differs in AH compared to EL and L3.

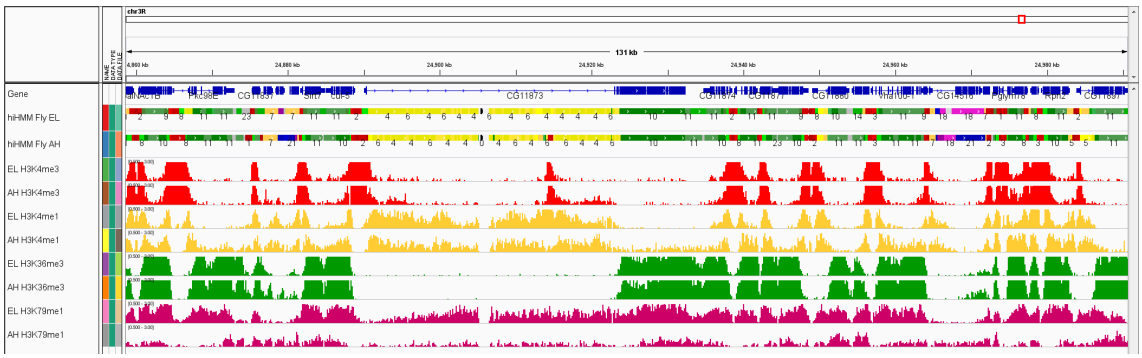

Figure 8.7: **IGV browser plot showing differences in active state composition between developmental stages in fly.** For each histone modification two fly developmental stages are shown, EL (above) and AH (below). H3K79me1 is coloured in purple, last two tracks. Active states are coloured red, yellow and green in the hiHMM chromatin state track (below gene annotations) and are marked by H3K4me3 (red signal), H3K4me1 (yellow signal) and H3K36me3 (green signal).

this problem by assuming a common state space definition across all the data-sets.

One distinguishing feature of hiHMM is that we can jointly infer emission values for genomes of differing lengths without the dominating influence of a longer genome on the result or the need to perform compensatory sub-sampling during training. This is a major limitation of existing methods that concatenate multiple samples for joint learning that is overcome using hiHMM.

Similar to Segway or the HMM approach of Kharchenko *et al.* (2011), hiHMM directly models continuous ChIP signal values and therefore alleviate the need of selecting a binary threshold cutoff.

Our real data analysis shows that the advantages of Model 1 become more evident when the data discrepancy across different conditions is large, for example, in the case of multiple species data. In terms of revealing interesting biology, the ability to infer sample specific emission matrix parameters (as in Model 1) allows for intuitive and detailed comparison of chromatin mark combinations between different species, or cell-types. This is evident from our 25 state analysis in fly vs worm in case study 1. In addition to the finding that fly and worm have different chromatin modifications compositions in heterochromatin, here we observed several previously unreported differences between the two species as described above, most notably the relative depletion of H3K23ac and enrichment of H3K79me1 in fly promoter states, as well as multiple differences in the transcription states. Model 1 was also useful in identifying the unexpected changes in the distribution of H3K79me1 during fly development. Many of these observations were not possible in previous studies that did not include these marks in their comparison (Fig. 8.8), and they may suggest different mechanisms of genetic regulation between the two species.

Jointly learned and shared emission matrix parameters (as generated by Model 2) provide an easily interpretable platform on which to compare multiple samples from the same genome without the confounding factor of different state definitions. Model 2 can also be used to identify species specific states, when the genomic coverage of states approach 0 in one species. Additionally, Model 2 enjoys better statistical properties such as faster convergence and shorter running time, so Model 2 would be a better choice when the discrepancy between genomes is expected to be relatively small as in the case of different developmental stages or environmental conditions in a single species.

By applying hiHMM to this newly compiled data set we were able to identify two previously uncharacterised chromatin states that we have named 'Repressed' states in the fly development analysis (State 17 and 18 in Fig. 8.4A). These two states combined constitute roughly 18% of the fly genome and are almost exclusively characterised by marks

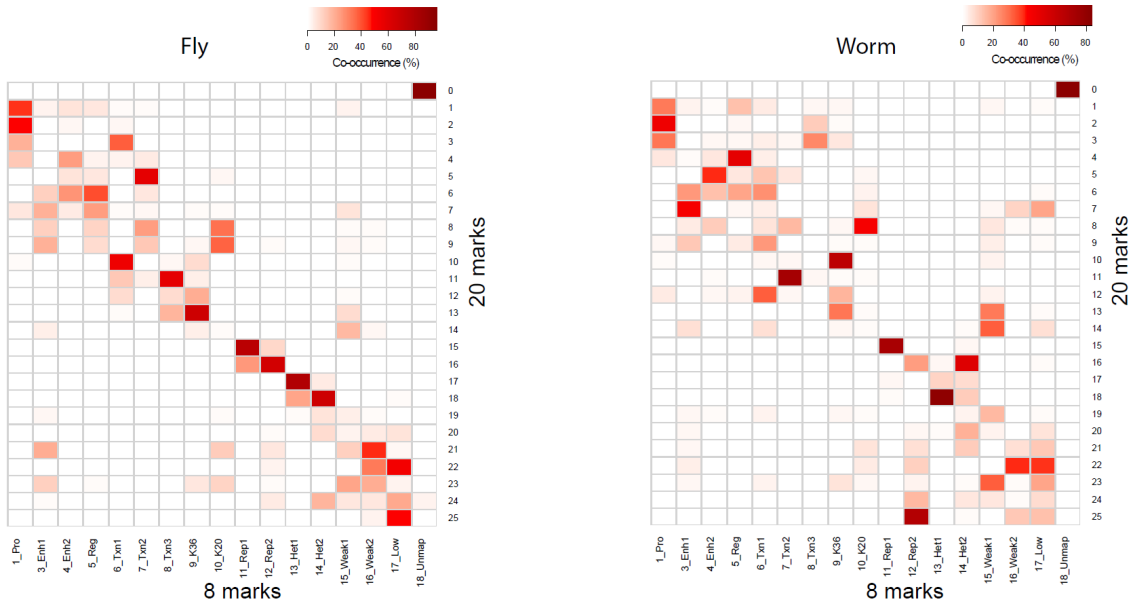

Figure 8.8: **Inter-study Co-occurrence matrices** Co-occurrence matrices showing the overlap between the 25 states generated by this study using 20 histone modification marks and the 17 state model generated by Ho et al. 2014 using a subset of 8 histone modification marks.

that were not profiled in previous studies, and so these states were missed.

In this paper, we demonstrated a variety of features of hiHMM that makes it useful for cross-sample joint chromatin state inference. We have implemented two models of hiHMM, each having advantages and limitations of interpretation and inference. The flexibility of using both learning models allows for a more comprehensive analysis during different applications and experimental designs.

Chapter 9

Conclusions and Future Challenges

The world of biomedical research is quickly becoming more data driven. The promise of this data for unravelling human disease makes it of utmost clinical importance. Many of the genetic systems that drive phenotypic variation remain undescribed, even though there are unprecedented amounts of biological data available. Data generation is no longer the bottleneck in biomedical discovery it once was, but data analysis and integration remain hindered by unsolved technical challenges and are very time consuming.

This thesis aimed to develop and apply methods that would facilitate the extraction of biomedically relevant information from the wealth of data produced by the genomics revolution. This included methods for constructing and analysing highly informative causal GRNs, extracting and integrating multiple data types and sharing biological insights across species.

9.1 Towards more complete causal GRNs

Many different types of experiment aim to resolve the key questions about genetic regulation – which molecular pathway controls a gene's expression, and how? One popular way to model the underlying systems is via gene regulatory network construction. The most common approaches for large scale GRN inference are based on correlating gene expression values across many samples. Comparing the results of these approaches to 'gold standard' causal regulatory edges as measured by perturbation experiments, showed that these methods do not perform very well in eukaryotic organisms (Marbach *et al.*, 2012). As such, while being a very well studied topic, GRN inference remains a difficult and unsolved problem for biomedical research.

We collected 187 microarray samples from two mouse organs and 2228 matched high quality perturbation data from the literature, and showed that correlation based algorithms have a very high false positive rate, rendering them essentially useless for causal GRN construction. Additional sources of data such as biological pathways and PPI networks has been shown to improve inference of causal phospho-protein networks (Hill *et al.*, 2016), however they did not contain discriminative power for causal GRN inference in our analysis. We therefore conclude that in mammalian systems it is preferable to utilise perturbation data directly and have a reliable high quality causal GRN.

To facilitate this, we developed the GEOracle tool to allow non-experts to very rapidly extract 'gold standard' perturbation data directly from the GEO microarray database. Using this tool we constructed in 8 hours what may be the largest organ specific causal GRN in a mammal, and qualitatively demonstrated its biomedical utility by extracting insights confirmed by the literature via the MURSS algorithm developed in this thesis. We have solved a critical component of the pipeline to make it feasible for biomedical researchers to easily harness the power of published perturbation data, if it is available.

In the future it is likely that correlation based GRN inference algorithms may somewhat improve by integrating directly relevant prior knowledge such as transcription factor bind-

ing sites. However based on the work in this thesis, it is unlikely that correlation based algorithms will be able to confidently distinguish true from false positive edges on their own. More exciting however is that perturbation data is becoming more widespread, high-throughput and much cheaper to generate. Indeed, genome-wide CRISPR based genetic knockdown screens are already being deployed (Chen *et al.*, 2015). Combining such powerful perturbation assays with single cell transcriptomic technology will create very high dimensionality perturbation datasets from which we will likely be able to extract high quality cell type-specific causal regulatory relationships. We will soon have many such data sets that will need to be integrated, across conditions, tissues, developmental time, and across species. This is a very promising area of future research that will likely lead to reliable, predictive and explanatory causal GRNs.

9.2 Towards large scale principled data integration

Vast resources are being dedicated to generating and analysing many types of genomic data in every biological context. But how can we best integrate the resulting knowledge to generate clinically or commercially relevant insights? There is potential to greatly reduce the time between data generation and understanding causative mechanisms with systems based integrative analyses.

As has been discussed in this thesis, many network and systems based approaches for integrative analyses have been applied to studying human diseases (Azuaje *et al.*, 2013; Barabasi *et al.*, 2011; He *et al.*, 2011b; Lage *et al.*, 2010, 2012; MacLellan *et al.*, 2012; Sperling, 2011; Zhang *et al.*, 2013). In that context, disease gene prioritisation is critical, as up to 80% of disease sequencing studies fail to find a causative mutation in a known disease gene (Taylor *et al.*, 2015). The logical next steps are to intelligently link more genes to each disease, and to link non-coding regulatory variation to their relevant genes.

In chapter four we recorded disease phenotypes from mouse perturbation experiments in the literature, and by applying the MURSS algorithm to a tissue-specific GRN, we iden-

tified potential key regulators for different eye disease phenotypes. We also implemented this GRN based strategy in the contexts of heart disease. In both case studies, some unexpected results were validated by recently published literature. This approach will become more powerful as disease gene databases and the GRNs we can construct become larger and more complete. Additionally, a more sophisticated pathway based analysis that incorporates additional data types like the potential for TF binding, tissue-specific genomic and epigenomic activity, and protein interactions and phosphorylation levels, should extend the power of this approach to prioritise candidate disease genes beyond the perturbation based GRN.

Non-network data like tissue-specific gene expression, disease-specific differential expression, or GO annotations, can be used to predict disease gene status directly, or can itself be propagated through a network structure to improve detection of the functional modules most relevant to the disease. We implemented this approach by propagating GO annotations and lens-specific gene expression from the developing mouse through a PPI network, and significantly improved our ability to classify human cataract causing genes. The approach of combining different types of functional data through network scaffolds to identify key modules would likely be further improved by explicitly modelling the relationships between the data types, as an alternative to the homogenous SVM approach used in this thesis.

Incorporating non-coding regulatory information into disease studies remains a largely unsolved problem, because unlike the genome, the epigenome is different in every cell type and disease context. Some aspects of genetic regulation we are only just beginning to understand, and no single experiment gives a complete picture of the regulatory state of the entire genome. It is therefore still very expensive and time-consuming to unravel comprehensive regulatory states in disease samples. As more cell type-specific epigenomes are constructed (for example by ENCODE (Dunham *et al.*, 2012)) we are provided with a valuable foundational resource with which to start looking at these questions.

An outstanding bioinformatics challenge is how to integrate tissue-specific gene regulatory information across different cell types and developmental time points. This is a crucial task

for the field, as performing every relevant assay in every tissue and experimental context will remain unfeasible for the foreseeable future. A potential solution lies in the intelligent use of structured tissue ontologies. Tissue ontologies (such as those provided by the e-Mouse Atlas Project (Richardson *et al.*, 2014)) describe the constituent tissues and cell types in each organ during development. If each publicly available data set was annotated to a specific entry in such an ontology, we could automatically borrow and integrate data from tissues and time points most similar to the one we are studying, greatly increasing the re-usability of published data.

The lack of consistent sample annotation, or metadata, remains a major hurdle to large-scale integration of publicly available data. We began to address this by constructing GEOracle, an online tool that semi-automatically processes large numbers of perturbation data sets from GEO, by mining and analysing the free-text metadata. If combined with automated ontology-based sample annotation and search strategies (Galeota and Pelizzola, 2016), further development of this approach could provide a means to rapidly integrate many related data sets, and greatly increase the pool of data from which we can extract reliable insights.

As more genomic and gene regulatory information becomes available we will see improved prioritisation of disease genes and elucidation of biological mechanism. Although the bioinformatics analysis can theoretically be semi-automated and completed in a matter of days, laboratory validation and *in vivo* studies can take months or years. While no novel laboratory validation of the predictions made in this thesis have been completed, a portion were validated by recent biomedical literature.

9.3 Towards cross-species analyses

The extreme difficulty with which we share data across species means this often not attempted in routine biomedical research. Integrating data representing the richness of biological diversity across the tree of life presents an opportunity to uncover discoveries

that are otherwise implausible to elucidate from studying one species alone. We urgently need the theoretical basis and bioinformatics tools to reuse data for cross-species analyses in a principled way, to understand biological mechanisms that are conserved, as well as those that are unique, across the many branches of life. In particular, we need methods to 'steal' from the knowledge 'rich' organisms including human and mouse, and transfer this data to the vast majority of knowledge 'poor' species. More researchers are understanding this and incorporating cross-species analyses into their studies (Claussnitzer *et al.*, 2014; Gerstein *et al.*, 2014; Ho *et al.*, 2014; Rittschof *et al.*, 2014; Zheng *et al.*, 2011), even though there has been very little investigation into the unique challenges such analyses create, and very few tools specifically designed to address them (Kristiansson *et al.*, 2013; Yang *et al.*, 2014).

The major obstacle to cross-species analyses is the complex-homology problem – which gene in one species matches to which gene in another species, and how functionally similar are they? This question becomes harder to answer as the species become more evolutionary divergent. Many studies ignore this issue and use a naïve homology mapping approach (Baker *et al.*, 2012; Kang *et al.*, 2014; Reimand *et al.*, 2007). Another common approach used by researchers today is to eliminate the complexity entirely (i.e. by using the BLAST reciprocal best hits approach) (Britto *et al.*, 2012; Gohin *et al.*, 2010; Labbé *et al.*, 2012), but while convenient, this throws away a lot of valuable homology information. The complexity of life is rich and fascinating, and depending on the biological question you are investigating this discard of data may be unacceptable. In order to begin to address this problem, we developed XGSA for cross-species gene set analysis, the first method for cross-species gene set analysis that utilises the complete complex-homology between species and alleviates the associated bias.

Gene set analyses are broadly useful to any biological problem where you can construct a set of related genes, and we demonstrated several varied applications of XGSA. First we identified conserved and species-specific molecular signatures in spinal cord regeneration and responses to social challenge, integrating data sets from mouse, fish, frog, lizard, insects and human. We also used XGSA to gain some clinical insight into human heart

disease, by identifying genetic signatures from mouse perturbation experiments that resemble human heart disease signatures. Finally, we applied XGSA to cross-species cell type identification, to confirm that the marker genes of Pax7 positive cells from lizard closely resemble those of skeletal muscle satellite cells in human and mouse.

A limitation of XGSA is that it only deals with binary gene sets, and not other structures such as ranked lists of genes which contain additional information about how the genes relate to each other. This is an important area for potential future research which would make the XGSA platform more powerful and broadly applicable.

In this thesis we also empirically evaluated and applied a recently developed method for cross-species chromatin state inference on epigenomic data sets from fly and worm, two model organisms separated by around 600 million years of evolution. We discovered previously unreported differences in the histone modification patterns that occur in these two species, raising interesting hypotheses about the evolutionary timeline of epigenomic mechanisms. The fact that most of the chromatin state patterns were consistent demonstrates the high conservation in the overall systems of genetic regulation in animals, strengthening the notion that we should be sharing knowledge across species. Further innovations in cross-species bioinformatics methodologies will undoubtedly help us understand many more aspects of biology in a timely manner.

9.4 An expanded suite of computational methods for biomedical discovery across species

Alongside theoretical and applied research, this thesis presented three methodological innovations that aid in analysing the regulation of genetic systems across species: GEOracle, MURSS and XGSA. While various distinct biological problems provided the basis for the development and application of these methods, this section describes a hypothetical pipeline that uses all three synergistically during the modern biomedical research process.

A common problem

A research group is searching for drugs that can increase heart regeneration after myocardial infarcts in humans. They perform a compound screen in an axolotl animal model to evaluate effects on heart regeneration, finding two compounds that positively affect organ regeneration, and want to know if and how these effects might translate to humans. Previous mouse studies have shown that these compounds are largely inactive in mammalian hearts, due to a lack of receptor expression in heart tissues.

The researchers perform transcriptome profiling of the regenerating axolotl hearts before and after these two compounds are administered and produce transcriptional signatures for the compounds. How can they translate these findings back to humans?

A novel solution

The researchers use XGSA to compare the compound transcriptome signatures to human gene set knowledge bases like GO and KEGG. This identifies the key tissues in which the changes are happening (muscle cells and satellite cells) and a mixture of GO terms involving some signaling pathways and differentiation.

The researchers want more detailed mechanistic insights for human tissues so they can generate hypotheses for further experiments and potentially find existing drugs that may activate the same combination of pathways. They construct large tissue specific GRNs in relevant human tissues (heart, muscle, stem cells) using GEOracle derived perturbation based data to integrate and validate mechanistic edges from a scaffold of existing network resources like TF networks and PPI data.

The researchers apply the MURSS algorithm to the tissue specific GRNs to find upstream regulators of the genes and pathways identified from the compound signatures. Several regulators are significant across different tissues and for both compound signatures, providing a small number of high confidence predicted regulators of the desired gene signatures

in the human heart.

The researches search the literature, including various drug-target databases and GEO for compounds that interfere with or alter the expression of those key regulators, revealing several promising candidate drugs for pre-clinical validation studies.

Bibliography

Abzhanov, A., Kuo, W. P., Hartmann, C., Grant, B. R., Grant, P. R., and Tabin, C. J. (2006). The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches. *Nature*, **442**(7102), 563–567.

Al-Zaidy, R. A. and Giles, C. L. (2015). Automatic Extraction of Data from Bar Charts. pages 1–4. ACM Press.

Alexa, A. and Rahnenfuhrer, J. (2010). *topGO: topGO: Enrichment analysis for Gene Ontology*. R package version 2.18.0.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3), 403–410.

Alvarez-Saavedra, M., Carrasco, L., Sura-Trueba, S., Demarchi Aiello, V., Walz, K., Neto, J. X., and Young, J. I. (2010). Elevated expression of MeCP2 in cardiac and skeletal tissues is detrimental for normal development. *Hum Mol Genet*, **19**(11), 2177–2190.

Amthor, H., Christ, B., Weil, M., and Patel, K. (1998). The importance of timing differentiation during limb muscle development. *Current Biology*, **8**(11), 642–652.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**(2), 166–169.

Asakura, A., Komaki, M., and Rudnicki, M. (2001). Muscle satellite cells are multipotential stem cells that exhibit myogenic, osteogenic, and adipogenic differentiation. *Differentiation; Research in Biological Diversity*, **68**(4-5), 245–253.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**(1), 25–29.

Attrill, H., Falls, K., Goodman, J. L., Millburn, G. H., Antonazzo, G., Rey, A. J., Marygold, S. J., and the FlyBase consortium (2016). FlyBase: establishing a Gene Group resource for *Drosophila melanogaster*. *Nucleic Acids Research*, **44**(D1), D786–D792.

Azuaje, F., Zhang, L., Jeanty, C., Puhl, S.-L., Rodius, S., and Wagner, D. R. (2013). Analysis of a gene co-expression network establishes robust association between Col5a2 and ischemic heart disease. *BMC Med Genomics*, **6**, 13.

Babiarz, J. E., Ravon, M., Sridhar, S., Ravindran, P., Swanson, B., Bitter, H., Weiser, T., Chiao, E., Certa, U., and Kolaja, K. L. (2012). Determination of the Human Cardiomyocyte mRNA and miRNA Differentiation Network by Fine-Scale Profiling. *Stem Cells and Development*, **21**(11), 1956–1965.

Baker, E. J., Jay, J. J., Bubier, J. A., Langston, M. A., and Chesler, E. J. (2012). GeneWeaver: a web-based system for integrative functional genomics. *Nucleic Acids Research*, **40**(D1), D1067–D1076.

Baker, M. (2011). Making sense of chromatin states. *Nature Methods*, **8**(9), 717–722.

Ballouz, S., Liu, J. Y., George, R. A., Bains, N., Liu, A., Oti, M., Gaeta, B., Fatkin, D., and Wouters, M. A. (2013). Gentrepid V2.0: a web server for candidate disease gene prediction. *BMC Bioinformatics*, **14**(1), 249.

Ballouz, S., Liu, J. Y., Oti, M., Gaeta, B., Fatkin, D., Bahlo, M., and Wouters, M. A. (2014). Candidate disease gene prediction using *Gentrepid* : application to a genome-wide association study on coronary artery disease. *Molecular Genetics & Genomic Medicine*, **2**(1), 44–57.

Bandyopadhyay, A., Tsuji, K., Cox, K., Harfe, B. D., Rosen, V., and Tabin, C. J. (2006). Genetic Analysis of the Roles of BMP2, BMP4, and BMP7 in Limb Patterning and Skeletogenesis. *PLoS Genetics*, **2**(12), e216.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Molecular Systems Biology*, **3**, 78.

Barabasi, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat Rev Genet*, **12**(1), 56–68.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Research*, **41**(D1), D991–D995.

Barriot, R., Breckpot, J., Thienpont, B., Brohée, S., Van Vooren, S., Coessens, B., Tranchevent, L.-C., Van Loo, P., Gewillig, M., Devriendt, K., and Moreau, Y. (2010). Collaboratively charting the gene-to-phenotype network of human congenital heart defects. *Genome Med*, **2**(3), 16.

Barth, A. S. (2005). Reprogramming of the Human Atrial Transcriptome in Permanent Atrial Fibrillation: Expression of a Ventricular-Like Genomic Signature. *Circulation Research*, **96**(9), 1022–1029.

Barzel, B. and Barabási, A.-L. (2013). Network link prediction by global silencing of indirect correlations. *Nature Biotechnology*, **31**(8), 720–725.

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2002). The Infinite Hidden Markov Model. In *Advances in Neural Information Processing Systems*, volume 14, pages 577–584. MIT Press.

Beck, T., Hastings, R. K., Gollapudi, S., Free, R. C., and Brookes, A. J. (2014). GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *European Journal of Human Genetics*, **22**(7), 949–952.

Beckmann, J. S., Estivill, X., and Antonarakis, S. E. (2007). Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nat Rev Genet*, **8**(8), 639–646.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, **57**(1), 289–300.

Berger, S. I., Ma'ayan, A., and Iyengar, R. (2010). Systems pharmacology of arrhythmias. *Sci Signal*, **3**(118), ra30.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, **28**(1), 235–242.

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L., and Lander, E. S. (2006). A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell*, **125**(2), 315–326.

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen, T. S., and Thomson, J. A. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, **28**(10), 1045–1048.

Berul, C. I., Maguire, C. T., Aronovitz, M. J., Greenwood, J., Miller, C., Gehrmann, J., Housman, D., Mendelsohn, M. E., and Reddy, S. (1999). DMPK dosage alterations result in atrioventricular conduction abnormalities in a mouse myotonic dystrophy model. *Journal of Clinical Investigation*, **103**(4), R1–R7.

Biesinger, J., Wang, Y., and Xie, X. (2013). Discovering and mapping chromatin states using a tree hidden Markov model. *BMC Bioinformatics*, **14**(Suppl 5), S4.

Biressi, S., Miyabara, E. H., Gopinath, S. D., M. Carlig, P. M., and Rando, T. A. (2014). A Wnt-TGF 2 axis induces a fibrogenic program in muscle stem cells from dystrophic mice. *Science Translational Medicine*, **6**(267), 267ra176–267ra176.

Blue, G. M., Kirk, E. P., Sholler, G. F., Harvey, R. P., and Winlaw, D. S. (2012). Congenital heart disease: current knowledge about causes and inheritance. *The Medical Journal of Australia*, **197**(3), 155–159.

Brack, A. S., Conboy, M. J., Roy, S., Lee, M., Kuo, C. J., Keller, C., and Rando, T. A. (2007). Increased Wnt Signaling During Aging Alters Muscle Stem Cell Fate and Increases Fibrosis. *Science*, **317**(5839), 807–810.

Brack, A. S., Conboy, I. M., Conboy, M. J., Shen, J., and Rando, T. A. (2008). A Temporal Switch from Notch to Wnt Signaling in Muscle Stem Cells Is Necessary for Normal Adult Myogenesis. *Cell Stem Cell*, **2**(1), 50–59.

Brack, A. S., Murphy-Seiler, F., Hanifi, J., Deka, J., Eyckerman, S., Keller, C., Aguet, M., and Rando, T. A. (2009). BCL9 is an essential component of canonical Wnt signaling that mediates the differentiation of myogenic progenitors during muscle regeneration. *Developmental Biology*, **335**(1), 93–105.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C. P., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics*, **29**(4), 365–371.

Breckpot, J., Thienpont, B., Bauters, M., Tranchevent, L.-C., Gewillig, M., Allegaert, K., Vermeesch, J. R., Moreau, Y., and Devriendt, K. (2012). Congenital heart defects in a novel recurrent 22q11.2 deletion harboring the genes CRKL and MAPK1. *American Journal of Medical Genetics Part A*, **158A**(3), 574–580.

Breitling, R., Armengaud, P., Amtmann, A., and Herzyk, P. (2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters*, **573**(1-3), 83–92.

Britto, R., Sallou, O., Collin, O., Michaux, G., Primig, M., and Chalmel, F. (2012). GPSy: a cross-species gene prioritization system for conserved biological processes–application in male gamete development. *Nucleic Acids Research*, **40**(W1), W458–W465.

Buckingham, M., Meilhac, S., and Zaffran, S. (2005). Building the mammalian heart from two sources of myocardial cells. *Nature Reviews Genetics*, **6**(11), 826–837.

Cai, C.-L., Liang, X., Shi, Y., Chu, P.-H., Pfaff, S. L., Chen, J., and Evans, S. (2003). Isl1 identifies a cardiac progenitor population that proliferates prior to differentiation and contributes a majority of cells to the heart. *Developmental Cell*, **5**(6), 877–889.

Cañestro, C., Albalat, R., Irimia, M., and Garcia-Fernàndez, J. (2013). Impact of gene gains, losses and duplication modes on the origin and diversification of vertebrates. *Seminars in Cell & Developmental Biology*, **24**(2), 83–94.

Canver, M. C., Smith, E. C., Sher, F., Pinello, L., Sanjana, N. E., Shalem, O., Chen, D. D., Schupp, P. G., Vinjamur, D. S., Garcia, S. P., Luc, S., Kurita, R., Nakamura, Y., Fujiwara, Y., Maeda, T., Yuan, G.-C., Zhang, F., Orkin, S. H., and Bauer, D. E. (2015). BCL11a enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature*, **527**(7577), 192–197.

Cappola, T. P., Li, M., He, J., Ky, B., Gilmore, J., Qu, L., Keating, B., Reilly, M., Kim, C. E., Glessner, J., Frackelton, E., Hakonarson, H., Syed, F., Hindes, A., Matkovich, S. J., Cresci, S., and Dorn, 2nd, G. W. (2010). Common variants in HSPB7 and FRMD4b associated with advanced heart failure. *Circ Cardiovasc Genet*, **3**(2), 147–154.

Chamberlain, A. A., Lin, M., Lister, R. L., Maslov, A. A., Wang, Y., Suzuki, M., Wu, B., Greally, J. M., Zheng, D., and Zhou, B. (2014). DNA methylation is developmentally regulated for genes essential for cardiogenesis. *J Am Heart Assoc*, **3**(3), e000976.

Chang, S., McKinsey, T. A., Zhang, C. L., Richardson, J. A., Hill, J. A., and Olson, E. N. (2004). Histone deacetylases 5 and 9 govern responsiveness of the heart to a subset of stress signals and play redundant roles in heart development. *Mol Cell Biol*, **24**(19), 8467–8476.

Charville, G. W., Cheung, T. H., Yoo, B., Santos, P. J., Lee, G. K., Shrager, J. B., and Rando, T. A. (2015). Ex Vivo Expansion and In Vivo Self-Renewal of Human Muscle Stem Cells. *Stem Cell Reports*, **5**(4), 621–632.

Chatr-aryamontri, A., Breitkreutz, B.-J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., O'Donnell, L., and al, e. (2012). The BioGRID interaction database: 2013 update. *Nucleic Acids Research*, **41**(D1), D816–D823.

Chen, H. and VanBuren, V. (2014). A provisional gene regulatory atlas for mouse heart development. *PLoS One*, **9**(1), e83364.

Chen, J. C. J. and Goldhamer, D. J. (2003). Skeletal muscle stem cells. *Reproductive biology and endocrinology: RB&E*, **1**, 101.

Chen, S., Sanjana, N. E., Zheng, K., Shalem, O., Lee, K., Shi, X., Scott, D. A., Song, J., Pan, J. Q., Weissleder, R., Lee, H., Zhang, F., and Sharp, P. A. (2015). Genome-wide CRISPR Screen in a Mouse Model of Tumor Growth and Metastasis. *Cell*, **160**(6), 1246–1260.

Chen, Y.-H., Xu, S.-J., Bendahhou, S., Wang, X.-L., Wang, Y., Xu, W.-Y., Jin, H.-W., Sun, H., Su, X.-Y., Zhuang, Q.-N., Yang, Y.-Q., Li, Y.-B., Liu, Y., Xu, H.-J., Li, X.-F., Ma, N., Mou, C.-P., Chen, Z., Barhanin, J., and Huang, W. (2003). KCNQ1 gain-of-function mutation in familial atrial fibrillation. *Science*, **299**(5604), 251–254.

Chowdhury, S., Erickson, S. W., MacLeod, S. L., Cleves, M. A., Hu, P., Karim, M. A., and Hobbs, C. A. (2011). Maternal genome-wide DNA methylation patterns and congenital heart defects. *PLoS One*, **6**(1), e16506.

Cirulli, E. T. and Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet*, **11**(6), 415–425.

Claussnitzer, M., Dankel, S. N., Klocke, B., Grallert, H., Glunk, V., Berulava, T., Lee, H., Oskolkov, N., Fadista, J., Ehlers, K., Wahl, S., Hoffmann, C., Qian, K., Rönn, T., Riess, H., Müller-Nurasyid, M., Bretschneider, N., Schroeder, T., Skurk, T., Horsthemke, B., , D. I. A. G. R. A. M., Spieler, D., Klingenspor, M., Seifert, M., Kern, M. J., Mejhert, N., Dahlman, I., Hansson, O., Hauck, S. M., Blüher, M., Arner, P., Groop, L., Illig, T., Suhre, K., Hsu, Y.-H., Mellgren, G., Hauner, H., and Laumen, H. (2014). Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms. *Cell*, **156**(1-2), 343–358.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*.

Darehzereshki, A., Rubin, N., Gamba, L., Kim, J., Fraser, J., Huang, Y., Billings, J., Mohammadzadeh, R., Wood, J., Warburton, D., Kaartinen, V., and Lien, C.-L. (2015). Differential regenerative capacity of neonatal mouse hearts after cryoinjury. *Developmental Biology*, **399**(1), 91–99.

Davidson, E. H. (2006). *The regulatory genome: gene regulatory networks in development and evolution*. Academic Press, Amsterdam.

Davidson, E. H. (2010). Emerging properties of animal gene regulatory networks. *Nature*, **468**(7326), 911–920.

Davis, S. and Meltzer, P. S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, **23**(14), 1846–1847.

De Jong, A. M., Maass, A. H., Oberdorf-Maass, S. U., Van Veldhuisen, D. J., Van Gilst, W. H., and Van Gelder, I. C. (2011). Mechanisms of atrial structural changes caused by stretch occurring before and during early atrial fibrillation. *Cardiovascular Research*, **89**(4), 754–765.

de la Pompa, J. L. and Epstein, J. A. (2012). Coordinating tissue interactions: Notch signaling in cardiac development and disease. *Dev Cell*, **22**(2), 244–254.

Deb, A. (2014). Cell-cell interaction in the heart via Wnt/β-catenin pathway after cardiac injury. *Cardiovasc Res*, **102**(2), 214–223.

Delgado-Olguin, P., Huang, Y., Li, X., Christodoulou, D., Seidman, C. E., Seidman, J. G., Tarakhovsky, A., and Bruneau, B. G. (2012). Epigenetic repression of cardiac progenitor gene expression by Ezh2 is required for postnatal cardiac homeostasis. *Nat Genet*, **44**(3), 343–347.

Deo, R. and Albert, C. M. (2012). Epidemiology and genetics of sudden cardiac death. *Circulation*, **125**(4), 620–637.

Dewey, F. E., Perez, M. V., Wheeler, M. T., Watt, C., Spin, J., Langfelder, P., Horvath, S., Hannenhalli, S., Cappola, T. P., and Ashley, E. A. (2011). Gene coexpression network topology of cardiac development, hypertrophy, and failure. *Circ Cardiovasc Genet*, **4**(1), 26–35.

Dhawan, J. and Rando, T. A. (2005). Stem cells in postnatal myogenesis: molecular mechanisms of satellite cell quiescence, activation and replenishment. *Trends in Cell Biology*, **15**(12), 666–673.

Dickel, D. E., Zhu, Y., Nord, A. S., Wylie, J. N., Akiyama, J. A., Afzal, V., Plajzer-Frick, I., Kirkpatrick, A., Göttgens, B., Bruneau, B. G., Visel, A., and Pennacchio, L. A. (2014). Function-based identification of mammalian enhancers using site-specific integration. *Nat Methods*, **11**(5), 566–571.

Dickerson, J. E., Zhu, A., Robertson, D. L., and Hentges, K. E. (2011). Defining the Role of Essential Genes in Human Disease. *PLoS ONE*, **6**(11), e27368.

Djordjevic, D., Yang, A., Zadoorian, A., Rungrugeecharoen, K., and Ho, J. W. K. (2014). How Difficult Is Inference of Mammalian Causal Gene Regulatory Networks? *PLOS ONE*, **9**(11), e111661.

Djordjevic, D., Deshpande, V., Szczesnik, T., Yang, A., Humphreys, D. T., Giannoulatou, E., and Ho, J. W. K. (2015). Decoding the complex genetic causes of heart diseases using systems biology. *Biophysical Reviews*, **7**(1), 141–159.

Djordjevic, D., Kusumi, K., and Ho, J. W. K. (2016). XGSA: A statistical method for cross-species gene set analysis. *Bioinformatics*, **32**(17), i620–i628.

Dominguez, A. A., Lim, W. A., and Qi, L. S. (2015). Beyond editing: repurposing CRISPR–Cas9 for precision genome regulation and interrogation. *Nature Reviews Molecular Cell Biology*, **17**(1), 5–15.

Dumas, J., Gargano, M. A., and Dancik, G. M. (2016). shinyGEO: a web-based application for analyzing gene expression omnibus datasets. *Bioinformatics*, **32**(23), 3679–3681.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B.-K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., Shoresh, N., Simon, J. M., Song, L., Trinklein, N. D., Altshuler, R. C., Birney, E., Brown, J. B., Cheng, C., Djebali, S., Dong, X., Dunham, I., Ernst, J., Furey, T. S., Gerstein, M., Giardine, B., Greven, M., Hardison, R. C., Harris, R. S., Herrero, J., Hoffman, M. M., Iyer, S., Kellis, M., Khatun, J., Kheradpour, P., Kundaje, A., Lassmann, T., Li, Q., Lin, X., Marinov, G. K., Merkel, A., Mortazavi, A., Parker, S. C. J., Reddy, T. E., Rozowsky, J., Schlesinger, F., Thurman, R. E., Wang, J., Ward, L. D., Whitfield, T. W., Wilder, S. P., Wu, W., Xi, H. S., Yip, K. Y., Zhuang, J., Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., Snyder, M., Pazin, M. J., Lowdon, R. F., Dillon, L. A. L., Adams, L. B., Kelly, C. J., Zhang, J., Wexler, J. R., Green, E. D., Good, P. J., Feingold, E. A., Bernstein, B. E., Birney, E., Crawford, G. E., Dekker, J., Elnitski, L., Farnham,

P. J., Gerstein, M., Giddings, M. C., Gingeras, T. R., Green, E. D., Guigó, R., Hardison, R. C., Hubbard, T. J., Kellis, M., Kent, W. J., Lieb, J. D., Margulies, E. H., Myers, R. M., Snyder, M., Stamatoyannopoulos, J. A., Tenenbaum, S. A., Weng, Z., White, K. P., Wold, B., Khatun, J., Yu, Y., Wrobel, J., Risk, B. A., Gunawardena, H. P., Kuiper, H. C., Maier, C. W., Xie, L., Chen, X., Giddings, M. C., Bernstein, B. E., Epstein, C. B., Shoresh, N., Ernst, J., Kheradpour, P., Mikkelsen, T. S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M. J., Durham, T., Ku, M., Truong, T., Ward, L. D., Altshuler, R. C., Eaton, M. L., Kellis, M., Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Batut, P., Bell, I., Bell, K., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H. P., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Li, G., Luo, O. J., Park, E., Preall, J. B., Presaud, K., Ribeca, P., Risk, B. A., Robyr, D., Ruan, X., Sammeth, M., Sandhu, K. S., Schaeffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T. J., Reymond, A., Antonarakis, S. E., Hannon, G. J., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., Gingeras, T. R., Rosenbloom, K. R., Sloan, C. A., Learned, K., Malladi, V. S., Wong, M. C., Barber, G. P., Cline, M. S., Dreszer, T. R., Heitner, S. G., Karolchik, D., Kent, W. J., Kirkup, V. M., Meyer, L. R., Long, J. C., Maddren, M., Raney, B. J., Furey, T. S., Song, L., Grasfeder, L. L., Giresi, P. G., Lee, B.-K., Battenhouse, A., Sheffield, N. C., Simon, J. M., Showers, K. A., Safi, A., London, D., Bhinge, A. A., Shestak, C., Schaner, M. R., Ki Kim, S., Zhang, Z. Z., Mieczkowski, P. A., Mieczkowska, J. O., Liu, Z., McDaniell, R. M., Ni, Y., Rashid, N. U., Kim, M. J., Adar, S., Zhang, Z., Wang, T., Winter, D., Keefe, D., Birney, E., Iyer, V. R., Lieb, J. D., Crawford, G. E., Li, G., Sandhu, K. S., Zheng, M., Wang, P., Luo, O. J., Shahab, A., Fullwood, M. J., Ruan, X., Ruan, Y., Myers, R. M., Pauli, F., Williams, B. A., Gertz, J., Marinov, G. K., Reddy, T. E., Vielmetter, J., Partridge, E., Trout, D., Varley, K. E., Gasper, C., Bansal, A., Pepke, S., Jain, P., Amrhein, H., Bowling, K. M., Anaya, M., Cross, M. K., King, B., Muratet, M. A., Antoshechkin, I., Newberry, K. M., McCue, K., Nesmith, A. S., Fisher-Aylor, K. I., Pusey, B., DeSalvo, G., Parker, S. L., Balasubramanian, S., Davis, N. S., Meadows, S. K., Eggleston, T., Gunter, C., Newberry, J. S., Levy, S. E., Absher, D. M., Mortazavi, A., Wong, W. H., Wold, B., Blow, M. J., Visel, A., Pennachio, L. A., Elnitski, L., Margulies, E. H., Parker, S. C. J., Petrykowska, H. M., Abyzov, A., Aken, B., Barrell, D., Barson, G., Berry, A., Bignell, A., Boychenko, V., Bussotti, G., Chrast, J., Davidson, C., Derrien, T., Despacio-Reyes, G., Diekhans, M., Ezkurdia, I., Frankish, A., Gilbert, J., Gonzalez, J. M., Griffiths, E., Harte, R., Hendrix, D. A., Howald, C., Hunt, T., Jungreis, I., Kay, M., Khurana, E., Kokocinski, F., Leng, J., Lin, M. F., Loveland, J., Lu, Z., Manthravadi, D., Mariotti, M., Mudge, J., Mukherjee, G., Notredame, C., Pei, B., Rodriguez, J. M., Saunders, G., Sboner, A., Searle, S., Sisu, C., Snow, C., Steward, C., Tanzer, A., Tapanari, E., Tress, M. L., van Baren, M. J., Walters, N (2012). An integrated encyclopedia of DNA

elements in the human genome. *Nature*, **489**(7414), 57–74.

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, **4**(8), 1184–1191.

Echeverri, K. and Tanaka, E. M. (2002). Mechanisms of muscle dedifferentiation during regeneration. *Seminars in Cell & Developmental Biology*, **13**(5), 353–360.

Eckalbar, W. L., Lasku, E., Infante, C. R., Elsey, R. M., Markov, G. J., Allen, A. N., Corneveaux, J. J., Losos, J. B., DeNardo, D. F., Huentelman, M. J., Wilson-Rawls, J., Rawls, A., and Kusumi, K. (2012). Somitogenesis in the anole lizard and alligator reveals evolutionary convergence and divergence in the amniote segmentation clock. *Developmental Biology*, **363**(1), 308–319.

Edgar, R. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, **30**(1), 207–210.

Edmondson, D. G. and Olson, E. N. (1989). A gene with homology to the myc similarity region of MyoD1 is expressed during myogenesis and is sufficient to activate the muscle differentiation program. *Genes & Development*, **3**(5), 628–640.

Ehrenberg, M. (2003). Systems Biology Is Taking Off. *Genome Research*, **13**(11), 2377–2380.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet*, **11**(6), 446–450.

Ernst, J. and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, **28**(8), 817–825.

Ernst, J. and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, **9**(3), 215–216.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**(7345), 43–49.

Fahed, A. C., Gelb, B. D., Seidman, J. G., and Seidman, C. E. (2013). Genetics of congenital heart disease: the glass half empty. *Circ Res*, **112**(4), 707–720.

Fakhro, K. A., Choi, M., Ware, S. M., Belmont, J. W., Towbin, J. A., Lifton, R. P., Khokha, M. K., and Brueckner, M. (2011). Rare copy number variations in congenital heart disease patients identify unique genes in left-right patterning. *Proceedings of the National Academy of Sciences*, **108**(7), 2915–2920.

Fisher, R. E., Geiger, L. A., Stroik, L. K., Hutchins, E. D., George, R. M., Denardo, D. F., Kusumi, K., Rawls, J. A., and Wilson-Rawls, J. (2012). A Histological Comparison of the Original and Regenerated Tail in the Green Anole, *Anolis carolinensis*. *The Anatomical Record: Advances in Integrative Anatomy and Evolutionary Biology*, **295**(10), 1609–1619.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., and al, e. (2012). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, **41**(D1), D808–D815.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, **7**(3-4), 601–620.

Friedrichs, M., Wirsdöerfer, F., Flohé, S. B., Schneider, S., Wuelling, M., and Vortkamp, A. (2011). BMP signaling balances proliferation and differentiation of muscle satellite cell descendants. *BMC Cell Biology*, **12**(1), 26.

Fröhlich, H., Sahin, Ö., Arlt, D., Bender, C., and Beißbarth, T. (2009a). Deterministic Effects Propagation Networks for reconstructing protein signaling networks from multiple interventions. *BMC Bioinformatics*, **10**(1), 322.

Fröhlich, H., Tresch, A., and Beißbarth, T. (2009b). Nested effects models for learning signaling networks from perturbation data. *Biom. J.*, **51**(2), 304–323.

Fukushima, K., Badlani, N., Usas, A., Riano, F., Fu, F., and Huard, J. (2001). The use of an antifibrosis agent to improve muscle recovery after laceration. *The American Journal of Sports Medicine*, **29**(4), 394–402.

Galeota, E. and Pelizzola, M. (2016). Ontology-based annotations and semantic relations in large-scale (epi)genomics data. *Briefings in Bioinformatics*, page bbw036.

Gao, G., Zhang, X.-F., Hubbell, K., and Cui, X. (2017). NR2f2 regulates chondrogenesis of human mesenchymal stem cells in bioprinted cartilage: NR2f2 Regulates Chondrogenesis of Human MSCs. *Biotechnology and Bioengineering*, **114**(1), 208–216.

Gelb, B. D. and Chung, W. K. (2014). Complex Genetics and the Etiology of Human Congenital Heart Disease. *Cold Spring Harb Perspect Med*, **4**(7).

George, R. M., Hahn, K. L., Rawls, A., Viger, R. S., and Wilson-Rawls, J. (2015). Notch signaling represses GATA4-induced expression of genes involved in steroid biosynthesis. *Reproduction*, **150**(4), 383–394.

Gerstein, M. B., Rozowsky, J., Yan, K.-K., Wang, D., Cheng, C., Brown, J. B., Davis, C. A., Hillier, L., Sisu, C., Li, J. J., Pei, B., Harmanci, A. O., Duff, M. O., Djebali, S., Alexander, R. P., Alver, B. H., Auerbach, R., Bell, K., Bickel, P. J., Boeck, M. E., Boley, N. P., Booth, B. W., Cherbas, L., Cherbas, P., Di, C., Dobin, A., Drenkow, J., Ewing, B., Fang, G., Fastuca, M., Feingold, E. A., Frankish, A., Gao, G., Good, P. J., Guigó, R., Hammonds, A., Harrow, J., Hoskins, R. A., Howald, C., Hu, L., Huang,

H., Hubbard, T. J. P., Huynh, C., Jha, S., Kasper, D., Kato, M., Kaufman, T. C., Kitchen, R. R., Ladewig, E., Lagarde, J., Lai, E., Leng, J., Lu, Z., MacCoss, M., May, G., McWhirter, R., Merrihew, G., Miller, D. M., Mortazavi, A., Murad, R., Oliver, B., Olson, S., Park, P. J., Pazin, M. J., Perrimon, N., Pervouchine, D., Reinke, V., Reymond, A., Robinson, G., Samsonova, A., Saunders, G. I., Schlesinger, F., Sethi, A., Slack, F. J., Spencer, W. C., Stoiber, M. H., Strasbourger, P., Tanzer, A., Thompson, O. A., Wan, K. H., Wang, G., Wang, H., Watkins, K. L., Wen, J., Wen, K., Xue, C., Yang, L., Yip, K., Zaleski, C., Zhang, Y., Zheng, H., Brenner, S. E., Graveley, B. R., Celniker, S. E., Gingeras, T. R., and Waterston, R. (2014). Comparative analysis of the transcriptome across distant species. *Nature*, **512**(7515), 445–448.

Gilbert, E. A. B., Payne, S. L., and Vickaryous, M. K. (2013). The Anatomy and Histology of Caudal Autotomy and Regeneration in Lizards. *Physiological and Biochemical Zoology*, **86**(6), 631–644.

Gill, R., Hitchins, L., Fletcher, F., and Dhoot, G. K. (2010). Sulf1a and HGF regulate satellite-cell growth. *Journal of Cell Science*, **123**(Pt 11), 1873–1883.

Glass, K., Huttenhower, C., Quackenbush, J., and Yuan, G.-C. (2013). Passing Messages between Biological Networks to Refine Predicted Interactions. *PLoS ONE*, **8**(5), e64832.

Glukhov, A. V., Fedorov, V. V., Kalish, P. W., Ravikumar, V. K., Lou, Q., Janks, D., Schuessler, R. B., Moazami, N., and Efimov, I. R. (2012). Conduction Remodeling in Human End-Stage Nonischemic Left Ventricular Cardiomyopathy. *Circulation*, **125**(15), 1835–1847.

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. (2007). The human disease network. *Proc Natl Acad Sci U S A*, **104**(21), 8685–8690.

Gohin, M., Bobe, J., and Chesnel, F. (2010). Comparative transcriptomic analysis of follicle-enclosed oocyte maturational and developmental competence acquisition in two non-mammalian vertebrates. *BMC Genomics*, **11**(1), 18.

Greenway, S. C., Pereira, A. C., Lin, J. C., DePalma, S. R., Israel, S. J., Mesquita, S. M., Ergul, E., Conta, J. H., Korn, J. M., McCarroll, S. A., Gorham, J. M., Gabriel, S., Altshuler, D. M., Quintanilla-Dieck, M. d. L., Artunduaga, M. A., Eavey, R. D., Plenge, R. M., Shadick, N. A., Weinblatt, M. E., De Jager, P. L., Hafler, D. A., Breitbart, R. E., Seidman, J. G., and Seidman, C. E. (2009). De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot. *Nat Genet*, **41**(8), 931–935.

Gudbjartsson, D. F., Holm, H., Gretarsdottir, S., Thorleifsson, G., Walters, G. B., Thorgeirsson, G., Gulcher, J., Mathiesen, E. B., Njølstad, I., Nyrnes, A., Wilsgaard, T., Hald, E. M., Hveem, K., Stoltenberg, C., Kucera, G., Stubblefield, T., Carter, S., Roden, D., Ng, M. C. Y., Baum, L., So, W. Y., Wong, K. S., Chan, J. C. N., Gieger, C., Wichmann, H.-E., Gschwendtner, A., Dichgans, M., Kuhlenbäumer, G., Berger, K., Ringelstein, E. B., Bevan, S., Markus, H. S., Kostulas, K., Hillert, J., Sveinbjörnsdóttir, S., Valdimarsson, E. M., Løchen, M.-L., Ma, R. C. W., Darbar, D., Kong, A., Arnar, D. O., Thorsteinsdottir, U., and Stefansson, K. (2009). A sequence variant in ZFHX3 on 16q22 associates with atrial fibrillation and ischemic stroke. *Nat Genet*, **41**(8), 876–878.

Guo, Y., Ma, L., Cristofanilli, M., Hart, R., Hao, A., and Schachner, M. (2011). Transcription factor Sox11b is involved in spinal cord regeneration in adult zebrafish. *Neuroscience*, **172**, 329–341.

Gusterson, R. J., Jazrawi, E., Adcock, I. M., and Latchman, D. S. (2003). The transcriptional co-activators CREB-binding protein (CBP) and p300 play a critical role in cardiac hypertrophy that is dependent on their histone acetyltransferase activity. *J Biol Chem*, **278**(9), 6838–6847.

Haas, A. R. and Tuan, R. S. (1999). Chondrogenic differentiation of murine C3h10t1/2 multipotential mesenchymal cells: II. Stimulation by bone morphogenetic protein-2 requires modulation of N-cadherin expression and function. *Differentiation*, **64**(2), 77–89.

Haldar, S. M., Lu, Y., Jeyaraj, D., Kawanami, D., Cui, Y., Eapen, S. J., Hao, C., Li, Y., Doughman, Y. Q., Watanabe, M., Shimizu, K., Kuivaniemi, H., Sadoshima, J., Margulies, K. B., Cappola, T. P., and Jain, M. K. (2010). Klf15 Deficiency Is a Molecular Link Between Heart Failure and Aortic Aneurysm Formation. *Science Translational Medicine*, **2**(26), 26ra26–26ra26.

Hamosh, A. (2004). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, **33**(Database issue), D514–D517.

Han, H., Shim, H., Shin, D., Shim, J. E., Ko, Y., Shin, J., Kim, H., Cho, A., Kim, E., Lee, T., Kim, H., Kim, K., Yang, S., Bae, D., Yun, A., Kim, S., Kim, C. Y., Cho, H. J., Kang, B., Shin, S., and Lee, I. (2015). TRRUST: a reference database of human transcriptional regulatory interactions. *Scientific Reports*, **5**(1).

Han, P., Hang, C. T., Yang, J., and Chang, C.-P. (2011). Chromatin remodeling in cardiovascular development and physiology. *Circ Res*, **108**(3), 378–396.

Hang, C. T., Yang, J., Han, P., Cheng, H.-L., Shang, C., Ashley, E., Zhou, B., and Chang, C.-P. (2010). Chromatin regulation by Brg1 underlies heart muscle development and disease. *Nature*, **466**(7302), 62–67.

He, A., Kong, S. W., Ma, Q., and Pu, W. T. (2011a). Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proc Natl Acad Sci U S A*, **108**(14), 5632–5637.

He, A., Ma, Q., Cao, J., von Gise, A., Zhou, P., Xie, H., Zhang, B., Hsing, M., Christodoulou, D. C., Cahan, P., Daley, G. Q., Kong, S. W., Orkin, S. H., Seidman, C. E., Seidman, J. G., and Pu, W. T. (2012). Polycomb repressive complex 2 regulates normal development of the mouse heart. *Circ Res*, **110**(3), 406–415.

He, D., Liu, Z.-P., and Chen, L. (2011b). Identification of dysfunctional modules and disease genes in congenital heart disease by a network-based approach. *BMC Genomics*, **12**, 592.

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., Ching, K. A., Antosiewicz-Bourget, J. E., Liu, H., Zhang, X., Green, R. D., Lobanenkov, V. V., Stewart, R., Thomson, J. A., Crawford, G. E., Kellis, M., and Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**(7243), 108–112.

Hershberger, R. E., Hedges, D. J., and Morales, A. (2013). Dilated cardiomyopathy: the complexity of a diverse genetic architecture. *Nature Reviews. Cardiology*, **10**(9), 531–547.

Hill, S. M., Heiser, L. M., Cokelaer, T., Unger, M., Nesser, N. K., Carlin, D. E., Zhang, Y., Sokolov, A., Paull, E. O., Wong, C. K., Graim, K., Bivol, A., Wang, H., Zhu, F., Afsari, B., Danilova, L. V., Favorov, A. V., Lee, W. S., Taylor, D., Hu, C. W., Long, B. L., Noren, D. P., Bisberg, A. J., Afsari, B., Al-Ouran, R., Anton, B., Arodz, T., Sichani, O. A., Bagheri, N., Berlow, N., Bisberg, A. J., Bivol, A., Bohler, A., Bonet, J., Bonneau, R., Budak, G., Bunescu, R., Caglar, M., Cai, B., Cai, C., Carlin, D. E., Carlon, A., Chen, L., Ciaccio, M. F., Cokelaer, T., Cooper, G., Creighton, C. J., Daneshmand, S.-M.-H., de la Fuente, A., Di Camillo, B., Danilova, L. V., Dutta-Moscato, J., Emmett, K., Evelo, C., Fassia, M.-K. H., Favorov, A. V., Fertig, E. J., Finkle, J. D., Finotello, F., Friend, S., Gao, X., Gao, J., Garcia-Garcia, J., Ghosh, S., Giaretta, A., Graim, K., Gray, J. W., Großeholz, R., Guan, Y., Guinney, J., Hafemeister, C., Hahn, O., Haider, S., Hase, T., Heiser, L. M., Hill, S. M., Hodgson, J., Hoff, B., Hsu, C. H., Hu, C. W., Hu, Y., Huang, X., Jalili, M., Jiang, X., Kacprowski, T., Kaderali, L., Kang, M., Kannan, V., Kellen, M., Kikuchi, K., Kim, D.-C., Kitano, H., Knapp, B., Komatsoulis, G., Koeppl, H., Krämer, A., Kursa, M. B., Kutmon, M., Lee, W. S., Li, Y., Liang, X., Liu, Z., Liu, Y., Long, B. L., Lu, S., Lu, X., Manfrini, M., Matos, M. R. A., Meerzaman, D., Mills, G. B., Min, W., Mukherjee, S., Müller, C. L., Neapolitan, R. E., Nesser, N. K., Noren, D. P., Norman, T., Oliva, B., Opiyo, S. O., Pal, R., Palinkas, A., Paull, E. O., Planas-Iglesias, J., Poglayen, D., Qutub, A. A., Saez-Rodriguez, J., Sambo, F., Sanavia, T., Sharifi-Zarchi, A., Slawek, J., Sokolov, A., Song, M., Spellman, P. T., Streck, A., Stolovitzky, G., Strunz, S., Stuart, J. M., Taylor, D., Tegnér, J., Thobe, K., Toffolo, G. M., Trifoglio, E., Unger, M., Wan, Q., Wang, H., Welch, L., Wong, C. K., Wu, J. J., Xue, A. Y., Yamanaka, R., Yan, C., Zairis, S., Zengerling, M., Zenil, H., Zhang, S., Zhang, Y., Zhu, F., Zi, Z., Mills, G. B., Gray, J. W., Kellen, M., Norman, T., Friend, S., Qutub, A. A., Fertig, E. J., Guan, Y., Song, M., Stuart, J. M., Spellman, P. T., Koeppl, H., Stolovitzky, G., Saez-Rodriguez, J., and Mukherjee, S. (2016). Inferring causal molecular networks: empirical assessment through a community-based effort. *Nature Methods*, **13**(4), 310–318.

Hirsinger, E., Duprez, D., Jouve, C., Malapert, P., Cooke, J., and Pourquié, O. (1997). Noggin acts downstream of Wnt and Sonic Hedgehog to antagonize BMP4 in avian somite patterning. *Development (Cambridge, England)*, **124**(22), 4605–4614.

Ho, J. W. K. (2012). Application of a systems approach to study developmental gene regulation. *Biophysical Reviews*, **4**(3), 245–253.

Ho, J. W. K., Stefani, M., dos Remedios, C. G., and Charleston, M. A. (2008). Dif-

ferential variability analysis of gene expression and its application to human diseases. *Bioinformatics*, **24**(13), i390–i398.

Ho, J. W. K., Jung, Y. L., Liu, T., Alver, B. H., Lee, S., Ikegami, K., Sohn, K.-A., Minoda, A., Tolstorukov, M. Y., Appert, A., Parker, S. C. J., Gu, T., Kundaje, A., Riddle, N. C., Bishop, E., Egelhofer, T. A., Hu, S. S., Alekseyenko, A. A., Rechtsteiner, A., Asker, D., Belsky, J. A., Bowman, S. K., Chen, Q. B., Chen, R. A.-J., Day, D. S., Dong, Y., Dose, A. C., Duan, X., Epstein, C. B., Ercan, S., Feingold, E. A., Ferrari, F., Garrigues, J. M., Gehlenborg, N., Good, P. J., Haseley, P., He, D., Herrmann, M., Hoffman, M. M., Jeffers, T. E., Kharchenko, P. V., Kolasinska-Zwierz, P., Kotwaliwale, C. V., Kumar, N., Langley, S. A., Larschan, E. N., Latorre, I., Libbrecht, M. W., Lin, X., Park, R., Pazin, M. J., Pham, H. N., Plachetka, A., Qin, B., Schwartz, Y. B., Shoresh, N., Stempor, P., Vielle, A., Wang, C., Whittle, C. M., Xue, H., Kingston, R. E., Kim, J. H., Bernstein, B. E., Dernburg, A. F., Pirrotta, V., Kuroda, M. I., Noble, W. S., Tullius, T. D., Kellis, M., MacAlpine, D. M., Strome, S., Elgin, S. C. R., Liu, X. S., Lieb, J. D., Ahringer, J., Karpen, G. H., and Park, P. J. (2014). Comparative analysis of metazoan chromatin organization. *Nature*, **512**(7515), 449–452.

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., and Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, **9**(5), 473–476.

Holterman, C. E., Le Grand, F., Kuang, S., Seale, P., and Rudnicki, M. A. (2007). Megf10 regulates the progression of the satellite cell myogenic program. *The Journal of Cell Biology*, **179**(5), 911–922.

Hon, G., Ren, B., and Wang, W. (2008). Chromasig: A probabilistic approach to finding common chromatin signatures in the human genome. *PLOS Computational Biology*, **4**(10), 1–16.

Hoogaars, W. M., Engel, A., Brons, J. F., Verkerk, A. O., de Lange, F. J., Wong, L. E., Bakker, M. L., Clout, D. E., Wakker, V., Barnett, P., Ravesloot, J. H., Moorman, A. F., Verheijck, E. E., and Christoffels, V. M. (2007). Tbx3 controls the sinoatrial node gene program and imposes pacemaker function on the atria. *Genes & Development*, **21**(9), 1098–1112.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, **37**(1), 1–13.

Hui, S. P., Sengupta, D., Lee, S. G. P., Sen, T., Kundu, S., Mathavan, S., and Ghosh, S. (2014). Genome Wide Expression Profiling during Spinal Cord Regeneration Identifies Comprehensive Cellular Responses in Zebrafish. *PLoS ONE*, **9**(1), e84212.

Hutchins, E. D., Markov, G. J., Eckalbar, W. L., George, R. M., King, J. M., Tokuyama, M. A., Geiger, L. A., Emmert, N., Ammar, M. J., Allen, A. N., Siniard, A. L., Corneveaux, J. J., Fisher, R. E., Wade, J., DeNardo, D. F., Rawls, J. A., Huentelman,

M. J., Wilson-Rawls, J., and Kusumi, K. (2014). Transcriptomic Analysis of Tail Regeneration in the Lizard Anolis carolinensis Reveals Activation of Conserved Vertebrate Developmental and Repair Mechanisms. *PLoS ONE*, **9**(8), e105004.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS ONE*, **5**(9), e12776.

Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*, **2**, 343–372.

Ishii, M. (2005). Combined deficiencies of Msx1 and Msx2 cause impaired patterning and survival of the cranial neural crest. *Development*, **132**(22), 4937–4950.

Jeong, J. Y., Kim, J. M., Rajesh, R. V., Suresh, S., Jang, G. W., Lee, K.-T., Kim, T. H., Park, M., Jeong, H. J., Kim, K. W., Cho, Y. M., and Lee, H.-J. (2013). Transcriptional Profiling of Differentially Expressed Genes in Porcine Satellite Cell. *Reproductive & Developmental Biology*, **37**(4), 233–245.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**(6096), 816–821.

Jopling, C., Sleep, E., Raya, M., Martí, M., Raya, A., and Izpisúa Belmonte, J. C. (2010). Zebrafish heart regeneration occurs by cardiomyocyte dedifferentiation and proliferation. *Nature*, **464**(7288), 606–609.

Julienne, H., Zoufir, A., Audit, B., and Arneodo, A. (2013). Human genome replication proceeds through four chromatin states. *PLoS computational biology*, **9**(10), e1003233.

Kacprowski, T., Doncheva, N. T., and Albrecht, M. (2013). NetworkPrioritizer: a versatile tool for network-based prioritization of candidate disease genes or other molecules. *Bioinformatics*, **29**(11), 1471–1473.

Kahn, E. B. and Simpson, S. B. (1974). Satellite cells in mature, uninjured skeletal muscle of the lizard tail. *Developmental Biology*, **37**(1), 219–223.

Kamisago, M., Sharma, S. D., DePalma, S. R., Solomon, S., Sharma, P., McDonough, B., Smoot, L., Mullen, M. P., Woolf, P. K., Wigle, E. D., Seidman, J., Jarcho, J., Shapiro, L. R., and Seidman, C. E. (2000). Mutations in Sarcomere Protein Genes as a Cause of Dilated Cardiomyopathy. *New England Journal of Medicine*, **343**(23), 1688–1696.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2013). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research*, **42**(D1), D199–D205.

Kang, H., Choi, I., Cho, S., Ryu, D., Lee, S., and Kim, W. (2014). gsGator: an integrated web platform for cross-species gene set analysis. *BMC bioinformatics*, **15**, 13.

Katagiri, T., Yamaguchi, A., Komaki, M., Abe, E., Takahashi, N., Ikeda, T., Rosen, V., Wozney, J. M., Fujisawa-Sehara, A., and Suda, T. (1994). Bone morphogenetic protein-2 converts the differentiation pathway of C2c12 myoblasts into the osteoblast lineage. *The Journal of Cell Biology*, **127**(6 Pt 1), 1755–1766.

Keating, M., Atkinson, D., Dunn, C., Timothy, K., Vincent, G. M., and Leppert, M. (1991). Linkage of a cardiac arrhythmia, the long QT syndrome, and the Harvey ras-1 gene. *Science*, **252**(5006), 704–706.

Kelder, T., van Iersel, M. P., Hanspers, K., Kutmon, M., Conklin, B. R., Evelo, C. T., and Pico, A. R. (2011). WikiPathways: building research communities on biological pathways. *Nucleic Acids Research*, **40**(D1), D1301–D1307.

Kemp, T., Sadusky, T., Saltisi, F., Carey, N., Moss, J., Yang, S., Sassoon, D., Goldspink, G., and Coulton, G. (2000). Identification of Ankrd2, a Novel Skeletal Muscle Gene Coding for a Stretch-Responsive Ankyrin-Repeat Protein. *Genomics*, **66**(3), 229–241.

Kharchenko, P. V., Alekseyenko, A. A., Schwartz, Y. B., Minoda, A., Riddle, N. C., Ernst, J., Sabo, P. J., Larschan, E., Gorchakov, A. A., Gu, T., Linder-Basso, D., Plachetka, A., Shanower, G., Tolstorukov, M. Y., Luquette, L. J., Xi, R., Jung, Y. L., Park, R. W., Bishop, E. P., Canfield, T. K., Sandstrom, R., Thurman, R. E., MacAlpine, D. M., Stamatoyannopoulos, J. A., Kellis, M., Elgin, S. C. R., Kuroda, M. I., Pirrotta, V., Karpen, G. H., and Park, P. J. (2011). Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature*, **471**(7339), 480–485.

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, **12**(4), 357–360.

Kimura, A., Harada, H., Park, J. E., Nishi, H., Satoh, M., Takahashi, M., Hiroi, S., Sasaoka, T., Ohbuchi, N., Nakamura, T., Koyanagi, T., Hwang, T. H., Choo, J. A., Chung, K. S., Hasegawa, A., Nagai, R., Okazaki, O., Nakamura, H., Matsuzaki, M., Sakamoto, T., Toshima, H., Koga, Y., Imaizumi, T., and Sasazuki, T. (1997). Mutations in the cardiac troponin I gene associated with hypertrophic cardiomyopathy. *Nat Genet*, **16**(4), 379–382.

King, J. A., Marker, P. C., Seung, K. J., and Kingsley, D. M. (1994). BMP5 and the Molecular, Skeletal, and Soft-Tissue Alterations in short ear Mice. *Developmental Biology*, **166**(1), 112–122.

King, M. and Wilson, A. (1975). Evolution at two levels in humans and chimpanzees. *Science*, **188**(4184), 107–116.

Kitano, H. (2002). Computational systems biology. *Nature*, **420**(6912), 206–210.

Klattenhoff, C. A., Scheuermann, J. C., Surface, L. E., Bradley, R. K., Fields, P. A., Steinhauser, M. L., Ding, H., Butty, V. L., Torrey, L., Haas, S., Abo, R., Tabebordbar, M., Lee, R. T., Burge, C. B., and Boyer, L. A. (2013). Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. *Cell*, **152**(3), 570–583.

Knippenberg, M., Helder, M., Zandieh Doulabi, B., Wuisman, P., and Klein-Nulend, J. (2006). Osteogenesis versus chondrogenesis by BMP-2 and BMP-7 in adipose stem cells. *Biochemical and Biophysical Research Communications*, **342**(3), 902–908.

Knopp, P., Figeac, N., Fortier, M., Moyle, L., and Zammit, P. S. (2013). Pitx genes are redeployed in adult myogenesis where they can act to promote myogenic differentiation in muscle satellite cells. *Developmental Biology*, **377**(1), 293–304.

Kohler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the Interactome for Prioritization of Candidate Disease Genes. *The American Journal of Human Genetics*, **82**(4), 949 – 958.

Konermann, S., Brigham, M. D., Trevino, A. E., Joung, J., Abudayyeh, O. O., Barcena, C., Hsu, P. D., Habib, N., Gootenberg, J. S., Nishimasu, H., Nureki, O., and Zhang, F. (2014). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*, **517**(7536), 583–588.

Koshiba-Takeuchi, K., Mori, A. D., Kaynak, B. L., Cebra-Thomas, J., Sukonnik, T., Georges, R. O., Latham, S., Beck, L., Henkelman, R. M., Black, B. L., Olson, E. N., Wade, J., Takeuchi, J. K., Nemer, M., Gilbert, S. F., and Bruneau, B. G. (2009). Reptilian heart development and the molecular basis of cardiac chamber evolution. *Nature*, **461**(7260), 95–98.

Kragl, M., Knapp, D., Nacu, E., Khattak, S., Maden, M., Epperlein, H. H., and Tanaka, E. M. (2009). Cells keep a memory of their tissue origin during axolotl limb regeneration. *Nature*, **460**(7251), 60–65.

Krajinovic, M., Pinamonti, B., Sinagra, G., Vatta, M., Severini, G. M., Milasin, J., Falaschi, A., Camerini, F., Giacca, M., and Mestroni, L. (1995). Linkage of familial dilated cardiomyopathy to chromosome 9. Heart Muscle Disease Study Group. *Am J Hum Genet*, **57**(4), 846–852.

Kramer, A., Green, J., Pollard, J., and Tugendreich, S. (2014). Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics*, **30**(4), 523–530.

Krauthammer, M., Kaufmann, C. A., Gilliam, T. C., and Rzhetsky, A. (2004). Molecular triangulation: Bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proceedings of the National Academy of Sciences*, **101**(42), 15148–15153.

Kristiansson, E., Österlund, T., Gunnarsson, L., Arne, G., Joakim Larsson, D. G., and Nerman, O. (2013). A novel method for cross-species gene expression analysis. *BMC Bioinformatics*, **14**(1), 70.

Kriventseva, E. V., Tegenfeldt, F., Petty, T. J., Waterhouse, R. M., Simao, F. A., Pozdnyakov, I. A., Ioannidis, P., and Zdobnov, E. M. (2015). OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Research*, **43**(D1), D250–D256.

Kuhn, T., Nagy, M., Luong, T., and Krauthammer, M. (2014). Mining images in biomedical publications: Detection and analysis of gel diagrams. *Journal of Biomedical Semantics*, **5**(1), 10.

Kumar, A., Velloso, C. P., Imokawa, Y., and Brockes, J. P. (2004). The Regenerative Plasticity of Isolated Urodele Myofibers and Its Dependence on Msx1. *PLoS Biology*, **2**(8), e218.

Labbé, R. M., Irimia, M., Currie, K. W., Lin, A., Zhu, S. J., Brown, D. D. R., Ross, E. J., Voisin, V., Bader, G. D., Blencowe, B. J., and Pearson, B. J. (2012). A comparative transcriptomic analysis reveals conserved features of stem cell pluripotency in planarians and mammals. *Stem Cells (Dayton, Ohio)*, **30**(8), 1734–1745.

Lachke, S. A., Ho, J. W. K., Kryukov, G. V., O'Connell, D. J., Aboukhalil, A., Bulyk, M. L., Park, P. J., and Maas, R. L. (2012). *iSyTE* : I ntegrated Sy stems T ool for E ye Gene Discovery. *Investigative Opthalmology & Visual Science*, **53**(3), 1617.

Lage, K., Møllg\a ard, K., Greenway, S., Wakimoto, H., Gorham, J. M., Workman, C. T., Bendsen, E., Hansen, N. T., Rigina, O., Roque, F. S., Wiese, C., Christoffels, V. M., Roberts, A. E., Smoot, L. B., Pu, W. T., Donahoe, P. K., Tommerup, N., Brunak, S., Seidman, C. E., Seidman, J. G., and Larsen, L. A. (2010). Dissecting spatio-temporal protein networks driving human heart development and related disorders. *Mol Syst Biol*, **6**, 381.

Lage, K., Greenway, S. C., Rosenfeld, J. A., Wakimoto, H., Gorham, J. M., Segrè, A. V., Roberts, A. E., Smoot, L. B., Pu, W. T., Pereira, A. C., Mesquita, S. M., Tommerup, N., Brunak, S., Ballif, B. C., Shaffer, L. G., Donahoe, P. K., Daly, M. J., Seidman, J. G., Seidman, C. E., and Larsen, L. A. (2012). Genetic and environmental risk factors in congenital heart disease functionally converge in protein networks driving heart development. *Proc Natl Acad Sci U S A*, **109**(35), 14035–14040.

Larson, J. L., Huttenhower, C., Quackenbush, J., and Yuan, G.-C. (2013). A tiered hidden Markov model characterizes multi-scale chromatin states. *Genomics*, **102**(1), 1–7.

Le, H.-S., Oltvai, Z. N., and Bar-Joseph, Z. (2010). Cross-species queries of large gene expression databases. *Bioinformatics*, **26**(19), 2416–2423.

Lee, E. J., Malik, A., Pokharel, S., Ahmad, S., Mir, B. A., Cho, K. H., Kim, J., Kong, J. C., Lee, D.-M., Chung, K. Y., Kim, S. H., and Choi, I. (2014). Identification of Genes Differentially Expressed in Myogenin Knock-Down Bovine Muscle Satellite Cells during Differentiation through RNA Sequencing Analysis. *PLoS ONE*, **9**(3), e92447.

Lee, S.-J. and McPherron, A. C. (2001). Regulation of myostatin activity and muscle growth. *Proceedings of the National Academy of Sciences*, **98**(16), 9306–9311.

Lee, S.-J., Huynh, T. V., Lee, Y.-S., Sebald, S. M., Wilcox-Adelman, S. A., Iwamori, N., Lepper, C., Matzuk, M. M., and Fan, C.-M. (2012). Role of satellite cells versus myofibers in muscle hypertrophy induced by inhibition of the myostatin/activin signaling pathway. *Proceedings of the National Academy of Sciences*, **109**(35), E2353–E2360.

Leinonen, R., Sugawara, H., Shumway, M., and on behalf of the International Nucleotide Sequence Database Collaboration (2011). The Sequence Read Archive. *Nucleic Acids Research*, **39**(Database), D19–D21.

Lepper, C., Partridge, T. A., and Fan, C.-M. (2011). An absolute requirement for Pax7-positive satellite cells in acute injury-induced skeletal muscle regeneration. *Development*, **138**(17), 3639–3646.

Levine, M. and Davidson, E. H. (2005). Gene regulatory networks for development. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(14), 4936–4942.

Li, T.-R. and White, K. P. (2003). Tissue-Specific Gene Expression and Ecdysone-Regulated Genomic Networks in Drosophila. *Developmental Cell*, **5**(1), 59–72.

Li, W., Chen, L., He, W., Li, W., Qu, X., Liang, B., Gao, Q., Feng, C., Jia, X., Lv, Y., Zhang, S., and Li, X. (2013). Prioritizing Disease Candidate Proteins in Cardiomyopathy-Specific Protein-Protein Interaction Networks Based on "Guilt by Association?? Analysis. *PLoS ONE*, **8**(8), e71191.

Li, X., McFarland, D. C., and Velleman, S. G. (2008). Extracellular matrix proteoglycan decorin-mediated myogenic satellite cell responsiveness to transforming growth factor-β1 during cell proliferation and differentiation. *Domestic Animal Endocrinology*, **35**(3), 263–273.

Li, X., Martinez-Fernandez, A., Hartjes, K. A., Kocher, J.-P. A., Olson, T. M., Terzic, A., and Nelson, T. J. (2014a). Transcriptional atlas of cardiogenesis maps congenital heart disease interactome. *Physiol Genomics*, **46**(13), 482–495.

Li, Y., Rivera, C. M., Ishii, H., Jin, F., Selvaraj, S., Lee, A. Y., Dixon, J. R., and Ren, B. (2014b). CRISPR Reveals a Distal Super-Enhancer Required for Sox2 Expression in Mouse Embryonic Stem Cells. *PLoS ONE*, **9**(12), e114485.

Liao, J., Hu, N., Zhou, N., Lin, L., Zhao, C., Yi, S., Fan, T., Bao, W., Liang, X., Chen, H., Xu, W., Chen, C., Cheng, Q., Zeng, Y., Si, W., Yang, Z., and Huang, W. (2014). Sox9 Potentiates BMP2-Induced Chondrogenic Differentiation and Inhibits BMP2-Induced Osteogenic Differentiation. *PLoS ONE*, **9**(2), e89025.

Lin, C.-C., Hsiang, J.-T., Wu, C.-Y., Oyang, Y.-J., Juan, H.-F., and Huang, H.-C. (2010). Dynamic functional modules in co-expressed protein interaction networks of dilated cardiomyopathy. *BMC Syst Biol*, **4**, 138.

Lin, Q., Schwarz, J., Bucana, C., and Olson, E. N. (1997). Control of mouse cardiac morphogenesis and myogenesis by transcription factor MEF2c. *Science (New York, N.Y.)*, **276**(5317), 1404–1407.

Liu, T., Rechtsteiner, A., Egelhofer, T. A., Vielle, A., Latorre, I., Cheung, M.-S., Ercan, S., Ikegami, K., Jensen, M., Kolasinska-Zwierz, P., Rosenbaum, H., Shin, H., Taing, S., Takasaki, T., Iniguez, A. L., Desai, A., Dernburg, A. F., Kimura, H., Lieb, J. D.,

Ahringer, J., Strome, S., and Liu, X. S. (2011). Broad chromosomal domains of histone modification patterns in C. elegans. *Genome Research*, **21**(2), 227–236.

Lokody, I. (2014). Genetic therapies: Correcting genetic defects with CRISPR-Cas9. *Nat Rev Genet*, **15**(2), 63.

Lopes, R., Korkmaz, G., and Agami, R. (2016). Applying CRISPR–Cas9 tools to identify and characterize transcriptional enhancers. *Nature Reviews Molecular Cell Biology*, **17**(9), 597–604.

Love, N. R., Chen, Y., Bonev, B., Gilchrist, M. J., Fairclough, L., Lea, R., Mohun, T. J., Paredes, R., Zeef, L. A., and Amaya, E. (2011). Genome-wide analysis of gene expression during Xenopus tropicalis tadpole tail regeneration. *BMC Developmental Biology*, **11**(1), 70.

Lozano-Velasco, E., Contreras, A., Crist, C., Hernández-Torres, F., Franco, D., and Aránega, A. E. (2011). Pitx2c modulates Pax3+/Pax7+ cell populations and regulates Pax3 expression by repressing miR27 expression during myogenesis. *Developmental Biology*, **357**(1), 165–178.

Lu, Y., Huggins, P., and Bar-Joseph, Z. (2009). Cross species analysis of microarray expression data. *Bioinformatics*, **25**(12), 1476–1483.

Lu, Y., Rosenfeld, R., Nau, G. J., and Bar-Joseph, Z. (2010). Cross Species Expression Analysis of Innate Immune Response. *Journal of Computational Biology*, **17**(3), 253–268.

Lyons, I., Parsons, L. M., Hartley, L., Li, R., Andrews, J. E., Robb, L., and Harvey, R. P. (1995). Myogenic and morphogenetic defects in the heart tubes of murine embryos lacking the homeo box gene Nkx2-5. *Genes & Development*, **9**(13), 1654–1666.

Maathuis, M. H., Colombo, D., Kalisch, M., and Bühlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nature Methods*, **7**(4), 247–248.

MacLellan, W. R., Wang, Y., and Lusis, A. J. (2012). Systems-based approaches to cardiovascular disease. *Nature Reviews Cardiology*, **9**(3), 172–184.

MacRae, C. A. (2010). The Genetics of Congestive Heart Failure. *Heart Failure Clinics*, **6**(2), 223–230.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, **461**(7265), 747–753.

Mar, J. C., Matigian, N. A., Mackay-Sim, A., Mellick, G. D., Sue, C. M., Silburn, P. A., McGrath, J. J., Quackenbush, J., and Wells, C. A. (2011). Variance of Gene Expression

Identifies Altered Network Constraints in Neurological Disease. *PLoS Genetics*, **7**(8), e1002207.

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences*, **107**(14), 6286–6291.

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Consortium, D., Kellis, M., Collins, J. J., and Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature methods*, **9**(8), 796–804.

Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., and Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature Methods*, **13**(4), 366–370.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006a). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, **7 Suppl 1**, S7.

Margolin, A. A., Wang, K., Lim, W. K., Kustagi, M., Nemenman, I., and Califano, A. (2006b). Reverse engineering cellular networks. *Nature protocols*, **1**(2), 662–671.

Markowetz, F., Bloch, J., and Spang, R. (2005). Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics*, **21**(21), 4026–4032.

Martínez-Fernandez, S., Hernández-Torres, F., Franco, D., Lyons, G. E., Navarro, F., and Aránega, A. E. (2006). Pitx2c overexpression promotes cell proliferation and arrests differentiation in myoblasts. *Developmental Dynamics*, **235**(11), 2930–2939.

Matkovich, S. J., Van Booven, D. J., Eschenbacher, W. H., and Dorn, 2nd, G. W. (2011). RISC RNA sequencing for context-specific identification of in vivo microRNA targets. *Circ Res*, **108**(1), 18–26.

McCroskery, S. (2005). Improved muscle healing through enhanced regeneration and reduced fibrosis in myostatin-null mice. *Journal of Cell Science*, **118**(15), 3531–3541.

McCulley, D. J. and Black, B. L. (2012). Transcription factor pathways and congenital heart disease. *Curr Top Dev Biol*, **100**, 253–277.

McFarlane, C., Hui, G. Z., Amanda, W. Z. W., Lau, H. Y., Lokireddy, S., XiaoJia, G., Mouly, V., Butler-Browne, G., Gluckman, P. D., Sharma, M., and Kambadur, R. (2011). Human myostatin negatively regulates human myoblast growth and differentiation. *AJP: Cell Physiology*, **301**(1), C195–C203.

Merlo, M., Pyxaras, S. A., Pinamonti, B., Barbati, G., Di Lenarda, A., and Sinagra, G. (2011). Prevalence and Prognostic Significance of Left Ventricular Reverse Remodeling in Dilated Cardiomyopathy Receiving Tailored Medical Treatment. *Journal of the American College of Cardiology*, **57**(13), 1468–1476.

Messina, D. N., Speer, M. C., Pericak-Vance, M. A., and McNally, E. M. (1997). Linkage of familial dilated cardiomyopathy with conduction defect and muscular dystrophy to chromosome 6q23. *Am J Hum Genet*, **61**(4), 909–917.

Meyer, A. and Van de Peer, Y. (2005). From 2r to 3r: evidence for a fish-specific genome duplication (FSGD). *BioEssays*, **27**(9), 937–945.

Meyer, P. E., Lafitte, F., and Bontempi, G. (2008). minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information. *BMC Bioinformatics*, **9**(1), 461.

Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T., and Thomas, P. D. (2016). PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Research*, **44**(D1), D336–342.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R. P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E. S., and Bernstein, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**(7153), 553–560.

Min, K.-D., Asakura, M., Liao, Y., Nakamaru, K., Okazaki, H., Takahashi, T., Fujimoto, K., Ito, S., Takahashi, A., Asanuma, H., Yamazaki, S., Minamino, T., Sanada, S., Seguchi, O., Nakano, A., Ando, Y., Otsuka, T., Furukawa, H., Isomura, T., Takashima, S., Mochizuki, N., and Kitakaze, M. (2010). Identification of genes related to heart failure using global gene expression profiling of human failing myocardium. *Biochemical and Biophysical Research Communications*, **393**(1), 55–60.

Miyamoto, S., Kawamura, T., Morimoto, T., Ono, K., Wada, H., Kawase, Y., Matsumori, A., Nishio, R., Kita, T., and Hasegawa, K. (2006). Histone acetyltransferase activity of p300 is required for the promotion of left ventricular remodeling after myocardial infarction in adult mice in vivo. *Circulation*, **113**(5), 679–690.

Mohamed, T. M. A., Abou-Leisa, R., Stafford, N., Maqsood, A., Zi, M., Prehar, S., Baudoin-Stanley, F., Wang, X., Neyses, L., Cartwright, E. J., and Oceandy, D. (2016). The plasma membrane calcium ATPase 4 signalling in cardiac fibroblasts mediates cardiomyocyte hypertrophy. *Nature Communications*, **7**, 11074.

Molkentin, J. D., Black, B. L., Martin, J. F., and Olson, E. N. (1995). Cooperative activation of muscle gene expression by MEF2 and myogenic bHLH proteins. *Cell*, **83**(7), 1125–1136.

Molkentin, J. D., Lin, Q., Duncan, S. A., and Olson, E. N. (1997). Requirement of the transcription factor GATA4 for heart tube formation and ventral morphogenesis. *Genes & Development*, **11**(8), 1061–1072.

Montgomery, R. L., Potthoff, M. J., Haberland, M., Qi, X., Matsuzaki, S., Humphries, K. M., Richardson, J. A., Bassel-Duby, R., and Olson, E. N. (2008). Maintenance of cardiac energy metabolism by histone deacetylase 3 in mice. *J Clin Invest*, **118**(11), 3588–3597.

Mora, A. and Donaldson, I. M. (2011). iRefR: an R package to manipulate the iRefIndex consolidated protein interaction database. *BMC Bioinformatics*, **12**(1), 455.

Myerburg, R. J. (2001). Sudden cardiac death: exploring the limits of our knowledge. *Journal of Cardiovascular Electrophysiology*, **12**(3), 369–381.

Myocardial Infarction Genetics Consortium, Kathiresan, S., Voight, B. F., Purcell, S., Musunuru, K., Ardissino, D., Mannucci, P. M., Anand, S., Engert, J. C., Samani, N. J., Schunkert, H., Erdmann, J., Reilly, M. P., Rader, D. J., Morgan, T., Spertus, J. A., Stoll, M., Girelli, D., McKeown, P. P., Patterson, C. C., Siscovick, D. S., O'Donnell, C. J., Elosua, R., Peltonen, L., Salomaa, V., Schwartz, S. M., Melander, O., Altshuler, D., Ardissino, D., Merlini, P. A., Berzuini, C., Bernardinelli, L., Peyvandi, F., Tubaro, M., Celli, P., Ferrario, M., Fetiveau, R., Marziliano, N., Casari, G., Galli, M., Ribichini, F., Rossi, M., Bernardi, F., Zonzin, P., Piazza, A., Mannucci, P. M., Schwartz, S. M., Siscovick, D. S., Yee, J., Friedlander, Y., Elosua, R., Marrugat, J., Lucas, G., Subirana, I., Sala, J., Ramos, R., Kathiresan, S., Meigs, J. B., Williams, G., Nathan, D. M., MacRae, C. A., O'Donnell, C. J., Salomaa, V., Havulinna, A. S., Peltonen, L., Melander, O., Berglund, G., Voight, B. F., Kathiresan, S., Hirschhorn, J. N., Asselta, R., Duga, S., Spreafico, M., Musunuru, K., Daly, M. J., Purcell, S., Voight, B. F., Purcell, S., Nemesh, J., Korn, J. M., McCarroll, S. A., Schwartz, S. M., Yee, J., Kathiresan, S., Lucas, G., Subirana, I., Elosua, R., Surti, A., Guiducci, C., Gianniny, L., Mirel, D., Parkin, M., Burtt, N., Gabriel, S. B., Samani, N. J., Thompson, J. R., Braund, P. S., Wright, B. J., Balmforth, A. J., Ball, S. G., Hall, A. S., , W. T. C. C. C., Schunkert, H., Erdmann, J., Linsel-Nitschke, P., Lieb, W., Ziegler, A., König, I., Hengstenberg, C., Fischer, M., Stark, K., Grosshennig, A., Preuss, M., Wichmann, H.-E., Schreiber, S., Schunkert, H., Samani, N. J., Erdmann, J., Ouwehand, W., Hengstenberg, C., Deloukas, P., Scholz, M., Cambien, F., Reilly, M. P., Li, M., Chen, Z., Wilensky, R., Matthai, W., Qasim, A., Hakonarson, H. H., Devaney, J., Burnett, M.-S., Pichard, A. D., Kent, K. M., Satler, L., Lindsay, J. M., Waksman, R., Knouff, C. W., Waterworth, D. M., Walker, M. C., Mooser, V., Epstein, S. E., Rader, D. J., Scheffold, T., Berger, K., Stoll, M., Huge, A., Girelli, D., Martinelli, N., Olivieri, O., Corrocher, R., Morgan, T., Spertus, J. A., McKeown, P., Patterson, C. C., Schunkert, H., Erdmann, E., Linsel-Nitschke, P., Lieb, W., Ziegler, A., König, I. R., Hengstenberg, C., Fischer, M., Stark, K., Grosshennig, A., Preuss, M., Wichmann, H.-E., Schreiber, S., Hólm, H., Thorleifsson, G., Thorsteinsdottir, U., Stefansson, K., Engert, J. C., Do, R., Xie, C., Anand, S., Kathiresan, S., Ardissino, D., Mannucci, P. M., Siscovick, D., O'Donnell, C. J., Samani, N. J., Melander, O., Elosua, R., Peltonen, L., Salomaa, V., Schwartz, S. M., and Altshuler, D. (2009). Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet*, **41**(3), 334–341.

Nandal, U. K., Kampen, A. H. C. v., and Moerland, P. D. (2016). compendiumdb: an R package for retrieval and storage of functional genomics data. *Bioinformatics*, **32**(18), 2856–2857.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search

for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**(3), 443–453.

Nie, J., Jiang, M., Zhang, X., Tang, H., Jin, H., Huang, X., Yuan, B., Zhang, C., Lai, J. C., Nagamine, Y., Pan, D., Wang, W., and Yang, Z. (2015). Post-transcriptional Regulation of Nkx2-5 by RHAU in Heart Development. *Cell Reports*, **13**(4), 723–732.

O'Connell, D. J., Ho, J. W. K., Mammoto, T., Turbe-Doan, A., O'Connell, J. T., Haseley, P. S., Koo, S., Kamiya, N., Ingber, D. E., Park, P. J., and Maas, R. L. (2012). A Wnt-bmp feedback circuit controls intertissue signaling dynamics in tooth organogenesis. *Science signaling*, **5**(206), ra4.

Odelberg, S. J., Kollhoff, A., and Keating, M. T. (2000). Dedifferentiation of mammalian myotubes induced by msx1. *Cell*, **103**(7), 1099–1109.

Odom, D. T. (2004). Control of Pancreas and Liver Gene Expression by HNF Transcription Factors. *Science*, **303**(5662), 1378–1381.

Olsen, C., Fleming, K., Prendergast, N., Rubio, R., Emmert-Streib, F., Bontempi, G., Haibe-Kains, B., and Quackenbush, J. (2014). Inference and validation of predictive gene networks from biomedical literature and gene expression data. *Genomics*, **103**(5-6), 329–336.

Ono, Y., Calhabeu, F., Morgan, J. E., Katagiri, T., Amthor, H., and Zammit, P. S. (2011). BMP signalling permits population expansion by preventing premature myogenic differentiation in muscle satellite cells. *Cell Death and Differentiation*, **18**(2), 222–234.

Ozeki, N., Jethanandani, P., Nakamura, H., Ziober, B. L., and Kramer, R. H. (2007). Modulation of satellite cell adhesion and motility following BMP2-induced differentiation to osteoblast lineage. *Biochemical and Biophysical Research Communications*, **353**(1), 54–59.

Padmanabhan, N., Jia, D., Geary-Joo, C., Wu, X., Ferguson-Smith, A. C., Fung, E., Bieda, M. C., Snyder, F. F., Gravel, R. A., Cross, J. C., and Watson, E. D. (2013). Mutation in folate metabolism causes epigenetic instability and transgenerational effects on development. *Cell*, **155**(1), 81–93.

Padovan-Merhar, O. and Raj, A. (2013). Using variability in gene expression as a tool for studying gene regulation: Characterizing gene regulation using expression variability. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, **5**(6), 751–759.

Palumbo-Zerr, K., Zerr, P., Distler, A., Fliehr, J., Mancuso, R., Huang, J., Mielenz, D., Tomcik, M., Fürnrohr, B. G., Scholtysek, C., Dees, C., Beyer, C., Krönke, G., Metzger, D., Distler, O., Schett, G., and Distler, J. H. W. (2015). Orphan nuclear receptor NR4a1 regulates transforming growth factor-β signaling and fibrosis. *Nature Medicine*, **21**(2), 150–158.

Parikh, J. R., Klinger, B., Xia, Y., Marto, J. A., and Blüthgen, N. (2010). Discovering causal signaling pathways through gene-expression patterns. *Nucleic Acids Research*, **38**(suppl 2), W109–W117.

Park, A., Won, S. T., Pentecost, M., Bartkowski, W., and Lee, B. (2014). CRISPR/Cas9 Allows Efficient and Complete Knock-In of a Destabilization Domain-Tagged Essential Protein in a Human Cell Line, Allowing Rapid Knockdown of Protein Function. *PLoS ONE*, **9**(4), e95101.

Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics*, **10**(10), 669–680.

Patterson, A. J., Chen, M., Xue, Q., Xiao, D., and Zhang, L. (2010). Chronic prenatal hypoxia induces epigenetic programming of PKCepsilon gene repression in rat hearts. *Circ Res*, **107**(3), 365–373.

Pauling, L. and Itano, H. A. (1949). Sickle cell anemia a molecular disease. *Science*, **110**(2865), 543–548.

Pérez de Castro, I., Aguirre-Portolés, C., Fernández-Miranda, G., Cañamero, M., Cowley, D. O., Van Dyke, T., and Malumbres, M. (2013). Requirements for Aurora-A in tissue regeneration and tumor development in adult mammals. *Cancer Research*, **73**(22), 6804–6815.

Peterkin, T., Gibson, A., and Patient, R. (2003). GATA-6 maintains BMP-4 and Nkx2 expression during cardiomyocyte precursor maturation. *The EMBO journal*, **22**(16), 4260–4273.

Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S., and Goldstein, D. B. (2013). Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genetics*, **9**(8), e1003709.

Pinna, A., Soranzo, N., and de la Fuente, A. (2010). From Knockouts to Networks: Establishing Direct Cause-Effect Relationships through Graph Analysis. *PLoS ONE*, **5**(10), e12912.

Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., and Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, **21**(3), 447–455.

Porrello, E. R., Mahmoud, A. I., Simpson, E., Hill, J. A., Richardson, J. A., Olson, E. N., and Sadek, H. A. (2011). Transient Regenerative Potential of the Neonatal Mouse Heart. *Science*, **331**(6020), 1078–1080.

Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., Gabaldón, T., Rattei, T., Creevey, C., Kuhn, M., Jensen, L. J., von Mering, C., and Bork, P. (2014). eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Research*, **42**(D1), D231–D239.

Raines, A. M., Magella, B., Adam, M., and Potter, S. S. (2015). Key pathways regulated by HoxA9,10,11/HoxD9,10,11 during limb development. *BMC Developmental Biology*, **15**(1).

Rajendran, R., Gopal, S., Masood, H., Vivek, P., and Deb, K. (2013). Regenerative potential of dental pulp mesenchymal stem cells harvested from high caries patient's teeth. *Journal of Stem Cells*, **8**(1), 25–41.

Ran, F. A., Hsu, P. D., Lin, C.-Y., Gootenberg, J. S., Konermann, S., Trevino, A. E., Scott, D. A., Inoue, A., Matoba, S., Zhang, Y., and Zhang, F. (2013). Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell*, **154**(6), 1380–1389.

Re'em-Kalma, Y., Lamb, T., and Frank, D. (1995). Competition between noggin and bone morphogenetic protein 4 activities may regulate dorsalization during Xenopus development. *Proceedings of the National Academy of Sciences of the United States of America*, **92**(26), 12141–12145.

Reimand, J., Kull, M., Peterson, H., Hansen, J., and Vilo, J. (2007). g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Research*, **35**(Web Server issue), W193–200.

Relaix, F. and Zammit, P. S. (2012). Satellite cells are essential for skeletal muscle regeneration: the cell on the edge returns centre stage. *Development*, **139**(16), 2845–2856.

Reshef, R., Maroto, M., and Lassar, A. B. (1998). Regulation of dorsal somitic cell fates: BMPs and Noggin control the timing and pattern of myogenic regulator expression. *Genes & Development*, **12**(3), 290–303.

Rhodes, D. R., Kalyana-Sundaram, S., Tomlins, S. A., Mahavisno, V., Kasper, N., Varambally, R., Barrette, T. R., Ghosh, D., Varambally, S., and Chinnaiyan, A. M. (2007). Molecular concepts analysis links tumors, pathways, mechanisms, and drugs. *Neoplasia (New York, N.Y.)*, **9**(5), 443–454.

Richardson, L., Venkataraman, S., Stevenson, P., Yang, Y., Moss, J., Graham, L., Burton, N., Hill, B., Rao, J., Baldock, R. A., and Armit, C. (2014). EMAGE mouse embryo spatial gene expression database: 2014 update. *Nucleic Acids Research*, **42**(D1), D835–D844.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**(7), e47–e47.

Rittschof, C. C., Bukhari, S. A., Sloofman, L. G., Troy, J. M., Caetano-Anollés, D., Cash-Ahmed, A., Kent, M., Lu, X., Sanogo, Y. O., Weisner, P. A., Zhang, H., Bell, A. M., Ma, J., Sinha, S., Robinson, G. E., and Stubbs, L. (2014). Neuromolecular responses to social challenge: Common mechanisms across mouse, stickleback fish, and honey bee. *Proceedings of the National Academy of Sciences*, **111**(50), 17929–17934.

Rivals, I., Personnaz, L., Taing, L., and Potier, M.-C. (2007). Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**(4), 401–407.

Roach, J. C., Glusman, G., Smit, A. F. A., Huff, C. D., Hubley, R., Shannon, P. T., Rowen, L., Pant, K. P., Goodman, N., Bamshad, M., Shendure, J., Drmanac, R., Jorde, L. B., Hood, L., and Galas, D. J. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, **328**(5978), 636–639.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–140.

Rodriguez-Esteban, R., Roberts, P. M., and Crawford, M. E. (2009). Identifying and classifying biomedical perturbations in text. *Nucleic Acids Research*, **37**(3), 771–777.

Roudier, F., Ahmed, I., Berard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., Caillieux, E., Duvernois-Berthet, E., Al-Shikhley, L., Giraut, L., Despres, B., Drevensek, S., Barneche, F., Derozier, S., Brunaud, V., Aubourg, S., Schnittger, A., Bowler, C., Martin-Magniette, M.-L., Robin, S., Caboche, M., and Colot, V. (2011). Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *The EMBO Journal*, **30**(10), 1928–1938.

Roux, J., Rosikiewicz, M., and Robinson-Rechavi, M. (2015). What to compare and how: Comparative transcriptomics for Evo-Devo: COMPARATIVE TRANSCRIPTOMICS FOR Evo-Devo. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, **324**(4), 372–382.

Rung, J. and Brazma, A. (2013). Reuse of public genome-wide gene expression data. *Nature Reviews Genetics*, **14**(2), 89–99.

Runyan, R. B. and Markwald, R. R. (1983). Invasion of mesenchyme into three-dimensional collagen gels: a regional and temporal analysis of interaction in embryonic heart tissue. *Dev Biol*, **95**(1), 108–114.

Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Ison, J., Keays, M., and al, e. (2013). ArrayExpress update — trends in database growth and links to data analysis tools. *Nucleic Acids Research*, **41**(D1), D987–D990.

Ryall, J. G., Dell'Orso, S., Derfoul, A., Juan, A., Zare, H., Feng, X., Clermont, D., Koulnis, M., Gutierrez-Cruz, G., Fulco, M., and Sartorelli, V. (2015). The NAD+-Dependent SIRT1 Deacetylase Translates a Metabolic Switch into Regulatory Epigenetics in Skeletal Muscle Stem Cells. *Cell Stem Cell*, **16**(2), 171–183.

Sadrieh, A., Domanski, L., Pitt-Francis, J., Mann, S. A., Hodkinson, E. C., Ng, C.-A., Perry, M. D., Taylor, J. A., Gavaghan, D., Subbiah, R. N., Vandenberg, J. I., and Hill, A. P. (2014). Multiscale cardiac modelling reveals the origins of notched T waves in long QT syndrome type 2. *Nature Communications*, **5**, 5069.

Sambasivan, R., Yao, R., Kissenpfennig, A., Van Wittenberghe, L., Paldi, A., Gayraud-Morel, B., Guenou, H., Malissen, B., Tajbakhsh, S., and Galy, A. (2011). Pax7-expressing satellite cells are indispensable for adult skeletal muscle regeneration. *Development*, **138**(17), 3647–3656.

Sandoval-Guzmán, T., Wang, H., Khattak, S., Schuez, M., Roensch, K., Nacu, E., Tazaki, A., Joven, A., Tanaka, E. M., and Simon, A. (2014). Fundamental Differences in Dedifferentiation and Stem Cell Recruitment during Skeletal Muscle Regeneration in Two Salamander Species. *Cell Stem Cell*, **14**(2), 174–187.

Sartor, M. A., Leikauf, G. D., and Medvedovic, M. (2009). LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics (Oxford, England)*, **25**(2), 211–217.

Sartori, R., Schirwis, E., Blaauw, B., Bortolanza, S., Zhao, J., Enzo, E., Stantzou, A., Mouisel, E., Toniolo, L., Ferry, A., Stricker, S., Goldberg, A. L., Dupont, S., Piccolo, S., Amthor, H., and Sandri, M. (2013). BMP signaling controls muscle mass. *Nature Genetics*, **45**(11), 1309–1318.

Satokata, I., Ma, L., Ohshima, H., Bei, M., Woo, I., Nishizawa, K., Maeda, T., Takano, Y., Uchiyama, M., Heaney, S., Peters, H., Tang, Z., Maxson, R., and Maas, R. (2000). Msx2 deficiency in mice causes pleiotropic defects in bone growth and ectodermal organ formation. *Nature Genetics*, **24**(4), 391–395.

Savu, O., Jurcut, R., Giusca, S., van Mieghem, T., Gussi, I., Popescu, B. A., Ginghina, C., Rademakers, F., Deprest, J., and Voigt, J.-U. (2012). Morphological and Functional Adaptation of the Maternal Heart During Pregnancy. *Circulation: Cardiovascular Imaging*, **5**(3), 289–297.

Schlesinger, J., Schueler, M., Grunert, M., Fischer, J. J., Zhang, Q., Krueger, T., Lange, M., Tönjes, M., Dunkel, I., and Sperling, S. R. (2011). The cardiac transcription network modulated by Gata4, Mef2a, Nkx2.5, Srf, histone modifications, and microRNAs. *PLoS Genet*, **7**(2), e1001313.

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., Talianidis, I., Flicek, P., and Odom, D. T. (2010). Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. *Science*, **328**(5981), 1036–1040.

Schmitt, B., Ringe, J., Häupl, T., Notter, M., Manz, R., Burmester, G.-R., Sittinger, M., and Kaps, C. (2003). BMP2 initiates chondrogenic lineage development of adult human mesenchymal stem cells in high-density culture. *Differentiation*, **71**(9-10), 567–577.

Schubert, M., Klinger, B., Klünemann, M., Garnett, M. J., Blüthgen, N., and Saez-Rodriguez, J. (2016). Perturbation-response genes reveal signaling footprints in cancer gene expression. *bioRxiv*, page 065672.

Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelsen, T. S., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G., and Zhang, F. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, **343**(6166), 84–87.

Shea, C. M., Edgar, C. M., Einhorn, T. A., and Gerstenfeld, L. C. (2003). BMP treatment of C3h10t1/2 mesenchymal stem cells induces both chondrogenesis and osteogenesis. *Journal of Cellular Biochemistry*, **90**(6), 1112–1127.

Shen, B., Wei, A., Whittaker, S., Williams, L. A., Tao, H., Ma, D. D., and Diwan, A. D. (2009). The role of BMP-7 in chondrogenic and osteogenic differentiation of human bone marrow multipotent mesenchymal stromal cells in vitro. *Journal of Cellular Biochemistry*, pages n/a–n/a.

Sheng, W., Qian, Y., Zhang, P., Wu, Y., Wang, H., Ma, X., Chen, L., Ma, D., and Huang, G. (2014). Association of promoter methylation statuses of congenital heart defect candidate genes with Tetralogy of Fallot. *J Transl Med*, **12**, 31.

Shiels, A., Bennett, T. M., and Hejtmancik, J. F. (2010). Cat-Map: putting cataract on the map. *Molecular Vision*, **16**, 2007–2015.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**(20), 3940–3941.

Smaill, B. H. and Hunter, P. J. (2010). Computer Modeling of Electrical Activation: From Cellular Dynamics to the Whole Heart. In D. C. Sigg, P. A. Iaizzo, Y.-F. Xiao, and B. He, editors, *Cardiac Electrophysiology Methods and Models*, pages 159–185. Springer US, Boston, MA. DOI: 10.1007/978-1-4419-6658-2_8.

Smemo, S., Campos, L. C., Moskowitz, I. P., Krieger, J. E., Pereira, A. C., and Nobrega, M. A. (2012). Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Human Molecular Genetics*, **21**(14), 3255–3263.

Smith, P. D., Sun, F., Park, K. K., Cai, B., Wang, C., Kuwako, K., Martinez-Carrasco, I., Connolly, L., and He, Z. (2009). SOCS3 Deletion Promotes Optic Nerve Regeneration In Vivo. *Neuron*, **64**(5), 617–623.

Smyth, G. K. (2005). Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 397–420. Springer, New York.

Snelling, S. J. B., Hulley, P. A., and Loughlin, J. (2010). BMP5 activates multiple signaling pathways and promotes chondrogenic differentiation in the ATDC5 growth plate model. *Growth Factors*, **28**(4), 268–279.

Soemedi, R., Wilson, I. J., Bentham, J., Darlay, R., Töpf, A., Zelenika, D., Cosgrove, C., Setchfield, K., Thornborough, C., Granados-Riveron, J., Blue, G. M., Breckpot, J., Hellens, S., Zwolinkski, S., Glen, E., Mamasoula, C., Rahman, T. J., Hall, D., Rauch, A., Devriendt, K., Gewillig, M., O' Sullivan, J., Winlaw, D. S., Bu'Lock, F., Brook, J. D., Bhattacharya, S., Lathrop, M., Santibanez-Koref, M., Cordell, H. J., Goodship, J. A., and Keavney, B. D. (2012). Contribution of global rare copy-number variants to the risk of sporadic congenital heart disease. *Am J Hum Genet*, **91**(3), 489–501.

Sohn, K.-A., Ho, J. W. K., Djordjevic, D., Jeong, H.-H., Park, P. J., and Kim, J. H. (2015). hiHMM: Bayesian non-parametric joint inference of chromatin state maps. *Bioinformatics (Oxford, England)*, **31**(13), 2066–2074.

Song, J. and Chen, K. C. (2014). *Spectacle: Faster and more accurate chromatin state annotation using spectral learning*. Cold Spring Harbor Laboratory Press.

Sonnhammer, E. L. L. and Ostlund, G. (2015). InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Research*, **43**(D1), D234–D239.

Sperling, S. R. (2011). Systems biology approaches to heart development and congenital heart disease. *Cardiovasc Res*, **91**(2), 269–278.

Starkey, J. D., Yamamoto, M., Yamamoto, S., and Goldhamer, D. J. (2011). Skeletal Muscle Satellite Cells Are Committed to Myogenesis and Do Not Spontaneously Adopt Nonmyogenic Fates. *Journal of Histochemistry & Cytochemistry*, **59**(1), 33–46.

Stennard, F. A., Costa, M. W., Lai, D., Biben, C., Furtado, M. B., Solloway, M. J., McCulley, D. J., Leimena, C., Preis, J. I., Dunwoodie, S. L., Elliott, D. E., Prall, O. W. J., Black, B. L., Fatkin, D., and Harvey, R. P. (2005). Murine T-box transcription factor Tbx20 acts as a repressor during heart development, and is essential for adult heart integrity, function and adaptation. *Development (Cambridge, England)*, **132**(10), 2451–2462.

Stern, S., Haverkamp, S., Sinske, D., Tedeschi, A., Naumann, U., Di Giovanni, S., Kochanek, S., Nordheim, A., and Knöll, B. (2013). The transcription factor serum response factor stimulates axon regeneration through cytoplasmic localization and cofilin interaction. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, **33**(48), 18836–18848.

Strasser, B. J. (1999). Perspectives: molecular medicine. "Sickle cell anemia, a molecular disease". *Science*, **286**(5444), 1488–1490.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**(43), 15545–15550.

Takács, R., Matta, C., Somogyi, C., Juhász, T., and Zákány, R. (2013). Comparative Analysis of Osteogenic/Chondrogenic Differentiation Potential in Primary Limb Bud-Derived and C3h10t1/2 Cell Line-Based Mouse Micromass Cultures. *International Journal of Molecular Sciences*, **14**(8), 16141–16167.

Takeuchi, J. K., Lou, X., Alexander, J. M., Sugizaki, H., Delgado-Olguín, P., Holloway, A. K., Mori, A. D., Wylie, J. N., Munson, C., Zhu, Y., Zhou, Y.-Q., Yeh, R.-F., Henkelman, R. M., Harvey, R. P., Metzger, D., Chambon, P., Stainier, D. Y. R., Pollard, K. S., Scott, I. C., and Bruneau, B. G. (2011). Chromatin remodelling complex dosage modulates transcription factor function in heart development. *Nat Commun*, **2**, 187.

Tan, N., Chung, M. K., Smith, J. D., Hsu, J., Serre, D., Newton, D. W., Castel, L., Soltesz, E., Pettersson, G., Gillinov, A. M., Van Wagoner, D. R., and Barnard, J. (2013). Weighted gene coexpression network analysis of human left atrial tissue identifies gene modules associated with atrial fibrillation. *Circ Cardiovasc Genet*, **6**(4), 362–371.

Tanaka, M., Chen, Z., Bartunkova, S., Yamasaki, N., and Izumo, S. (1999). The cardiac homeobox gene Csx/Nkx2.5 lies genetically upstream of multiple genes essential for heart development. *Development (Cambridge, England)*, **126**(6), 1269–1280.

Taylor, J. C., Martin, H. C., Lise, S., Broxholme, J., Cazier, J.-B., Rimmer, A., Kanapin, A., Lunter, G., Fiddy, S., Allan, C., Aricescu, A. R., Attar, M., Babbs, C., Becq, J., Beeson, D., Bento, C., Bignell, P., Blair, E., Buckle, V. J., Bull, K., Cais, O., Cario, H., Chapel, H., Copley, R. R., Cornall, R., Craft, J., Dahan, K., Davenport, E. E., Dendrou, C., Devuyst, O., Fenwick, A. L., Flint, J., Fugger, L., Gilbert, R. D., Goriely, A., Green, A., Greger, I. H., Grocock, R., Gruszczyk, A. V., Hastings, R., Hatton, E., Higgs, D., Hill, A., Holmes, C., Howard, M., Hughes, L., Humburg, P., Johnson, D., Karpe, F., Kingsbury, Z., Kini, U., Knight, J. C., Krohn, J., Lamble, S., Langman, C., Lonie, L., Luck, J., McCarthy, D., McGowan, S. J., McMullin, M. F., Miller, K. A., Murray, L., Németh, A. H., Nesbit, M. A., Nutt, D., Ormondroyd, E., Oturai, A. B., Pagnamenta, A., Patel, S. Y., Percy, M., Petousi, N., Piazza, P., Piret, S. E., Polanco-Echeverry, G., Popitsch, N., Powrie, F., Pugh, C., Quek, L., Robbins, P. A., Robson, K., Russo, A., Sahgal, N., van Schouwenburg, P. A., Schuh, A., Silverman, E., Simmons, A., Sørensen, P. S., Sweeney, E., Taylor, J., Thakker, R. V., Tomlinson, I., Trebes, A., Twigg, S. R. F., Uhlig, H. H., Vyas, P., Vyse, T., Wall, S. A., Watkins, H., Whyte, M. P., Witty, L., Wright, B., Yau, C., Buck, D., Humphray, S., Ratcliffe, P. J., Bell, J. I., Wilkie, A. O. M., Bentley, D., Donnelly, P., and McVean, G. (2015). Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nature Genetics*, **47**(7), 717–726.

Tester, D. J., Medeiros-Domingo, A., Will, M. L., Haglund, C. M., and Ackerman, M. J. (2012). Cardiac Channel Molecular Autopsy: Insights From 173 Consecutive Cases of Autopsy-Negative Sudden Unexplained Death Referred for Postmortem Genetic Testing. *Mayo Clinic Proceedings*, **87**(6), 524–539.

Thakore, P. I., D'Ippolito, A. M., Song, L., Safi, A., Shivakumar, N. K., Kabadi, A. M., Reddy, T. E., Crawford, G. E., and Gersbach, C. A. (2015). Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nature Methods*, **12**(12), 1143–1149.

Thienpont, B., Zhang, L., Postma, A. V., Breckpot, J., Tranchevent, L.-C., Van Loo, P., Møllgård, K., Tommerup, N., Bache, I., Tümer, Z., van Engelen, K., Menten, B., Mortier, G., Waggoner, D., Gewillig, M., Moreau, Y., Devriendt, K., and Larsen, L. A. (2010). Haploinsufficiency of TAB2 Causes Congenital Heart Defects in Humans. *The American Journal of Human Genetics*, **86**(6), 839–849.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**(22), 4673–4680.

Tranchevent, L.-C., Barriot, R., Yu, S., Van Vooren, S., Van Loo, P., Coessens, B., De Moor, B., Aerts, S., and Moreau, Y. (2008). ENDEAVOUR update: a web re-

source for gene prioritization in multiple species. *Nucleic Acids Res*, **36**(Web Server issue), W377–W384.

Tranchevent, L.-C., Capdevila, F. B., Nitsch, D., De Moor, B., De Causmaecker, P., and Moreau, Y. (2011). A guide to web tools to prioritize candidate genes. *Briefings in Bioinformatics*, **12**(1), 22–32.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, **25**(9), 1105–1111.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, **28**(5), 511–515.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol*, **32**(4), 381–386.

Trent, R. J. (2012). *Molecular and Cellular Therapies*. Elsevier BV.

Tucker, N. R. and Ellinor, P. T. (2014). Emerging Directions in the Genetics of Atrial Fibrillation. *Circulation Research*, **114**(9), 1469–1482.

Udali, S., Guarini, P., Moruzzi, S., Choi, S.-W., and Friso, S. (2013). Cardiovascular epigenetics: from DNA methylation to microRNAs. *Mol Aspects Med*, **34**(4), 883–901.

Uno, Y., Nishida, C., Takagi, C., Ueno, N., and Matsuda, Y. (2013). Homoeologous chromosomes of Xenopus laevis are highly conserved after whole-genome duplication. *Heredity*, **111**(5), 430–436.

Usas, A. and Huard, J. (2007). Muscle-derived stem cells for tissue engineering and regenerative therapy. *Biomaterials*, **28**(36), 5401–5406.

Van Noorden, R., Maher, B., and Nuzzo, R. (2014). The top 100 papers.

Wada, M. R., Inagawa-Ogashiwa, M., Shimizu, S., Yasumoto, S., and Hashimoto, N. (2002). Generation of different fates from multipotent muscle stem cells. *Development (Cambridge, England)*, **129**(12), 2987–2995.

Wagner, A. (2001). How to reconstruct a large genetic network from n gene perturbations in fewer than n(2) easy steps. *Bioinformatics (Oxford, England)*, **17**(12), 1183–1197.

Wang, H., Noulet, F., Edom-Vovard, F., Le Grand, F., and Duprez, D. (2010). Bmp Signaling at the Tips of Skeletal Muscles Regulates the Number of Fetal Muscle Progenitors and Satellite Cells during Development. *Developmental Cell*, **18**(4), 643–654.

Wang, H., Yang, H., Shivalila, C. S., Dawlaty, M. M., Cheng, A. W., Zhang, F., and Jaenisch, R. (2013a). One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell*, **153**(4), 910–918.

Wang, T., Wei, J. J., Sabatini, D. M., and Lander, E. S. (2014a). Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, **343**(6166), 80–84.

Wang, V. Y., Hoogendoorn, C., Frangi, A. F., Cowan, B. R., Hunter, P. J., Young, A. A., and Nash, M. P. (2013b). Automated Personalised Human Left Ventricular FE Models to Investigate Heart Failure Mechanics. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, O. Camara, T. Mansi, M. Pop, K. Rhode, M. Sermesant, and A. Young, editors, *Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges*, volume 7746, pages 307–316. Springer Berlin Heidelberg, Berlin, Heidelberg. DOI: 10.1007/978-3-642-36961-2_35.

Wang, Y., Penfold, C. A., Hodgson, D. A., Gifford, M. L., and Burroughs, N. J. (2014b). Correcting for link loss in causal network inference caused by regulator interference. *Bioinformatics*.

Wang, Y. X. and Rudnicki, M. A. (2011). Satellite cells, the engines of muscle repair. *Nature Reviews. Molecular Cell Biology*, **13**(2), 127–133.

Wang, Z., Monteiro, C. D., Jagodnik, K. M., Fernandez, N. F., Gundersen, G. W., Rouillard, A. D., Jenkins, S. L., Feldmann, A. S., Hu, K. S., McDermott, M. G., Duan, Q., Clark, N. R., Jones, M. R., Kou, Y., Goff, T., Woodland, H., Amaral, F. M. R., Szeto, G. L., Fuchs, O., Schüssler-Fiorenza Rose, S. M., Sharma, S., Schwartz, U., Bausela, X. B., Szymkiewicz, M., Maroulis, V., Salykin, A., Barra, C. M., Kruth, C. D., Bongio, N. J., Mathur, V., Todoric, R. D., Rubin, U. E., Malatras, A., Fulp, C. T., Galindo, J. A., Motiejunaite, R., Jüschke, C., Dishuck, P. C., Lahl, K., Jafari, M., Aibar, S., Zaravinos, A., Steenhuizen, L. H., Allison, L. R., Gamallo, P., de Andres Segura, F., Dae Devlin, T., Pérez-García, V., and Ma'ayan, A. (2016). Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nature Communications*, **7**, 12846.

Wellcome Trust Case Control Consortium, Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D. F., Giannoulatou, E., Holmes, C., Marchini, J. L., Stirrups, K., Tobin, M. D., Wain, L. V., Yau, C., Aerts, J., Ahmad, T., Andrews, T. D., Arbury, H., Attwood, A., Auton, A., Ball, S. G., Balmforth, A. J., Barrett, J. C., Barroso, I., Barton, A., Bennett, A. J., Bhaskar, S., Blaszczyk, K., Bowes, J., Brand, O. J., Braund, P. S., Bredin, F., Breen, G., Brown, M. J., Bruce, I. N., Bull, J., Burren, O. S., Burton, J., Byrnes, J., Caesar, S., Clee, C. M., Coffey, A. J., Connell, J. M. C., Cooper, J. D., Dominiczak, A. F., Downes, K., Drummond, H. E., Dudakia, D., Dunham, A., Ebbs, B., Eccles, D., Edkins, S., Edwards, C., Elliot, A., Emery, P., Evans, D. M., Evans, G., Eyre, S., Farmer, A., Ferrier, I. N., Feuk, L., Fitzgerald, T., Flynn, E., Forbes, A., Forty, L., Franklyn, J. A., Freathy, R. M., Gibbs, P., Gilbert, P., Gokumen, O., Gordon-Smith, K., Gray, E., Green, E., Groves, C. J., Grozeva, D., Gwilliam, R., Hall, A., Hammond, N., Hardy, M., Harrison, P., Hassanali, N., Hebaishi, H., Hines, S., Hinks, A., Hitman, G. A., Hocking, L., Howard, E., Howard, P., Howson, J. M. M., Hughes, D., Hunt, S., Isaacs, J. D.,

Jain, M., Jewell, D. P., Johnson, T., Jolley, J. D., Jones, I. R., Jones, L. A., Kirov, G., Langford, C. F., Lango-Allen, H., Lathrop, G. M., Lee, J., Lee, K. L., Lees, C., Lewis, K., Lindgren, C. M., Maisuria-Armer, M., Maller, J., Mansfield, J., Martin, P., Massey, D. C. O., McArdle, W. L., McGuffin, P., McLay, K. E., Mentzer, A., Mimmack, M. L., Morgan, A. E., Morris, A. P., Mowat, C., Myers, S., Newman, W., Nimmo, E. R., O'Donovan, M. C., Onipinla, A., Onyiah, I., Ovington, N. R., Owen, M. J., Palin, K., Parnell, K., Pernet, D., Perry, J. R. B., Phillips, A., Pinto, D., Prescott, N. J., Prokopenko, I., Quail, M. A., Rafelt, S., Rayner, N. W., Redon, R., Reid, D. M., Renwick, Ring, S. M., Robertson, N., Russell, E., St Clair, D., Sambrook, J. G., Sanderson, J. D., Schuilenburg, H., Scott, C. E., Scott, R., Seal, S., Shaw-Hawkins, S., Shields, B. M., Simmonds, M. J., Smyth, D. J., Somaskantharajah, E., Spanova, K., Steer, S., Stephens, J., Stevens, H. E., Stone, M. A., Su, Z., Symmons, D. P. M., Thompson, J. R., Thomson, W., Travers, M. E., Turnbull, C., Valsesia, A., Walker, M., Walker, N. M., Wallace, C., Warren-Perry, M., Watkins, N. A., Webster, J., Weedon, M. N., Wilson, A. G., Woodburn, M., Wordsworth, B. P., Young, A. H., Zeggini, E., Carter, N. P., Frayling, T. M., Lee, C., McVean, G., Munroe, P. B., Palotie, A., Sawcer, S. J., Scherer, S. W., Strachan, D. P., Tyler-Smith, C., Brown, M. A., Burton, P. R., Caulfield, M. J., Compston, A., Farrall, M., Gough, S. C. L., Hall, A. S., Hattersley, A. T., Hill, A. V. S., Mathew, C. G., Pembrey, M., Satsangi, J., Stratton, M. R., Worthington, J., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W., Parkes, M., Rahman, N., Todd, J. A., Samani, N. J., and Donnelly, P. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, **464**(7289), 713–720.

Werner, P., Latney, B., Deardorff, M. A., and Goldmuntz, E. (2016). MESP1 Mutations in Patients with Congenital Heart Defects. *Human Mutation*, **37**(3), 308–314.

Westerhoff, H. V. and Palsson, B. O. (2004). The evolution of molecular biology into systems biology. *Nat Biotechnol*, **22**(10), 1249–1252.

Wills, Q. F., Livak, K. J., Tipping, A. J., Enver, T., Goldson, A. J., Sexton, D. W., and Holmes, C. (2013). Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat Biotechnol*, **31**(8), 748–752.

Winbanks, C. E., Chen, J. L., Qian, H., Liu, Y., Bernardo, B. C., Beyer, C., Watt, K. I., Thomson, R. E., Connor, T., Turner, B. J., McMullen, J. R., Larsson, L., McGee, S. L., Harrison, C. A., and Gregorevic, P. (2013). The bone morphogenetic protein axis is a positive regulator of skeletal muscle mass. *The Journal of Cell Biology*, **203**(2), 345–357.

Wire, B. (2017). Illumina Introduces the NovaSeq Series—a New Architecture Designed to Usher in the $100 Genome.

Wong, D., Teixeira, A., Oikonomopoulos, S., Humburg, P., Lone, I., Saliba, D., Siggers, T., Bulyk, M., Angelov, D., Dimitrov, S., Udalova, I. A., and Ragoussis, J. (2011). Extensive characterization of NF-κB binding uncovers non-canonical motifs and advances the interpretation of genetic functional traits. *Genome Biology*, **12**(7), R70.

Wu, Y., Liang, D., Wang, Y., Bai, M., Tang, W., Bao, S., Yan, Z., Li, D., and Li, J. (2013). Correction of a genetic disease in mouse via use of CRISPR-Cas9. *Cell Stem Cell*, **13**(6), 659–662.

Xiao, Y., Gong, Y., Lv, Y., Lan, Y., Hu, J., Li, F., Xu, J., Bai, J., Deng, Y., Liu, L., Zhang, G., Yu, F., and Li, X. (2015). Gene Perturbation Atlas (GPA): a single-gene perturbation repository for characterizing functional mechanisms of coding and non-coding genes. *Scientific Reports*, **5**, 10889.

Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., Kong, L., Gao, G., Li, C.-Y., and Wei, L. (2011). KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Research*, **39**(suppl), W316–W322.

Xu, H., Morishima, M., Wylie, J. N., Schwartz, R. J., Bruneau, B. G., Lindsay, E. A., and Baldini, A. (2004). Tbx1 has a dual role in the morphogenesis of the cardiac outflow tract. *Development (Cambridge, England)*, **131**(13), 3217–3227.

Xu, M., Wu, X., Li, Y., Yang, X., Hu, J., Zheng, M., and Tian, J. (2014). CITED2 mutation and methylation in children with congenital heart disease. *J Biomed Sci*, **21**, 7.

Yang, Z., Alwatban, A., Everson, R., and Yang, Z. R. (2014). Multi-Scale Gaussian Mixtures for Cross-species Study. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1.

Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Nuhn, M., Parker, A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Birney, E., Harrow, J., Muffato, M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S. J., Cunningham, F., Aken, B. L., Zerbino, D. R., and Flicek, P. (2016). Ensembl 2016. *Nucleic Acids Research*, **44**(D1), D710–D716.

Yilmaz, A., Engeler, R., Constantinescu, S., Kokkaliaris, K. D., Dimitrakopoulos, C., Schroeder, T., Beerenwinkel, N., and Paro, R. (2015). Ectopic expression of Msx2 in mammalian myotubes recapitulates aspects of amphibian muscle dedifferentiation. *Stem Cell Research*, **15**(3), 542–553.

Yu, L., Han, M., Yan, M., Lee, E.-C., Lee, J., and Muneoka, K. (2010). BMP signaling induces digit regeneration in neonatal mice. *Development*, **137**(4), 551–559.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B. D., Shen, Y., Pervouchine, D. D., Djebali, S., Thurman, R. E., Kaul, R., Rynes, E., Kirilusha, A., Marinov, G. K., Williams, B. A., Trout, D., Amrhein, H., Fisher-Aylor, K., Antoshechkin, I., DeSalvo, G., See, L.-H., Fastuca, M., Drenkow, J., Zaleski, C., Dobin, A., Prieto, P., Lagarde, J., Bussotti, G., Tanzer, A., Denas, O., Li, K., Bender, M. A., Zhang, M., Byron, R., Groudine, M. T., McCleary, D., Pham, L.,

Ye, Z., Kuan, S., Edsall, L., Wu, Y.-C., Rasmussen, M. D., Bansal, M. S., Kellis, M., Keller, C. A., Morrissey, C. S., Mishra, T., Jain, D., Dogan, N., Harris, R. S., Cayting, P., Kawli, T., Boyle, A. P., Euskirchen, G., Kundaje, A., Lin, S., Lin, Y., Jansen, C., Malladi, V. S., Cline, M. S., Erickson, D. T., Kirkup, V. M., Learned, K., Sloan, C. A., Rosenbloom, K. R., Lacerda de Sousa, B., Beal, K., Pignatelli, M., Flicek, P., Lian, J., Kahveci, T., Lee, D., James Kent, W., Ramalho Santos, M., Herrero, J., Notredame, C., Johnson, A., Vong, S., Lee, K., Bates, D., Neri, F., Diegel, M., Canfield, T., Sabo, P. J., Wilken, M. S., Reh, T. A., Giste, E., Shafer, A., Kutyavin, T., Haugen, E., Dunn, D., Reynolds, A. P., Neph, S., Humbert, R., Scott Hansen, R., De Bruijn, M., Selleri, L., Rudensky, A., Josefowicz, S., Samstein, R., Eichler, E. E., Orkin, S. H., Levasseur, D., Papayannopoulou, T., Chang, K.-H., Skoultchi, A., Gosh, S., Disteche, C., Treuting, P., Wang, Y., Weiss, M. J., Blobel, G. A., Cao, X., Zhong, S., Wang, T., Good, P. J., Lowdon, R. F., Adams, L. B., Zhou, X.-Q., Pazin, M. J., Feingold, E. A., Wold, B., Taylor, J., Mortazavi, A., Weissman, S. M., Stamatoyannopoulos, J. A., Snyder, M. P., Guigo, R., Gingeras, T. R., Gilbert, D. M., Hardison, R. C., Beer, M. A., and Ren, B. (2014). A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, **515**(7527), 355–364.

Zaidi, S., Choi, M., Wakimoto, H., Ma, L., Jiang, J., Overton, J. D., Romano-Adesman, A., Bjornson, R. D., Breitbart, R. E., Brown, K. K., Carriero, N. J., Cheung, Y. H., Deanfield, J., DePalma, S., Fakhro, K. A., Glessner, J., Hakonarson, H., Italia, M. J., Kaltman, J. R., Kaski, J., Kim, R., Kline, J. K., Lee, T., Leipzig, J., Lopez, A., Mane, S. M., Mitchell, L. E., Newburger, J. W., Parfenov, M., Pe'er, I., Porter, G., Roberts, A. E., Sachidanandam, R., Sanders, S. J., Seiden, H. S., State, M. W., Subramanian, S., Tikhonova, I. R., Wang, W., Warburton, D., White, P. S., Williams, I. A., Zhao, H., Seidman, J. G., Brueckner, M., Chung, W. K., Gelb, B. D., Goldmuntz, E., Seidman, C. E., and Lifton, R. P. (2013). De novo mutations in histone-modifying genes in congenital heart disease. *Nature*, **498**(7453), 220–223.

Zhang, B., Gaiteri, C., Bodea, L.-G., Wang, Z., McElwee, J., Podtelezhnikov, A. A., Zhang, C., Xie, T., Tran, L., Dobrin, R., Fluder, E., Clurman, B., Melquist, S., Narayanan, M., Suver, C., Shah, H., Mahajan, M., Gillis, T., Mysore, J., MacDonald, M. E., Lamb, J. R., Bennett, D. A., Molony, C., Stone, D. J., Gudnason, V., Myers, A. J., Schadt, E. E., Neumann, H., Zhu, J., and Emilsson, V. (2013). Integrated Systems Approach Identifies Genetic Nodes and Networks in Late-Onset Alzheimer's Disease. *Cell*, **153**(3), 707–720.

Zhang, J., Ma, X., Wang, H., Ma, D., and Huang, G. (2014). Elevated methylation of the RXRA promoter region may be responsible for its downregulated expression in the myocardium of patients with TOF. *Pediatr Res*, **75**(5), 588–594.

Zhang, Q.-J., Chen, H.-Z., Wang, L., Liu, D.-P., Hill, J. A., and Liu, Z.-P. (2011). The histone trimethyllysine demethylase JMJD2a promotes cardiac hypertrophy in response to hypertrophic stimuli in mice. *J Clin Invest*, **121**(6), 2447–2456.

Zhao, P. and Hoffman, E. P. (2004). Embryonic myogenesis pathways in muscle regeneration. *Developmental Dynamics*, **229**(2), 380–392.

Zheng, W., Wang, Z., Collins, J. E., Andrews, R. M., Stemple, D., and Gong, Z. (2011). Comparative Transcriptome Analyses Indicate Molecular Homology of Zebrafish Swimbladder and Mammalian Lung. *PLoS ONE*, **6**(8), e24019.

Zhou, N., Li, Q., Lin, X., Hu, N., Liao, J.-Y., Lin, L.-B., Zhao, C., Hu, Z.-M., Liang, X., Xu, W., Chen, H., and Huang, W. (2016). BMP2 induces chondrogenic differentiation, osteogenic differentiation and endochondral ossification in stem cells. *Cell and Tissue Research*, **366**(1), 101–111.

Zhu, N., Wang, H., Wang, B., Wei, J., Shan, W., Feng, J., and Huang, H. (2016). A Member of the Nuclear Receptor Superfamily, Designated as NR2f2, Supports the Self-Renewal Capacity and Pluripotency of Human Bone Marrow-Derived Mesenchymal Stem Cells. *Stem Cells International*, **2016**, 1–11.

Zhu, Y., Davis, S., Stephens, R., Meltzer, P. S., and Chen, Y. (2008). GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. *Bioinformatics*, **24**(23), 2798–2800.

Zinman, G. E., Naiman, S., Kanfi, Y., Cohen, H., and Bar-Joseph, Z. (2013). ExpressionBlast: mining large, unstructured expression databases. *Nature Methods*, **10**(10), 925–926.