EVOLUTIONARILY CONSERVED REGULATORY PROGRAMS

by

Tae-Jun Andrew Kwon

M.Sc., The University of British Columbia, 2003
B.Sc., The University of British Columbia, 2000

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Genetics)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

August 2011

© Tae-Jun Andrew Kwon, 2011
Abstract

Despite the diversity of metazoans, common biochemical systems and structures can be found in distinct taxonomic groups. The development and formation of metazoan tissues and structures has been well researched, but their regulatory mechanisms are not understood well. To this end, we implemented bioinformatics tools regulatory mechanism analysis and applied them to study regulatory program conservation with an emphasis on muscle development.

We first performed a genome-wide scan for muscle-specific cis-regulatory modules (CRMs) using three computational prediction programs. Based on the predictions, 339 candidate CRMs were tested in cell culture with NIH3T3 fibroblasts and C2C12 myoblasts for capacity to direct selective reporter gene expression to differentiated myotubes. A subset of 19 CRMs validated as functional in the assay. The rate of predictive success reveals striking limitations of computational CRM discovery. Motif-based methods performed no better than predictions based only on sequence conservation. Analysis of the properties of the functional sequences relative to inactive sequences identifies nucleotide sequence composition and ChIP-Seq evidence as important characteristics to incorporate in future methods for improved predictive specificity.

In studying the transcriptional regulation, motif enrichment analysis of co-expressed genes is often employed to determine mediating transcription factors. We built oPOSSUM-3, a web-based software system for identification of enriched transcription factor binding sites (TFBS) and TFBS families in DNA sequences of co-expressed genes and sequences generated from high-throughput methods, such as ChIP-Seq. Validation of the system with known sets of published data demonstrates the capacity for oPOSSUM-3 to identify mediating TFs for co-regulated genes.
Studies have shown that TF binding profiles tend to be highly conserved over long evolutionary distances. In large-scale public genome annotation projects, such as modENCODE, transcriptional regulation data is compiled for comparative genomics research. Using the oPOSSUM-3 system and published data, we performed comparative analyses of the regulatory programs across evolutionarily divergent species, including human, fruit fly, and nematode, and examined the extent of conservation in major regulatory programs.

The thesis research provides new approaches to computational analysis of DNA sequences and insights into the analysis of transcription regulation across the phylogenetic spectrum.
Preface

Chapter 2 is based on work conducted in collaboration with Alice Chou. I collected the pre-annotated regulatory regions from literature, and developed the software and the protocols for the genome-wide computational predictions. Alice Chou designed the experimental protocols for the validation of the computational predictions, and we jointly performed the validation experiments. David J. Arenillas designed the oligonucleotides used for PCR in the experiments. I performed the computational analysis of the results and wrote the whole manuscript. Dr. Wasserman supervised the work and revised the manuscript, as well as contributing to the development of the ideas. A version of chapter 2 has been submitted for publication and is currently undergoing requested revisions.

Chapter 3 is based on work conducted in collaboration with David J. Arenillas and Rebecca Worsley Hunt. I was responsible for designing the TFBS clustering algorithm and preparing the TFBS clusters used by the oPOSSUM system, as well as designing routines for nematode operon analysis. David J. Arenillas and I jointly participated in the software development. Rebecca Worsley Hunt was responsible for studying the effects of sequence composition on the results produced by the oPOSSUM system. I wrote the manuscript, with contributions from Rebecca Worsley Hunt for the sections pertaining to her work. David J. Arenillas produced two figures that were used in the manuscript. Dr. Wasserman supervised the work and revised the manuscript, as well as contributing to the development of the ideas. A version of chapter 3 has been submitted for publication and is currently under review.

The research described in Chapter 4 was the product of my efforts. Dr. Wasserman supervised the work and revised the manuscript, as well as contributing to the development of the ideas.
Table of Contents

Abstract........................................................................................................................................... ii
Preface.............................................................................................................................................. iv
List of Tables .................................................................................................................................... ix
List of Figures ................................................................................................................................... xi
List of Abbreviations ........................................................................................................................ xiv
Acknowledgements .......................................................................................................................... xvi

1 Introduction ..................................................................................................................................... 1
   1.1 Metazoan Tissue Development ............................................................................................... 3
       1.1.1 Tissue and Structure Conservation among Metazoans .................................................... 4
   1.2 Gene Regulation ...................................................................................................................... 5
       1.2.1 Promoter and Regulatory Regions .................................................................................... 6
       1.2.2 Transcription Factors .................................................................................................... 7
       1.2.3 MicroRNA ...................................................................................................................... 8
       1.2.4 Chromatin and Epigenetics ........................................................................................... 9
   1.3 Analysis of Gene Expression ................................................................................................... 10
       1.3.1 DNA Microarrays ........................................................................................................... 11
       1.3.2 Serial Analysis of Gene Expression (SAGE) ................................................................ 12
       1.3.3 RNA-Seq ....................................................................................................................... 12
       1.3.4 Differential Gene Expression Analysis .......................................................................... 13
   1.4 Experimental Methods for Identification of Regulatory Regions ......................................... 14
       1.4.1 Earlier Methods ............................................................................................................. 14
       1.4.2 High Throughput Methods ........................................................................................... 15
   1.5 Comparative Genomics .......................................................................................................... 18
       1.5.1 Homology: Orthologs and Paralogs .............................................................................. 19
       1.5.2 Sequence Alignment ...................................................................................................... 20
       1.5.3 Evolutionary Constraint Measures ................................................................................. 20
   1.6 Public Repositories of Genomic Data ...................................................................................... 21
       1.6.1 Public Genome Databases ............................................................................................ 21
4.2 Methods ................................................................................................................................. 139
4.2.1 Input Gene Collections ........................................................................................................ 139
4.2.2 List of Known TFs Involved in Each Regulatory Program ..................................................... 142
4.2.3 TFBS Over-Representation Analysis Using oPOSSUM-3 ..................................................... 143
4.2.4 Analysis Tools ..................................................................................................................... 144
4.3 Results .................................................................................................................................... 145
4.3.1 Cilia Genes .......................................................................................................................... 147
4.3.2 Nfe2l2 / Nrf2 Target Genes .................................................................................................. 149
4.3.3 Muscle Genes ...................................................................................................................... 155
4.4 Discussion .................................................................................................................................. 187

5 Discussion and Conclusions ...................................................................................................... 190
5.1 Identification of Tissue-Specific cis-Regulatory Modules Through Computational Prediction Programs ........................................................................................................ 191
5.2 TFBS Over-Representation Analysis for the Identification of Main Regulatory TFs in Co-expressed Genes .............................................................................................................. 193
5.3 Utility of Evolution in Identifying Conserved Regulation Programs ........................................ 194
5.4 Future Directions ...................................................................................................................... 196

References ...................................................................................................................................... 200
List of Tables

Table 2-1. List of some of the published CRM prediction programs. ........................................... 40

Table 2-2. Number of candidate regions........................................................................................................ 55

Table 2-3. List of genomic regions validated as driving muscle-specific expression. ........ 60

Table 2-4. Overrepresented TFBS in the validated regions vs. non-responding regions ranked by Fisher p-values. ........................................................................................................ 62

Table 2-5. Sequence composition characteristics. .................................................................................................. 64

Table 2-6. GC and AT skews of the responding regions vs. non-responding regions. ....... 64

Table 2-7. The distribution of the regions in the muscle set according to the evidence source for muscle expression. .................................................................................................................. 68

Table 2-8. Sequence conservation based on phastCons scores (28-way Placental Mammals). ........................................................................................................ 69

Table 2-9. Comparison of the five CRM prediction programs................................................. 71

Table 2-10. Distances to the nearest annotated transcription start sites (Ensembl v61). ....... 77

Table 2-11. Regions associated with CpG islands................................................................................. 77

Table 2-12. Comparison of the mean phyloP scores in the three region sets for profiles with at least 2-fold increase in phyloP scores for predicted TFBS positions vs. non-TFBS positions. .................................................................................................................. 81

Table 2-13. Regions overlapping MyoD ChIP-Seq peaks in C2C12 cells................................. 82
Table 3-1. Search region level distances for human/mouse, fly and nematode in oPOSSUM-3

Table 3-2. oPOSSUM-3 results for the muscle reference gene set (human).

Table 3-3. oPOSSUM-3 results for the cilia gene set in nematodes.

Table 3-4. oPOSSUM-3 results for Nrf2/Nfe2L2 ChIP-Seq data set, using JASPAR CORE vertebrate profiles.

Table 3-5. oPOSSUM-3 results for FoxA2 ChIP-Seq data set, using JASPAR CORE vertebrate profiles.

Table 4-1. Target genes of Nfe2l2 and its orthologs cnc and skn-1.

Table 4-2. Muscle gene sets analyzed.

Table 4-3. List of TFs available in JASPAR 2010 used in this study.

Table 4-4. Gene ontology analysis of target genes for Nfe2l2 and its orthologs using DAVID.

Table 4-5. Gene ontology analysis on Meissner/Stormo-muscle gene sets.

Table 4-6. Gene ontology analysis on Schnorrer-larval gene sets.

Table 4-7. Gene ontology analysis on Schnorrer-adult genes.
List of Figures

Figure 1-1. Differentiation of human tissues................................................................. 4

Figure 1-2. Overview of transcriptional regulation...................................................... 6

Figure 1-3. cis-Regulatory module examples.............................................................. 8

Figure 1-4. Transcription factor binding site (TFBS) and position-specific scoring matrix (PSSM).................................................................................................................. 28

Figure 2-1. Biases in the location of wells with successful reporter assays................. 48

Figure 2-2. The overview of the region selection pipeline.......................................... 53

Figure 2-3. Reporter expression for known muscle-specific and non-muscle specific enhancer regions in myotubes vs. myoblasts......................................................... 56

Figure 2-4. Selection of clones for differential expression analysis............................ 57

Figure 2-5. Comparison of the reporter expression levels between background regions and non-background regions in the validated positive set.................................................... 58

Figure 2-6. Logos of the 5 muscle TFBSs (SRF, MEF2A, Myf, TEAD, and SP1) used for CRM prediction and 2 additional TFBSs (RREB1, NHLH1) found to be overrepresented in the validated set......................................................................................................................... 63

Figure 2-7. Dinucleotide frequencies in responding regions vs. non-responding regions..... 66

Figure 2-8. Examples of Positive Regions......................................................................... 70

Figure 2-9. ROC Analysis of the conservation filter and the 4 CRM Prediction Programs.. 73
Figure 2-10. Density plots for phyloP (46-way All) scores of the predicted binding site positions for the three region sets. ............................................................... 79

Figure 2-11. Phylogenetic depth analysis of TFBSs in responding and non-responding regions using phyloP (46-ay, hg19). ............................................................... 80

Figure 2-12. Histone modifications in the responding and non-responding regions. .......... 84

Figure 3-1. Overview of the main analysis types available in oPOSSUM-3. ...................... 92

Figure 3-2. oPOSSUM-specific JASPAR PENDING collection. ........................................ 94

Figure 3-3. Defining conserved regions.................................................................................. 97

Figure 3-4. TF structural families and TFBS clusters............................................................... 99

Figure 3-5. TFBS cluster analysis.......................................................................................... 101

Figure 3-6. oPOSSUM system provisions for species with operon structures. ............... 103

Figure 3-7. The build process for oPOSSUM 3 gene-based analysis................................. 105

Figure 3-8. The range of enrichment scores increases as the number of foreground sequences increases.......................................................................................... 107

Figure 3-9. oPOSSUM analysis pipeline............................................................................... 109

Figure 3-10. Web interface. ................................................................................................. 110

Figure 3-11. Relationship between TF profile GC content and enrichment statistics........ 126

Figure 3-12. Fisher scores vs. Z-scores from oPOSSUM analysis on sequence-based data. 128

Figure 4-1. Overview of the conserved regulatory program analysis............................... 138
Figure 4-2. Fisher vs. Z score plot for oPOSSUM TCA results on human muscle reference and validated regions from Chapter 2. .......................................................... 146

Figure 4-3. Score distributions for Rfx profiles from oPOSSUM TCA performed on cilia genes collected from the Ciliome DB and Geremek et al. ................................................. 148

Figure 4-4. oPOSSUM TCA results on Nfe2l2, cnc and skn-1 target genes............... 151

Figure 4-5. Logos of Nfe2l2 orthologs in flies and nematodes ............................... 153

Figure 4-6. Enrichment of Mef2A motif in each gene set, as measured by the distance between the combined score of Mef2A motif and the median score of all the other motifs. 161

Figure 4-7. Enrichment of Myf/Myf6 motifs in each gene set, as measured by the distance between the combined score of Myf/Myf6 and the median score of all the other motifs. ... 166

Figure 4-8. Enrichment of Srf motifs in each gene set, as measured by the distance between the combined score of Srf and the median score of all the other motifs. For control, results for cilia genes, Nfe2l2/Nrf2 genes, and randomly selected genes are also included. ............... 171

Figure 4-9. Enrichment of Egr1 motifs in each gene set, as measured by the distance between the combined score of Egr1 and the median score of all the other motifs. For control, results for cilia genes, Nfe2l2/Nrf2 genes, and randomly selected genes are also included........... 176

Figure 4-10. oPOSSUM TCA results on Mef2 target genes. ........................................... 180

Figure 4-11 oPOSSUM TCA results on Biniou target genes. ........................................... 182
List of Abbreviations

- 3C: Chromosome Conformation Capture
- aCSA: anchored Combination Site Analysis
- aCTCA: anchored Combination TFBS Cluster Analysis
- API: Application Programming Interface
- AUC: Area Under Curve
- CAGE: Cap Analysis Gene Expression
- ChIP: Chromatin Immunoprecipitation
- COG: Cluster of Orthologous Group
- CRM: cis-Regulatory Modules
- DPE: Downstream Promoter Element
- EMSA: Electrophoretic Mobility Shift Assay
- ENCODE: ENCyclopedia Of DNA Elements
- FDR: False Discovery Rate
- GFP: Green Fluorescent Protein
- GO: Gene Ontology
- HMM: Hidden Markov Model
- IC: Information Content
- Inr: Initiator Region
- IUPAC: International Union of Pure and Applied Chemistry
- KOG: euKaryotic Orthologous Group
- LCR: Locus Control Region
- LRA: Logistic Regression Analysis
- MEME: Multiple EM for Motif Elicitation
- modENCODE: model organism ENCODE
- MSA: Multiple Sequence Alignment
- PBM: Protein Binding Microarray
- PCR: Polymerase Chain Reaction
• PFM: Position Frequency Matrix
• PSSM: Position-Specific Scoring Matrix
• PWM: Position Weight Matrix
• QT-PCR: Quantitative – Polymerase Chain Reaction
• ROC: Receiver Operating Characteristic
• RT-PCR: Reverse Transcriptase – Polymerase Chain Reaction
• SAGE: Serial Analysis of Gene Expression
• SAM: Significance Analysis of Microarrays
• SELEX: Systematic Evolution of Ligands by Exponential Enrichment
• SSA: Single Site Analysis
• TCA: TFBS Cluster Analysis
• TF: Transcription Factor
• TFBS: Transcription Factor Binding Site
• TSS: Transcription Start Site
• UTR: UnTranslated Region
• VSN: Variance Stabilization Normalization
• WTSS: Whole Transcriptome Shotgun Sequencing
Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Wyeth W. Wasserman, who has provided me with years of patient and careful guidance in my endeavours. I would also like to thank my thesis supervisory committee members, Dr. Holger Hoos, Dr. Don Moerman and Dr. Francis Ouellett, for their mentorship and support.

I am also grateful to both the current and past members of the Wasserman group, whom I have shared much stimulating interactions with and received immeasurable support from. Many thanks to Dr. Shannan Ho Sui, Dr. Debra Fulton, Dr. Jochen Brumm, Alice Chou, Dr. Elodie Portales-Casamar, Dr. David Martin, David Arenillas, Rebecca Worsley Hunt, Warren Cheung and Jonathan Lim. In particular, I gratefully recognize Drs. Ho Sui and Dr. Fulton as the key contributors to past versions of oPOSSUM, as the discussions of the earlier implementations informed a significant part of my research. Also, I couldn’t have made it this far without the administrative assistance from Dora Pak.

I would like to acknowledge the National Sciences and Engineering Research Council for funding my research. I would like to thank Dr. Rob Holt and Andrea McLeod for assistance with the plasmid sequencing, Dr. Catherine Pallen, Dr. Elizabeth M. Simpson and Dr. Blair Leavitt for the use of their laboratories. “The work was supported by grants from the National Institutes of Health (USA) 1R01GM084875, the Canadian Institutes for Health Research, the National Science and Engineering Research Council (NSERC) and Genome Canada (Pleiades Promoter Project).”

Finally, I would like to give special thanks to my parents, Chung-Kun and Sook-Ja Kwon, and my sister Jung-Eun Kwon and my brother Peter Kwon, whose unwavering support and encouragement have allowed me to complete this long journey.
1 Introduction

An astounding diversity of life exists in nature, with each organism featuring complex structures ranging from sub-cellular organelles to multi-cellular organs that enable them to function within an environmental niche. Despite the diversity, common structures and mechanisms can be found in species from widely separated branches of the evolutionary tree. The study of evolutionarily conserved mechanisms has been a fundamental tool to advance our understanding of biological systems.

Within the metazoan branch of the tree of life, the study of development can reveal biomedically relevant findings. A metazoan develops from a single-celled zygote to a multicellular adult consisting of specialized organs and tissues. At the cellular level, the developmental program can be considered a series of transitions from pluripotent stem cells to progressively more restricted tissue-specific progenitors cells, eventually producing fully differentiated cells that characterize tissues and organs. The genes controlling the process have been repeatedly shown to be responsible for human disease and disability. With the potential of stem cell therapy looming on the horizon, a fuller understanding of these master control genes stands out as a high priority for biological research. As the master control genes are fundamental components of development, they often exhibit strong patterns of conservation. A well-studied example of master control genes is the Hox gene family, which is vital for correct body plan development and organogenesis of metazoans and is highly conserved (Carroll 1995).

While the developmental stages of tissue development in metazoans have been well established, the underlying regulatory controls behind the specialization process are not well understood. The gene regulatory network is the fundamental mechanism by which a cell controls the activity of its genes. For multicellular metazoans, such as worms, flies and humans, maintaining precise spatial and temporal control of transcription is vital for correct tissue development and specialization (Look 1997;
Gellon and McGinnis 1998; Lee et al. 2004). The regulatory network governing gene transcription is a complex input-output system involving a number of players, including (but not limited to) the cis-regulatory regions of genes and the corresponding trans-acting transcription factors (TFs). Depending on the type and the number of TFs that bind to the target regulatory regions, the expression pattern of a gene can be controlled in a specific manner. Additional complexities arise from the widely dispersed regulatory region locations and presence of multiple alternative promoters. Thus, deciphering just one component of this network, transcriptional regulation, has been a strong focus of decades of research and yet remains a major challenge.

Understanding how this regulation takes place can be a key to understanding how it may contribute to human genetic diseases. Known diseases that are caused by faulty regulatory controls are growing in number. The earliest discoveries of regulation-related diseases include certain forms of thalassemias, blood disorders caused by incorrect levels of ω- and β-globin chains in hemoglobin (Kioussis et al. 1983; Driscoll et al. 1989). In some patients, genomic translocations result in the removal of the cis-acting locus control region (LCR) for the β-globin locus. Such links between regulatory disruptions and phenotypes have been found for many tissues and organs. In patients with Van Buchem disease, abnormal bone density results in enlargement of skull and mandible leading to other serious complications (Wergedal 2003). The disease can be caused by the removal of the cis-acting ECR5 element from the locus of sclerotin (SOST), a gene encoding a negative regulator of bone formation (Loots et al. 2005). The genetic eye disease aniridia is caused by disruption of the trans-acting PAX6 master regulator of transcription (Hanson et al. 1993). The onset and rate of progression of acquired immunodeficiency syndrome (AIDS) has been tied to the presence of certain alleles of common polymorphisms in the regulatory region of the CCR5 gene. Examples of regulatory region-related diseases are plentiful, and have been the subject of recent reviews (Kleinjan and Lettice 2008).
There is clear biomedical importance to developing a full understanding of the human gene regulatory network. Due to its complexity, however, much of the progress has been largely made through the study of model organisms. The popular metazoan model organisms offer different advantages and technical convenience. The laboratory mouse is expensive to study, but can provide exquisite relevance to the study of human gene regulation. Within the more distant metazoans, both flies and worms have been extensively used to study gene regulation, but the capacity to interpret the specific properties of human genes based on the research findings has been lacking. The considerable evolutionary distances between vertebrates, nematodes and insects makes interpretation challenging, but key findings in both worms and flies have revealed critical insights into diseases and conditions ranging from diabetes (Zhang et al. 2009) to limb malformations (Goodman 2002). Improving the capacity to use regulatory insights from insects and nematodes could accelerate our understanding of the human transcription regulatory network. The main goal of the research outlined in this thesis is to develop bioinformatics methods for the identification of cis-regulatory regions governing gene transcription, and identify the important regulatory modules that are conserved and shared across evolutionarily distant species.

1.1 Metazoan Tissue Development

Metazoans with bilateral symmetry, including worms, flies, and humans, form three primary tissue layers: ectoderm, mesoderm, and endoderm (Nance et al. 2005; Solnica-Krezel 2005). After the zygote formation and subsequent cell divisions, early embryos in general go through the gastrulation process to produce the three layers (Figure 1-1). Ectoderm is the source of epidermis and neurons. Endoderm develops into the digestive system, while mesoderm turns into muscle, blood, connective tissue etc. At each developmental stage, specific sets of regulatory controls are exerted to ensure that cells differentiate correctly. The most well known regulatory system involved in the development process involves the Hox family of transcriptional regulators. Hox genes are homeodomain-containing transcription factors that specify the anterior-posterior
axis and segment identity in the early embryo stages of metazoans (Carroll 1995). These genes are responsible for the development of the correct body plan and organogenesis, and are highly conserved in most metazoan species.

Figure 1-1. Differentiation of human tissues.
Gastrulation leads to formation of ectoderm, mesoderm, and endoderm. (Source: NCBI. http://www.ncbi.nih.gov/About/primer/images/layoutdiff8.gif)

1.1.1 Tissue and Structure Conservation among Metazoans

The metazoan evolutionary tree is well established, serving as a useful guide to important organogenesis events. By looking at the ‘oldest’ (farthest from humans in branch length) species with a tissue type of interest, one can analyze the commonalities in the tissue between the two species and determine the extent of conservation. Intermediate species can provide insight into the additional complexities in terms of
both the tissue structure and the regulatory controls behind the structure. A good example is vertebrate striated muscle myogenesis, a highly structured process in which mononucleate myoblasts fuse together to form multinucleate myotubes, which then develop into various classes of myofibres (Valdez et al. 2000). This differentiation process requires complex transcriptional regulation controls, which in vertebrate skeletal muscle involves two major TF families, MyoD and MEF2 (Braun et al. 1994; Rudnicki and Jaenisch 1995; Naya and Olson 1999). Muscle is present in most branches of the metazoans, from jellyfish to vertebrates. When one examines nematode muscle structure and the regulatory controls involved, one finds both commonalities and pronounced differences with vertebrates (Dichoso et al. 2000; Castanon and Baylies 2002; Fukushige et al. 2006).

Some cells exhibit key structures critical for their role within an organism. Such structures can appear in different tissues, but contain similar protein components. Cilium is a cellular structure that is used in a variety of roles, from cell motility to sensory detection. This structure has been adapted for different purposes over evolution, but the general structure and its components are highly conserved, not only between the different cell types but also between multiple species (Li et al. 2004; Blacque et al. 2005). As in muscle, the structural conservation is paralleled by retention of regulatory controls. The RFX transcription factors perform regulatory roles for cilia gene expression across vertebrates and nematodes (Emery et al. 1996; Efimenko et al. 2005).

### 1.2 Gene Regulation

In the simplest model of transcriptional regulation, as proposed for lactose-regulated transcription in bacteria (Reznikoff 1992), the regulatory organization of a gene consists of a basal promoter, a coding region, and the flanking sequence containing TF binding sites (TFBSs). When these sites are bound by the appropriate sequence-specific TFs, a complex forms to promote the recruitment and positioning of RNA polymerase at
the promoter region to initiate transcription (Hahn 2004). However, there is added complexity at many levels that is not represented by this simple model (Figure 1-2).

**Figure 1-2. Overview of transcriptional regulation.**

Groups of transcription factors (TFs) bind to sets of transcription factor binding sites (TFBSs) within or adjoining genes to activate or repress gene expression. Within the figure are examples of numerous regulatory regions, including regions proximal to the promoter, distal regions both 5’ and 3’ of the transcription start site (TSS). (reproduced with permission;(Wasserman and Krivan 2003))

![Diagram of transcriptional regulation](image)

1.2.1 **Promoter and Regulatory Regions**

The core promoter, which extends ~34 bp upstream and downstream of a transcription start site (TSS), can contain all or a subset of a TATA-box, initiator region (Inr) and downstream promoter element (DPE). The general transcription factors bind to the core promoter to form the pre-initiation complex, which positions the RNA polymerase II for proper transcription initiation. The proximal promoter region located to either side of the core promoter usually contains some transcription factor binding sites (TFBS) that participate in the regulation process. Regulatory regions can be found not
only at the 5’ upstream sequences but also within the introns and 3’ downstream sequences. The regulatory regions are spread over much larger distances as well, possibly hundreds of kilobases away from the exons of regulated genes in vertebrates (Loots et al. 2000). Some genes have been found to be under the control of regulatory regions that are located in introns of an adjacent gene (McBride and Kleinjan 2004) or even on other chromosomes (Morris et al. 1999). The mediator complex acts as the intermediary between the basal transcription machinery and the TFs bound to the regulatory regions. Evidence continues to build indicating that most genes have multiple alternative promoters that can be used in a selective manner. Depending on promoter organization, the alternative promoters may produce transcripts encoding identical or different proteins. Nematodes have a distinguishing feature from many other metazoans in that they contain operons in their genome, where multiple genes are transcribed as a single unit before being processed into separate mRNAs (Blumenthal and Gleason 2003).

1.2.2 Transcription Factors

There are estimated to be 1,500 TFs encoded by human genes and over 900 TFs by nematode genes (Reece-Hoyes et al. 2005; Fulton et al. 2009). TFs can be classified according to the structural class of their DNA binding domains, which tend to be highly conserved. For certain DNA binding domains, the DNA sequences bound exhibit class-specific DNA sequence. TFs act in a multitude of combinations, some binding as monomers, some as homodimers or multimers, or through forming hetero-complexes with different TFs. The interactions between TFs allows for a portion of observed cell-type specific expression, for instance through combinations of different TFs binding to clusters of cis-regulatory elements cooperatively (Arnone and Davidson 1997). These clusters are often referred to as cis-regulatory modules (CRMs). For each gene, specific combinations of TFBSs form units of regulatory control. This model allows for a combinatorial control of the gene, where only a limited number of TFs can create an
exponential number of unique combinations, each conferring a specific regulation to the gene (Figure 1-3).

**Figure 1-3. cis-Regulatory module examples.**

CYP1A2 has a CRM in a distal enhancer located at -2,352 to -2,094 relative to the TSS. IGF1 has a CRM in its promoter with more than 1 factor binding at the first two sites, while ALDOB has a CRM located in the first intron.

1.2.3 MicroRNA

Discoveries of new RNA classes and their roles in gene regulation provided further insight into gene regulatory mechanisms. One prominent discovery is the role that microRNA plays in gene regulation. MicroRNAs are short noncoding RNAs that act as posttranscriptional regulators (Brennecke and Cohen 2003; Ambros 2004). Processed microRNAs have an average length of 22 nucleotides, and most hybridize to target sequences within the 3’ UTR of target transcripts. The interaction of a microRNA with a target sequence in an mRNA promotes the degradation of the mRNA. Cohen and colleagues suggested that animal microRNAs confer robustness to developmental gene expression programs to ensure tissue specificity (Stark et al. 2005). Lim et al. have
shown that transfecting human cells with microRNAs that are preferentially expressed in a particular tissue type shifts the expression profile of the cells to that of the tissue, and that a significant portion of those genes with shifted (reduced) expression have the transfected microRNA target sequence in their 3’ UTRs (Lim et al. 2005). Genes with microRNA target sequences tend to be highly expressed at the developmental stage prior to the microRNA expression, while those genes that show the same spatio-temporal expression pattern as a particular microRNA lack the target sequences for the microRNA, potentially reflecting selective avoidance (Farh et al. 2005).

1.2.4 Chromatin and Epigenetics

The study of epigenetics—the heritable changes in gene function and expression without any change to the gene sequence itself—has provided researchers with better insight into additional layers of regulation (Olins and Olins 2003). Two important means of epigenetic control are histone modification and nucleosome remodeling, which in turn lead to changes in chromatin conformation (Fischle et al. 2003). Acetylation of histone tail residues tends to be associated with open chromatin and widely spaced nucleosomes. On the other hand, methylation of histone tail residues have been associated with both open and closed chromatin with tightly packed nucleosomes, depending on which specific residues undergo modification. With open chromatin, TFs have better access to the TFBSs on DNA for binding; conversely with closed chromatin, the compactness hinders TFs from binding to the TFBS. Acetylated histones facilitate better access and stability for the TFs binding to CRMs by opening up the chromatin, while methylated histones that lead to closed chromatin act to close the conformation reducing access. Thus, actively expressed genes tend to be associated with open chromatin structure and silenced genes tend to be associated with heterochromatin. It is important to note that the chromatin conformation changes result in spatial chromatin re-organization, which means that one needs to examine the chromatin in three-dimensional space in order to fully capture the details of the conformation effects. The recent development of the chromosome conformation
capture (3C) technology and its higher-throughput derivatives has provided researchers with a much-needed method of determining the distribution of enhancers and promoters in 3D space (McBride and Kleinjan 2004).

Another important regulatory mechanism coupled with chromatin conformation is the control provided by insulators - DNA sequence elements that block the influence of regulatory events from adjacent chromatin domains. There are two different types of insulators: 1) those that prevent enhancers from exerting their control on inappropriate genes, and 2) those that act as barriers against spreading of chromatin condensation to allow the adjacent genes to continue with their expression. Insulator elements have been identified in yeast, fruit flies and vertebrates (Golovnin et al. 1999; Bell et al. 2001; Zhan et al. 2001; West and Fraser 2005). In vertebrates, CTCF has been identified as the main enhancer-blocking protein that binds to the insulator elements (Brasset and Vaury 2005).

1.3 Analysis of Gene Expression

Gene expression can roughly be evaluated in two parts: RNA and protein. RNA abundance is determined by a series of processes, including transcription initiation, processing and stability. In addition to the rate of translation by ribosomes, protein abundance reflects protein processing and stability. Because of the relative ease with which cellular mRNA levels can be measured compared to protein levels, mRNA expression profiles are often used as the primary measurement of gene expression. This common practice is performed in spite of the fact that mRNA levels and protein levels often can correlate poorly, due to the actions of post-translational regulatory mechanisms (Gygi et al. 1999). For study of the transcription regulatory network, RNA measurements are more directly relevant than protein measurements. In the remainder of this chapter, I will often refer to the term ‘gene’ as being inclusive of both the transcribed RNA and the translated protein.
For this thesis, I evaluated and incorporated into my research several gene expression measurement data sets that were generated with the methodologies outlined in the following sections.

### 1.3.1 DNA Microarrays

DNA microarray technology was developed as a way of parallelizing traditional northern blotting experiments for detecting cellular mRNA levels. Microarrays are glass or silicon chips that contain arrays of spots, to each of which a unique DNA probe is affixed. In a microarray experiment, RNA extracted from the biological sample to be tested is converted to fluorescently labeled cDNA, which is then hybridized to the microarray. Spots that bind the labeled cDNA can be detected using a laser scanner, and the strength of fluorescence is used to quantify the gene expression level.

There are two common types of microarray technologies: 1) cDNA microarrays, which requires two samples with two different fluorescent dyes, and 2) oligonucleotide microarrays, which requires only one sample as they provide absolute measurement levels for each one (Bowtell 1999). It is also possible to classify the microarrays based on the types of probes used. In traditional microarrays, probes are designed to identify the known or predicted genes only. On the other hand, in tiling arrays, short fragments of DNA designed to cover the contiguous regions of a genome (which can be extended to cover the entire genome) are used as probes, regardless of gene annotation. The resolution of tiling arrays is determined by the probe lengths and spacing. This allows an unbiased reporting of transcription, allowing transcriptome profiling in the given genomic region.

Microarray measurements can be affected by multiple sources of variation over the course of experiment, including dye integration, sample preparation, hybridization and image artifacts. To account for these variations, measurement values must be normalized statistically based on both technical and biological replicates before they can be analyzed for biological interpretation. Numerous normalization methods have
been developed, including Robust Multi-Chip Analysis (RMA) and MAS5 (Lim et al. 2007).

1.3.2 Serial Analysis of Gene Expression (SAGE)

Serial analysis of gene expression (SAGE) is another method to quantitatively measure mRNA abundance in a given biological sample (Velculescu et al. 1995). The central idea behind SAGE is that a short (10-25bp) sequence tag from a transcript would be obtained and used to uniquely identify the gene. In a SAGE experiment, mRNA is isolated from a sample and short fragments are obtained from the 3’ end of transcripts using pre-defined restriction enzymes, and these fragments are concatenated using DNA ligase. The concatemers are sequenced, and the obtained sequences are computationally analyzed to identify each constituent tag, which are then mapped back to the genome to identify the genes involved. By counting the number of observations for each unique tag, quantitative measures of mRNA levels are obtained. Improved versions of SAGE have since been developed, including Long-SAGE and SuperSAGE, which employ longer tags to improve the confidence of the gene identification (Saha et al. 2002; Matsumura et al. 2005). The SAGE and aforementioned microarray methods are rapidly being supplanted by RNA-Seq.

1.3.3 RNA-Seq

RNA-Seq, also known as whole transcriptome shotgun sequencing (WTSS), is a transcriptome profiling technology based on high-throughput sequencing (Wang et al. 2009). RNA-Seq measures the cellular RNA content by sequencing reverse transcribed cDNAs. As with ChIP-Seq (explained in section 1.4.2.1), the deep coverage provided by the next generation sequencing methods allows researchers to examine the transcriptome with greater resolution than was possible with previous methods. The counts obtained are highly quantitative, and provide the added benefit of information
about alternative splice forms of RNA, allelic differences in RNA production and mutations.

1.3.4 Differential Gene Expression Analysis

Gene expression profiling measurements are often used to detect genes that exhibit differential expression across samples, such as drug-treated and untreated cells or multiple tissue types (Lockhart et al. 1996). The identification of genes that display a selective pattern of expression in such studies is a fundamental challenge in applied genomics and has therefore been the topic of research by hundreds of research groups. Thorough reviews and useful textbooks have been published about the analysis methods (Dudoit et al. 2000; Butte 2002; Cui and Churchill 2003). I will focus on a few key points here. Statistical tests such as t-tests can be used to perform the comparisons, but such analyses are often complicated by the fact that there are many genes to be analyzed but only a few observations available, leading to multiple testing and variance estimation challenges. Furthermore, no assumption of normal distribution of expression measurements can be made, which is a requirement for t-tests. Significance analysis of microarrays (SAM) is a widely used method that makes no normal distribution assumption and instead calculates the p-values based on permutations of data (Tusher et al. 2001). SAM estimates the false discovery rate (FDR), which is the proportion of data likely to have been incorrectly identified as being significant. SAM can be used for differential expression analysis on both pairwise and multiple biological samples. Another complication in expression analysis is the dependence of significance of differential expression on signal intensities. The same fold change in expression between two samples observed at high intensities is more likely to be significant than the one observed at low intensities. Variance stabilization normalization (VSN) was developed to compensate for this effect, by transforming the data so that the variance is approximately uniform at all intensity values (Huber et al. 2002).
1.4 Experimental Methods for Identification of Regulatory Regions

In order to understand how a given gene is regulated, it is important to be able to first locate the controlling regulatory regions and identify the responsible TFBSs. Numerous experimental methods have been utilized for this purpose. Researchers traditionally have relied on these methods to 1) identify the TF-DNA binding and interactions and 2) validate the effects of cis-regulatory regions on gene expression. While earlier methods involved individually isolating genomic regions of interest or tests using synthetically generated oligonucleotides, advances in high throughput methods now allow researchers to search for regulatory elements on a genomic scale.

1.4.1 Earlier Methods

Some of the data utilized in this thesis derives from older publications that identified TFBS and cis-regulatory regions using arduous laboratory methods. Protein-DNA binding can be detected through the use of DNA binding assays, of which there are several variations relevant to the thesis. Except for chromatin immunoprecipitation, the following techniques are most commonly performed in vitro using cells or cell extracts from tissue culture.

- Electrophoretic mobility shift assays (EMSA) detect protein-DNA complexes by measuring the retardation of protein-bound DNA in gel electrophoresis compared to unbound DNA (Garner and Revzin 1981). An antibody that targets the protein of interest can be added to create a larger complex for better separation, resulting in ‘supershifting’ of the complex.
- DNase I footprinting is used to elucidate the actual protein-bound DNA sequences given a larger DNA region of interest (Brenowitz et al. 1986). A dsDNA fragment is labeled at one end. With protein bound to the labeled DNA, appropriate conditions are used for DNase I digestion to obtain on average one cut per DNA molecule. The resulting product is separated using gel
electrophoresis, revealing a banding pattern of the DNA. At the site of protein binding, the banding pattern is disrupted compared to DNA digested in the absence of protein.

• Systematic evolution of ligands by exponential enrichment (SELEX) is used to determine the sequence of DNA bound by a protein of interest (Tuerk and Gold 1990). Starting with a pool of dsDNA of random sequences, those molecules bound by the protein are purified (using one of a variety of techniques) and amplified using PCR. This process is repeated until the stringency conditions are satisfied. The resulting DNA is sequenced.

• Chromatin immunoprecipitation (ChIP) is a method of determining genomic regions bound by a protein of interest *in vivo* (Collas 2010). After cross-linking proteins to the DNA, the chromatin is sheared. The protein-bound sheared DNA fragments are immunoprecipitated and the associated DNAs are determined (by sequencing or PCR).

• Reporter gene assays are employed when one needs to determine if a potential *cis*-regulatory region of interest can drive gene expression. Reporter genes are chosen based on their ability to produce measurable phenotypes, including green fluorescent protein (GFP), which fluoresces under ultraviolet light, and luciferase, which gives off light when it reacts with luciferin. A reporter construct is produced by fusing the putative regulatory region, a minimal promoter if not included in the former, and the reporter gene. The construct is introduced to a cell or transgenic organism and the resulting reporter gene expression measured.

### 1.4.2 High Throughput Methods

The development of DNA microarray technology along with the availability of genome sequences has brought significant advances in the field of gene expression analysis. This technology has been applied to DNA-protein binding analysis as well, bringing considerable improvements in regulatory region identification. It became possible to
identify functional noncoding elements on a genomic scale in the lab. The debut of high-throughput sequencing technologies has brought new advances in the analysis methods, already starting to supplant the microarray-based methods.

1.4.2.1 ChIP-chip and ChIP-seq

By combining chromatin immunoprecipitation (ChIP) with tiled DNA microarrays (chip), ChIP-chip analysis can identify in vivo DNA-protein interactions and chromatin modifications on a genomic scale (Aparicio et al. 2004). In this method, chromatin immunoprecipitation (as explained in section 1.4.1) is followed by tiling DNA microarray analysis of the purified protein-bound DNA fragments. ChIP-chip experiments have been used to identify various types of regulatory regions, including enhancers, silencing elements, and insulators (Buck and Lieb 2004). They have also been used to examine the distribution of histone modifications and their effects on gene regulation (Huebert et al. 2006). The development of this technology has allowed researchers to embark on efforts to catalogue and identify how the distribution of these functional elements and epigenetic marks vary according to cell types and physiological conditions.

However, ChIP-chip technology suffers from a number of limitations. The availability of high-quality antibodies used for the immunoprecipitation is a limiting factor. Furthermore, there is a high cost for the experiments due to the expense of tiled DNA microarrays. Because ChIP-chip uses tiling arrays, it requires large sets of arrays with narrow probe spacing to achieve high resolution and precision, but the costs of such arrays are limiting. Due to the challenges of the hybridization procedures, the microarray experiments require multiple replications that add to the expense. The DNA fragments obtained from the chromatin immunoprecipitation step are hundreds of base pairs long, as the sonication process for breaking up DNA can only achieve a minimum size of ~200 bp. Because tiling arrays are used, the resolution of experiments depends on the coverage provided by the probes. As with microarray experiments for gene expression, the statistical analysis of the data generated remains a challenge.
ChIP-Seq utilizes a different detection method than ChIP-chip. Microarrays are supplanted by high-throughput sequencing of the purified DNA fragments (Johnson et al. 2007). ChIP-Seq offers better resolution and precision than ChIP-chip. ChIP-Seq can produce highly precise results, specifying TFBSs within tens of bases from the actual protein-binding site. With ChIP-Seq, protein-DNA binding affinity can be quantified with the tag densities at the binding sites, making it easier to compare the relative affinities at different binding sites. One can increase the sensitivity of ChIP-Seq as required by increasing the sequencing depth, which is not possible with ChIP-chip. Since the introduction of this technology, there has been a sharp increase in the number of regulatory region data sets generated (Jothi et al. 2008; Cao et al. 2010).

1.4.2.2 Cap Analysis Gene Expression (CAGE)

Cap analysis gene expression makes use of the principles underlying SAGE for identification of transcriptional starting points (TSPs) (Shiraki et al. 2003). SAGE in its original form cannot be used to identify the 5’ TSPs, as the tags used are taken from the 3’ ends of transcripts. For CAGE, full-length cDNAs are selected using “biotinylated cap-trapper” technology (Carninci et al. 1996). Linker DNA oligonucleotides containing restriction enzyme sites and biotin at the 5’ ends are attached at the 5’ ends of cDNAs. The modified linker-cDNAs are separated using magnetic beads, purified and amplified with PCR. The linker-cDNAs are then cleaved with restriction enzymes, resulting in 32 bp 5’ fragments (20 bp without the linker). The fragments are again separated using magnetic beads, and are subsequently cut from biotin, concatenated and sequenced as for the SAGE method. CAGE has been used by the FANTOM 2 and FANTOM 3 Consortiums for transcript analysis in mammalian genomes (Okazaki et al. 2002; Carninci et al. 2005). An extension of this technique called deepCAGE, which combines CAGE with deep sequencing, is being used by the FANTOM4 Consortium to study the dynamics of TSPs during cellular differentiation process (de Hoon and Hayashizaki 2008).
1.4.2.3 Protein Binding Microarray (PBM)

Protein binding microarrays (PBMs) are a rapid, high-throughput method based on DNA microarrays for DNA-protein binding identification in vitro (Berger et al. 2006). In this method, purified epitope-tagged proteins are added to microarrays spotted with double-stranded DNAs, and the protein-bound DNA probes are identified by fluorescently labeled antibodies to the epitope. Alternatively, proteins directly labeled with fluorophore can be used, removing the antibody treatment step (Bulyk 2007). The dsDNA probes can be constructed from primer extension, self-hairpinning, or PCR of genomic regions. After the labeling step, the microarrays are scanned for fluorescence, and the normalized values are used to identify the DNA sequences to which the protein binds. As a control, another set of microarrays is stained with SybrGreen I, which is specific for dsDNA. Since PBM is an in vitro method, no prior knowledge of cellular conditions for in vivo TF binding is required. PBMs can also provide comprehensive data on the effects of DNA sequence variants on protein binding preference. Berger et al. determined 168 mouse homeodomain TF binding profiles (Berger et al. 2008) and Noyes et al. evaluated the 84 Drosophila homeodomain TF binding profiles using PBMs (Noyes et al. 2008).

1.5 Comparative Genomics

The biological knowledge gained from studying model organisms brings meaningful benefits to understanding human physiology. Comparative genomics facilitates this transfer, in part, through the analysis of genomic elements that are under evolutionary constraint across diverged species. The identification of orthologous genes is a vital part of this process, as it provides a framework to infer functionality of experimentally undetermined genes. The availability of multiple genomes of organisms at diverse branches of the phylogenetic tree offers an unprecedented opportunity to reconstruct the shared history of evolution of functional genomic elements. As the goal of this thesis
is to compare and identify the conserved regulatory programs, comparative genomics is the central strategy behind the bioinformatics framework that I implement.

1.5.1 Homology: Orthologs and Paralogs

Gene evolution can take place via: 1) speciation with modification, 2) gene duplication followed by speciation with modification, 3) gene loss, 4) horizontal gene transfer, and 5) fusion, fission and other gene rearrangements (Koonin 2005). Genes that share common descent are referred to as homologs, which can be divided into orthologs and paralogs. Orthologs are genes separated only by speciation events, whereas the separation of paralogs includes duplication events. However, homologs cannot always be classified into these classes, due to complex mixtures of speciation, gene duplication, gene rearrangements, gene loss and horizontal gene transfers.

On a formal level, orthologs are more consistently defined than paralogs. Orthologs are “genes derived from a single ancestral gene in the last common ancestor of the compared species” (Koonin 2005). It is important to distinguish between the orthologs and paralogs, because a common, but not required, property of orthologs is that they perform equivalent functions in the compared species, whereas paralogs are more likely to perform related but distinct roles. This does not mean that all orthologs from the compared species have one-to-one relationships, as a lineage-specific gene duplication event can take place subsequent to a speciation event. Inparalogs result from a duplication event following a speciation event, and outparalogs result from a duplication event preceding a speciation event. The inparalogs derived from a single ancestral gene are still orthologs of that gene. Moreover, gene rearrangements can bring more complications to the orthology classification, as different parts of a gene in one species may be orthologous to multiple genes in another species.
1.5.2 Sequence Alignment

An early method for the identification of homology is pair-wise sequence alignment of protein or nucleic acid sequences between two species. Two main alignment classes were developed: (i) local alignment methods, which identify conserved segments within longer sequences (Smith and Waterman 1981) and (ii) global alignment methods, which align two sequences across the entire length (Needleman and Wunsch 1970). Heuristic derivatives of the classic methods, such as BLAST, have dominated bioinformatics in the genome era (Altschul et al. 1990). Recently developed pair-wise alignment programs include LAGAN (Brudno et al. 2003a) and ORCA (developed in Wasserman lab).

Multiple sequence alignment methods have been available for decades (Chenna et al. 2003). With the release of genomes for multiple species, much effort has been made to improve the performance of the multiple sequence alignment (MSA) algorithms. The most widely utilized MSA algorithms are based on the “glocal” alignment strategy, which is a hybrid of local and global alignment methods, combined with a progressive alignment approach (Brudno et al. 2003b). With the glocal strategy, sequences are first processed with a local aligner, and the identified local alignment blocks are concatenated to form longer alignments. Starting with the sequences from the two most closely related species according to a given phylogenetic tree, pairwise alignment is performed, after which the sequence from the next species in the order of the tree is added, resulting in a progressive construction of multiple alignment. A review of the current alignment strategies is given in (Kumar and Filipski 2007).

1.5.3 Evolutionary Constraint Measures

With multiple sequence alignments and a species distance tree available, one can quantify the observed evolutionary constraint on either the individual or fixed-size blocks of nucleotides. Examples of widely used evolutionary constraint measures include phastCons and PhyloP from the PHAST package (Hubisz et al. 2011). PhastCons...
models the evolutionary process based on the phylogenetic hidden Markov model (phylo-HMM) (Fan et al. 2007). With this statistical model, phastCons can account for multiple substitutions at each nucleotide site and unequal substitution rates between different pairs of bases when estimating the probability that each nucleotide is under evolutionary constraint. PhyloP, which computes the p-values of conservation based on a neutral model of evolution, differs from phastCons in that it considers each nucleotide individually, ignoring the effects from the neighbouring bases. PhastCons is more suitable for detecting conserved elements, whereas PhyloP is better suited for evaluating signatures at particular nucleotides or classes of nucleotides. Scores from both methods are available from the UCSC Genome Browser for multiple species, and are used in several stages of the thesis research that follows.

1.6 Public Repositories of Genomic Data

The development of technologies that have enabled researchers to sequence entire genomes of organisms has brought revolutionary changes to the field of biology. With the sequence information stored in public databases, researchers can now retrieve the sequences of genomic regions of their interest within seconds. The availability of multiple genomes has enabled detailed analysis of homology among different genes. The increasing amount of data accumulated by researchers around the world is added back to the public repositories as genomic annotations, further promoting the sharing of knowledge and accelerating research. As with many other bioinformatics projects, the work performed in this thesis heavily relies on the public genomic data.

1.6.1 Public Genome Databases

The main public repositories such as NCBI and Ensembl offer the latest versions of the genomes of species that span diverse taxonomic groups, along with the annotations for functional genomic elements and comparative genomics data. As of version 59, Ensembl contains genomic annotations for 56 species from various taxonomic groups (Flicek et
al. 2010). Other public genome analysis systems such as the UCSC Genome Bioinformatics Site provide rich computational and experimental annotations (Kent et al. 2002). There are species-specific public databases as well, including Saccharomyces Genome Database (SGD), WormBase, FlyBase, and Mouse Genome Informatics (MGI) (Cherry et al. 1998; Tweedie et al. 2009; Harris et al. 2010). These databases contain detailed annotations that are specific to the respective species that might otherwise be missing from more general repositories. Each public repository adds a unique set of features and tools, offering the users with a wealth of choice in their research.

Available data can be readily obtained from these genome annotation systems, and open-source software facilitates data mining of the information for computational biologists. BioMart offers a convenient, centralized gateway to multiple repositories, providing the users with powerful data mining tools in its interface (Haider et al. 2009). Another benefit provided by Ensembl is the comprehensive application programming interface (API) based on BioPerl, an open source library of bioinformatics analysis modules for the Perl programming language (Stajich et al. 2002). This greatly facilitates computational analysis of the data available at Ensembl, as it offers a well defined and very flexible set of tools for easy manipulation of the data. For regulatory region analysis, the TFBS Perl module library allows for easy manipulation of TFBS data (Lenhard and Wasserman 2002). The UCSC Genome Browser allows users direct access to the data contained in their databases, and its Galaxy suite of bioinformatics tool set is emerging as another popular alternative for genome data mining (Goecks et al. 2010).

### 1.6.2 Ortholog Databases

The identification of orthologs across species is essential for comparative genomics. There have been numerous large-scale efforts at building a public repository for multi-species ortholog annotation, which started with single-cell organisms and then expanded to multiple vertebrate species.
The earliest effort at building an orthology database was the Clusters of Orthologous Groups (COGs) (Tatusov et al. 2001). The central clustering algorithm behind the system is based on best reciprocal BLAST matches among three species. Briefly, the protein sequence collections (proteomes) to be clustered are filtered to remove ‘promiscuous’ domains, which are widespread and generally repetitive in nature. The filtered proteomes are then searched for consistent best reciprocal hits among three species using gapped BLAST, which are classified as seed clusters. The seed clusters are analyzed case by case for false-positives, only retaining those that share conserved protein domains (“core” domains). The sequences containing promiscuous domains are returned to the set and assessed for the presence of the core domains. Initially, COGs was built for unicellular, mainly prokaryotic organisms, but the system was extended to include eukaryotic species (Tatusov et al. 2003). This eukaryotic version has been named KOGs, for Eukaryotic Orthologous Groups.

The InParanoid database was constructed to identify true orthologs and to separate paralogs into inparalogs and outparalogs (O’Brien et al. 2005). By focusing on clustering of the inparalogs, one-to-one and one-to-many orthology relationships could be captured. Starting with the proteomes from Ensembl and UniProt, pair-wise similarity scores are calculated using pair-wise sequence alignments from BLAST. Ortholog clusters of non-overlapping sequences are constructed by first establishing the sequences from two species with the best reciprocal scores. For these ortholog pairs, additional inparalogs are added separately for each species. Inparalogs are determined based on the assumption that sequences from the same species that are more similar to the main ortholog than to any sequences in the other species would belong to the same ortholog cluster. The Inparanoid database first started with a handful of organisms only, but has gone through regular expansions. The latest version 7.0 (updated June 2009) includes 100 species with a total of 1,687,023 sequences.
The HomoloGene system from the National Center for Biotechnology Information (NCBI) differs from the previous two databases in that it makes use of a taxonomic tree based on sequence similarity when assigning protein sequences to orthologous groups (Wheeler et al. 2006). The sequences from closer organisms are compared first using BLASTP, with the order of organism comparison following the order of the tree traversal to its root. The system also takes chromosomal synteny into account when assigning ortholog groups. Species-specific paralogs are determined by identifying sequences that are closer within a given species than other species. As of HomoloGene Release 64, there are 20 species with ortholog annotation. In human, there are 18,876 HomoloGene groups identified.

Ensembl Compara is built based on best reciprocal hits from pairwise species comparisons using BLASTP, but the results are processed with a unique collection of inference tools (Vilella et al. 2009). Multiple alignment of sequences clustered together based on BLASTP results is used to create gene trees, and together with a species taxonomy tree, the clustered sequences are re-processed to create higher quality groupings. Orthologs and paralogs are classified based on the constructed species and gene trees. As part of the Ensembl database collection, Compara is updated with every Ensembl release.

### 1.6.3 Transcription Factor and Regulatory Region Databases

Computational analysis of regulatory regions depends heavily on the availability of reliable TFBS profiles (details on TFBS profile computation are given in section 1.7.1). Because experimental data relevant to regulatory region analysis including validated TFBS sequences have been produced by many different research groups, they are dispersed in a disorganized manner. To alleviate this problem, a number of public repositories of TFBS profile data have been set up. The two most widely used databases are JASPAR and TRANSFAC (Sandelin et al. 2004; Matys et al. 2006). JASPAR, which recently has gone through a significant expansion (Portales-Casamar et al. 2010),
contains high quality motif models for vertebrates, insects, nematodes, plants and yeast. The TFBS matrix models in JASPAR are annotated with the target TF, species, taxonomy group, TF structural classification, and links to the original experiments in which the sequence data was obtained. While TRANSFAC contains a higher number of TFBS profiles than JASPAR, the database is redundant and users are constrained by the licensing rules of the commercial provider. Therefore within this thesis the JASPAR collection will be used exclusively.

There are additional types of regulatory region repositories. The Open REGulatory ANNOtation (ORegAnno) system is a database for the curation of known regulatory elements based on open access principles, where annotation collected from literature is first cross-referenced with various genome databases and PubMed and then recorded (Montgomery et al. 2006). PAZAR, another open access database for cataloguing regulatory regions, allows users to maintain individual regulatory region collections and distribute them within a larger ‘information mall’ format (Portales-Casamar et al. 2009). DBD is a database of TF protein sequences organized by DNA-binding domains (Wilson et al. 2008). The TF structural classification used by JASPAR is in part based on the categories distinguished within DBD. Drosophila DNase I Footprint Database and REDfly are important sources of experimental data for TFBSs in fruit flies (Bergman et al. 2005; Gallo et al. 2006).

1.7 Bioinformatics Approaches for Regulatory Region Analysis

Because experimental methods of regulatory region identification are expensive in terms of both time and cost, much effort has been made by the bioinformatics research community to develop reliable computational prediction method for cis-regulatory regions. Unlike the success seen with prediction methods for protein coding regions, searching the genome sequence for TFBSs has been, in large part, frustrating because the short lengths of TFBSs lead to unacceptably high false positive rates. However, with the growing availability of high-throughput data from ChIP-chip and ChIP-Seq
experiments for TF identification, combined with other sources of data, it is becoming more feasible to create useful computational prediction methods for regulatory regions. The data generated from these experiments can provide some of the necessary information for delineating functional regulatory regions, which cannot be deciphered from genomic sequences alone.

1.7.1 TFBS Representation

The *de facto* method for modeling a TFBS is based upon the consensus of the known binding sites for the given TF. Binding sites are aligned and the frequency of nucleotides in each column of the alignment is used to specify the key bases that the TF recognizes. The simplest model used to represent a TFBS is the consensus sequence, where each nucleotide position is recorded by a letter in the International Union of Pure and Applied Chemistry (IUPAC) ambiguity codes (Cornish-Bowden 1985). While this depiction is simple and compact, it cannot quantitatively represent the distribution of the nucleotides in each column of the alignment. As TFs can generally tolerate some variability from consensus in most positions of the binding site, the loss of this information is a significant disadvantage.

1.7.1.1 Matrix Model

The most common method of modeling a TFBS is the position specific scoring matrix (PSSM), frequently called the position weight matrix (PWM). The matrix model reflects the likelihood of finding a particular base at each position of the binding region (Stormo 2000). A well-defined matrix produces scores that directly correlate with the binding energies between the TF and its binding site (Stormo and Fields 1998). The process of constructing a PSSM given a set of TFBSs is given in Figure 1-4. The first step is to create a position frequency matrix (PFM) from the alignment of binding sites, where the number of each nucleotide is enumerated at each position. A PSSM is calculated from a
PFM by dividing the nucleotide empirical probabilities by the expected background probabilities and converting the quotient to log scale.

The PSSM value of base $b$ at column $i$, $W_{b,i}$, is calculated by:

$$ W_{b,i} = \log_2 \frac{p(b,i)}{p(b)} \quad (1) $$

where $p(b,i)$ is the corrected probability of base $b$ found at column $i$, and $p(b)$ is the background probability of base $b$. The corrected probability $p(b,i)$ is computed by:

$$ p(b,i) = \frac{f_{b,i} + s(b)}{N + \sum_{b \in \{A,C,G,T\}} s(b)} \quad (2) $$

where $f_{b,i}$ is the frequency of base $b$ at column $i$ and $N$ is the sum of all binding site sequences in the alignment, and $s(b)$ is a pseudocount correction function used to correct for small sample sizes. There are a number of pseudocount correction functions available, ranging from simply adding a small non-zero constant to fitting a Dirichlet prior to all known PFM (Nishida et al. 2009).

A PFM is often visualized as a sequence logo, where the frequency of nucleotides observed in each column is used to calculate information content (in bits). Each nucleotide occurrence at each position is scaled according to the total information content in that column. The probability values in (2) can be used to calculate the information content at each column by:

$$ D_i = 2 + \sum_{b \in \{A,C,G,T\}} p_{b,i} \log_2 p_{b,i} \quad (3) $$

where $D_i$ is the information content at position $i$. 
Figure 1-4. Transcription factor binding site (TFBS) and position-specific scoring matrix (PSSM).

A PSSM is constructed by counting the base frequencies at each binding site position and converting them to log odds ratios. The PSSM is used to scan the sequence segment containing the binding site in question, which will score higher than the rest of the sequence. The sequence ‘CTACGTATTAT’ receives the score of 14.0 in this example.

To search for TFBSs, sequences are scanned with the PSSM for the TFBS of interest, and the score at each sequence position is calculated by summing the PSSM values. However, because these binding sites are so short (7 - 10 nucleotides long on average), the likelihood of finding potential sites on any given sequence is high, resulting in many spurious hits. For example, Fickett has shown that the binding model for MEF2 results in one MEF2 binding site prediction per 5000 base pairs in the human genome (Fickett 1996). This has led to the TFBS Prediction Futility Conjuncture, which states that individual TFBS predictions by themselves generally provide little information on the identification of binding sites actually functioning in vivo (Wasserman and Sandelin 2004). Approaches for addressing this problem are discussed in sections 1.7.3 and 1.7.4.
1.7.1.2 Hidden Markov Models

Although widely used, the matrix model suffers from a number of limitations. Because the model relies on a fixed length matrix, it cannot easily represent a TFBS that contains a variable spacer within the site. Also, because the matrix model treats each column independently, it cannot represent any inter-dependence among different positions of a binding site. Hidden Markov models (HMMs) have been employed by a number of researchers to create a modeling framework that can overcome these limitations (Tokovenko et al. 2009). A HMM can be depicted as a directed network of states. A state can emit symbols (such as nucleotides) with a pre-defined emission probability, and the transition between any two states has an associated transition probability. A specific path through the states produces a sequence of symbols, and this path would be taken with a probability defined as the product of all emission probabilities for the symbols and the transition probabilities between the states. While using HMMs can overcome the limitations imposed by the matrix model, PSSMs still remain the predominant form of TFBS representation due to their simplicity, the ease of computation and the fact that relatively few TF classes tolerate the variability of spacing and site organization that HMMs address. Within the thesis only PSSMs are used, but in the long-term HMM-based methods may replace PSSMs.

1.7.2 Motif Discovery

The generation of PSSMs requires a set of binding site sequences for a TF. In many cases however, one is faced with a situation where the putative target sequences are too long for accurate alignment of the binding sites. As many experimental techniques for TFBS identification cannot provide the resolution necessary for direct alignment of the binding site sequences, one is faced with the need to discern the binding sites from sequences much longer than the sites themselves. In these situations, it is necessary to perform motif discovery using pattern discovery algorithms. There are two major categories of pattern discovery algorithms, word-based and probabilistic (Das and Dai
Word-based algorithms count the occurrences of a nucleotide sequence ('word') with a given length in the input sequence set and compare it against its frequency in the background set, whereas probabilistic methods search for the most over-represented patterns using some form of directed or random selection. As finding the optimal pattern is a NP-hard problem, heuristic methods such as Gibbs sampling are employed (Lawrence et al. 1993). Many software packages have been implemented based on variations of these approaches, such as MEME, AlignAce and Weeder; Tompa et al. performed a comprehensive benchmark of 13 motif discovery tools and discussed the issues involved in the selection of the correct approaches for the evaluation of their performances (Tompa et al. 2005). Within the thesis, the MEME motif discovery package is used. MEME (Multiple EM for Motif Elicitation), one of the most widely used motif discovery tools, identifies ungapped motifs that occur repeatedly in the input sequence set through expectation maximization and maximum likelihood methods (Bailey et al. 2006).

1.7.3 Phylogenetic Footprinting

The search for functional regulatory elements in the metazoan genomes is severely hampered by the fact that one needs to detect short motifs of several base pairs within large genomic regions that can range in kilobases. Efforts into reducing the search space to enhance the specificity of regulatory region predictions have gained a powerful tool in the form of phylogenetic footprinting (Wasserman et al. 2000). With this strategy, one restricts TFBS search space to conserved orthologous genomic regions in different species only. Functional regulatory regions are often under evolutionary selective pressure compared to the background sequence. If the expression pattern of a gene is conserved among a group of species, the CRM that confers the pattern is likely to be retained as well. Mutations leading to significant changes in the TFBSs of a CRM can destroy its functionality, as the correct binding of a TF is dependent on the TFBS sequence. By employing phylogenetic footprinting with closely related species, one can
determine which non-coding regions of the genome are conserved, and limit one’s search for the regulatory regions to these conserved segments.

While orthologous coding regions are relatively well conserved over long evolutionary distances, the same is not usually true of non-coding regulatory regions. Still, phylogenetic footprinting using human and mouse alignments has been shown to reduce the search space by up to 80%, and up to 70% of functional binding sites have been located in the conserved regions (Dermitzakis and Clark 2002; Lenhard et al. 2003). While there have been genome-wide ChIP-chip studies for transcription factors that have questioned the extent of functional regulatory element conservation, it is important to note that not all binding sites detected through ChIP experiments function as cis-regulatory elements themselves (Odom et al. 2007).

However, the presence of functional regulatory motifs in non-conserved genomic regions is a valid concern; because the TFBS are short and TFs have degeneracy in their binding requirement, the likelihood of a TFBS being lost or a new TFBS arising by chance as species diverge is relatively high (Dermitzakis and Clark 2002; Dermitzakis et al. 2003). Although the functional motifs are under evolutionary constraint, it has been shown that substantial TFBS turnover occurs among related species. The short lengths of motifs mean that it is necessary to distinguish conserved motifs due to evolutionary constraints from those that are conserved by chance alone. Distinguishing between the two classes of conserved motifs requires numerous species with varying levels of evolutionary divergence. With the increasing availability of multiple genomes, large scale multiple alignment-based conservation scores such as phastCons and phyloP are now available for major taxonomic groups. While applying phylogenetic footprinting can improve the specificity of regulatory element detection drastically, sensitivity does suffer in the process. Determining the appropriate application of phylogenetic footprinting is an ill-defined process, dependent upon the conservation scoring methods and species being studied.
1.7.4 *cis*-Regulatory Module Detection Methods

One way of improving TFBS prediction is to develop models that incorporate the interactions between elements required for functional CRMs. Instead of scanning the genome for single TFBS, CRM detection methods look for the presence of clusters of TFBSs for groups of TFs known to work in concert (Wasserman and Krivan 2003). By limiting predictions to such TFBS clusters within a short (often arbitrarily determined based on past empirical observations) distance, these algorithms have been shown to reduce the false positive rate of TFBS predictions. There has been a steady development of various CRM detection programs over the last few years (Table 2-1). One of the earliest CRM detection programs used logistic regression analysis (LRA) to distinguish between CRM and non-CRM sequences (Wasserman and Fickett 1998). Some methods are based on hidden Markov models (Cister, ClusterBuster), while others are based on computation of the statistical significance of sets of non-overlapping potential TFBSs (MSCAN) (Frith et al. 2003; Alkema et al. 2004a). The latest methods incorporate phylogenetic footprinting to improve their specificity (Warner et al. 2008; He et al. 2009). The reliability of a subset of these methods will be explored within the thesis.

1.7.5 Regulome

As model organisms such as worms and flies are distant from humans in the evolutionary tree, it is difficult or impossible to apply phylogenetic footprinting directly for TFBS studies. To overcome this problem, a strategy known as Regulog analysis has been introduced (Alkema et al. 2004b). Algorithms following this strategy search for TF models common to a list of genes in one species with similar expression patterns and compare them to the orthologous genes in another species. Given a candidate TF binding model and a set of orthologous gene sequences, one penalizes the candidate TF if putative binding sites are not conserved between closely related species and rewards the candidate if the TFBSs are conserved over long evolutionary distances.
There has been some previous work that employs Regulog-like analyses to relate gene regulation between distant model organisms and humans. GuhaThakurta et al. analyzed the worm genome to identify the motifs involved in muscle gene regulation, and transferred the predicted motif models over to vertebrates and evaluated TFBS overrepresentation in the vertebrate muscle genes (GuhaThakurta et al. 2004). In order to find those genes that are involved in ciliary and basal body biogenesis and function, Li et al. have compared the genomes of Chlamydomonas, which are ciliated, and Arabidopsis, which is not flagellated (Li et al. 2004; Blacque et al. 2005). By selecting genes that are unique to Chlamydomonas and looking for orthologous genes in humans with similar regulatory mechanisms, they were able to identify a flagellar and basal body proteome in human, including a gene that is responsible for Bardet-Biedl syndrome. Zhang et al. has applied this idea to study the transcription factor FOXO target genes in worms (Xuan and Zhang 2005). The FOXO family of TFs belongs to a larger family known as Forkhead, which is well conserved, even with vertebrates. Zhang et al. perform a Regulog-style analysis to find human genes that are orthologous to the worm genes and are also under the regulatory control of FOXO. Thus far, the implementation of Regulog methods for metazoan species has been limited to specific biological processes using carefully selected species.

### 1.8 Expansion of Regulatory Region Data

Until recently, studies of functional elements in noncoding regions had been hampered by a lack of high quality large-scale data. Data were obtained from low-throughput experiments such as gene reporter assays. Often, the only regulatory region data for a given species were limited to a few TFs and a small number of genes. Advances in high-throughput methods have brought renewed focus on the genome-wide search for functional noncoding regions such as CRMs.
1.8.1 ENCODE Data

The ENCODE (ENCyclopedia Of DNA Elements) Consortium was established with the purpose of identifying and annotating all functional elements in the human genome (Birney et al. 2007). It was launched as a pilot project in 2003, with the goal of identifying the functional elements in 1% of the human genome. In the pilot phase, there was an emphasis on the development of high-throughput methods to identify and catalogue functional elements. In this phase, experimental techniques such as tiling arrays, tag sequencing, and QT-PCR were employed, along with the associated computational methods for analysis of the generated data. In 2007, following the initial trial with the pilot phase, the project was expanded to cover the entire human genome. The data generated by the consortium are continuously being released to the public.

1.8.2 modENCODE Data

Two of the most widely used model organisms are Drosophila melanogaster (fruit fly) and Caenorhabditis elegans (nematode). These species have served as an invaluable platform for genetic experiments, where their simpler structures, ease of genetic manipulation and shorter life cycles allowed researchers to perform studies that are intractable in vertebrates. Despite their importance as experimental platforms, our understanding of their genomes is still rudimentary. While fruit flies and nematodes contain comparable numbers of TFs as vertebrates (900 to 1000 vs. ~1,500 TFs in humans), there had not been systematic studies of the gene regulatory networks and functional noncoding elements in these species (Vaquerizas et al. 2009). In part to overcome these limitations, the model organism ENCODE (modENCODE) consortium was launched in 2007 as an offshoot of the ENCODE project, with the goal of comprehensive annotation of functional elements in fruit fly and nematode genomes (Celniker et al. 2009).
The modENCODE consortium is currently working on the identification of the transcripts, TFBSs, chromatin signatures, and DNA replication mechanisms of fruit flies and nematodes. Various experimental techniques are being employed, including (but not limited to) ChIP-chip, ChIP-Seq, RNA-Seq, RT-PCR and mass spectrometry. By generating cell- and tissue-specific data for transcripts, the associated TFBSs, small RNA binding sites and structures, DNA replication origins and other chromatin marks and interaction sites, the consortium hopes to understand the details of the functional genomic landscapes of the model organisms. The data generated is being rapidly released in coordination with major public databases such as UCSC, FlyBase and WormBase, facilitating the use of the data for further research and understanding the inner workings of the model organisms. The modENCODE data is used in the later portions of the thesis.

1.9 Thesis Overview

In this thesis, I develop and evaluate bioinformatics methods for the identification of cis-regulatory regions governing gene transcription. The work initiates with the study of skeletal muscle regulatory sequences in mammals, including both computational prediction and laboratory testing to evaluate the performance of the methods. The work progresses to the engineering of software for the identification of enriched TFBS patterns across sets of genes; software designed to work across the popular animal model organisms. The software serves as the framework for the identification of important regulatory modules in co-expressed gene sets that are conserved and shared across evolutionarily distant species. The work delivers useful methods to the scientific community for the characterization of regulatory programs for experimentally defined gene sets, while offering insight into severe limitations of sequence analysis for regulatory sequence annotation.
In Chapter 2, I describe the computational prediction and experimental validation of skeletal muscle cis-regulatory modules. Predictions are made by scanning the human genome for skeletal muscle-specific CRMs using three computational prediction tools. A subset of predictions was validated in a C2C12 murine cell culture model using a dual reporter expression assay. Of the 339 candidate CRMs tested, only 19 were able to drive muscle-specific expression. The low success rate illustrates the limitations of the existing computational CRM discovery methods. Further analysis of the functional sequences indicates a bias in mononucleotide sequence composition for functional sequences, suggesting avenues for future improvement of the CRM prediction methods.

In Chapter 3, I describe the oPOSSUM-3 system, a major redesign of a popular regulatory motif over-representation analysis software. Given a set of genes thought to be regulated by a common set of TFs, the system determines which motifs are over-represented in the gene sequences relative to background. The new system provides significant improvements and new features that make it more suitable for genome-scale data analysis. The system incorporates phylogenetic footprinting based on phastCons conservation scores calculated upon multiple sequence alignments. The adoption of the JASPAR 2010 TFBS profile database results in a 5-fold increase in the number of profiles incorporated, rendering the system for the study of TF combinations implemented for the previous version of oPOSSUM unsuitable for practical usage due to the heavy computational costs. A new system was implemented to identify sets of motifs that occur in combination. Recognizing the increasing use of ChIP-Seq procedures, the new system was designed to support sequence-based analysis, complementing the classic gene-based analysis methods. The performance of the system is evaluated on reference gene collections, revealing the influence of sequence composition on the results. Procedures to mitigate these effects are developed and shown to improve performance.
In Chapter 4, I describe the comparative analysis of the regulatory programs among evolutionarily divergent species and report on the extent of conservation in the major regulatory programs. Multiple genes known to be under common regulatory controls (namely, muscle, cilia, and Nfe2l2-responding genes) and their orthologs in human, fruit flies and nematodes were compared for regulatory signatures using oPOSSUM-3, and both the commonalities and differences found among the three species are discussed. While a number of significant differences do exist among the three species, orthologs of key TFs that function as main regulatory switches in each gene set could be identified in each species despite their evolutionary distances.

In Chapter 5, I present a summary of the results from the previous chapters and a discussion of what future research directions should be taken to improve upon the findings provided by this thesis, in light of the current developments and advances being made in the genomics field.
2 Validation of skeletal muscle cis-regulatory module predictions reveals nucleotide composition bias in functional enhancers

2.1 Introduction

A regulatory network represents the complex interplay between regulatory proteins and biochemical processes that govern when and where genes are expressed. Two important components of a regulatory network are cis-regulatory modules (CRM), composed of functionally interacting clusters of transcription factor binding sites (TFBS) sufficient to confer a pattern of expression upon a promoter, and the corresponding trans-acting transcription factors (TFs) that bind to a CRM to regulate transcription initiation. The multiple TFBS that constitute a CRM allow for combinatorial control of expression; a limited number of TFs can participate in an exponential number of combinations with each potentially conferring specific patterns of gene activity (Arnone and Davidson 1997).

CRMs can be situated almost anywhere relative to the structure of a gene: both near and far (even exceptionally far) from the promoter region(s) at which transcription initiates. While there are indications of quantitative orientation effects in some cases (Shimizu et al. 2002), CRMs are generally thought to be active in either direction relative to a gene promoter. Linear distance in primary sequence is no indication of the three dimensional distance (or orientation) within the nucleus. Regulatory regions can be located in introns of an adjacent gene (McBride and Kleinjan 2004; Ilnytska et al. 2009), can skip over intervening genes (Ling et al. 2004) and there are suggestions that CRMs can act on genes located on different chromosomes (McBride and Kleinjan 2004; Miele and Dekker 2008). Reflecting these properties, the discovery of CRMs stands out as a significant challenge for both computational and experimental research.
In multicellular organisms, maintaining precise spatial and temporal control of transcription in various cell types is vital for correct tissue development and specialization (Frech et al. 1997; Look 1997; Lee et al. 2004). One of the most widely studied “programs” of tissue development is the regulation of skeletal muscle differentiation. Myogenesis is a structured process, in which mononucleate myoblasts fuse together to form multinucleate myotubes, which then develop into classes of myofibres (Arnold and Winter 1998). C2C12 cells provide a popular model for this process, with an easily triggered switch between the growth and differentiation phases (Gramolini and Jasmin 1999). Any tissue differentiation process requires complex transcriptional regulation controls. For skeletal muscle, differentiated cell gene expression involves at least two major TF families, the myogenin family and the MADS family (Braun et al. 1994; Rudnicki and Jaenisch 1995; Naya and Olson 1999). In many differentiation processes, multiple proteins within a homology-based family can participate in the regulatory control of gene expression at overlapping temporal stages of the process. Skeletal muscle differentiation follows this model; thus the myogenin family may equally refer to Myogenin, MyoD, and Myf-2 while the MADS set encompasses both Srf and multiple members of the Mef2 gene family. Dozens of muscle-specific CRMs have been identified (Wasserman and Fickett 1998; Angus et al. 2001; Zheng et al. 2002), usually based on reporter gene assays in the C2C12 cell culture myogenesis model.

Aided by the relatively plentiful set of skeletal muscle CRMs, much effort has been made by the bioinformatics research community to develop predictive algorithms for CRM discrimination. Multiple CRM detection programs have been developed, which look for clusters of TFBS specific to the TFs known to be involved in the cell type of interest. An original discriminative method to distinguish between CRM and non-CRM sequences based on a logistic regression analysis (LRA) procedure has been followed by a plethora of more advanced approaches (Table 2-1) (Wasserman and Fickett 1998). For example, MSCAN makes use of motif-specific p-values to compute the statistical significance of sets of non-overlapping potential TFBSs (Alkema et al. 2004), while ClusterBuster is based on a hidden Markov model that incorporates heuristics to improve predictive
performance (Frith et al. 2003). None of the methods are sufficiently reliable for direct genome annotation; the specificity of predictions is sufficiently low that laboratory validation is essential to distinguish functional CRMs. The overall performance of the methods and the properties that differentiate the functional CRMs from the false candidates remain to be determined.

**Table 2-1. List of some of the published CRM prediction programs.**

<table>
<thead>
<tr>
<th>Name</th>
<th>Method</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRA</td>
<td>Logistic regression analysis</td>
<td>Wasserman and Fickett</td>
</tr>
<tr>
<td>MSCAN</td>
<td>Motif-specific p-values</td>
<td>Johansson et al.</td>
</tr>
<tr>
<td>MCAST</td>
<td>Hidden Markov model</td>
<td>Bailey and Noble</td>
</tr>
<tr>
<td>Cister</td>
<td>Hidden Markov model</td>
<td>Frith et al.</td>
</tr>
<tr>
<td>COMET</td>
<td>Hidden Markov model</td>
<td>Frith et al.</td>
</tr>
<tr>
<td>Cluster-Buster</td>
<td>Hidden Markov model</td>
<td>Frith et al.</td>
</tr>
</tbody>
</table>

In some cases, the prediction of CRMs has been coupled with phylogenetic footprinting under the premise that sequence conservation of known CRMs and TFBS is indicative of function and therefore a conservation filter will improve the positive predictive value of the CRM prediction methods (Wasserman and Fickett 1998; Blanchette and Tompa 2002; Sinha et al. 2006; Warner et al. 2008). It is often the case that the regulatory sequences display evidence of evolutionary selective pressure compared to the background rates of sequence change in non-functional sequence (Gumucio et al. 1992; Shelton et al. 1997). If the expression pattern of a gene is conserved between two species in the same taxonomy class, the CRM that confers the pattern is likely to be retained as well (although the individual TFBS within the CRM may be altered). By applying phylogenetic footprinting to the analysis of closely related species (i.e. 50-100 million years of separation for vertebrates), it becomes possible to concentrate predictions within a subset of regions in the conserved segments of genes. Improved specificity is balanced against the reduced sensitivity imposed by any filter.
Once predictions of regulatory sequences have been made, laboratory validation is required to confirm regulatory function. One of the most widely used methods for validating computational predictions of regulatory regions are reporter gene assays in a cell culture model system (Chalfie et al. 1994). A fusion construct of the predicted regulatory sequence and a reporter gene with a basal promoter in a plasmid is transiently transfected into cells, and the reporter gene activity is measured to determine the regulatory impact that the tested sequence exerts. It is feasible to conduct larger-scale experiments to investigate functional properties of panels of candidate CRMs and promoters within cells. Cooper et al performed a large screen of promoter activity in 16 cell lines on all predicted promoters in the 1% of the human genome targeted for in depth annotation by the ENCODE Project (Cooper et al. 2006). Similarly, relatively large-scale in vivo enhancer studies have been performed using highly conserved (human to fish) sequences driving reporter gene expression in transgenic mouse embryos, leading to the identification of 75 forebrain-specific enhancers (Pennacchio et al. 2006). Kappen et al. analyzed the regulatory controls for lsl, a LIM/homeodomain transcription factor, by inserting randomly sheared 8-10 kb fragments from the lsl genomic locus into reporter constructs and testing for expression both in vitro and in vivo (Kappen and Salbaum 2009). Using a single copy insertion mouse transgenesis procedure, the Pleiades Promoter Project evaluated over 100 candidate regulatory sequences for the capacity to direct selective patterns of reporter gene expression in the developed brain (Portales-Casamar et al. 2010). The development of higher-throughput approaches to verify enhancer and promoter function has been a focus of recent efforts to annotate vertebrate genomes.

The properties of skeletal muscle CRMs have been widely studied, but relatively few novel functional CRMs have been described since CRM prediction methods have emerged. To quantify the performance of CRM prediction methods requires a new body of reference data. We generated predictions of CRMs with three published methods and assessed the predictive benefit of sequence conservation and annotation of the expression patterns of proximal genes. We employed LRA, MSCAN, and ClusterBuster to scan the human genome for putative skeletal muscle regulatory regions, and tested a
subset for the capacity to drive reporter gene expression in a selective manner in the C2C12 cell skeletal muscle differentiation assay. We compare the reporter gene expression in immature myoblasts against expression in mature myotubes, as well as in a fibroblast cell line. Based on the outcomes of the analysis, we define additional properties of sequence composition that are predictive of function and establish a new reference collection for the continuing development of predictive methods.
2.2 Methods

2.2.1 Candidate CRM Region Selection

2.2.1.1 Human Genome Search Regions

Promoter regions are identified following the procedure described for the oPOSSUM database (Ho Sui et al. 2005). The oPOSSUM database contains the set of genes identified as being in one-to-one human and mouse ortholog pairs based on annotations in Ensembl v. 41 and UCSC hg18/mm8 whole genome alignments. For each ortholog pair, 10 kb upstream and downstream of a TSS is searched for CRMs. All noncoding regions are included in the search, including intergenic regions, introns, and untranslated regions (UTR) of exons; protein coding portions of exons are excluded. Any noncoding region that constitutes a portion of a coding exon in an alternative transcript is removed from the selection process. All alternative transcription start sites (TSS) supported by either human or mouse Fantoms3 CAGE evidence were identified and 50 bp on either side of each TSS was excluded in order to avoid core promoter regions (Kawaji et al. 2006).

2.2.1.2 Muscle cis-Regulatory Module Prediction

CRM prediction tools were used to search for muscle-specific regulatory modules within the specified genome sequences. Logistic Regression Analysis (LRA), MSCAN, and ClusterBuster were applied to the human genomic sequence regions specified above (Wasserman and Fickett 1998; Frith et al. 2003; Alkema et al. 2004). The input TFBS motif models were taken from JASPAR, a database of transcription factor binding site profiles (Sandelin et al. 2004). The models used were MEF2A (MA0052), SRF (MA0083), MYF (MA0055), TEAD (MA0090), and SP1 (MA0079); TFs with described key roles in muscle-specific gene expression. Predicted CRMs composed entirely of SP1 TFBS were excluded.
2.2.1.3 Conservation Analyses

The candidate regions were analyzed for conservation based on phastCons scores (generated with 28 placental mammal genomes) obtained from the UCSC Genome Annotation system (Karolchik et al. 2008). For a region to be classified as conserved, the presence of at least one sub-region with phastCons scores of 0.7 or greater over 20 bp is required. For each region, both the mean and the maximum phastCons scores were calculated and sub-regions with phastCons scores over 0.7 were extracted and the ratio of the length of these sub-regions over the total length of the region calculated. For phylogenetic depth evaluation, three sets of human phyloP scores (generated with 46 vertebrates, 46 placental mammals and 46 primates; database version hg19) were obtained from the UCSC Genome Annotation system.

2.2.2 Experimental Validation

2.2.2.1 Cell Culture

Mouse C2C12 myoblasts (ATCC CRL-1772; American Type Culture Collection; Manassas, VA, USA) and mouse NIH-3T3 fibroblasts (ATCC CRL-1658; American Type Culture Collection; Manassas, VA, USA) were maintained in Dulbecco’s modified Eagle’s medium, supplemented with 10% (v/v) heat inactivated fetal bovine serum, 100U/ml penicillin, and 100µg/ml streptomycin. The cultures were grown at 37°C and 5% CO₂. Differentiation of myoblasts into myotubes was induced by transferring C2C12 cells to differentiating media consisting of 2% (v/v) horse serum, 100U/ml penicillin, and 100µg/ml streptomycin. The media and reagents for cell culture were obtained from Gibco-Invitrogen (GIBCO-Invitrogen Canada, Canadian Life Technologies, Burlington, ON, Canada).
2.2.2.2 Plasmids and Cloning

Prime3 was used to design the flanking primers for each predicted CRM for PCR (Rozen and Skaletsky 2000). After performing PCR with the designed primers (synthesized by Invitrogen Corporation (Carlsbad, CA, USA)), 20 ng of each PCR product was pooled, which were then purified using the PCR purification kit (NEB, Mississauga, ON, Canada) and subcloned into the pGL-3 promoter luciferase vector (Promega; Fisher Scientific, Nepean, ON, Canada) via Kpn I and Bgl II restriction enzymes sites. Restriction digest was performed overnight at 37 °C. Post-digestion, the vector was dephosphorylated with calf intestinal alkaline phosphatase (NEB, Mississauga, ON, Canada). The restriction enzyme-digested PCR products and the vector were gel-purified using QIAquick gel extraction kit (Qiagen Inc. Mississauga, ON, Canada) and ligated using T4 DNA ligase (NEB, Mississauga, ON, Canada).

A set of control clones and a sample of the library were prepared. Constructs were transformed into sub-cloning efficient DH5α chemically competent E. coli cells (GIBCO Invitrogen Canada, Canadian Life Technologies, Burlington, ON, Canada) via heatshock at 42 °C and plated on LB agar plates containing 100 μg/ml of Ampicillin for preliminary bacterial colony screening. Colonies were picked and inoculated overnight in 3ml LB broth with ampicillin. Plasmids were prepared using QIAprep Spin Miniprep Kit (Qiagen Inc. Mississauga, ON, Canada). Sequence confirmation was performed by the CMMT/CFRI DNA Sequencing Core Facility.

2.2.2.3 High-throughput Screening of Clone Libraries

Large-scale transformation, colony picking, miniprep, and sequencing reactions with the constructs were performed (Genome Science Centre, Vancouver, BC, Canada). 1 μl of ligation mix was transformed by electroporation into E. coli DH10B T1 resistant cells (Invitrogen). Transformed cells were recovered using 1 ml of SOC medium and plated onto 22 cm × 22 cm agar plates (Genetix) containing 100 μg/ul ampicillin. Bacterial colonies were picked from the agar plates and arrayed into 384-well microtiter plates (Genetix) using a QPIX automated colony 15 picker (Genetix). Plasmid preparations
were performed via an alkaline lysis protocol. DNA sequencing reactions were prepared using a Biomek FX workstation (Beckman-Coulter) and performed using BigDye 3.1 (Applied Biosystems). Analysis of the resulting sequences to the target DNA regions was performed with AlignX from the Vector NTI software (Invitrogen).

### 2.2.2.4 DNA Concentration Measurement and Normalization

Concentration of the plasmid products was quantified using Picogreen assays (GIBCO-Invitrogen Canada, Canadian Life Technologies, Burlington, ON, Canada) via fluorescence measurement with a POLARstar Omega microplate reader (BMG Labtech; Fisher Scientific, Nepean, ON, Canada). All DNA samples were normalized to 100ng/μl per well.

### 2.2.2.5 Transfection and Reporter Gene Assays

Two sets of C2C12 myoblasts and one set of NIH-3T3 fibroblasts were seeded in 96-well plates at a density of 6000 cells per well. The myoblasts were divided into two sets so that one set could be harvested as myoblasts, while the other set could be differentiated into myotubes prior to harvest. After 24 h (at 70 % confluency) in growth media, the cells were transfected with 200ng of a pGL3-promoter firefly luciferase plasmid construct and 20ng renilla phRL-TK internal control luciferase plasmid (Promega, Madison, WI) using Lipofectamine 2000 according to the manufacturer’s protocol (GIBCO-Invitrogen Canada, Canadian Life Technologies, Burlington, ON, Canada). At 24 h post-transfection, the myoblast C2C12 set and the NIH-3T3 fibroblasts were harvested and luciferase activity measured using the Dual-Luciferase Reporter Assay System (Promega, Madison, WI) and a POLARstar Omega microplate reader (BMG Labtech; Fisher Scientific, Nepean, ON, Canada). The final set of C2C12 myoblasts was switched to differentiating media 24 h after transfection, and incubated for 96 h for differentiation into myotubes. For each clone, duplicate transfections (technical replicates) were performed. The reporter gene activity assays were carried out in two phases. In phase 1, all plasmid constructs were tested in the three cell types. In the second phase, only myotube and myoblast activities were assessed.
2.2.2.6 Technical Challenges

Compared to the initial expectations of the authors, the number of successfully validated muscle-specific CRMs was quite low. One significant contribution to the low success rate was technical failures in the transfection assays. It has been reported that the outer columns and rows in 96-well plates can exhibit lower expression values compared to the inner wells (Malo et al. 2006). For this study, no special precautions were taken against such effects because it was expected that any depressed expression would be normalized across the experiments, as each well location was treated independently for analysis. The ultimate measure was the relative expression observed at a position between the myoblasts and myotubes. However, a large fraction of undetectable expression was observed in the outer wells (Figure 2-1).

The data analysis was made more complex by the stochastic nature of the laboratory procedure. Because the plasmids from the clones are picked and transferred randomly to the 96 well plates, the number of replicates for each insert sequence of interest is variable. In this study the replicates range from 1 to 12, but the median was 2.
Figure 2-1. Biases in the location of wells with successful reporter assays.

a) Distribution of validated regions across rows and columns. The number of validated regions in the middle rows and columns tend to be higher compared to others.

b) Distribution of failed assays across the 96-well plate. The vertical axis represents the count of wells that did not show any significant reporter expression. The 4 corner wells clearly have the highest number of such inactivity.
2.2.3 Validated Region Selection

The following terminology will be used when discussing the experimental data:

- **Clone**: a single clone bacterial colony with a homogeneous insert sequence
- **Plasmid prep**: plasmid extraction from a single bacterial culture
- **Insert sequence**: the block of genomic region introduced into a plasmid
- **Insert plasmid**: the vector plasmid containing a sequence of interest
- **Clone replicates**: replicated experiments using the plasmids from the same clone but from different plasmid preps (i.e. independent DNA preparation)
- **Insert replicates**: experiments using plasmids recovered from different bacterial clones but sharing the same insert sequence
- **Technical replicates**: replicated experiments using plasmids from the same DNA preparation

All statistical analyses were done using the R software (R Development Core Team 2008). The ratios of firefly luciferase expression values over the renilla luciferase expression values were calculated to measure the relative increase of the firefly luciferase activity over the renilla luciferase activity (the internal control for transfection efficiency). Clones that did not produce both firefly and renilla luciferase expression values above the minimum threshold of 1000 luminescence relative light units (LRUs) were marked as failed transfections and removed from subsequent analyses. This heuristic filter was applied to exclude spurious expression ratio measurements, as the ratio of two small values can result in a disproportionately high value, and the VSN procedure intended to mitigate this effect was not sufficient (Huber et al. 2002). For those clones where only the firefly luciferase values were above this threshold, the renilla luciferase value was set to the threshold level. This step was designed to minimize the occurrence of large ratios even when the firefly luciferase expression values are near the threshold. The threshold of 1000 LRUs is higher than the median machine background level, which was found to be below 250 LRUs. While this conservative heuristic filter may result in a decrease in sensitivity, the trade-off was
deemed acceptable in order to avoid situations where spurious measurements are accepted as false positive results. The expression ratios from the two technical replicates for each clone were averaged, excepting the cases where a replicate transfection failed the minimum expression threshold filter (in such cases the single replicate value was used). The expression ratios obtained for each cell type were normalized using the VSN package. Each clone was treated as an independent sample even though there were in some cases insert replicates. The stochastic variation in the number of insert replicates would otherwise have complicated the analysis. Differential expression between 1) fibroblasts and myotubes, 2) myoblasts and myotubes, and 3) fibroblasts and myotubes groups were determined using the SAM package (Tusher et al. 2001), applying a false discovery rate maximum of 0.05. The two sets of clones selected from phase 1 and phase 2 at the FDR of 0.05 were combined and grouped according to the insert sequence. For each sequence, the number of clones that were identified as showing differential expression was counted, and those sequences with only one supporting clone and/or less than 50 % of the available clones identified as positive were excluded from the final set.

2.2.4 Data Analysis

2.2.4.1 MyoD ChIP-Seq Data

The ChIP-Seq peak locations for MyoD binding regions in the mouse genome were obtained from http://www.cs.washington.edu/homes/ruzzo/papers/DevCell/2010a/, the companion web resource to the reference publication (Cao et al. 2010).
2.2.4.2 Histone Modification ChIP-Seq Data

C2C12 cell ChIP-Seq peak locations for H3K4me1/2/3, H3K9me3, H3K9Ac, H3K18Ac, H3K27me3, and H4K12Ac annotated by Asp et al. were downloaded from the NCBI GEO database (GSE25308; (Asp et al. 2011)).

2.2.4.3 TFBS Profile Similarity Comparison

MatrixAligner was used to calculate the profile similarity of two TFBSs (Sandelin et al. 2003). This program generates scores from 0 to 2, where a score of 2 indicates complete identity between two matrices being compared.
2.3 Results

The experimental procedures and analyses presented in this paper consist of four main components: i) computational prediction of muscle-specific CRMs within the human genome; ii) validation of predictions using reporter gene assays and cell culture; iii) assessing performance of CRM prediction tools on the experimentally tested regions; and iv) analysis of the properties of newly validated muscle-specific CRMs relative to the properties of non-active sequences.

2.3.1 Region Selection

The overall region selection process is illustrated in Figure 2-2. Three sets of genomic sequences were identified for the study of skeletal muscle CRM predictions: (i) background regions randomly selected from conserved regions for control (background set); (ii) predicted skeletal muscle CRM regions proximal to skeletal muscle-expressed genes (muscle set); and (iii) predicted skeletal muscle CRM regions proximal to genes with no observed link to skeletal muscle (non-muscle set). Prediction of CRMs was performed for the muscle and non-muscle sets, while the background sequences were randomly selected from conserved intergenic regions, which may or may not contain predicted CRMs. The sets are further described below.
**Background:** A set of 200 regions was selected randomly from intergenic regions within the oPOSSUM conserved sequences (Methods) with high regulatory potential scores (Kolbe et al. 2004). The scores are intended to reflect consistency with the pattern of sequence identity in genome sequence alignments observed in known CRMs, and minimally reflect regions of greater sequence conservation. It is important to note that regions selected from conserved regions of the genome are likely to have distinct properties from regions randomly selected from the whole genome.

**Muscle:** Gene expression profiles associated with elevated expression concurrent with C2C12 myoblast-to-myotube differentiation were identified from the literature. Distinct sources of annotated skeletal muscle genes follow. Moran et al. performed gene expression profiling using Affymetrix oligonucleotide arrays, and identified 108 genes...
up-regulated in differentiated myotubes using one-way nested analysis of variance (Moran et al. 2002). Tomczak et al. profiled expression using Affymetrix GeneChips, from which they identified 447 genes up-regulated in myotubes through hierarchical cluster analysis with CAGED 1.1 software (Tomczak et al. 2004). In a complementary study, Blais et al. performed ChIP-chip analysis that identified 198 regions bound by MyoD, myogenin or Mef2 (Blais et al. 2005). Kislinger et al. examined global proteome changes by tracking the abundance of 1865 proteins through gel-free tandem mass spectrometry in both myoblasts and myotubes, of which 80 were identified as up-regulated in myotubes (Kislinger et al. 2005). The superset of the skeletal muscle genes arising from these studies was compiled. We previously annotated a list of 28 CRMs in 24 human genes for which at least one regulatory region responsible for skeletal muscle expression had been confirmed experimentally; we refer to this collection as the muscle reference set (a portion of this set was described in (Wasserman and Fickett 1998); listed in Supplemental Lists). Combining the superset with the muscle reference set yields 610 unique skeletal muscle-selective genes based on C2C12 experimental data.

Three CRM prediction programs were applied to the 610 sequences and 2,167 putative CRMs were recorded. A total of 518 candidate regions was predicted by more than one program, and high-quality primers for the same experimental PCR settings could be designed for 271 of them using Primer3. Further 220 primers could be designed for 400 randomly selected putative CRMs that were predicted by one program only. In the end, 384 candidate regions were selected for PCR amplification. For the muscle reference set, albeit highly circular due to the use of most of the sequences in parameter training for the published methods, we assessed the number of known CRMs detected by each program: Cluster-Buster detected 16, LRA detected 13, and MSCAN detected 10.

**Non-Muscle:** The set of genes represented in the oPOSSUM database excluding the 610 muscle genes were similarly scanned with the three prediction tools. These candidate CRMs were screened to remove any overlap with CRMs included in the muscle set or the background set.
2.3.2 Validation of CRM Activity in Cell Culture

**Library construction and properties:** The above sets of predicted CRMs and background regions were inserted into luciferase reporter gene plasmids and prepared as clone libraries. As the clone recovery process was stochastic, only a subset of the regions was recovered from each library and the number of insert replicates for each candidate CRM was variable. In the end, 355 unique insert sequences were present in 672 tested plasmids, of which 339/355 were successfully aligned to the intended PCR regions. The specific number of recovered candidate regions from each collection (88 *background*, 196 *muscle* or 55 *non-muscle*) is given in Table 2-2; their locations are listed in Supplemental Lists.

**Table 2-2. Number of candidate regions.**
The number of candidate regions that were actually cloned and tested is lower than the number of candidate regions computationally identified and allocated due to the random selection process.

<table>
<thead>
<tr>
<th>Region Set</th>
<th># Samples Allocated</th>
<th># Cloned and Tested</th>
<th># Viable Clones</th>
<th># Validated as Positive</th>
<th># Validated / # Viable Clones (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>192</td>
<td>88</td>
<td>55</td>
<td>4</td>
<td>7.3 %</td>
</tr>
<tr>
<td>Non-muscle</td>
<td>96</td>
<td>55</td>
<td>37</td>
<td>4</td>
<td>10.8 %</td>
</tr>
<tr>
<td>Muscle</td>
<td>384</td>
<td>198</td>
<td>186</td>
<td>11</td>
<td>5.9 %</td>
</tr>
</tbody>
</table>

**Validation of the assay:** Before proceeding with the experimental validation of the 339 sequences, we first assessed the performance of the dual luciferase reporter assay with known CRMs from muscle-expressed genes (desmin, TN-I) (Li *et al.* 1993; Gao *et al.* 1998; Mullen and Barton 2000) and non-muscle expressed genes (PAH) (Chen *et al.* 2002). The activity of the pGL3-promoter plasmid served as a negative control, while the PAH sequence was anticipated to function equivalently in differentiated and undifferentiated C2C12 cells. Two independent plasmid preparations were assessed with transfections performed in triplicate. The expression of the reporter gene driven
by the muscle CRMs was elevated 5-fold (DES) and 15-fold (TNI) in myotubes relative to myoblasts, while the non-muscle CRM (PAH) was unchanged (Figure 2-3).

**Figure 2-3. Reporter expression for known muscle-specific and non-muscle specific enhancer regions in myotubes vs. myoblasts.**

Only those constructs with known muscle-specific enhancers show marked increase in expression over the empty vectors. Expression ratios are calculated by dividing firefly luciferase reporter expression values by renilla luciferase reporter expression values. The error bars represent the standard error.

**Reporter Expression Analysis:** An overview of the clone production process is presented in Figure 2-4. The subset of plasmids that selectively directed myotube expression (2-fold increase or elevated based on SAM analysis) in phase 1 was advanced for further analysis in phase 2. In addition, single rows from each of the seven plates used in phase 1 were advanced, in order to assess the reproducibility of results. This selection process resulted in 204 plasmids being advanced. Independent plasmid preparations were used in the second round.
Figure 2-4. Selection of clones for differential expression analysis.
The selection is divided into 2 phases, where the clones selected for Phase 2 are a subset of all clones tested in Phase 1. Phase 1 and Phase 2 samples are from different plasmid preparations.

While individual predicted CRM inserts exhibit higher expression than the background controls, the mean expression ratios of the two sets are not significantly different based on a t-test (Figure 2-5). Reporter gene expression increases from myoblasts to myotubes are similar between the two groups, although the predicted CRM inserts exhibit higher variability. Both firefly and renilla raw reporter expression values were lower for the background controls. This characteristic was observed for both phase 1 and phase 2 (independent plasmid preparations).
Figure 2-5. Comparison of the reporter expression levels between background regions and non-background regions in the validated positive set.

a) Ratios of firefly and renilla luciferase reporter expression levels of the background regions vs. non-background regions in the validated positive set (myotubes). The difference between the two region sets is not significant.

b) Difference in expression ratios between myoblasts and myotubes. While non-background regions show higher spread in expression level differences than the background regions, the average values are not significantly different between the two.
c) Firefly and renilla luciferase reporter expression levels (log2 scale) of the background regions vs. non-background regions in the validated positive set. Non-background regions show significantly higher expression levels than background regions. The difference between the mean firefly activity and renilla activity is also higher for the non-background regions. (BG = Background; Non-BG = Non-Background)

Using the analysis criteria described in Methods, a set of 19 novel insert sequences was identified as driving selective myotube expression (relative to myoblasts and fibroblasts) (Table 2-3). These 19 CRMs are hereafter referred to as the validated positive regions. Of the 19 CRMs, 11 were derived from the muscle gene insert set, 4 from the non-muscle set and 4 from the conserved regions control group. Application of the CRM prediction tools to the 4 functional sequences from the control group resulted in 1 putative CRM being identified by ClusterBuster.
Table 2-3. List of genomic regions validated as driving muscle-specific expression.

Gene names were chosen for their proximity to the regions of interest (UCSC hg18).
‘Positive Wells’ column shows the number of replicates that were classified as positive out of all replicates for the sequence. Columns ‘C.Buster,’ ‘LRA,’ and ‘MSCAN’ indicate programs which predicted a CRM in the given region.

<table>
<thead>
<tr>
<th>Set</th>
<th>Coordinates</th>
<th>Gene Name</th>
<th>Positive Wells</th>
<th>Prediction Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chr</td>
<td>Prediction</td>
<td>Insert</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BG 11</td>
<td>12021584</td>
<td>116201218-116201584</td>
<td>APOA4; APOC3</td>
<td>4/5</td>
</tr>
<tr>
<td>BG 11</td>
<td>1721765</td>
<td>1721407-1721765</td>
<td>HCCA2</td>
<td>3/6</td>
</tr>
<tr>
<td>BG 11</td>
<td>6609164</td>
<td>66008803-66009164</td>
<td>DPP3</td>
<td>2/2</td>
</tr>
<tr>
<td>BG 11</td>
<td>71615709</td>
<td>71615350-71615709</td>
<td>INPPL1</td>
<td>2/2</td>
</tr>
<tr>
<td>NM 15</td>
<td>83189491</td>
<td>83185168-83189491</td>
<td>ALPK3</td>
<td>2/2</td>
</tr>
<tr>
<td>NM 18</td>
<td>40637579</td>
<td>40637176-40637555</td>
<td>SETBP1</td>
<td>3/4</td>
</tr>
<tr>
<td>NM 22</td>
<td>22517062</td>
<td>22516591-22517062</td>
<td>DERL3; SLC2A11</td>
<td>10/11</td>
</tr>
<tr>
<td>NM 22</td>
<td>22884017</td>
<td>22883584-22883974</td>
<td>CAIN1</td>
<td>2/4</td>
</tr>
<tr>
<td>M 1</td>
<td>12088282</td>
<td>119250477-119250882</td>
<td>TBX15</td>
<td>2/2</td>
</tr>
<tr>
<td>M 1</td>
<td>199612457</td>
<td>199612030-199612429</td>
<td>TNNT2</td>
<td>2/4</td>
</tr>
<tr>
<td>M 1</td>
<td>22087970</td>
<td>144878638-144878867</td>
<td>ZEB2</td>
<td>2/2</td>
</tr>
<tr>
<td>M 2</td>
<td>88148315</td>
<td>88147915-88148315</td>
<td>SMYD1</td>
<td>4/4</td>
</tr>
<tr>
<td>M 4</td>
<td>37728957</td>
<td>37728510-37728799</td>
<td>TBC1D1</td>
<td>2/4</td>
</tr>
<tr>
<td>M 6</td>
<td>42106867</td>
<td>42106432-42106831</td>
<td>CCND3</td>
<td>2/2</td>
</tr>
<tr>
<td>M 6</td>
<td>7127817</td>
<td>7127465-7127684</td>
<td>RREB1</td>
<td>2/4</td>
</tr>
<tr>
<td>M 9</td>
<td>35678364</td>
<td>35677888-35678364</td>
<td>TPM2</td>
<td>6/9</td>
</tr>
<tr>
<td>M 14</td>
<td>104259861</td>
<td>104259446-104259775</td>
<td>INF2; ADSS1</td>
<td>2/3</td>
</tr>
<tr>
<td>M 19</td>
<td>3326884</td>
<td>3326530-3326884</td>
<td>NFIC</td>
<td>4/4</td>
</tr>
<tr>
<td>M 19</td>
<td>54185225</td>
<td>54184834-54185225</td>
<td>GYS1</td>
<td>2/3</td>
</tr>
</tbody>
</table>
2.3.3 Properties of the Validated Positive Set

To identify defining characteristics of the positive regions compared to the non-responding regions, the validated set was subjected to analyses based on sequence and conservation properties.

**Overrepresented TFBS:** The oPOSSUM analysis method was applied to identify TF binding motifs over-represented in the 19 functional CRMs relative to the inactive inserts (Table 2-4). The top-scoring TFBS in the myotube-directing background regions (Table 2-4a) are those of RREB1, dorsal and NHLH1 (Pan and Courey 1992; Brown and Baer 1994; Thiagalingam et al. 1996). We could not find any direct link between muscle development and RREB1 in the literature. It is possible that the enrichment is just an artefact due to the RREB1 motif having high information content, which results in infrequent binding site predictions compared to most other motifs. If the foreground regions contain even just a few sites, it can result in high over-representation scores compared to the background set. Dorsal, a Rel TF, is involved in early stages of fly development; however, we could not find any direct role that Rel TFs play in vertebrate muscle development. As there could be other contributing motifs beyond the 5 muscle-linked motifs used the initial CRM prediction methods, further oPOSSUM analyses were carried out: 1) comparing the validated regions from the non-background sets as the test set against the remainder of the non-responding regions of the non-background sets as the control set (Table 2-4b), and 2) comparing the entire set of validated regions against all inactive inserts (Table 2-4c). Because all non-background regions are CRM predictions made with the 5 muscle motifs, these motifs were expected to be prevalent in both the positive and non-responding regions. Unexpected motifs present in the positive regions but absent in the non-responding regions could contribute to increased expression in myotubes. If this were the case, these additional motifs would be expected to be overrepresented in the above oPOSSUM analyses. Comparison of the non-background validated regions against the non-responding regions returned MEF2A and NHLH1 as being the most over-represented TFs. MEF2A is one of the five muscle TFs used to make the CRM predictions. As not all predictions necessarily contain
MEF2A hits, this result indicates the importance of this binding site being present for functional muscle-specific CRM relative to the other four muscle motifs. It suggests that MEF2 binding sites are potential master sites for a subset of active CRMs. NHLH1 is a bHLH TF, the same TF class as the myogenin family, sharing a similar binding profile with this group (normalized score of 1.73 using the MatrixAligner program, as explained in the Methods section). The over-representation of this motif may be an indirect marker for Myf, one of the five profiles contributing to the CRM predictions. Comparison of all validated regions against all non-responding regions again returns NHLH1, RREB1 and MEF2A as the most over-represented motifs (Figure 2-6).

Table 2-4. Overrepresented TFBS in the validated regions vs. non-responding regions ranked by Fisher p-values.

<table>
<thead>
<tr>
<th>TF</th>
<th>TF Class</th>
<th>Ctrl gene hits</th>
<th>Ctrl gene non-hits</th>
<th>Target gene hits</th>
<th>Target gene non-hits</th>
<th>Z-score</th>
<th>Fisher score</th>
</tr>
</thead>
<tbody>
<tr>
<td>RREB1</td>
<td>ZnF-C2H2</td>
<td>8</td>
<td>43</td>
<td>3</td>
<td>1</td>
<td>13.29</td>
<td>2.23E-02</td>
</tr>
<tr>
<td>DI</td>
<td>REL</td>
<td>26</td>
<td>25</td>
<td>4</td>
<td>0</td>
<td>5.27</td>
<td>8.04E-02</td>
</tr>
<tr>
<td>NHLH1</td>
<td>bHLH</td>
<td>16</td>
<td>35</td>
<td>3</td>
<td>1</td>
<td>7.03</td>
<td>1.14E-01</td>
</tr>
</tbody>
</table>

b) Non-background regions included in the validated set vs. the rest of the non-responding non-background regions.

<table>
<thead>
<tr>
<th>TF</th>
<th>TF Class</th>
<th>Ctrl gene hits</th>
<th>Ctrl gene non-hits</th>
<th>Target gene hits</th>
<th>Target gene non-hits</th>
<th>Z-score</th>
<th>Fisher score</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEF2A</td>
<td>MADS</td>
<td>116</td>
<td>92</td>
<td>13</td>
<td>2</td>
<td>4.48</td>
<td>1.55E-02</td>
</tr>
<tr>
<td>NHLH1</td>
<td>bHLH</td>
<td>120</td>
<td>88</td>
<td>12</td>
<td>3</td>
<td>7.98</td>
<td>7.36E-02</td>
</tr>
<tr>
<td>Fos</td>
<td>bZIP</td>
<td>181</td>
<td>27</td>
<td>15</td>
<td>0</td>
<td>4.10</td>
<td>1.35E-01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TF</th>
<th>TF Class</th>
<th>Ctrl gene hits</th>
<th>Ctrl gene non-hits</th>
<th>Target gene hits</th>
<th>Target gene non-hits</th>
<th>Z-score</th>
<th>Fisher score</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHLH1</td>
<td>bHLH</td>
<td>136</td>
<td>123</td>
<td>15</td>
<td>4</td>
<td>9.79</td>
<td>2.07E-02</td>
</tr>
<tr>
<td>RREB1</td>
<td>ZnF-C2H2</td>
<td>50</td>
<td>209</td>
<td>8</td>
<td>11</td>
<td>11.49</td>
<td>2.49E-02</td>
</tr>
<tr>
<td>MEF2A</td>
<td>MADS</td>
<td>125</td>
<td>134</td>
<td>13</td>
<td>6</td>
<td>3.38</td>
<td>7.16E-02</td>
</tr>
</tbody>
</table>
Sequence Composition: The next property examined was the sequence composition of the validated regions. Specifically, we analyzed both the single and dinucleotide composition characteristics of these regions to see if any significant biases could be found compared to the non-responding regions. The Wilcoxon rank sum test was performed to determine if any of these region sets showed significantly different composition characteristics. As shown in Table 2-5a, both the muscle validated regions have higher G/C mononucleotide frequency compared to the non-responding regions; the significance of these differences is supported by the rank sum tests for which most p-values were below 0.05.
Table 2-5. Sequence composition characteristics.
P-values were calculated using Wilcoxon rank sum tests.

a) GC content of the responding regions vs. non-responding regions.

<table>
<thead>
<tr>
<th></th>
<th>Responders</th>
<th>Non-Responders</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Muscle Validated</td>
<td>0.54</td>
<td>0.51</td>
<td>4.35E-02</td>
</tr>
<tr>
<td>Muscle Reference</td>
<td>0.58</td>
<td></td>
<td>2.87E-06</td>
</tr>
<tr>
<td>Pleiades Curated All</td>
<td>0.55</td>
<td>0.51</td>
<td>1.67E-02</td>
</tr>
<tr>
<td>Pleiades Curated Human</td>
<td>0.56</td>
<td></td>
<td>3.40E-03</td>
</tr>
</tbody>
</table>

b) MyoD ChIP-Seq peaks compared against the non-responding regions from this study.

<table>
<thead>
<tr>
<th></th>
<th>MyoD ChIP-Seq</th>
<th>Non-Responders</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Myotube</td>
<td>0.54</td>
<td>0.51</td>
<td>1.23E-06</td>
</tr>
<tr>
<td>Myoblast</td>
<td>0.54</td>
<td></td>
<td>6.60E-07</td>
</tr>
</tbody>
</table>

We also calculated the G/C and A/T skews in these sequences, but no significant differences in these two measures could be found between the responding regions and the non-responding regions (Table 2-6).

Table 2-6. GC and AT skews of the responding regions vs. non-responding regions.
The skew values were calculated by: GC Skew = (|G| - |C|) / (|G| + |C|), and AT Skew = (|A| - |T|) / (|A| + |T|).

<table>
<thead>
<tr>
<th></th>
<th>GC Skew</th>
<th>AT Skew</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Responders</td>
<td>Non-Responders</td>
<td>p-value</td>
<td>Responders</td>
<td>Non-Responders</td>
<td>p-value</td>
</tr>
<tr>
<td>Muscle Validated</td>
<td>0.122</td>
<td>0.099</td>
<td>0.13</td>
<td>0.091</td>
<td>0.101</td>
<td>0.64</td>
</tr>
<tr>
<td>Muscle Reference</td>
<td>0.109</td>
<td>0.18</td>
<td>0.18</td>
<td>0.116</td>
<td>0.084</td>
<td>0.49</td>
</tr>
<tr>
<td>Pleiades Curated All</td>
<td>0.113</td>
<td>0.26</td>
<td>0.26</td>
<td>0.094</td>
<td>0.084</td>
<td>0.82</td>
</tr>
<tr>
<td>Pleiades Curated Human</td>
<td>0.123</td>
<td>0.08</td>
<td>0.08</td>
<td>0.096</td>
<td>0.084</td>
<td>0.76</td>
</tr>
</tbody>
</table>
To further characterize the sequence compositional differences between the responding and the non-responding regions, we analyzed the dinucleotide compositions of the sequences (Figure 2-7a). Differences were found in the frequencies of AA, CC and GG, where the responding regions have higher frequencies of CC and GG dinucleotides, and the non-responding regions have higher frequencies of AA dinucleotides. This gives further support to the difference in the G/C compositional characteristics of the responding vs. non-responding regions. The CpG dinucleotide was not enriched in the validated regions and analysis presented below suggests that the enrichment properties are not related to the well-known properties of CpG islands (this point will be further explored below).
Figure 2-7. Dinucleotide frequencies in responding regions vs. non-responding regions.

a) Muscle Regulatory Regions. Muscle Validated = 19 validated muscle regions in this study. Muscle Reference = 28 muscle reference regions from literature. Muscle Non-Responders = all regions that were tested in this study and found not to drive gene expression.
b) Pleiades Curated Regulatory Regions. Pleiades Curated All = 1341 curated regulatory regions from all species. Pleiades Curated Human = 631 curated regulatory regions in humans only. Non-Responders = all regions that were tested and found not to drive gene expression.

To examine whether such biases are present in CRM regions for other tissue types, we performed composition analyses of the curated brain-specific CRM collection from the Pleiades Promoter Project (Portales-Casamar et al. 2010). The goal of this project is to construct human mini-promoters that drive gene expression in specific brain regions. As part of this project, the authors compiled a set of regulatory sequences from 296 genes shown to act as brain-specific CRMs in literature, which they deposited into the PAZAR database (Portales-Casamar et al. 2007). They also performed in vivo expression
studies of their mini-promoter constructs to identify the regulatory sequences that can drive gene expression specifically in brain regions. Because this data set is comprised of sequences from a number of different species, we performed our analysis on both the entire set and human sequence subset. For the control set to compare against, we use the non-coding human sequences that were tested in the project and found to have no effect on driving gene expression in the mouse brain. Similar to the muscle sequences, the Pleiades brain CRMs display elevated G/C content compared to non-functional sequences (Table 2-6). As for the dinucleotide frequencies, while not as pronounced in the muscle responders vs. non-responders, for the Pleiades regions, TA, TG and TT dinucleotides occur more frequently in the non-responders (Figure 2-7b).

**Evidence of Muscle Expression:** We attempted to assess if myotube expression data was predictive for functional CRM activity in the results (Table 2-7), however there were insufficient numbers from most datasets to determine significance.

**Table 2-7. The distribution of the regions in the muscle set according to the evidence source for muscle expression.**

For each muscle gene, there can be multiple candidate regions selected.

<table>
<thead>
<tr>
<th>Evidence Source</th>
<th># Genes</th>
<th># Included in Candidate Regions</th>
<th># Regions Included in the Positive Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blais</td>
<td>46</td>
<td>58</td>
<td>4</td>
</tr>
<tr>
<td>Emili</td>
<td>80</td>
<td>105</td>
<td>2</td>
</tr>
<tr>
<td>Moran</td>
<td>108</td>
<td>69</td>
<td>3</td>
</tr>
<tr>
<td>Tomczak</td>
<td>447</td>
<td>596</td>
<td>8</td>
</tr>
</tbody>
</table>

**Sequence Conservation:** In selecting candidate skeletal muscle CRMs, we did not incorporate phylogenetic footprinting (sequence conservation). This exclusion allows for a retrospective assessment of the impact of conservation-based filters on the specificity and sensitivity of the predictions. Table 2-8 gives the comparison of the sequence conservation characteristics of the validated regions versus the non-responding regions. While the average lengths of the PCR amplified inserts averaged ~400 bp, only a portion of each sequence may be conserved. Global measures may therefore fail to reflect the presence of a locally conserved putative CRM. To alleviate
this potential problem, we measured both the mean and the maximum sequence conservation scores for each region and then calculated the mean of these values for each region set. For the validated regions from the non-background sets, we observe both higher mean and maximum phastCons scores compared to the non-responding regions. This observation supports the validity of the widely used approach of applying conservation filters when making predictions for functional genomic regions. While some genome-wide ChIP-Seq studies for TFs have suggested that the conservation of the TF binding regions are limited, it is important to recognize that the TF binding by itself does not necessarily indicate cis-regulatory function. Figure 2-8 presents examples of positive regions from each of the background, non-muscle and muscle sets, representing the different conservation characteristics observed. The performances of the three methods with and without sequence conservation filter are summarized in Table 2-9.

Table 2-8. Sequence conservation based on phastCons scores (28-way Placental Mammals).

Column ‘Mean Score’ refers to the overall mean of the mean scores for each region in each set, while ‘Avg. Max Score’ refers to the mean of the highest score for each region in each set. ‘Conserved Region Make-Up’ lists the summed conserved region lengths (identified as sub-regions with phastCons score over 0.7) divided by the sum of the lengths of all regions in each set.

<table>
<thead>
<tr>
<th>Region Set</th>
<th>Mean Score</th>
<th>Avg. Max Score</th>
<th>Conserved Region Make-Up</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Positives</td>
<td>0.20</td>
<td>0.84</td>
<td>18.9 %</td>
</tr>
<tr>
<td>Positives from Background Set</td>
<td>0.12</td>
<td>0.74</td>
<td>7.9 %</td>
</tr>
<tr>
<td>Positives from Non-Background Sets</td>
<td>0.22</td>
<td>0.87</td>
<td>21.2 %</td>
</tr>
<tr>
<td>All Non-Responding Regions</td>
<td>0.17</td>
<td>0.77</td>
<td>15.1 %</td>
</tr>
<tr>
<td>Non-Responding Regions from Background Set</td>
<td>0.22</td>
<td>0.87</td>
<td>18.4 %</td>
</tr>
<tr>
<td>Non-Responding Regions from Non-Background Sets</td>
<td>0.16</td>
<td>0.75</td>
<td>14.4 %</td>
</tr>
</tbody>
</table>
**Figure 2-8. Examples of Positive Regions.**

Muscle TFBS hits (threshold of 80%) and the phastCons conservation profile for the region are shown as well. When square brackets are shown, they indicate the original CRM prediction.

a) Positive sequence from the muscle set. The muscle-specific TFBS are located in regions of high sequence conservation.

b) Positive sequence from the non-muscle. This sequence showed the most consistent increase in reporter expression, with all 12 replicates determined as significantly up-regulated in muscle.
c) Positive sequence from the background set. Despite the clear cluster of muscle-specific TFBS located in the region of high sequence conservation, none of the CRM prediction tools could classify this as a muscle CRM.

Table 2-9. Comparison of the five CRM prediction programs.

The five programs were tested on the 278 successfully cloned sequences and on the previously collected muscle reference regions. For the column ‘Programs’, CF refers to the application of additional conservation filter with the same parameters as used in Table 4.

<table>
<thead>
<tr>
<th>Programs</th>
<th>Validated Regions</th>
<th>Non-Responding Regions</th>
<th>Muscle Reference Regions (28 regions)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Background (4 regions)</td>
<td>Non-background (15 regions)</td>
<td>Background (51 regions)</td>
</tr>
<tr>
<td>Cluster-Buster</td>
<td>1</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>Cluster-Buster + CF</td>
<td>0</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>LRA</td>
<td>0</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>LRA + CF</td>
<td>0</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>MSCAN</td>
<td>0</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>MSCAN + CF</td>
<td>0</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>
We performed receiver operating characteristic (ROC) analysis for Cluster-Buster, LRA, and MSCAN using the ROCR package in R (Sing et al. 2005). The ROC analysis was performed both with and without a conservation filter applied, based on the maximum phastCons scores (Figure 2-9). For true positive regions, we included both the insert sequences that were validated through the reporter expression assays or literature-derived known muscle reference regions exclusive of the skeletal muscle training sets used in the development of LRA and ClusterBuster (see Supplemental Lists). For negative examples, all predicted and tested CRM regions (muscle and non-muscle sets) that did not respond to the reporter assays were used. An ROC curve based solely on the conservation filter was also generated. While adding the conservation filter improved the prediction performance for all methods tested, the conservation filter-only results exhibited the best performance, with the AUC of 0.76. However, it is noted that a large percentage of the non-responding regions come from the predictions by the three methods (211 of 295 regions, or 71.5%); as such, this high AUC of 0.76 is not independent of the contributions by the prediction programs. The findings confirm utility for incorporating sequence conservation into the prediction of cis-regulatory modules.
Figure 2-9. ROC Analysis of the conservation filter and the 4 CRM Prediction Programs.

The conservation filter analysis is based on phastCons (28-way Placental Mammals). The AUC value is given in the lower right corner of each graph.

a) Cluster-Buster
b) LRA

LRA ROC analysis

LRA with Conservation Filter ROC analysis
c) MSCAN

MSCAN with Conservation Filter ROC analysis

MSCAN ROC analysis

AUC = 0.57

AUC = 0.51
Promoters and CpG Islands: We examined the distances of the regions to the nearest Ensembl-annotated TSSs (Table 2-10). While there was much variability in the distances, with some regions located more than 100 kb away, the responding regions were in general located closer to the TSSs than the non-responding regions (median of ~1 kb vs. ~12.5 kb). In order to determine if the responding regions are associated with CpG islands, 1 kb upstream and downstream from each region were searched for UCSC-annotated CpG islands (Table 2-11). While a higher proportion of the responding regions were associated with CpG islands compared to non-responding regions, the difference did not have statistical significance (p-value of 0.13 obtained with Fisher test). Only 2 regions from the validated set were associated with CpG islands (10.5 %), while 10 regions from the reference set overlapped with CpG islands in their flanking sequences (35.7 %). This increase is likely due to the fact that the reference set regions are more proximal to the TSSs (median distance of 122.5 bp) than the validated set regions (median distance of 4,606 bp), and CpG islands are also typically in the vicinity of TSSs (Saxonov et al. 2006). As evident in Table 2-11 and Figure 2-7, the muscle
reference regions do display elevated CpG frequency and CpG island proximity, consistent with experimental bias in early promoter analysis for regions proximal to transcription start sites.

Table 2-10. Distances to the nearest annotated transcription start sites (Ensembl v61).
All units are in base pairs.

<table>
<thead>
<tr>
<th>Region</th>
<th>Min.</th>
<th>Median</th>
<th>Mean</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Muscle Validated</td>
<td>100</td>
<td>4,606</td>
<td>14,699</td>
<td>122,235</td>
</tr>
<tr>
<td>Muscle Reference</td>
<td>8</td>
<td>122.5</td>
<td>1,506</td>
<td>15,000</td>
</tr>
<tr>
<td>Combined Responders</td>
<td>8</td>
<td>928</td>
<td>6,839.4</td>
<td>122,235</td>
</tr>
<tr>
<td>Non-Responders</td>
<td>1</td>
<td>12,536</td>
<td>80,198.7</td>
<td>810,495</td>
</tr>
</tbody>
</table>

Table 2-11. Regions associated with CpG islands.
1 kb upstream and downstream from each region was searched. CpG island annotations are from the UCSC Genome Browser (hg19).

<table>
<thead>
<tr>
<th>Region</th>
<th># with CpG</th>
<th># without CpG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Muscle Validated</td>
<td>2 / 19 (10.5 %)</td>
<td>17 / 19 (89.5 %)</td>
</tr>
<tr>
<td>Muscle Reference</td>
<td>10 / 28 (35.7 %)</td>
<td>18 / 28 (64.3 %)</td>
</tr>
<tr>
<td>Combined Responders</td>
<td>12 / 47 (25.5 %)</td>
<td>35 / 47 (74.5 %)</td>
</tr>
<tr>
<td>Non-Responders</td>
<td>45 / 269 (16.7 %)</td>
<td>224 / 269 (83.3 %)</td>
</tr>
</tbody>
</table>

Phylogenetic Depth: Cheng et al. performed ChIP-chip analysis GATA1 binding regions in mouse erythroid cells and observed that most of the GATA1 binding regions contained the canonical binding site motifs (Cheng et al. 2008). However, they determined that the GATA1 binding motifs in regions associated with high enhancer activity were more evolutionarily conserved compared to those motifs in regions with no identifiable enhancer activities. To evaluate whether this observation holds for the muscle regulatory regions identified in this study, we searched the three region sets for binding site hits with all available vertebrate profiles from the JASPAR CORE collection. When overlapping sites for the same motif were found, only the highest scoring site was kept.
We first calculated the average phyloP scores for the predicted binding site and non-binding site positions in each of the three region sets. The scores for the responding regions are spread over a larger range than the non-responding regions, which also showed the lowest mean score (Figure 2-10). We identified the TFs with predicted binding sites that exhibited at least 2-fold increase in their phyloP scores over the non-binding site positions in each region set, and compared the mean phyloP scores in the predicted sites for these TFs among the three region sets (Table 2-12). The average phyloP scores were significantly higher for the validated and reference sets than for the non-responding set, as confirmed by t-tests. Included in the TFs with 2-fold increase in the validated and reference sets are some of the known muscle-specific TFs, such as MEF2A, Myf, SRF and PBX1 (Stankunas et al. 2008). We combined the list of the TFs with at least 2-fold increase in either the validated or the reference set, and compared the ratios of the phyloP scores for the predicted sites for these TFs and the non-binding site positions in each region set (Figure 2-11). The ratios are significantly higher for the responding regions (confirmed with t-test; p-value of 2.4x10^-4 for all species; Table 2-12B). When phyloP scores are calculated using more closely related species, the p-values become more significant. Non-responding regions do not show such a trend in scores (confirmed with t-test; p-value of 0.76 for all species and primates only).
Figure 2-10. Density plots for phyloP (46-way All) scores of the predicted binding site positions for the three region sets.

The scores for the responding regions (validated and reference) are spread out over a larger range.
Figure 2-11. Phylogenetic depth analysis of TFBSs in responding and non-responding regions using phyloP (46-ay, hg19).

The x-axis grouping indicates the species used to calculate the phyloP scores. TFBSs were searched using all vertebrate profiles from the JASPAR CORE collection using the threshold of 0.8. TFs with at least 2-fold increase in phyloP scores (46wayAll) in the TFBS positions over the non-TFBS positions in the responding regions were identified. The score ratios for these TFs were compared among the three region sets.

a) Average phyloP scores for the predicted TFBS positions and non-TFBS positions in each region set.

b) Ratios of phyloP scores for the TFBS positions and non-TFBS positions in each region set.
Table 2-12. Comparison of the mean phyloP scores in the three region sets for profiles with at least 2-fold increase in phyloP scores for predicted TFBS positions vs. non-TFBS positions.

a) Mean phyloP scores for TFBS positions in each region set.

<table>
<thead>
<tr>
<th>PhyloP species</th>
<th>Validated</th>
<th>Reference</th>
<th>Non-Responders</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>1.15</td>
<td>1.77</td>
<td>0.49</td>
</tr>
<tr>
<td>Placental</td>
<td>0.78</td>
<td>1.51</td>
<td>0.33</td>
</tr>
<tr>
<td>Primates</td>
<td>0.32</td>
<td>0.45</td>
<td>0.21</td>
</tr>
</tbody>
</table>

b) P-values from t-test results between each region set (alternative hypothesis: true difference in means is not equal to 0)

<table>
<thead>
<tr>
<th>PhyloP Species</th>
<th>Validated vs. Non-Responders</th>
<th>Reference vs. Non-Responders</th>
<th>Validated vs. Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>2.39E-04</td>
<td>8.14E-09</td>
<td>2.06E-03</td>
</tr>
<tr>
<td>Placental</td>
<td>2.12E-05</td>
<td>7.05E-09</td>
<td>1.28E-07</td>
</tr>
<tr>
<td>Primates</td>
<td>2.86E-07</td>
<td>9.58E-08</td>
<td>8.43E-09</td>
</tr>
</tbody>
</table>

c) TFBS profiles with greater than 2-fold increase in average phyloP scores between predicted TFBS and non-TFBS positions

<table>
<thead>
<tr>
<th>Region Set</th>
<th>TF Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validated</td>
<td>Sox5, HOXA5, Prrx2, RELA, Myf, FOXF2, NR2F1, MEF2A, NFYA, Arnt::Ahr, ZEB1, Foxa2, Nkx3-2, CREB1, NHLH1, PBX1, TLX1::NFIC, Myc, GABPA, Ar</td>
</tr>
<tr>
<td>Reference</td>
<td>Myf, TEAD1, MEF2A, ZEB1, SRF, NFIL3, PBX1, Pdx1, ARID3A, Sox5, CREB1, Pax5, FOXI1, IRF1, Sox2</td>
</tr>
<tr>
<td>Common (Intersection)</td>
<td>Sox5, Myf, MEF2A, ZEB1, CREB1, PBX1</td>
</tr>
</tbody>
</table>

**MyoD ChIP-Seq Evidence:** Cao *et al.* performed a genome-wide ChIP-Seq binding assay for MyoD in C2C12 myoblasts and myotubes (Cao *et al.* 2010). We compared the MyoD peak locations with our three region sets to determine the extent of overlap with each. As we were testing the regulatory effects of human genomic sequences in murine C2C12 cells, we first performed a lift-over of the regions to the mouse genome using the Galaxy service, transitioning from human assembly hg19 to mouse assembly mm9 (Goecks *et al.* 2010). Successfully mapped regions were then compared with the MyoD peak locations (Table 2-13). Whereas only 15.6% of the successfully mapped non-responding regions overlapped with the MyoD peaks in myotubes, 58.1% of the responding regions overlapped with the peaks, which is a significant increase (p-value = 3.0x10^-8 with Fisher exact test). These observations support the use of ChIP-Seq assay results for achieving improved specificity in the identification of functional regulatory elements.
Table 2-13. Regions overlapping MyoD ChIP-Seq peaks in C2C12 cells.

P-values were calculated using Fisher exact tests.

<table>
<thead>
<tr>
<th></th>
<th>Responding</th>
<th>Myotubes</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Responding</td>
<td>27.9 % (12 / 43)</td>
<td>58.1 % (25 / 43)</td>
<td>4.3E-03</td>
</tr>
<tr>
<td>Non-responding</td>
<td>11.6 % (23 / 199)</td>
<td>15.6 % (31 / 199)</td>
<td>1.5E-01</td>
</tr>
<tr>
<td>P-value</td>
<td>8.4E-03</td>
<td>3.0E-08</td>
<td></td>
</tr>
</tbody>
</table>

**Histone Modifications**: Chromatin conformation changes through histone modifications play an important role in the regulation of gene expression. Acetylation of histone tail residues lead to open chromatin, allowing TFs better access to enhancer regions. Histone methylation has been associated with both activation and repression depending on which residues are modified. In order to examine the epigenetic changes associated with myotube formation, Fischer *et al.* performed a ChIP-chip study of major histone modifications (H4ac, H3ac, H3K4me2/3) in C2C12 cells, from which they observed that H3K4me2 (when combined with other acetylation in the same region) and H4ac were frequently associated with elevated expression (Fischer *et al.* 2008). Asp *et al.* performed a more comprehensive ChIP-Seq study in C2C12 cells, where they identified the locations of H3K4me1/2/3, H3K9me3, H3K9Ac, H3K18Ac, H3K27me3, H4K12Ac, and PolII in both myoblasts and myotubes (Asp *et al.* 2011). Using the subsets of the validated, reference and non-responders that were successfully mapped to the mouse genome, we examined the extent of overlap between these regions and the modified histone peaks for H3K4me1/2/3, H3K9Ac, H3K18Ac, H3K27me3 and H4K12Ac (Figure 2-12). We observed increases in the proportion of the combined set of responding regions that overlapped with H3K4me2 (p-value = 3.5x10^-6 with Fisher exact test), H3Kme3 (p-value = 2.7x10^-5), and H3K18ac (p-value = 3.4x10^-3) peaks in myotubes compared to myoblasts, whereas the non-responding regions did not show as large increases (p-values for H3K4me2: 6.5x10^-3, H3Kme3: 1.5x10^-1, and H3K18ac: 3.7x10^-1). As the responding regions act as enhancers in myotubes, such observations point to these histone marks as being activating. While not statistically significant, increases in H3K12Ac peak overlaps were observed in responding regions (p-value = 7.8x10^-2), whereas decreases were observed in non-responding regions (p-value =
1.7x10^{-3}). Differences between the properties of the validated and reference regions were observed in some cases (e.g. H4K12Ac), which may reflect the previously mentioned promoter proximity of the later set. There is some decrease in the number of responding regions with H3K27me3 marks from myoblasts to myotubes (p-value = 5.8x10^{-2}), whereas there is little change in the non-responding regions (p-value = 7.2x10^{-1}). H3K27me3 histone marks are known to play important roles in repression of muscle-specific genes in proliferating cells. If we look at the reference set and the validated set separately, we observe that most only the reference regions show a decrease in the H3K27me3 peaks with borderline statistical significance (p-value = 5.1x10^{-2}).
Figure 2-12. Histone modifications in the responding and non-responding regions. Proportion of the regions that overlap with ChIP-Seq peaks from Asp et al. are displayed. (MB = Myoblasts, MT = Myotubes)
2.4 Discussion

We generated genome-wide predictions of muscle-specific CRMs using three CRM prediction programs, including Cluster-Buster, LRA and MSCAN. Based on the predictions, 339 candidate sequences were tested for CRM activity using promoter-reporter gene assays in a cell culture model of skeletal muscle development, of which 278 were successfully transfected into cells and had reporter expression measurements taken. The validation process revealed 19 myotube-restricted promoter-enhancing sequences. In addition to the known enrichment for sequence conservation of functional CRMs, phylogenetic depth analysis revealed that the individual TFBSs display even higher sequence conservation than the surrounding sequence. The active CRMs exhibited elevated G/C mononucleotide content indicating the value for including sequence composition measures in the implementation of future methods. Comparison of the ChIP-Seq results for MyoD and histone modification marks in C2C12 cells with the identified enhancing sequences further supports their recognized utility in the detection of active, functional CRMs.

The performance of the CRM prediction programs used in this paper was not sufficient for genome annotation. The poor performance is likely reflective of the incomplete information presented for the prediction – the primary sequence and sequence conservation data does not convey information about the three dimensional properties of the nucleus nor the epigenetic state of the chromatin (Francastel et al. 2000; Gregory 2001; Georgel 2002; Fischle et al. 2003). As evidenced by the significant increase in the proportion of responding regions that overlap with MyoD and histone modification peaks from ChIP-Seq studies, incorporating the results from ChIP-Seq assays for the relevant TFs, co-activating proteins or histone modification marks can improve the specificity of the predictions. In order for such data to be useful, data needs to be generated for each tissue type analyzed, as CRMs are anticipated to be differentially marked when activated. At this time, there is an insufficient amount of such large scale data available to make this a feasible strategy for many tissue types, but more complete
data may become available as the costs of experiments come down and sensitivity increases. Ultimately an intersection of computational and experimental methods will be required for the highest quality annotation of CRMs.

A fundamental question arising out of the work reported here is why methods that appeared to be doing well for skeletal muscle CRM discovery failed to demonstrate strong predictive capacity in application here. One key reason may be driven by selection bias for laboratory studies. The reports of CRMs from individual gene studies may in many cases have been influenced by the identification of muscle-related motifs in the available genomic sequences. Due to the selective publication of those sequences showing positive expression, the relative importance of motif enrichment may have been over emphasized. Another key limitation is that most of the methods generate sufficiently high false prediction rates that the reliability of any specific set of predictions is unlikely to be high. The results here demonstrate the driving need for experimental validation of computational predictions whenever feasible.

One striking observation emerging from this study is the enrichment of G/C mononucleotides in the CRMs, observed both in the new muscle set as well as the brain-directing CRMs from the Pleiades Project (Portales-Casamar et al. 2010). The potential contribution of compositional properties to regulatory regions has been previously explored, including a statistical method for CRM prediction (Bina et al. 2009) and a recent approach from Evans to classify CRM-containing regions into compositional subsets of genome sequences (Evans 2010). These approaches and the data presented here are independent of the long-recognized role of CpG islands in demarcating promoter-containing regions and the influence of CpG enrichment on motif over-representation (Hannenhalli and Levy 2001; Kielbasa et al. 2010). While there have been prediction methods released, such as Stubb, EMMA and PhylCRM, that directly incorporate phylogenetic footprinting in order to reduce the false positive rate of their predictions (Sinha et al. 2003; Warner et al. 2008; He et al. 2009), the joint incorporation of nucleotide composition properties and sequence conservation remains to be explored.
The outcomes of this paper include both a novel set of 19 skeletal muscle-directing CRMs for use in future machine learning procedures and the specific call for the inclusion of nucleotide composition properties in the next generation of tools.
3 oPOSSUM-3: Advanced Analysis of Regulatory Motif Over-representation across Genes or ChIP-Seq Datasets

3.1 Introduction

The properties of cells within an organism are defined by a complex interplay between proteins, RNA, and the genome, which can be conceptualized as the gene regulatory network. Two important components of the gene regulatory network are the DNA-binding trans-acting transcription factors (TFs) and their corresponding transcription factor binding sites (TFBS) in the DNA. Sets of proximal TFBSs that are sufficient to cooperatively mediate TF regulated patterns of expression constitute cis-regulatory modules (CRM). CRMs are the scaffold for combinatorial TF interactions, enabling a limited number of sequence-specific DNA binding TFs to participate in an exponential number of combinations, each potentially conferring specific patterns of gene activity (Arnone and Davidson 1997).

In studying gene regulation within a cell or tissue, researchers are commonly confronted with the need to analyze sets of genes sharing a common characteristic, such as co-expression, as they seek to infer properties of the gene regulatory network. A significant insight into the regulatory network structure is obtained when the mediating TFs for the observed expression patterns are identified. A key strategy in genome biology for determining such TFs is to determine the sequence motifs that are over-represented in the cis-regulatory regions relative to some control. The successful predecessors to oPOSSUM-3, oPOSSUM (Ho Sui et al. 2005) and oPOSSUM-2 (Ho Sui et al. 2007), were developed to identify statistically over-represented, predicted TFBS in co-regulated gene sets. Two complementary scoring methods measured the over-representation: (i) Z-scores based on normal approximation to the binominal distribution that measures the change in the relative number of TFBS motifs in the foreground target gene set compared to the background set, and (ii) Fisher scores based on a one-tailed Fisher exact probability assessing the number of genes with the
TFBS motifs in the foreground set versus the background set. Using the JASPAR
database as the source of DNA binding profiles, the original oPOSSUM was designed to
identify over-represented TFBSs, later referred to as Single Site Analysis (SSA). The
original system also incorporated a conservation filter using phylogenetic footprinting
based on pair-wise alignments of orthologous sequences from human and mouse. In
oPOSSUM-2, an additional analysis method called the Combination Site Analysis (CSA)
was introduced, to identify over-represented proximal pairs of TFBSs. Separate
oPOSSUM-2 implementations were released for two additional model organisms (C.
*elegans* and *S. cerevisiae*). The nematode oPOSSUM-2 database was based on alignments
between *C. elegans* and *C. briggsae*. The oPOSSUM-2 yeast system did not incorporate
conservation filters as the compact nature of the yeast genome results in dramatically
reduced search space and noise compared to larger genomes. oPOSSUM-2 was found to
perform well in an independent assessment (*Meng et al. 2010*). The oPOSSUM software
is one of the most highly cited dedicated tools for TFBS motif over-representation
analysis (as assessed by Google Scholar citation counts), perhaps due to the ease of use
and power of the approach. On average, excluding automated internet search software,
340 unique users work with oPOSSUM-2 each month.

Since the release of the oPOSSUM system, a plethora of TFBS over-representation
analysis tools have been introduced. TOUCAN2, a workbench system for regulatory
sequence analysis implemented by Aerts *et al.*, contains features for identifying over-
represented TFBS in proximal promoters of co-regulated genes (*Aerts et al. 2005*). D
Defrance and Touzet developed the TFM-Explorer (*Defrance and Touzet 2006*), which
assesses conservation of spatial arrangements of regulatory elements. Promoter
Analysis Pipeline by Chang *et al.* includes TFBS identification in gene sets as a
component of the workbench, using non-redundant profiles from public databases
(*Chang et al. 2007*). Piechota *et al.* developed the cREMaG database, which attempts to
correct for the confounding influence of variable information content of TFBS profiles,
distinguishes between constitutive and inducible transcriptional forms of genes, and
reports the presence of CpG islands (*Piechota et al. 2010*). Many of the methods provide
web-based user interfaces, some of which are maintained.
Since the implementation of these approaches, dramatic changes have emerged for regulatory sequence studies. First, comprehensive multi-species sequence comparison measures are widely available in the form of phastCons and phyloP scores from the UCSC genome databases (Kent et al. 2002; Hubisz et al. 2011). Phylogenetic footprinting, used by many TFBS enrichment programs, when performed with only pair-wise sequence alignments places an undue weight upon organism selection. Multi-species conservation measures are expected to be more robust (Kumar and Filipski 2007). While the proliferation of large-scale regulatory sequence profiling methods such as ChIP-Seq has sparked an ongoing debate on whether conservation filtering places an undue constraint on detection of functional TFBSs, it is important to note that genomic regions that bind with TFs are not necessarily functional cis-regulatory sequences, and conservation still serves as a useful filter for making computational analysis feasible. Second, there has been a major update to the JASPAR database, an open-source, non-redundant, curated repository of TFBS profiles (Portales-Casamar et al. 2010). The new version of JASPAR contains a significant increase in non-vertebrate profiles, permitting the extension of regulatory analysis software to many non-vertebrate species such as insects. Third, large-scale sequence profiling methods have resulted in an explosion of the number of potential regulatory sequences to be analyzed (Johnson et al. 2007; Schmidt et al. 2010; Malhotra et al. 2010). These experiments produce sets of TF bound and control sequences, in which the foreground target (TF-bound) sets purportedly contain regulatory signatures of interest whereas the background (control) sets lack those features. Such data is optimal for TFBS over-representation analysis, creating strong demand for a new generation of software that allows analysis from both a sequence-based perspective and a gene-based perspective.

Here we describe oPOSSUM-3, a system that capitalizes upon the aforementioned research developments. The new system features a panel of upgraded and novel approaches to regulatory sequence analysis, including Single-Site Analysis (SSA) and anchored Combination-Site Analysis (aCSA) (Figure 3-1). A novel extension of the system addresses the challenge imposed by homologous TFs with highly similar (or identical) binding specificity. The TFBS Cluster Analysis (TCA) and anchored
Combination TFBS Cluster Analysis (aCTCA) present results focused on TFBS sequence patterns rather than individual profile names. This new approach to regulatory sequence motif over-representation analysis has been assessed against reference sets of co-regulated genes and large-scale ChIP-Seq sequence collections. Assessments against reference cases confirm the high performance of oPOSSUM-3 for the identification of mediating TFs. The new system should maintain the oPOSSUM service as a popular resource for motif over-representation analysis.
Figure 3-1. Overview of the main analysis types available in oPOSSUM-3.

The input for oPOSSUM can be either gene-based, which makes use of pre-computed results based on annotated genomic information, or sequenced-based, where the user supplies the input sequences (e.g. ChIP-seq results) for analysis. Two main types of analysis are available: individual site analysis and anchored combination analysis methods. For each method, users select either single site or TFBS cluster-based analysis. In summary, the methods include: 1) Single Site Analysis (SSA), 2) TFBS Cluster Analysis (TCA), 3) anchored Combination Site Analysis (aCSA) and 4) anchored TFBS Cluster Analysis (aCTCA). SSA and aCSA are represented in this figure.
3.2 Methods

3.2.1 Nomenclature

Throughout the manuscript we refer to TFs and genes using a capitalized first letter followed by lower case letters; all suffix characters are capitalized (e.g. Mef2A).

3.2.2 Data Sources

The gene and transcript annotations and genomic sequences were retrieved from Ensembl v57 (Flicek et al. 2010). For *C. elegans*, operon annotations were retrieved from Wormbase WS200 (Harris et al. 2010). The phastCons scores were retrieved from the UCSC Genome Browser, based on the following score sets: 1) for human (hg19), phastCons46wayPlacental, 2) for mouse (mm9), phastCons30wayPlacental, 3) for fruit fly (dm3), phastCons15way, and 4) for nematode (ce6), phastCons6way. As UCSC ce6 database for *C. elegans* is based on WS190, any genomic regions that had insertions or deletions between WS190 and WS200 were excluded. All known genes in the Ensembl databases were included in the database. For TFBS profiles, the 2010 release of the JASPAR database was used. All profiles from the CORE and PBM collections were included. A custom profile collection (referred to as PENDING) was implemented to include profiles of interest for our analysis, but not as yet included in JASPAR (Figure 3-2).
Figure 3-2. oPOSSUM-specific JASPAR PENDING collection.

PFMs for daf-19, Rfx1_1, Rfx1_2 and Nfe2l2 are included.

A. daf-19

\[
\begin{align*}
A & : [1 \ 0 \ 3 \ 6 \ 0 \ 0 \ 21 \ 0 \ 6 \ 1 \ 1 \ 20 \ 22 \ 0] \\
C & : [0 \ 1 \ 4 \ 3 \ 21 \ 17 \ 0 \ 0 \ 0 \ 1 \ 15 \ 0 \ 0 \ 21] \\
G & : [21 \ 0 \ 0 \ 3 \ 0 \ 0 \ 0 \ 0 \ 16 \ 20 \ 0 \ 1 \ 0 \ 0] \\
T & : [0 \ 21 \ 15 \ 10 \ 1 \ 5 \ 1 \ 22 \ 0 \ 0 \ 6 \ 1 \ 0 \ 1]
\end{align*}
\]

B. Rfx1_1

\[
\begin{align*}
A & : [5 \ 10 \ 6 \ 3 \ 11 \ 13 \ 3 \ 4 \ 4 \ 11 \ 0 \ 3 \ 30 \ 32 \ 0 \ 11 \ 11] \\
C & : [5 \ 4 \ 4 \ 2 \ 10 \ 1 \ 25 \ 14 \ 2 \ 3 \ 0 \ 16 \ 0 \ 0 \ 32 \ 9 \ 3] \\
G & : [6 \ 8 \ 22 \ 7 \ 2 \ 17 \ 1 \ 7 \ 0 \ 17 \ 32 \ 3 \ 1 \ 0 \ 0 \ 7 \ 5] \\
T & : [8 \ 6 \ 0 \ 20 \ 9 \ 1 \ 3 \ 7 \ 6 \ 1 \ 0 \ 10 \ 1 \ 0 \ 0 \ 4 \ 10]
\end{align*}
\]

C. Rfx1_2

\[
\begin{align*}
A & : [5 \ 10 \ 6 \ 3 \ 11 \ 13 \ 3 \ 4 \ 9 \ 4 \ 11 \ 0 \ 3 \ 30 \ 32 \ 0 \ 11 \ 11] \\
C & : [5 \ 4 \ 4 \ 2 \ 10 \ 1 \ 25 \ 14 \ 3 \ 2 \ 3 \ 0 \ 16 \ 0 \ 0 \ 32 \ 9 \ 3] \\
G & : [6 \ 8 \ 22 \ 7 \ 2 \ 17 \ 1 \ 7 \ 2 \ 0 \ 17 \ 32 \ 3 \ 1 \ 0 \ 0 \ 7 \ 5] \\
T & : [8 \ 6 \ 0 \ 20 \ 9 \ 1 \ 3 \ 7 \ 2 \ 10 \ 1 \ 0 \ 10 \ 1 \ 0 \ 0 \ 4 \ 10]
\end{align*}
\]

D. Nfe2l2

\[
\begin{align*}
A & : [288 \ 0 \ 0 \ 430 \ 14 \ 58 \ 31 \ 345 \ 3 \ 3 \ 361] \\
C & : [24 \ 2 \ 0 \ 0 \ 304 \ 13 \ 325 \ 2 \ 3 \ 418 \ 15] \\
G & : [112 \ 0 \ 419 \ 0 \ 75 \ 12 \ 31 \ 49 \ 424 \ 6 \ 17] \\
T & : [6 \ 428 \ 11 \ 0 \ 37 \ 347 \ 43 \ 34 \ 0 \ 3 \ 37]
\end{align*}
\]

3.2.3 Calculation of TFBS Motif Over-representation

Two statistical measures are used to determine the TFBS motifs that are over-represented in the foreground set versus the background set, representing two different models for the occurrences of TFBSs.
3.2.3.1 Z-scores

The Z-score is used to determine whether the rate of occurrence of a given TFBS in the foreground sequence set differs significantly from the expected rate calculated from the background set based on a simple binomial distribution model. For a given TFBS, let $X$ denote the number of predicted binding site nucleotides in the foreground sequence set, and $B$ the number of predicted binding site nucleotides in the background set. If $n$ is the total number of nucleotides in the foreground set, and $N$ the total number for the background set, the expected value of $X$ is given by $\mu = BC$, where $C = n/N$. The probability of success is given by $P = B/N$, and the standard deviation is $\sigma = \sqrt{nP(1-P)}$. Then, if $x$ is the observed number of binding site nucleotides in the foreground set, using the normal approximation to the binomial distribution with a continuity correction of 0.5, the Z-score can be calculated as $Z=(x-\mu-0.5)/\sigma$.

3.2.3.2 Fisher scores

The Fisher probability test is used to determine the probability of a non-random association between the foreground sequence set and a given TFBS by comparing the proportion of foreground sequences containing a given TFBS with the proportion of background sequences with that site. The Fisher probability is calculated using a hypergeometric probability distribution, which describes sampling without replacement from a finite population consisting of two types of elements. The Fisher scores are obtained by taking the negative logarithm of the probabilities (natural logarithm is used). In contrast to the Z-score, only the presence or absence of a TFBS in a given sequence is considered; the number of occurrences of a TFBS is not included in the probability calculation. The Fisher exact probability calculation was implemented in the R statistics package (R Development Core Team 2008).
3.2.3.3 False discovery rate calculation

False discovery rate (FDR) algorithm by Benjamini and Hochberg was used to calculate the adjusted p-values for both the Z-scores and Fisher scores (Benjamini and Hochberg 2007). Z-scores were first converted to p-values using the standard normal table, whereas for Fisher scores, the p-values obtained during Fisher exact probability calculation were used. All calculations were performed in the R statistics package (p.adjust for FDR and pnorm for the standard normal table).

3.2.4 Single Site Analysis (SSA)

SSA is the original over-representation analysis procedure developed for oPOSSUM. Individual TFBS hits within the user-specified search regions are counted for both the foreground gene/sequence set and the background set. For gene-based analysis, the search regions are determined by selecting the conserved regions located within a given distance from the gene transcription start sites. Both the Z-scores and Fisher scores are calculated for each TFBS based on the counts, which are then ranked accordingly. For a detailed explanation of the oPOSSUM SSA, please refer to (Ho Sui et al. 2005). The major difference in the SSA implementation between the previous versions of oPOSSUM and version 3 is the replacement of the pairwise sequence alignment-based phylogenetic footprinting procedure with phastCons scores. Conserved regions are defined as those genomic regions with phastCons scores over given thresholds (three sets of conserved regions are determined using thresholds of 0.6, 0.65 and 0.7) and TFBS searches are restricted to motifs within or overlapping (by at least 1 bp) these regions (Figure 3-3).
**Figure 3-3. Defining conserved regions.**

During the oPOSSUM DB build, the phastCons sequence conservation scores from UCSC are retrieved for the pre-defined search region near the transcription start site of each gene. Sub-regions with phastCons scores above the pre-defined threshold are marked as conserved regions, and the TFBS searches are restricted to these sub-regions only.

![Diagram of phastCons scores](image)

**3.2.5 Anchored Combination Site Analysis (aCSA)**

In many cases, given a gene set of interest, the user has beforehand knowledge of a TF that is central to the observed co-regulation. In order to focus the intensive computational analysis of motif combinations, the aCSA method restricts the over-representation assessment to sequence regions proximal to predicted TFBSs for a user-specified TF. Given an anchoring TF and an inter-binding site distance parameter value, pairings of the anchoring TFBS and all available secondary TFBS within the inter-binding site distance are counted and over-representation scores calculated. It is possible to observe self-interactions, when TFBS for the anchor TF frequently occur in clusters.
3.2.6 Gene-based vs. Sequence-based

Increasingly researchers focus on sets of sequences likely to be bound by a TF, rather than a list of genes. In order to facilitate the analysis of sequence-based data, such as those from ChIP-Seq studies, new functions within oPOSSUM-3 were implemented. The sequence-based oPOSSUM systems take as input a foreground sequence set and a background sequence set. The supplied sequences are searched for TFBS motifs, and over-representation scores calculated. There is no conservation filtering applied to the input sequences; the entire set of submitted sequences is screened for TFBS motifs.

3.2.7 TFBS Clustering

Transcription factors can be classified according to structural characteristics of DNA binding domains. It is often the case that TFs belonging to the same structural classes share similar DNA binding profiles, such that distinct TFs of the same class may bind to similar sequence patterns. If a given class contains numerous TFs with almost identical consensus sequences, oPOSSUM-3 analysis results can be dominated by clusters of profiles that are nearly identical to one another. In such cases, it is useful to condense the redundant results to allow the user to identify independent enriched profiles. It is not suitable to combine the results simply based on the TF class, as the extent of the binding sequence similarity is variable among the different classes. While some classes are defined by a characteristic consensus sequence, other classes such as zinc fingers have low profile similarities among the member TFs. Thus, it is necessary to divide each structural class into clusters based on profile similarity. Figure 3-4 illustrates the idea behind TFBS clustering and its application to oPOSSUM-3.
Figure 3-4. TF structural families and TFBS clusters.

The two analysis methods, Single Site Analysis (SSA) and anchored Combination Site Analysis (aCSA), introduce a variation based on TFBS clusters – groupings of TFs exhibiting highly similar binding site specificity. As many TFs cannot be distinguished in their binding specificity, this process promotes the consideration of all members of a functionally equivalent group – allowing users to focus on TFs likely to be active in a cell or condition relevant to their research. In the figure, the outer shapes and colour represent the structural class, while the inner shapes symbolize the binding specificity of a cluster. The different shading within the shapes denotes individual TFs.

Profiles in JASPAR 2010 have been annotated for TF structural class and family. Based on these classifications, the profiles in the given family are subject to a refined clustering process in oPOSSUM-3. First, using the MatrixAligner similarity scoring program (Sandelin et al. 2003), a pairwise similarity score table is calculated for the entire set of profiles in JASPAR. Two thresholds are set: 1) cluster score threshold $T$, which is the MatrixAligner score above which the two matrices being compared are deemed to be similar, and 2) radius margin $R$, which is a secondary score threshold used to determine whether those TFs at the boundary of the cluster join the cluster or not (Figure 3-5A). The process is based on tree traversal, with nodes being the profiles and the edges being the similarity scores. One profile within a given family is randomly chosen to act as the seed node for the cluster, and a tree is constructed between this
seed profile and all other profiles in the family. The nodes are traversed in sequence, and the traversed nodes are added to the cluster if: 1) the similarity scores between the parent node and the child node are lower than the cluster score threshold T, and 2) the average score S between the cluster member nodes (the parent nodes that have already been included in the cluster) and the child node is below T. If the child node in question qualifies for condition 1 but not condition 2, it can still be included in the cluster if S exceeds T by less than R. A pseudocode of the clustering process is given in Figure 3-5C. From the JASPAR 2010 database, 250 profiles from the CORE collection and 184 from the PBM collection were analyzed, along with the four profiles from the custom PENDING collection. A cluster score threshold of 1.8 and a radius margin of 0.1 were used. These values were selected empirically, based on the distribution of pairwise similarity scores among all available JASPAR profiles.

When a TFBS cluster-based analysis is performed, any overlapping TFBS hits that belong to the same TFBS cluster are merged together to form a single cluster hit. Only the merged cluster hits are counted for TFBS over-representation calculations (Figure 3-5B).
Figure 3-5. TFBS cluster analysis.

(A) TFBS clustering process. Binding site profiles are compared, and those exhibiting similarity are placed in a common group using a two-step process. (B) Individual TFBS hits are grouped together according to the clustering results and the overrepresentation analysis is carried out on the merged hits. Each color represents the binding sites for a cluster of TFs. Those overlapping binding sites for profiles within a common cluster are consolidated into a single site for over-representation analysis. Highly similar profiles are often offset by 1-2 bp, sharing a common core pattern. (C) TFBS clustering algorithm pseudocode.
3.2.8 Species-Specific Implementation Details

The amount of sequence analyzed for the pre-computation of putative TFBS locations for the gene-based over-representation variants was adjusted to reflect the intergenic distances (unpublished observations) of the target organism. For human and mouse, 10,000 bp upstream and 10,000 bp downstream from the Ensembl-annotated TSS were searched for TFBS hits. For fruit fly, 3,000 bp upstream and 3,000 bp downstream were searched and for nematode, 1,500 bp in each direction. For yeast, only 1,000 bp upstream of TSSs and downstream to the 3’ end of each gene were searched. In gene-based pre-calculations for invertebrates, exons were excluded as in general they are highly conserved and thus the conservation filter employed to reduce false positive TFBSs is rendered ineffective. As opposed to vertebrates and insects, nematodes exhibit some operon structures of gene organization (Blumenthal and Gleason 2003). An
operon consists of multiple adjacent genes that are transcribed as a single unit, which is then spliced into separate mRNAs for translation. When analyzing genes that are part of an annotated operon, oPOSSUM-3 accounts for the operon structures by restricting the search space to the regions flanking the annotated start position of the first gene in the operon. TFBS predictions in the search region of the first gene are deemed to apply to all other genes in the operon (Figure 3-6).

**Figure 3-6. oPOSSUM system provisions for species with operon structures.**

Nematode operon annotation is retrieved from Wormbase, a central repository for nematodes. If a gene is found to be a member of an operon, the search region is changed to that belonging to the most 5’ gene.

While JASPAR now includes distinct divisions for nematode and insect TFBS profiles, coverage is not optimal. For insects, the homeodomain family heavily dominates the set of profiles (90 of 125). For worms, there are currently only four profiles in the database. Thus, while human and mouse oPOSSUM-3 databases are built with vertebrate profiles only, fruit fly and nematode versions are built with all available metazoan profiles in JASPAR 2010.

The yeast gene-based oPOSSUM-3 database differs from the metazoan versions of oPOSSUM-3 in that it does not use conservation filtering, as the compact promoter regions analyzed more readily allow detection.
3.2.9 Build Process for oPOSSUM-3

Figure 3-7 outlines the build process for gene-based versions of oPOSSUM-3. The system imports the gene and transcript annotation data from Ensembl database, and the phastCons conservation scores from the UCSC database. These data are combined to define the conserved segments in each gene. TFBS profiles from JASPAR database are used to search for putative TFBSs within or overlapping the conserved regions, and the resulting counts (both single site and cluster sites) are stored in the oPOSSUM-3 database for pre-determined search region lengths. The pre-computation facilitates faster analysis for the user if the pre-determined search regions are used for analysis. If the user chooses custom search region lengths, TFBS counts are computed at the time of the analysis. At present the computation is performed on a high-performance computer cluster. The sequence-based oPOSSUM-3 foregoes this pre-computation pipeline, and all calculations are done at the time of analysis.
Figure 3-7. The build process for oPOSSUM 3 gene-based analysis.

The system incorporates data from the gene annotation data from Ensembl, TFBS profiles from JASPAR, and multi-species conservation information based on phastCons scores from UCSC Genome Browser. The oPOSSUM database incorporates the data from these sources to pre-compute TFBS profile hits for gene-based analysis.

3.2.10 Data Sources for ChIP-Seq Based Analyses

ChIP-seq data sets for Nrf2 (mouse embryonic fibroblast) and FoxA2 (mouse liver) were obtained in their final processed form from their respective authors (Robertson et al. 2008; Malhotra et al. 2010). The cMyc and Sox2 ChIP-Seq data sets (mouse embryonic stem cells (mES)) (Chen et al. 2008) were obtained as BED files from Gene Expression Omnibus (GEO, accession GSE11431) (Edgar et al. 2002) and processed into peaks (putative TF bound regions) using the FindPeaks algorithm (Fejes et al. 2008) with the following settings: -control -dist_type 1 200 -subpeaks 0.6 -trim 0.2 - duplicatefilter. All NCBI36/mm8 data sets were converted to build NCBI37/mm9 using the UCSC Utility: Batch Coordinate Conversion (liftOver). A control data set from mES cells was obtained from GEO (accession GSE18776) (Peng et al. 2009) to be used as alternative background sequences.
3.2.11 ChIP-Seq Related Analysis

Data analyses and plots were derived using the R statistical package (R Development Core Team 2008).

3.2.11.1 Selection of foreground data sets

The FoxA2, cMyc and Sox2 bound regions were identified using a ratio of peak height to width, and the top ranked regions were selected (1000 for Foxa2 and cMyc; 1200 for Sox2). Nrf2 had undergone stringent filtering by the authors and had a strong signal in the data thus all 1256 sequences were used.

3.2.11.2 Selection of background data sets

Background datasets were extracted from the control data of the related ChIP-Seq experiments, or in the case of cMyc from an alternative mES cell control dataset (Peng et al. 2009). All backgrounds were selected to match the mononucleotide GC composition of the foreground, except where we tested the impact of using different background GC compositions. The default background size for our analyses was 2-fold greater than the matching foreground, except for Nrf2 where the fibroblast background was limited to the same size as the foreground in order to obtain sequences with GC composition distribution similar to the foreground. Data analyses and plots were derived using the R statistical package (R Development Core Team 2008).
### 3.2.11.3 Thresholding by score

The ranges of the two enrichment scores are dependent on foreground dataset size, therefore thresholding results by score was empirically determined for each foreground size, yet selected to be generally true across all datasets. A foreground with ~200 sequences is thresholded at $Z$-score $\geq 20$ or Fisher score $\geq 7$, while a foreground with ~1000 sequences is thresholded at $Z$-score $\geq 50$ or Fisher score $\geq 35$ (Figure 3-8).

**Figure 3-8. The range of enrichment scores increases as the number of foreground sequences increases.**

We performed SSA on a series of foregrounds ranging from 100-1500 sequences and matching GC composition backgrounds. The impact of foreground size on the two enrichments scores is linear, as shown in (A) $Z$-score and (B) Fisher score. The plots display scores for i) the primary TF – Sox2, ii) the TF near an empirical threshold (from 1200 sequence dataset, $Z$-score=50 and Fisher score=35) – Sox5, and iii) the median TF – Nr1h2::Rxra.
3.2.11.4 GC composition calculation

Sequence composition of ChIP-Seq sequences was determined at the mononucleotide level, using counts of A, G, C, and T. TF profile composition was determined similarly by counting the frequency of the nucleotides in all the sequences contributing to the profile. All backgrounds used in testing were selected to reflect the sequence GC composition distribution of the foreground data set.

3.2.12 Analysis Parameters

All analyses were performed with the default parameters, unless otherwise stated. Default settings restrict oPOSSUM-3 results to those profiles with information content of at least 8 bits, and to putative TFBSs whose score is at least 80% of the optimal score. For gene-based analyses, the default search region distances were used for each species (search region level 3; Table 3-1).

Table 3-1. Search region level distances for human/mouse, fly and nematode in oPOSSUM-3.

oPOSSUM-3 searches for the presence of motifs in noncoding regions meeting a specified conservation threshold and situated within a selected search region.

<table>
<thead>
<tr>
<th>Search Region Level</th>
<th>Human, Mouse</th>
<th>Fruit Fly</th>
<th>Nematode</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Upstream (bp)</td>
<td>Downstream (bp)</td>
<td>Upstream (bp)</td>
</tr>
<tr>
<td>1</td>
<td>10,000</td>
<td>10,000</td>
<td>3,000</td>
</tr>
<tr>
<td>2</td>
<td>10,000</td>
<td>5,000</td>
<td>3,000</td>
</tr>
<tr>
<td>3</td>
<td>5,000</td>
<td>5,000</td>
<td>2,000</td>
</tr>
<tr>
<td>4</td>
<td>5,000</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>5</td>
<td>2,000</td>
<td>2,000</td>
<td>1,000</td>
</tr>
<tr>
<td>6</td>
<td>2,000</td>
<td>0</td>
<td>1,000</td>
</tr>
</tbody>
</table>
3.3 System Walk-Through

oPOSSUM-3 is available as a web service at http://opossum.cisreg.ca. The main page of oPOSSUM-3 lists each of the analysis methods offered as well as the species supported, allowing the user to select the appropriate combination. Figure 3-9 graphically depicts the oPOSSUM-3 analysis pipeline for both the gene-based and sequence-based variants, while Figure 3-10 depicts the user interface.

Figure 3-9. oPOSSUM analysis pipeline.
Dashed lines indicate optional stages.
Figure 3-10. Web interface.
(A) oPOSSUM home page listing the various analyses and organisms available. (B) Four steps for initiating a Single Site Analysis (SSA) with user chosen parameters.
3.3.1 Input for Gene-Based Analysis

For a gene-based analysis, the system takes as input the list of identifiers (IDs) of the genes to be analyzed. While the default input ID type is an Ensembl gene ID, the system accepts nine alternatives, including official gene symbols and Uniprot IDs. For the background gene set, the user can: (i) choose the entire gene set from the oPOSSUM-3 database; (ii) specify a number of genes to be randomly chosen; or (iii) supply a list of gene IDs.

The gene-based analysis can be performed in either default or custom modes. In default mode, the user is restricted to pre-defined search regions, sequence conservation levels and TFBS score thresholds. While in custom mode, the user can specify the parameter values within ranges allowed by the system. The default mode is faster as it makes use of pre-computed TFBS counts from the database, whereas the custom mode requires the system to compile binding site frequencies. For SSA, the user may select profiles from species-dependent subsets of JASPAR CORE, JASPAR PBM or oPOSSUM-3-specific PENDING collections. The PENDING collection contains profiles that are not available from the 2010 release of JASPAR. The user specifies whether to use the entire set or a selected subset of the profiles. For aCSA, the user must select an anchor profile. For the TFBS cluster-based versions (both TCA and aCTCA), instead of selecting individual TFBS profiles, the user selects the TF families of interest. The system will then include all TFBS clusters that belong to the selected TF families.

3.3.2 Input for Sequence-Based Analysis

The input parameter selection for the sequence-based methods is simpler than gene-based, as there are no search region or conservation level settings to specify. The user must supply fasta formatted sequences for both foreground and background data sets to be scanned for TFBS hits. oPOSSUM-3 provides a link to the Galaxy service for users needing to generate fasta files from sequence coordinates (Goecks et al. 2010). oPOSSUM-3 provides background sets for users lacking a matched background for their
data. Lastly, the user must select TFBS profiles to be included in the analysis, or provide a set of custom profiles.

3.3.3 Results Output

The results are generated and returned to the user in a table format, along with a summary of the user-specified parameters and the nucleotide compositions of the foreground and background sequences used in the analysis. In gene-based SSA and TCA, the results table is generated and returned to the user immediately, whereas for gene-based aCSA and aCTCA, as well as for all sequence-based methods, a link to the results is emailed to the user after computations are complete. If needed, the user can download the results as a tab-delimited text file for further analysis or record keeping. By default, the results in the table are ranked by descending Z-score and the user can specify the number of motifs to report based on one score (default is all results). The table can be re-sorted by the TF name, TF class and family, binding profile information content (IC), Z-score or Fisher score. The TF names are linked to the corresponding entries from the JASPAR web site. As a profile’s IC can affect its rate of occurrence, those profiles with IC lower than 9 or greater than 19 are highlighted to inform the user of extreme cases. The TFBS counts are linked to a separate TFBS details page that specifies the sequences and locations. For sequence-based analysis, the report specifies the overall GC mononucleotide content for both the background and foreground sequence sets. For SSA and aCSA, the results table reports the GC content of each TFBS and highlights the extreme values (i.e. those that are below 0.33 or above 0.66). For TFBS cluster-based versions, the results table entries are based on TFBS cluster names instead of individual profiles. The TFBS cluster hit coordinates and the corresponding sequences represent the merged TFBS hits. Each TFBS cluster name is linked to a separate page with the summary information on the cluster, such as TF class, family, and the member profiles that constitute the cluster.
3.4 Results

3.4.1 Application to Reference Collections

To validate the performance of the oPOSSUM-3 system, each of the analysis methods was tested using either sets of co-expressed genes or sequences identified in ChIP-Seq studies. The gene sets were restricted to cases in which there was prior knowledge of TF(s) responsible for co-expression. The ChIP-Seq sequences, by the nature of the ChIP method, are already self-restricted to a primary TF (the target of the antibody used in the experiments).

3.4.2 Skeletal Muscle Reference Gene Set

Skeletal muscle-specific genes are known to be regulated by a core set of TFs including (but not limited to) Mef2A, Myf, Srf, Tead1 and Sp1, with the first three performing prominent roles (Braun et al. 1994; Rudnicki and Jaenisch 1995; Wasserman and Fickett 1998; Naya and Olson 1999). From literature, a collection of 25 human genes regulated by muscle-specific enhancers was prepared (refer to supplemental data). This set was analyzed with the full set of oPOSSUM-3 gene-based methods for human, and the results are shown in Table 3-2. While rankings differ depending on the type of analysis performed and the scoring method used, the core muscle-specific TFs of Mef2A, Myf and Srf are included in the top 10 ranks. SSA (Table 3-2a) is able to identify the muscle factors Mef2A and Myf as candidate TFs for regulating the muscle gene set, along with Nkx2-5, which is also a known muscle-specific TF (Durocher et al. 1997). TCA (Table 3-2b) yields clusters c113 (Nkx2-5), C58 (Srf) and C130 (Mef2A) when ranked by Z-scores, and c58 (Srf) and c34 (Egr1) when ranked by Fisher scores. Egr1 is a muscle-related TF (Wada et al. 2003; Irrcher and Hood 2004). When run with Mef2A as the anchoring TF, aCSA (Table 3-2c) reports known muscle regulatory TF candidate pairings with Mef2A, including Sp1, Anrta::Ahr and Klf4 (Niermann et al. 2003; Yoshida et al. 2010). Similarly for aCTCA (Table 3-2d), Mef2A pairs with ETS domain TFs (a
known muscle regulatory TF family) in addition to clusters containing Nkx2-5, Arnt::Ahr and Klf4.

Table 3-2. oPOSSUM-3 results for the muscle reference gene set (human).

Known muscle TFs are bolded. FDR (false discovery rate) column lists the adjusted p-values calculated using the Benjamini & Hochberg algorithm. (Background: 5000 random, Conservation: 70%, Matrix Score Threshold: 80%, Search region: +/- 5000 bp)

a) Single Site Analysis, using JASPAR CORE vertebrate profiles. Mef2A is ranked 6<sup>th</sup> in the Z-score ranking (score = 16.6), but first in the Fisher score ranking. Nkx2-5 is a known muscle regulatory TF involved in cardiac muscle development. The bHLH TFs Nhlh1 and Myf have similar binding profiles.

<table>
<thead>
<tr>
<th>Name</th>
<th>Class:Family</th>
<th>Z-score</th>
<th>FDR</th>
<th>Name</th>
<th>Class:Family</th>
<th>Fisher score</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tbp</td>
<td>Beta-sheet:TATA-binding</td>
<td>24.0</td>
<td>6.0E-126</td>
<td>Mef2A</td>
<td>Other Alpha-Helix:MADS</td>
<td>14.7</td>
<td>4.8E-05</td>
</tr>
<tr>
<td>Nkx2-5</td>
<td>HTH:Homeo</td>
<td>21.8</td>
<td>5.5E-104</td>
<td>Sox5</td>
<td>Other Alpha-Helix:HMG</td>
<td>10.5</td>
<td>1.2E-03</td>
</tr>
<tr>
<td>Arid3 A</td>
<td>HTH:Arid</td>
<td>20.4</td>
<td>1.4E-91</td>
<td>Tcfcp 2</td>
<td>Other:CP2</td>
<td>10.1</td>
<td>1.2E-03</td>
</tr>
<tr>
<td>Prrx2</td>
<td>HTH:Homeo</td>
<td>20.3</td>
<td>1.3E-90</td>
<td>Nhlh1</td>
<td>Zipper-Type:HLH</td>
<td>10.0</td>
<td>1.2E-03</td>
</tr>
<tr>
<td>Nobox</td>
<td>HTH:Homeo</td>
<td>16.7</td>
<td>1.3E-61</td>
<td>Myf</td>
<td>Zipper-Type:HLH</td>
<td>9.8</td>
<td>1.2E-03</td>
</tr>
<tr>
<td>Mef2A</td>
<td>Other Alpha-Helix:MADS</td>
<td>16.6</td>
<td>1.3E-60</td>
<td>Creb1</td>
<td>Zipper-Type:LeuZip</td>
<td>8.8</td>
<td>2.6E-03</td>
</tr>
<tr>
<td>Pdx1</td>
<td>HTH:Homeo</td>
<td>16.6</td>
<td>1.3E-60</td>
<td>Ebf1</td>
<td>Zipper-Type:HLH</td>
<td>8.8</td>
<td>2.6E-03</td>
</tr>
<tr>
<td>Nfil3</td>
<td>Zipper-Type:LeuZip</td>
<td>14.8</td>
<td>1.3E-60</td>
<td>Myb</td>
<td>HTH:Myb</td>
<td>8.5</td>
<td>2.8E-03</td>
</tr>
<tr>
<td>Gfi</td>
<td>Zinc-coord:ZnF</td>
<td>14.1</td>
<td>3.1E-44</td>
<td>Zfx</td>
<td>Zinc-coord:ZnF</td>
<td>8.5</td>
<td>2.8E-03</td>
</tr>
<tr>
<td>Foxq1</td>
<td>Winged HTH:Forkhead</td>
<td>13.1</td>
<td>1.0E-38</td>
<td>Insm1</td>
<td>Zinc-coord:ZnF</td>
<td>8.0</td>
<td>3.7E-03</td>
</tr>
</tbody>
</table>
b) TFBS Cluster Analysis. TFs of interest are shown in parentheses under the cluster name. C58 (Srf) ranked 6\textsuperscript{th} (score = 18.1) and C130 (Mef2A) 8\textsuperscript{th} (score = 16.4) in the Z-score ranking, while C143 (Myf) ranked 7\textsuperscript{th} (score = 9.9) and C58 (Srf) ranked 9\textsuperscript{th} (score = 9.1) in the Fisher score ranking. C58 contains the Srf profile from the JASPAR PBM collection.

<table>
<thead>
<tr>
<th>Ordered by Z-scores</th>
<th>Ordered by Fisher scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Class:Family</td>
</tr>
<tr>
<td>C113 (Nkx2-5)</td>
<td>HTH:Homeo</td>
</tr>
<tr>
<td>C1</td>
<td>Beta-sheet:TATA-binding</td>
</tr>
<tr>
<td>C132</td>
<td>Beta-sheet:TATA-binding</td>
</tr>
<tr>
<td>C160</td>
<td>Other Alpha-Helix:HMG</td>
</tr>
<tr>
<td>C14</td>
<td>HTH:Arid</td>
</tr>
<tr>
<td>C58 (Srf)</td>
<td>Other Alpha-Helix:MADS</td>
</tr>
<tr>
<td>C161</td>
<td>Other Alpha-Helix:HMG</td>
</tr>
<tr>
<td>C130 (Mef2A)</td>
<td>Other Alpha-Helix:MADS</td>
</tr>
<tr>
<td>C15</td>
<td>Helix-Turn-Heli::Arid</td>
</tr>
<tr>
<td>C169</td>
<td>HTH:Homeo</td>
</tr>
</tbody>
</table>
c) Anchored Combination Site Analysis. (Anchor: Mef2A, Maximum Inter-Binding Distance: 100 bp)

<table>
<thead>
<tr>
<th>Name</th>
<th>Class:Family</th>
<th>Z-score</th>
<th>FDR</th>
<th></th>
<th>Name</th>
<th>Class:Family</th>
<th>Fisher score</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zeb1</td>
<td>Zinc-coord.:ZnF</td>
<td>34.1</td>
<td>1.4E-252</td>
<td></td>
<td>Zeb1</td>
<td>Zinc-coord.:ZnF</td>
<td>17.6</td>
<td>2.0E-06</td>
</tr>
<tr>
<td>Znf354C</td>
<td>Zinc-coord.:ZnF</td>
<td>31.8</td>
<td>4.8E-221</td>
<td></td>
<td>Arnt:Ahrr</td>
<td>Zipper-Type:HLH</td>
<td>17.2</td>
<td>2.0E-06</td>
</tr>
<tr>
<td>Gfi</td>
<td>Zinc-coord.:ZnF</td>
<td>31.8</td>
<td>5.0E-221</td>
<td></td>
<td>Sp1</td>
<td>Zinc-coord.:ZnF</td>
<td>16.7</td>
<td>2.2E-06</td>
</tr>
<tr>
<td>Mzf1_1-4</td>
<td>Zinc-coord.:ZnF</td>
<td>30.3</td>
<td>2.4E-200</td>
<td></td>
<td>Klf4</td>
<td>Zinc-coord.:ZnF</td>
<td>16.3</td>
<td>2.4E-06</td>
</tr>
<tr>
<td>Gata1</td>
<td>Zinc-coord.:GATA</td>
<td>30.1</td>
<td>7.4E-198</td>
<td></td>
<td>Tcfcp21</td>
<td>Other:CP2</td>
<td>15.1</td>
<td>6.4E-06</td>
</tr>
<tr>
<td>Nkx2-5</td>
<td>HTH:Homeo</td>
<td>28.9</td>
<td>1.5E-182</td>
<td></td>
<td>Mzf1_1-4</td>
<td>Zinc-coord:Znf</td>
<td>14.8</td>
<td>7.2E-06</td>
</tr>
<tr>
<td>SPIB</td>
<td>Winged HTH:Ets</td>
<td>27.8</td>
<td>4.0E-169</td>
<td></td>
<td>Zfx</td>
<td>Zinc-coord:ZnF</td>
<td>13.5</td>
<td>2.3E-05</td>
</tr>
<tr>
<td>HoxA5</td>
<td>HTH:Homeo</td>
<td>27.0</td>
<td>1.9E-159</td>
<td></td>
<td>Hif1A::Arnt</td>
<td>Zipper-Type:HTH</td>
<td>13.3</td>
<td>2.4E-05</td>
</tr>
<tr>
<td>Prrx2</td>
<td>HTH:Homeo</td>
<td>26.6</td>
<td>3.6E-155</td>
<td></td>
<td>Pparg::Rxra</td>
<td>Zinc-coord:NucRecept</td>
<td>12.3</td>
<td>5.9E-05</td>
</tr>
<tr>
<td>YY1</td>
<td>Zinc-coord:ZnF</td>
<td>25.4</td>
<td>1.5E-141</td>
<td></td>
<td>Znf354C</td>
<td>Zinc-coord:ZnF</td>
<td>12.1</td>
<td>6.4E-05</td>
</tr>
</tbody>
</table>
Anchored Combination TFBS Cluster Analysis. (Anchor: Mef2A, Maximum Inter-Binding Distance: 100 bp) C72 contains ZEB1, which placed first in the aCSA results.

<table>
<thead>
<tr>
<th>Name</th>
<th>Class:Family</th>
<th>Z-score</th>
<th>FDR</th>
<th>Name</th>
<th>Class:Family</th>
<th>Fisher score</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>C113 (Nkx2-5)</td>
<td>HTH:Homeo</td>
<td>46.1</td>
<td>0.0E+00</td>
<td>C30</td>
<td>Zinc-coord.:ZnF</td>
<td>21.4</td>
<td>8.6E-08</td>
</tr>
<tr>
<td>C55</td>
<td>Winged HTH:Ets</td>
<td>37.7</td>
<td>0.0E+00</td>
<td>C72</td>
<td>Zinc-coord.:ZnF</td>
<td>18.5</td>
<td>7.9E-07</td>
</tr>
<tr>
<td>C159</td>
<td>Other Alpha-Helix:HMG</td>
<td>37.7</td>
<td>0.0E+00</td>
<td>C140</td>
<td>Zipper-Type:HLH</td>
<td>17.9</td>
<td>9.5E-07</td>
</tr>
<tr>
<td>C72</td>
<td>Zinc-coord.:ZnF</td>
<td>34.9</td>
<td>2.7E-265</td>
<td>C104</td>
<td>Zinc-coord.:ZnF</td>
<td>17.5</td>
<td>1.2E-06</td>
</tr>
<tr>
<td>C66</td>
<td>Zinc-coord.:ZnF</td>
<td>34.2</td>
<td>1.7E-254</td>
<td>C51</td>
<td>Zipper-Type:HTH</td>
<td>16.9</td>
<td>1.6E-06</td>
</tr>
<tr>
<td>C41</td>
<td>Other Alpha-Helix:HMG</td>
<td>33.8</td>
<td>1.6E-249</td>
<td>C107</td>
<td>Other:CP2</td>
<td>16.0</td>
<td>3.2E-06</td>
</tr>
<tr>
<td>C75</td>
<td>Zinc-coord.:ZnF</td>
<td>32.9</td>
<td>7.8E-237</td>
<td>C75</td>
<td>Zinc-coord.:ZnF</td>
<td>14.9</td>
<td>8.2E-06</td>
</tr>
<tr>
<td>C30</td>
<td>Zinc-coord.:ZnF</td>
<td>32.8</td>
<td>9.2E-235</td>
<td>C156</td>
<td>Winged HTH:E2F</td>
<td>14.2</td>
<td>1.4E-05</td>
</tr>
<tr>
<td>C87</td>
<td>Zinc-coord.:ZnF</td>
<td>32.3</td>
<td>1.5E-228</td>
<td>C88</td>
<td>Zinc-coord.:ZnF</td>
<td>13.9</td>
<td>1.7E-05</td>
</tr>
<tr>
<td>C2</td>
<td>Zinc-coord:GATA</td>
<td>30.7</td>
<td>2.0E-206</td>
<td>C33</td>
<td>Zinc-coord.:ZnF</td>
<td>13.7</td>
<td>1.9E-05</td>
</tr>
</tbody>
</table>

### 3.4.3 Cilia Gene Set for Nematodes

Inglis *et al.* curated a collection of genes known to be involved in cilia function and structure, making the set available within Ciliome DB (Inglis *et al.* 2006). In vertebrates, RFX TFs are the key regulators of cilia gene expression. The RFX TFs bind to a regulatory motif termed the X-box. The regulation is conserved between mammals and nematodes; X-box sequences are associated with orthologous cilia genes (Efimenko *et al.* 2005). A nematode-specific RFX motif has been computed from these sequences, but these sequences were not experimentally validated in a rigorous manner. There are two RFX profiles in the research literature, differing in the distance between two half sites (Emery *et al.* 1996). These RFX profiles were added to oPOSSUM-3’s custom collection of TFBS profiles and were used to analyze 531 nematode cilia genes. The gene set was
obtained from Ciliome DB by combining the nematode genes from Avidor-Reiss, Blacque SAGE and Efimenko data sets and selecting for those genes that were included in at least 2 of the studies. The results are listed in Table 3-3. RFX profiles received the highest scores in both SSA and TCA. The high scores obtained using the vertebrate RFX profiles in nematode genomic sequences support the conservation of the TFBS sequences across large evolutionary distances.

**Table 3-3. oPOSSUM-3 results for the cilia gene set in nematodes.**

TFs known to be involved in cilia gene regulation are bolded. The scores for daf-19 profile are not shown, as the profile was built based on gene sets included in these analyses, making the results circular. FDR (false discovery rate) column lists the adjusted p-values calculated using the Benjamini & Hochberg algorithm.

a) Single Site Analysis, using JASPAR CORE collection and the custom PENDING collection. Two variants of Rfx1 profiles are from TRANSFAC (Rfx1_1 and Rfx1_2), which were placed into the custom PENDING collection for analysis. Rfx1_1 obtained lower scores than Rfx1_2.

<table>
<thead>
<tr>
<th>Name</th>
<th>Class:Family</th>
<th>Z-score</th>
<th>FDR</th>
<th>Name</th>
<th>Class:Family</th>
<th>Fisher score</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rfx1_2</td>
<td>HTH:RFX</td>
<td>25.0</td>
<td>0.0E+0</td>
<td>Brca1</td>
<td>Other:Other</td>
<td>28.2</td>
<td>3.1E-35</td>
</tr>
<tr>
<td>Gabpa</td>
<td>Winged HTH:Ets</td>
<td>15.3</td>
<td>0.0E+0</td>
<td>Mrr</td>
<td>HTH:Homeo</td>
<td>28.0</td>
<td>1.5E-26</td>
</tr>
<tr>
<td>Nfya</td>
<td>Other Alpha-Helix:NFY CCAAT</td>
<td>13.9</td>
<td>3.8E-224</td>
<td>Caup</td>
<td>HTH:Homeo</td>
<td>27.2</td>
<td>1.5E-26</td>
</tr>
<tr>
<td>Tfap2A</td>
<td>Zipper-Type:HLH</td>
<td>12.5</td>
<td>1.7E-214</td>
<td>Ara</td>
<td>HTH:Homeo</td>
<td>27.0</td>
<td>6.9E-26</td>
</tr>
<tr>
<td>Rfx1_1</td>
<td>HTH:RFX</td>
<td>11.0</td>
<td>1.4E-141</td>
<td>Hlf</td>
<td>Zinc-coord:GATA</td>
<td>27.0</td>
<td>8.7E-26</td>
</tr>
</tbody>
</table>
b) TFBS Cluster Analysis. C154 contains RFX profiles from TRANSFAC, while C155 consists of RFX profiles from the JASPAR PBM collection.

<table>
<thead>
<tr>
<th>Name</th>
<th>Class:Family</th>
<th>Z-score</th>
<th>FDR</th>
<th>Name</th>
<th>Class:Family</th>
<th>Fisher score</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>C154</td>
<td>HTH:RFX</td>
<td>35.8</td>
<td>5.8E-279</td>
<td>C154</td>
<td>HTH:RFX</td>
<td>30.0</td>
<td>1.8E-28</td>
</tr>
<tr>
<td>C155</td>
<td>HTH:RFX</td>
<td>22.6</td>
<td>6.6E-112</td>
<td>C55</td>
<td>Winged HTH:Ets</td>
<td>29.7</td>
<td>1.8E-28</td>
</tr>
<tr>
<td>C37</td>
<td>Zinc-coord:ZnF</td>
<td>13.7</td>
<td>2.0E-41</td>
<td>C3</td>
<td>Zinc-coord:GATA</td>
<td>28.1</td>
<td>4.3E-27</td>
</tr>
<tr>
<td>C23</td>
<td>Other Alpha-Helix:NFY CCAAT</td>
<td>12.3</td>
<td>1.6E-33</td>
<td>C155</td>
<td>HTH:RFX</td>
<td>27.6</td>
<td>1.1E-26</td>
</tr>
<tr>
<td>C139</td>
<td>Zipper-type:HLH</td>
<td>12.0</td>
<td>9.4E-32</td>
<td>C113</td>
<td>HTH:Homeo</td>
<td>26.8</td>
<td>5.2E-26</td>
</tr>
</tbody>
</table>

### 3.4.4 Results for ChIP-Seq Reference Data Sets

We used oPOSSUM-3 to analyze ChIP-Seq datasets, of which we describe two results here. The Nrf2 data from Malhotra et al. contains relatively few ChIP-Seq regions (1256 sequences) derived from studies of mouse embryonic fibroblasts (Malhotra et al. 2010). Nrf2, also known as Nfe2L2, is a stress activated TF linked to the regulation of detoxification enzymes. We analyzed the Nrf2 data with oPOSSUM-3 (Table 3-4). The Nrf2 profile is consistently found to be a top-scoring motif in all analyses, based on Z-score rankings. The Ap1 motif, known to overlap and coordinate with the Nrf2 binding sequence (Friling et al. 1992), was enriched in the SSA results (by Z-score), and clustered with the Nrf2 profile (cluster C11) in the TCA and aCTCA analyses.
Table 3-4. oPOSSUM-3 results for Nrf2/Nfe2L2 ChIP-Seq data set, using JASPAR CORE vertebrate profiles.

Nrf2/Nfe2L2 and Ap1, a TF with known Nrf2-related biological functions, are bolded. FDR (false discovery rate) column lists the adjusted p-values calculated using the Benjamini & Hochberg algorithm. (Matrix Score Threshold: 80%, JASPAR CORE vertebrate profiles, minimum information content = 8 bits)

a) Single Site Analysis. Nrf2/Nfe2L2 ranks first by both scores. Ap1, ranks second by Z-score.

<table>
<thead>
<tr>
<th>Name</th>
<th>Class:Family</th>
<th>Z-score</th>
<th>FDR</th>
<th>Name</th>
<th>Class:Family</th>
<th>Fisher score</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nfe2L2</td>
<td>Zipper-Type:LeuZip</td>
<td>147.8</td>
<td>0.0E+00</td>
<td>Nfe2L2</td>
<td>Zipper-Type:LeuZip</td>
<td>208.0</td>
<td>5.4E-89</td>
</tr>
<tr>
<td>Ap1</td>
<td>Zipper-Type:LeuZip</td>
<td>49.8</td>
<td>0.0E+00</td>
<td>E2f1</td>
<td>Winged HTH:E2f</td>
<td>12.2</td>
<td>2.9E-04</td>
</tr>
<tr>
<td>Esr2</td>
<td>Zinc-coord:NucRecept</td>
<td>27.9</td>
<td>1.8E-169</td>
<td>Esr2</td>
<td>Zinc-coord:NucRecept</td>
<td>9.31</td>
<td>3.5E-03</td>
</tr>
<tr>
<td>Myf</td>
<td>Zipper-Type:HLH</td>
<td>19.0</td>
<td>1.0E-79</td>
<td>Spi1</td>
<td>Winged HTH:Ets</td>
<td>8.43</td>
<td>6.3E-03</td>
</tr>
<tr>
<td>Esr1</td>
<td>Zinc-coord:NucRecept</td>
<td>18.6</td>
<td>4.3E-76</td>
<td>Myf</td>
<td>Zipper-Type:HLH</td>
<td>7.79</td>
<td>8.2E-03</td>
</tr>
</tbody>
</table>

b) TFBS Cluster Analysis. Nrf2/Nfe2L2 and Ap1 are both in cluster C11, which ranks first by Z-score and fourth by Fisher score.

<table>
<thead>
<tr>
<th>Name</th>
<th>Class:Family</th>
<th>Z-score</th>
<th>FDR</th>
<th>Name</th>
<th>Class:Family</th>
<th>Fisher score</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>C11(Nfe2L2)</td>
<td>Zipper-type:LeuZip</td>
<td>75.0</td>
<td>0.0E+00</td>
<td>C61</td>
<td>Winged HTH:E2f</td>
<td>12.2</td>
<td>8.6E-04</td>
</tr>
<tr>
<td>C143</td>
<td>Zipper-type:HLH</td>
<td>19.9</td>
<td>4.4E-87</td>
<td>C143</td>
<td>Zipper-type:HLH</td>
<td>7.8</td>
<td>2.4E-02</td>
</tr>
<tr>
<td>C88</td>
<td>Zinc-coord:ZnF</td>
<td>18.6</td>
<td>1.4E-76</td>
<td>C88</td>
<td>Zinc-coord:ZnF</td>
<td>7.8</td>
<td>2.4E-02</td>
</tr>
<tr>
<td>C61</td>
<td>Winged HTH:E2f</td>
<td>18.0</td>
<td>3.1E-71</td>
<td>C11(Nfe2L2)</td>
<td>Zipper-type:LeuZip</td>
<td>6.1</td>
<td>9.3E-02</td>
</tr>
<tr>
<td>C16</td>
<td>Ig-fold:Stat</td>
<td>14.1</td>
<td>4.2E-44</td>
<td>C83</td>
<td>Zinc-coord:ZnF</td>
<td>5.8</td>
<td>1.0E-01</td>
</tr>
</tbody>
</table>
c) Anchored Combination Site Analysis. Nrf2/Nfe2L2 associates most frequently with itself and Ap1, by Z-score. (Anchor: Nrf2/Nfe2L2, Maximum inter-binding distance = 100 bp)

<table>
<thead>
<tr>
<th>Name</th>
<th>Class:Family</th>
<th>Z-score</th>
<th>FDR</th>
<th>Name</th>
<th>Class:Family</th>
<th>Fisher score</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ap1</td>
<td>Zipper-Type:LeuZip</td>
<td>299.1</td>
<td>0.0E+00</td>
<td>Znf354C</td>
<td>Zinc-coord:ZnF</td>
<td>213.0</td>
<td>3.6E-91</td>
</tr>
<tr>
<td>Nfe2L2</td>
<td>Zipper-Type:LeuZip</td>
<td>277.9</td>
<td>0.0E+00</td>
<td>Spi</td>
<td>Winged HTH:Ets</td>
<td>210.0</td>
<td>3.6E-90</td>
</tr>
<tr>
<td>SpiB</td>
<td>Winged HTH:Ets</td>
<td>259.4</td>
<td>0.0E+00</td>
<td>HoxA5</td>
<td>HTH:Homeo</td>
<td>206.0</td>
<td>9.9E-89</td>
</tr>
<tr>
<td>HoxA5</td>
<td>HTH:Homeo</td>
<td>250.6</td>
<td>0.0E+00</td>
<td>Ap1</td>
<td>Zipper-Type:LeuZip</td>
<td>206.0</td>
<td>9.9E-89</td>
</tr>
<tr>
<td>Zeb1</td>
<td>Zinc-coord:ZnF</td>
<td>228.5</td>
<td>0.0E+00</td>
<td>Zeb1</td>
<td>Zinc-coord:ZnF</td>
<td>202.0</td>
<td>4.3E-87</td>
</tr>
</tbody>
</table>

d) Anchored Combination TFBS Cluster Analysis. Nrf2/Nfe2L2 associates most frequently with cluster C11, which contains Nrf2/Nfe2L2 and Ap1. (Anchor: Nrf2/Nfe2L2, Maximum inter-binding distance = 100 bp)

<table>
<thead>
<tr>
<th>Name (Nfe2L2)</th>
<th>Class:Family</th>
<th>Z-score</th>
<th>FDR</th>
<th>Name (Nfe2L2)</th>
<th>Class:Family</th>
<th>Fisher score</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>C11</td>
<td>Zipper-type:LeuZip</td>
<td>191.7</td>
<td>0.0E+00</td>
<td>C11</td>
<td>Zipper-type:LeuZip</td>
<td>40.6</td>
<td>4.0E-16</td>
</tr>
<tr>
<td>C55</td>
<td>Winged HTH:Ets</td>
<td>187.6</td>
<td>0.0E+00</td>
<td>C24</td>
<td>Winged HTH:Runt</td>
<td>16.7</td>
<td>4.8E-06</td>
</tr>
<tr>
<td>C113</td>
<td>HTH:Homeo</td>
<td>183.3</td>
<td>0.0E+00</td>
<td>C87</td>
<td>Zinc-coord:ZnF</td>
<td>15.8</td>
<td>7.8E-06</td>
</tr>
<tr>
<td>C57</td>
<td>Winged HTH:Forkhead</td>
<td>141.6</td>
<td>0.0E+00</td>
<td>C88</td>
<td>Zinc-coord:ZnF</td>
<td>12.7</td>
<td>1.3E-04</td>
</tr>
<tr>
<td>C41</td>
<td>Other Alpha-Helix:HMG</td>
<td>128.1</td>
<td>0.0E+00</td>
<td>C143</td>
<td>Zipper-type:HLH</td>
<td>12.0</td>
<td>1.9E-04</td>
</tr>
</tbody>
</table>
For contrast, the FoxA2 data from Robertson et al. shows a large number of TF bound regions (>10,000 sequences) using mouse liver as the source (Robertson et al. 2008). We took a subset of 1,000 sequences for analysis (see Methods). FoxA2 (also known as Hnf3α), a TF linked to differentiation, is a member of the forkhead-box family of TF proteins for which the JASPAR 2010 database has eight strong IC profiles. Using the sequence-based SSA we recovered FoxA2 as the top ranking TF, while three additional forkhead-box family profiles were ranked in the top five enriched TFBSs (Table 3-5). The other prominently enriched profile represents the Hnf4 hepatocyte nuclear factor, which is predominantly expressed in the liver. The TCA and aCTCA analyses complement the SSA results, with the clusters C57, representing forkhead-box binding profiles, and C19, containing Hnf4A, ranking either first and second (TCA) or first and fifth (aCTCA) by Z-score.
Table 3-5. oPOSSUM-3 results for FoxA2 ChIP-Seq data set, using JASPAR CORE vertebrate profiles.

FoxA2 is bolded. FDR (false discovery rate) column lists the adjusted p-values calculated using the Benjamini & Hochberg algorithm. (Matrix Score Threshold: 80%, JASPAR CORE vertebrate profiles, minimum information content = 8 bits)

a) Single Site Analysis Results. Forkhead-box binding profiles rank the highest by both Z-score and Fisher score. Hnf4A, a hepatocyte related TF, ranks fifth by both scores.

<table>
<thead>
<tr>
<th>Name</th>
<th>Class:Family</th>
<th>Z-score</th>
<th>FDR</th>
<th>Name</th>
<th>Class:Family</th>
<th>Fisher score</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>FoxA2</td>
<td>Winged HTH:Forkhead</td>
<td>128.6</td>
<td>0.0E+00</td>
<td>FoxD1</td>
<td>Winged HTH:Forkhead</td>
<td>106.0</td>
<td>1.7E-54</td>
</tr>
<tr>
<td>FoxF2</td>
<td>Winged HTH:Forkhead</td>
<td>117.3</td>
<td>0.0E+00</td>
<td>FoxA2</td>
<td>Winged HTH:Forkhead</td>
<td>96.3</td>
<td>6.7E-50</td>
</tr>
<tr>
<td>FoxD1</td>
<td>Winged HTH:Forkhead</td>
<td>107.2</td>
<td>0.0E+00</td>
<td>FoxF2</td>
<td>Winged HTH:Forkhead</td>
<td>80.3</td>
<td>1.1E-45</td>
</tr>
<tr>
<td>FoxA1</td>
<td>Winged HTH:Forkhead</td>
<td>100.9</td>
<td>0.0E+00</td>
<td>FoxA1</td>
<td>Winged HTH:Forkhead</td>
<td>33.5</td>
<td>4.4E-43</td>
</tr>
<tr>
<td>Hnf4A</td>
<td>Zinc-coord:NucRecept</td>
<td>66.9</td>
<td>0.0E+00</td>
<td>Hnf4A</td>
<td>Zinc-coord:NucRecept</td>
<td>15.1</td>
<td>2.1E-28</td>
</tr>
</tbody>
</table>

b) TFBS Cluster Analysis. Cluster C57, containing forkhead-box binding profiles, ranks highest by both scores. Cluster C19, containing the Hnf4A hepatocyte-related TF profile, ranks second by both scores.

<table>
<thead>
<tr>
<th>Name</th>
<th>Class:Family</th>
<th>Z-score</th>
<th>FDR</th>
<th>Name</th>
<th>Class:Family</th>
<th>Fisher score</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>C57</td>
<td>Winged HTH:Forkhead</td>
<td>75.6</td>
<td>0.0E+00</td>
<td>C57</td>
<td>Winged HTH:Forkhead</td>
<td>9.0</td>
<td>2.1E-02</td>
</tr>
<tr>
<td>C19</td>
<td>Zinc-coord:NucRecept</td>
<td>56.3</td>
<td>0.0E+00</td>
<td>C19</td>
<td>Zinc-coord:NucRecept</td>
<td>3.1</td>
<td>1.0E+00</td>
</tr>
<tr>
<td>C9</td>
<td>Zipper-type:LeuZip</td>
<td>40.4</td>
<td>0.0E+00</td>
<td>C7</td>
<td>Zinc-coord:Loop-Sheet-Helix</td>
<td>2.6</td>
<td>1.0E+00</td>
</tr>
<tr>
<td>C41</td>
<td>Other Alpha-Helix:HMG</td>
<td>37.7</td>
<td>0.0E+00</td>
<td>C111</td>
<td>HTH:Other Homeo::NF1-CCAAT</td>
<td>0.7</td>
<td>1.0E+00</td>
</tr>
<tr>
<td>C17</td>
<td>Zinc-coord:NucRecept</td>
<td>37.5</td>
<td>1.1E-306</td>
<td>C41</td>
<td>Other Alpha-Helix:HMG</td>
<td>0.5</td>
<td>1.0E+00</td>
</tr>
</tbody>
</table>
c) Anchored Combination Site Analysis. FoxA2 associates closely with other forkhead-box binding profiles, and the profile of Hnf4A, a hepatocyte-related TF. (Anchor: FoxA2, Maximum inter-binding distance = 100 bp)

<table>
<thead>
<tr>
<th>Name</th>
<th>Class:Family</th>
<th>Z-score</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>FoxA1</td>
<td>Winged HTH:Forkhead</td>
<td>285.6</td>
<td>0.0E+00</td>
</tr>
<tr>
<td>FoxD1</td>
<td>Winged HTH:Forkhead</td>
<td>249.6</td>
<td>0.0E+00</td>
</tr>
<tr>
<td>FoxF2</td>
<td>Winged HTH:Forkhead</td>
<td>244.5</td>
<td>0.0E+00</td>
</tr>
<tr>
<td>Hnf4A</td>
<td>Zinc-coord:NucRecept</td>
<td>238.2</td>
<td>0.0E+00</td>
</tr>
<tr>
<td>FoxA2</td>
<td>Winged HTH:Forkhead</td>
<td>223.3</td>
<td>0.0E+00</td>
</tr>
</tbody>
</table>

Ordered by Z-scores

<table>
<thead>
<tr>
<th>Name</th>
<th>Class:Family</th>
<th>Fisher score</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>FoxD1</td>
<td>Winged HTH:Forkhead</td>
<td>125.0</td>
<td>6.0E-53</td>
</tr>
<tr>
<td>FoxA2</td>
<td>Winged HTH:Forkhead</td>
<td>104.0</td>
<td>4.0E-44</td>
</tr>
<tr>
<td>Zeb1</td>
<td>Zinc-coord:ZnF</td>
<td>101.0</td>
<td>5.3E-43</td>
</tr>
<tr>
<td>Hnf4A</td>
<td>Zinc-coord:NucRecept</td>
<td>99.3</td>
<td>2.1E-42</td>
</tr>
<tr>
<td>FoxF2</td>
<td>Winged HTH:Forkhead</td>
<td>99.1</td>
<td>2.1E-42</td>
</tr>
</tbody>
</table>

d) Anchored Combination TFBS Cluster Analysis. Cluster C57, which contains forkhead-box binding profiles, ranked fifth by both scores for association with FoxA2. Cluster C19, containing the Hnf4A hepatocyte-related TF profile, ranks first by Z-score and fourth by Fisher score. (Anchor: FoxA2, Maximum inter-binding distance = 100 bp)

<table>
<thead>
<tr>
<th>Name</th>
<th>Class:Family</th>
<th>Z-score</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>C19 (Hnf4A)</td>
<td>Zinc-coord:NucRecept</td>
<td>183.2</td>
<td>0.0E+00</td>
</tr>
<tr>
<td>C41</td>
<td>Other Alpha-Helix:HMG</td>
<td>172.8</td>
<td>0.0E+00</td>
</tr>
<tr>
<td>C113</td>
<td>HTH:Homeo</td>
<td>162.5</td>
<td>0.0E+00</td>
</tr>
<tr>
<td>C17</td>
<td>Zinc-coord:NucRecept</td>
<td>159.3</td>
<td>0.0E+00</td>
</tr>
<tr>
<td>C57 (FoxA2)</td>
<td>Winged HTH:Forkhead</td>
<td>155.4</td>
<td>0.0E+00</td>
</tr>
</tbody>
</table>

Ordered by Z-scores

<table>
<thead>
<tr>
<th>Name</th>
<th>Class:Family</th>
<th>Fisher score</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>C55</td>
<td>Winged HTH:Ets</td>
<td>10.2</td>
<td>4.7E-03</td>
</tr>
<tr>
<td>C72</td>
<td>Zinc-coord:ZnF</td>
<td>9.8</td>
<td>4.7E-03</td>
</tr>
<tr>
<td>C113</td>
<td>HTH:Homeo</td>
<td>8.0</td>
<td>1.6E-02</td>
</tr>
<tr>
<td>C19 (Hnf4A)</td>
<td>Zinc-coord:NucRecept</td>
<td>8.0</td>
<td>1.6E-02</td>
</tr>
<tr>
<td>C57 (FoxA2)</td>
<td>Winged HTH:Forkhead</td>
<td>6.7</td>
<td>4.3E-02</td>
</tr>
</tbody>
</table>
3.4.5 GC Composition Effect on Over-Representation Results

The genome is comprised of regions that are distinct by the nature of their nucleotide composition. Promoters for housekeeping genes in particular are often found to be rich in G and C nucleotides or in CpG islands (Yamashita et al. 2005), while tissue specific promoters are often found to be relatively more AT-rich (Roider et al. 2009). These broad composition differences are reflected among the sequences derived from ChIP experiments, such as ChIP-Seq. The %GC content of the published ChIP-Seq experiments we examined ranged from ~30-57% GC (see Supplemental Methods; unpublished observations). The mean and variance of the %GC composition distribution differed between experiments for different TFs. The binding profiles of the TFs themselves also varied in GC content. Given a set of GC rich sequences to analyze, TFs with GC rich profiles will have a greater probability of having a motif present than AT rich profiles. As over-representation analysis is designed to report TFBS predictions that exist to a greater extent in a set of chosen sequences than in a background set of sequences, we examined the effect of GC composition on motif enrichment scoring using backgrounds with varying GC content (i.e. high, low, and neutral GC composition) against several ChIP-Seq data sets.

We used mouse derived ChIP-Seq data for several TFs with varying GC composition in their profiles: Sox2 (ES cells) – 36% GC, Nrf2 (fibroblast) – 43% GC, FoxA2 (liver) – 46% GC, and cMyc (ES cells) – 61% GC. For each ChIP-Seq dataset we analyzed a subset of the TF bound regions (foreground data) against background sequence sets with a variety of %GC distributions. The over-representation Z-score for each TF was then plotted against the %GC composition of the TF’s profile. As illustrated with Nrf2 in Figure 3-11, composition differences between the foreground and background sequences will affect the over-representation scores of TFs. When the background %GC composition is lower than that of the foreground sequences, the TFs with low GC profiles are reported as under-represented (low Z-score) in the foreground sequences relative to the background sequences, while the TFs with high GC profiles are over-represented (high Z-score) in the foreground sequences versus the background
sequences. The opposite trend occurs when the background %GC composition is higher than the foreground sequences; in this case the TFs with low GC profiles are reported as over-represented. A background with the same GC composition distribution (mean and variance) as the foreground data set removes bias arising from GC composition differences, and controls against TFBSs being unduly assigned high or low over-representation scores.

**Figure 3-11. Relationship between TF profile GC content and enrichment statistics.**

The percentage of G and C nucleotides in the TF profile models are plotted against the motif enrichment Z-scores. The three panels represent analysis results for the same 1,256 Nrf2 ChIP-seq regions (GC composition avg. 43%) compared to three background sets of different GC composition: (A) elevated background GC (avg. 51% GC); (B) low background GC (avg. 37% GC); (C) background with GC composition matched to the distribution of the ChIP-Seq regions (avg. 43% GC). The GC composition of the background used in B is that of the control associated with the Nrf2 ChIP-Seq data. The plotted Z-scores represent the enrichment of TFs in 1256 ChIP-Seq Nrf2-bound regions. While the Nrf2 profile can be distinguished in two cases (B and C), it is lost against the elevated GC background shown in (A). Detection of the Ap1 profile, representing a TF with a known Nrf2-related biological function, is sensitive to background selection. Most TFs that would be ranked highly by Z-score when the background is not corrected for GC composition (seen in A and B) are not ranked highly when the background is matched to the foreground distribution of GC composition (C).
3.4.6 Z-Score vs. Fisher Score for ChIP-Seq Based Analyses

oPOSSUM-3 uses two over-representation scores—a Z-score and a Fisher score. In combination, the two scores inform the user whether enrichment is i) due to a high frequency of TFBSs in some of the sequences and absence in others (Z-score high rank, Fisher score low rank; Figure 3-12, TF: Ap1), ii) due to a majority of regions carrying a TFBS but at low frequency per sequence (Fisher score high rank, Z-score low rank, or iii) due to both types of representation being elevated (Z-score and Fisher score both rank high). Generally, for sequence-based analyses on ChIP-Seq data we find that ordering by either score presents similar ranks. To measure the agreement between the Z- and Fisher scores, we took the top 10 motifs ranked by each of the two types of scores and counted the fraction of TF motifs that were common to both lists. Using results from SSA, we calculated 80% agreement for either the cMyc or FoxA2 datasets, and 90% agreement for the Sox2 dataset. The Chip-Seq target TF ranked in the top four enriched motifs for both measures. For the most stringent results, users may use both over-representation scores in combination when selecting motifs for further analysis (i.e. motifs scoring above both score thresholds).
Figure 3-12. Fisher scores vs. Z-scores from oPOSSUM analysis on sequence-based data.

The Fisher and Z-scores in each panel represent the respective enrichment statistic for 116 motifs in selected ChIP-Seq regions per TF of interest, (A) Nrf2 (1256 regions), (B) FoxA2 (1000 regions), (C) cMyc (1000 regions), and (D) Sox2 (1200 regions). The dotted lines (Z-score=50 and Fisher score=35) indicate empirically selected general threshold values for datasets with ~1200 foreground sequences, thus TF profiles in the upper-right quadrant are those favoured by both scoring measures. Oct4 and Sox2 share similar binding specificities and are known to interact.
3.4.7 Assessing Impact of Data Set Size on ChIP-Seq Based Analyses

For some foreground sets, extreme composition properties can make it difficult to obtain natural background sets of comparable size. To determine the impact of background set size on results, we assessed the impact of background to foreground ratios on results. Datasets from ChIP-Seq analysis of the Sox2 and cMyc TFs were used for the assessment. To establish a baseline for the variation between rankings, two different background sets of the same size were randomly selected from the associated control sequence pools. SSA analysis was performed for each. The Spearman correlation between motif ranks for all lists of length M (M = [4, 116]) was >0.83 for Sox2 and >0.80 for cMyc. Thus the selection of background sequences has only a subtle influence on rankings. We then generated SSA scores using 2-, 3-, and 4-fold background to foreground ratios. The Spearman correlation between the ranked scores, regardless of both M and the ratio, was at least 0.80. We conclude that using background data sets larger than the foreground is not critical as it does not strongly influence the rankings or the top motifs.
3.5 Discussion

The oPOSSUM-3 system incorporates new classes of data, providing researchers interested in the study of the gene regulatory network with a range of options to identify TFs that may be acting in a biological context. Several key advances in the implementation of oPOSSUM-3 make the system suitable for analysis of emerging high-throughput data. Using phastCons multi-species conservation scores obviates the restrictive pairwise sequence alignment phylogenetic footprinting procedures of past releases. This change enables the construction of oPOSSUM services for any species represented in the Ensembl genome annotation database and the UCSC Genome Browser. The new anchored approach to combination site analysis (aCSA) allows for computationally tractable identification of interacting TFs. Motivated by the highly similar profiles for structurally related TFs emerging from large-scale studies, the clustering of highly similar binding site profiles allows a more focused report to users. Most importantly, the new sequence-based methods enable the oPOSSUM system to function as an analysis tool for large-scale ChIP-Seq sequence sets. Taken as a whole, oPOSSUM-3 is a powerful tool for biologists seeking insight into gene regulatory networks.

While running analyses on oPOSSUM-3 is relatively straightforward, interpretation of results requires some consideration, as expected for any over-representation analysis. First, as shown in the analysis of the effects of background GC-content, one should be aware of nucleotide composition differences between the foreground target set and the background set as such differences can bias the TFBS profile enrichment scores. oPOSSUM-3 reports the GC-content of the foreground and background sets used in each analysis to allow users to make such assessments. A second consideration is the selection of the enrichment scoring method. While in some cases the profile of a mediating TF scores high on both Z-score and Fisher score rankings, empirical observations indicate that the two metrics can differ substantially. The differences reflect properties of the frequency of predicted TFBS. Fisher scores indicate the
number of sequences or genes containing a predicted TFBS, while Z-scores reflect the frequency of the predicted TFBSs.

There are several opportunities to enhance oPOSSUM-3 in the future. The oPOSSUM system should be linked to a curated database of TFs such as DBD or TFCat (Wilson et al. 2008; Fulton et al. 2009). By linking the systems, we can indicate which genes in the foreground target set are encoding TFs, and therefore which genes may be candidates for participation in regulatory feedback loops. Based on similar considerations, the regulatory ncRNA genes should be reported. The visualization can be improved, possibly to allow users the option to generate graphical plots of Z-scores vs. Fisher scores, to complement the tabular reports. Finally, the system can be modified to incorporate automated methods to correct for biases in the gene or sequence foreground sets that are not represented in the default background set.

Ultimately, we expect the oPOSSUM-3 system to be a convenient and useful tool for transcriptional regulation analysis, providing researchers with insight into the transcription factors acting on their sets of genes or sequences. The user-friendly interface provides researchers with access to a powerful bioinformatics tool.
4 Leveraging Evolution for the Identification of Regulatory Programs Directing Gene Transcription

4.1 Introduction

Across species from different taxonomic groups, conserved development patterns are found despite divergent evolutionary history. Both tissues and biochemical structures can remain strikingly similar despite millions of years of divergence. In metazoans, one can take the example of the gastrulation process that creates the three primary tissue layers of ectoderm, mesoderm and endoderm. The mesoderm develops into organ structures such as muscle, blood and connective tissue, while ectoderm differentiates into nervous system and epidermis. The structures and functions of orthologous tissues depend on core sets of genes with critical tissue-specific functions. Across species, the gene sets include some conserved members and some branch-specific members. The transcriptional mechanisms governing development and differentiation of tissues are often maintained, despite divergence of the memberships of the gene batteries. In light of the conservation of regulatory controls, it should be feasible to infer the regulatory mechanisms governing critical programs of human gene expression when the regulatory mechanisms are known in a model organism, even over long evolutionary distances.

4.1.1 Muscle Regulatory Mechanisms

A good example of a conserved tissue is muscle. Common components of muscle can be found from jellyfish to humans, including contractile machineries, neuromuscular junctions and attachment structures (Castanon and Baylies 2002; Seipel and Schmid 2005; Hooper and Thuma 2005). In vertebrates, muscle can be divided into three distinct classes: striated (skeletal), cardiac and smooth. In simpler invertebrates, such distinctions are less pronounced and important differences exist in the biomolecular structures. For example, while striated muscle cells are multinucleated in vertebrates
and insects, they are mononucleated in nematodes and jellyfish. Despite the differences, many regulatory mechanisms governing muscle gene expression are retained. The following are the summaries of key mediating factors for myogenesis in vertebrates and invertebrates, followed by a summarization of differences in the muscle regulatory mechanisms between the two.

4.1.2 TFs Involved in Vertebrate Muscle Development

The role that myogenic regulatory factor (MRF) and Mef2 families of TFs play in vertebrate skeletal muscle development is well documented (Olson et al. 1995; Rescan 2001; Shen et al. 2003). The MRF family of TFs, including MyoD, Myf4, Myf5, and Myf6, share the basic helix-loop-helix domain (bHLH) and its members bind to E box motifs. Myf5 is active in progenitors to trunk and intercostal muscle, while MyoD acts in progenitors to body wall and limb muscle (Berkes and Tapscott 2005). MyoD and Myf5 act as specification factors and are expressed in early stages of muscle development, whereas Myf4 and Myf6 function as differentiation factors that are expressed in later stages of muscle maturation. MRF TFs will be referred to as the Myf family in the rest of this chapter. Mef2A, which is characterized by its MADS domain, participates in muscle development by potentiating the activities of MRFs. Serum response factor (Srf), another TF containing a MADS domain, is involved in many stages of mesoderm development and is especially critical for skeletal muscle development (Li et al. 2005). Homeodomain TFs known as ‘Six’ family are also critical for muscle development (Kawakami et al. 2000; Berkes and Tapscott 2005). Other TFs known to be involved in muscle development include LIM domain TFs, Pbx-Meis family of TFs, Sp1, Tead and Ap1.
4.1.3 TFs Involved in Invertebrate Muscle Development

Similar roles for some orthologs of the vertebrate TFs have been observed in invertebrate muscle development. *Drosophila* has several well-defined muscle groups that can be categorized into heart, visceral and somatic. Nematodes have body wall muscle, pharyngeal muscle and several minor muscles; body wall muscle is most analogous to vertebrate skeletal muscle (Fukushige *et al.* 2006). *Drosophila* and *C. elegans* have only one member for each of the MRF and Mef2 families. In *Drosophila*, Mef2 is important for the differentiation of heart, visceral, body wall, pharyngeal and alary muscles. *Drosophila* Mef2 protein sequence shares greater than 85% sequence identity with mammalian Mef2, and importantly, the fly TF binds with sequence specificity to the mammalian Mef2 binding site and has the capacity to activate transcription in both *Drosophila* and mammalian cells (Olson *et al.* 1995). The *Drosophila* ortholog of MyoD is *nautilus*, which is involved in muscle differentiation, but evidence suggests it is required for a smaller subset of muscles compared to the role of MRFs in vertebrates (Balagopalan *et al.* 2001). For *Drosophila* heart and visceral muscle development, *tinman* and *bagpipe* are other important regulatory TFs (Jagla *et al.* 2001). *Notch*, *Wg*, and *Dpp* also serve as key TFs in the early stages of fly muscle differentiation (Baylies and Michelson 2001). Hlh-1, the nematode ortholog of MyoD, has been shown to be solely expressed in body wall muscle (Fukushige *et al.* 2006). Ceh-22, a NK-2 homeodomain TF, is involved in pharyngeal muscle expression in nematode (Okkema and Fire 1994; Okkema *et al.* 1997).
4.1.4 Differences between Muscle Regulatory TFs of Vertebrates and Invertebrates

One important difference between the vertebrate and invertebrate muscle development is the role of Twist, a bHLH TF that has important regulatory roles in mesoderm development (Baylies and Michelson 2001; Te and Reggiani 2002; Castanon and Baylies 2002). For vertebrates, Twist performs a negative role, acting as a repressor of myogenesis, along with Id, MyoR and Mist-1. Twist expression is no longer detected after the specification of mesoderm. However, in Drosophila, twist is required for correct body wall muscle patterning. Twist is highly expressed in muscle progenitor cells, promoting myogenesis, along with combinations of Kr, eve and slouch. The nematode twist ortholog hlh-8 shares 63 % sequence identity with Drosophila twist, and it has been shown to be required for the development of non-striated muscle such as sex muscle and defecation muscle but not for striated muscle. Another difference exists in nematodes: the ortholog of Mef2 does not have discernible effect on muscle differentiation, even though it shares 94% amino acid sequence identity across the MADS domain with human Mef2A (Dichoso et al. 2000).

4.1.5 Ciliary Gene Regulation

Conservation of regulatory programs has been observed for additional cellular structures. Avidor-Reiss et al. identified 187 ciliary genes found from nematodes to vertebrates (Avidor-Reiss et al. 2004). Li et al. found 688 genes involved in the flagellar and basal body proteome (FABP) that are conserved between Chlamydomonas and humans; similar numbers of conserved genes were reported with flies, worms and mice (Li et al. 2004). Li et al. found the X-box motif to be enriched within the upstream regions of nematode ciliary genes, indicating a strong conservation for the ciliary specification by RFX TFs. The RFX family of TFs that bind to the highly conserved X-box motifs are key regulators of ciliary gene expression in metazoans (Chu et al. 2010).
4.1.6 Motif Enrichment Analysis

Computational approaches to the study of cis-acting sequences have become powerful tools for the inference of regulatory mechanisms governing gene co-expression. As introduced in preceding chapters, the oPOSSUM-3 system, which incorporates collections of TF binding site position-specific scoring matrix models, identifies those TF motifs which are enriched in a set of co-expressed genes relative to a control gene set.

4.1.7 Regulog Analysis

Key regulatory controls behind major tissue and cellular structures are conserved, in some cases, among species from diverse taxonomic groups. Computational identification of mediating regulatory mechanisms from comparative analysis should be successful in such instances. Previous studies using bacterial genome comparisons have demonstrated one approach, Regulog analysis, to be successful. In such an approach, a group of genes co-expressed in one organism can be analyzed for the presence of a common regulatory motif and then orthologous genes in another organism are predicted to remain under the same regulatory program based on the retention of the regulatory sequences. The work in bacteria benefitted from the availability of dozens of genome sequences and the compact promoter regions subject to sequence analysis. Because of the considerable evolutionary distance between the well-characterized model organisms (e.g. flies or worms) and humans, TFBS analysis in the same manner is not feasible. Comparative studies within the nematodes have emerged for the targeted study of conserved regulatory programs. GuhaThakurta et al. implemented a regulog-like approach to identify muscle regulatory motifs, while Li et al. used comparative genome analysis to study FABP (GuhaThakurta et al. 2004). Only now is data describing the relationship between TFs and regulatory target genes becoming sufficiently deep to develop general approaches to the study of regulatory conservation across widely diverged branches of the tree of life.
4.1.8 Motivation and Experimental Approach

TFBS over-representation analysis can be performed to identify key TF mediators of regulation given a list of co-expressed genes. The oPOSSUM-3 system for TFBS over-representation analysis was developed to take advantage of the recent advances in the genomics field, including multiple alignment scores for a variety of species from different taxonomic groups, availability of high-throughput sequencing data, and increased annotation of TFBS profiles. As oPOSSUM-3 service has been implemented for human, drosophila and nematode, we can determine the over-represented motifs for genes enriched in a given tissue for each species. By comparing across species, we can assess the extent of regulatory program conservation among the species compared (Figure 4-1).

In this chapter, I perform the comparative analysis of regulatory programs among vertebrates, insects and nematodes, and examine the extent of conservation among major programs, including the development regulation of muscle, cilia genes, and Nfe2l2-regulated genes. Ultimately the analysis demonstrates a capacity for conserved regulatory program analysis to identify the mediating TFs acting across widely diverged species.
Figure 4-1. Overview of the conserved regulatory program analysis.
4.2 Methods

4.2.1 Input Gene Collections

Each gene collection was collected from the research literature, either from individual gene studies or from larger-scale experiments to define whole sets. To construct orthologous gene sets across humans, flies and worms, the sets of genes orthologous to the originating set identified using Ensembl Compara (Vilella et al. 2009). The gene lists are included in Supplementary Materials available on the accompanying CD.

4.2.1.1 Cilia Genes

Cilia genes were retrieved from Ciliome DB (Inglis et al. 2006). From the database, combined data from Avidor-Reiss, Li, Ostrowski, Pazour, Smith and Stolc was downloaded (Ostrowski et al. 2002; Avidor-Reiss et al. 2004; Li et al. 2004; Smith et al. 2005; Stolc et al. 2005; Pazour et al. 2005). The orthologs annotated in the database were used to map among humans, flies and nematodes. From the combined list, 243 human, 204 fly and 208 nematode genes were retrieved. These gene lists are labeled as ‘Ciliome (species).’ The set of 164 human cilia genes from Geremek et al. appear with label ‘Geremek-cilia (species)’ (Geremek et al. 2010).

4.2.1.2 Nfe2l2/Nrf2

The Nfe2l2 transcription factor mediates cellular responses to oxidative conditions in vertebrates. Malhotra et al. published a ChIP-seq study of Nfe2l2 binding sites (also known as Nrf2) in mouse and a corresponding set of genes displaying expression induced by this TF (Malhotra et al. 2010). The modENCODE data archive includes ChIP-Seq identified the binding sites of cnc (fly) and skn-1 (nematode), which are orthologs of Nfe2l2 (Gerstein et al. 2010; Roy et al. 2010; Niu et al. 2011). The most proximal genes were listed for each bound region.
Table 4-1. Target genes of Nfe2l2 and its orthologs cnc and skn-1.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Label</th>
<th>Source/Author/Accession Numbers</th>
<th>Method</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouse Nrf2 induced genes</td>
<td>Nrf2-induced</td>
<td>Malhotra et al.</td>
<td>ChIP-seq and microarray</td>
<td>645</td>
</tr>
<tr>
<td>Mouse Nrf2 integrated genes</td>
<td>Nrf2-integrated</td>
<td>Malhotra et al.</td>
<td>ChIP-seq and microarray</td>
<td>244</td>
</tr>
<tr>
<td>Fly cnc target genes</td>
<td>cnc</td>
<td>modENCODE Gerstein et al.</td>
<td>ChIP-seq</td>
<td>457</td>
</tr>
<tr>
<td>Nematode skn-1 target genes</td>
<td>skn-1</td>
<td>modENCODE Niu et al.</td>
<td>ChIP-seq</td>
<td>2050</td>
</tr>
</tbody>
</table>

4.2.1.3 Muscle Gene Sets

For GNF muscle expression sets (Su et al, 2004), the available microarray data sets were grouped into target and background sets. Four muscle expression gene sets were prepared: skeletal, total smooth, cardiac and general/combined. For skeletal and heart muscle, the correspondingly labelled gene expression data were used. For total smooth muscle, tongue and uterus expression data were grouped with smooth muscle data, as these sources consist predominantly of smooth muscle. For the general/combined set, all muscle types were combined. The background set consists of most tissues, excluding tissues with sizeable smooth muscle content, adipocytes and nervous system. The latter two tissues were excluded because they are known to share overlapping gene expression profiles with muscle. The values were normalized using the MAS5 algorithm, and SAM analysis was used to identify the genes that were differentially expressed in the target set compared to the background set (FDR = 0.05) (Lim, Wang, Lefebvre, & Califano, 2007; Tusher, Tibshirani, & Chu, 2001). From the ranked list of genes, the top 300 genes were selected for each type.
Table 4-2. Muscle gene sets analyzed.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Label</th>
<th>Source</th>
<th>Method</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human GNF muscle expression: skeletal, cardiac, smooth and general/combined</td>
<td>GNF-skeletal, GNF-heart GNF-smooth, GNF-combined</td>
<td>Su et al. (GEO ID: GDS596)</td>
<td>Affymetrix array</td>
<td>300 (each)</td>
</tr>
<tr>
<td>Nematode muscle</td>
<td>Meissner/Stormo muscle</td>
<td>Meissner et al. GuhaThakurta et al.</td>
<td>RNAi, SAGE, Literature</td>
<td>183</td>
</tr>
<tr>
<td>Nematode muscle-enriched</td>
<td>Meissner-enriched</td>
<td>Meissner et al.</td>
<td>SAGE</td>
<td>548</td>
</tr>
<tr>
<td>Fly Mef2 ChIP-chip, all stages</td>
<td>Mef2-all</td>
<td>Sandmann et al.</td>
<td>ChIP-chip, Microarray</td>
<td>211</td>
</tr>
<tr>
<td>Fly Mef2 ChIP-chip, late stage</td>
<td>Mef2-late</td>
<td>Sandmann et al.</td>
<td>ChIP-chip, Microarray</td>
<td>211</td>
</tr>
<tr>
<td>Fly Biniou ChIP-chip, all stages</td>
<td>Biniou-all</td>
<td>Jacobsen et al.</td>
<td>ChIP-chip, Microarray</td>
<td>146</td>
</tr>
<tr>
<td>Fly Biniou ChIP-chip, late stage</td>
<td>Biniou-late</td>
<td>Jacobsen et al.</td>
<td>ChIP-chip, Microarray</td>
<td>52</td>
</tr>
<tr>
<td>Fly larval body wall muscle</td>
<td>Schnorrer-Larval</td>
<td>Schnorrer et al.</td>
<td>RNAi</td>
<td>190</td>
</tr>
<tr>
<td>Fly adult IFM</td>
<td>Schnorrer-Adult</td>
<td>Schnorrer et al.</td>
<td>RNAi</td>
<td>202</td>
</tr>
<tr>
<td>Human muscle validated</td>
<td>Muscle-validated</td>
<td>Kwon et al. (Chapter 2)</td>
<td>Literature, reporter assays</td>
<td>43</td>
</tr>
</tbody>
</table>

GuhaThakurta et al. and Meissner et al. collected separate lists of known muscle-enriched genes in nematodes (GuhaThakurta et al. 2004; Meissner et al. 2009). Meissner et al. performed Long SAGE analysis in nematodes, comparing a prepared muscle SAGE library against intestinal and pan neuronal libraries, or alternatively against a whole-embryo library. Meissner et al. also performed an RNAi experiment of the nematode muscle transcriptome, revealing genes with a role in embryonic muscle formation. From these studies, the following gene lists were prepared:

- Combined list of Meissner RNAi-identified genes, Meissner collection of known genes and GuhaThakurta collection of known genes (Meissner/Stormo muscle)
• Genes enriched (3-fold) in muscle compared to the whole embryo, as identified by Meissner et al. through their SAGE data (Meissner-enriched)

Fly Mef2 and Biniou gene sets are based on the ChIP-chip analysis of TF binding during fly embryogenesis (Sandmann et al. 2006; Jakobsen et al. 2007). The authors identified three temporal patterns of TF binding: early, continuous, and late. Genes from all stages combined (Mef2/Biniou-all) and late stages (Mef2/Biniou-late) are analyzed in this study. Schnorrer et al. performed RNAi screens of larval body wall muscle and adult indirect flight muscle in fly (Schnorrer et al. 2010).

In the rest of this chapter, the gene sets will be referred by the labels given in Table 4-2. The species will be given in brackets after the gene list label (e.g. GNF-skeletal (human)).

4.2.2 List of Known TFs Involved in Each Regulatory Program

From literature, a list of TFs with known regulatory roles associated with each respective gene collection was prepared.

The muscle TFs were classified according to the muscle subtype they are known to regulate. In addition to the muscle TFs outlined section 4.1.1, a number of cardiac and smooth muscle TFs were recorded. GATA4 has important regulatory roles in cardiac muscle development (Qian et al. 2008). Other TFs with associated cardiac muscle regulatory roles include Nfat family of TFs, Nkx2.5, HIF-1, TEF-1 and ETS-domain TFs (Spangenburg and Booth 2003). Krueppel-like factor 4 (Klf4) participates in multiple tissue differentiation regulation, including cardiac and smooth muscle development (Yoshida et al. 2010).
Binding profiles used in this study that are available in JASPAR 2010 are listed in Table 4-3.

Table 4-3. List of TFs available in JASPAR 2010 used in this study.

<table>
<thead>
<tr>
<th>Type</th>
<th>JASPAR TF Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cilia</td>
<td>daf-19, RFX1_1, RFX1_2, Rfx3_1, Rfx4_1, Rfxdc2_1, Rfx3_2, Rfx4_2,</td>
</tr>
<tr>
<td></td>
<td>Rfxdc2_2</td>
</tr>
<tr>
<td>Nfe2l2/Nrf2</td>
<td>Nfe2l2</td>
</tr>
<tr>
<td>Skeletal Muscle</td>
<td>Sp1, Runx1, Egr1, Cf2-II, Mef2A, Srf, Myf, twi, Pparg, Rxra, Ap1,</td>
</tr>
<tr>
<td></td>
<td>Nkx2-5, Lhx3, Lim1/3, Tead1, Nkx3-2, tin, Sx4/6</td>
</tr>
<tr>
<td>Cardiac Muscle</td>
<td>Smad3, Egr1, Zeb1, Cf2-II, Klf4, Mef2A, Srf, Gata1/3/4/5/6, Rxra,</td>
</tr>
<tr>
<td></td>
<td>twi, Arnt, Hif1A, Ap1, Nkx2-5, Lhx3, Lim1/3, slou, Pax4/5/6, Elk1/4,</td>
</tr>
<tr>
<td></td>
<td>Ets1, Nfatc2</td>
</tr>
<tr>
<td>Smooth/Visceral Muscle</td>
<td>FoxA2, slp1, FoxF2, Egr1, Zeb1, Cf2-II, Klf4, Mef2A, Srf, Gata1/3/4/5/6, Myf, twi, Ap1, Nkx2-5, Nkx3-1, Nkx3-2, Lhx3, Lim1/3, slou, Tead1, tin, Sx4, ceh-22, hap, Pbx, Nfatc2</td>
</tr>
</tbody>
</table>

Note: The Nhhl1 profile is very similar to those of the MRF family of TFs. Likewise, both TBP and Tbp_1 profiles share similarity with a subportion of the binding site profile of MEF2A.

4.2.3 TFBS Over-Representation Analysis Using oPOSSUM-3

The oPOSSUM-3 system identifies transcription factor binding site motifs that are enriched within DNA sequences from a foreground set of genes relative to a control background set. The following oPOSSUM-3 runtime parameters were used. All oPOSSUM analyses were run with the default conservation level of 0.7. Single site analysis (SSA) and TFBS cluster analysis (TCA) were performed iteratively for each of six standard search regions (Table 3-1). In this chapter, all results are from the oPOSSUM-3 TCA option unless otherwise stated. The reported TFBSs are based on TFBS clusters used by oPOSSUM. The JASPAR collections of binding site profiles were used, including CORE, PBM and PENDING (custom in-house collection). For profiles in the PENDING collection, refer to (Figure 3-2).
oPOSSUM analyses employ two sets of motif enrichment measures: Z-scores and Fisher scores. These scores are not normalized, and it is often the case that a TF of interest may score highly in one measure but not in the other. In order to compare the enrichment of a particular TF across different gene sets, it is necessary to combine both scoring measures for overall enrichment assessment. To this end, we employ the following normalization and score combining procedure.

1. Normalize the scores by:

\[ S_{\text{normalized}} = \frac{S_i - S_{\text{min}}}{S_{\text{max}} - S_{\text{min}}} \]

where \( S_i \) is the score to be normalized, \( S_{\text{max}} \) is the maximum score, and \( S_{\text{min}} \) is the minimum score.

2. Calculate the median normalized scores for both scoring measures.

3. For each TF, calculate the Euclidean distance between the median point from step 2 and the normalized Z and Fisher scores, as long as at least one of the two scores are greater than the median. Thus, the combined score represents the separation of the profile in question from the median score obtained by the other profiles. The score will be referred to as the distance to median.

4.2.4 Analysis Tools

All statistical calculations were performed in R (R Development Core Team 2008). Gene ontology analysis was performed with DAVID (Huang et al. 2009). Violin plots were generated for specific sets of TFs of interest using the combined scores obtained across all search region settings using the vioplot package in R. Ensembl v61 orthology annotation was used for identification of orthologs. MEME-ChIP version 4.6.0 was used for motif discovery using ChIP-seq peak sequences, with ‘zero or one per sequence’ setting, minimum width of 6, and maximum width of 15 bp (Bailey et al. 2006).
4.3 Results

Motif over-representation analysis using oPOSSUM-3 was performed on orthologous gene sets listed in section 4.2.1, including gene sets linked to muscle, cilia and detoxification genes. Due to the depth of data and existing knowledge, the following survey of regulatory program conservation focuses on muscle gene expression across worms, flies and humans. The motif enrichment scores, measured using both Z- and Fisher-procedures, were calculated across six search region parameter settings. An example of an oPOSSUM-3 analysis is given in Figure 4-2. The combined set of muscle reference regions and the newly validated regions from Chapter 2 was analyzed using the default oPOSSUM analysis parameters, and the results are illustrated as a Fisher vs. Z score plot. Recalling the previous chapter, Z-scores and Fisher scores represent different methods of identifying over-represented TFs. Z-scores reflect the relative rate of TFBS occurrence in the given gene set compared to that of the background gene set. Fisher scores compare the fraction of genes containing a given TFBS between the test gene set and the background gene set. As explained in Methods, the results are based on oPOSSUM TCA, and the member TFBS names are used as shorthand for the reported TFBS clusters in the results.
Figure 4-2. Fisher vs. Z score plot for oPOSSUM TCA results on human muscle reference and validated regions from Chapter 2.

Analysis was performed at the default search region level of 3. Dark blue labels denote TFs associated with skeletal muscle development, and light blue labels denote TFs associated with other types of muscle development. Grey dotted lines demarcate the top 25 ranked TFs. Red labels denote TFs that scored above the threshold for one or both measures, but have no previously known muscle association.

In order to facilitate the comparison of results across different gene sets and species, the scores were normalized and combined using the procedure outlined in section 4.2.3. Violin plots were generated for specific sets of TFs of interest using the combined scores obtained across all search region settings. For complete sets of Fisher vs. Z score plots and comparative violin plots, refer to the Supplemental Materials CD.
The following summaries of the results should be considered as the reader goes through the rest of this chapter:

• For cilia genes, RFX motifs were most enriched in gene collections for all three species.
• For Nfe2l2 target genes, flies diverge from the other species. As motif discovery analysis on the available ChIP-Seq data does not reveal that the fly ortholog of Nfe2l2 has a different binding profile from the vertebrate proteins, further investigation is needed to explain the observations.
• In muscle, the known muscle regulators from the Myf and MADS families were observed for both humans and flies, whereas in nematodes, Mef2 and Srf of the MADS family were absent, in agreement with the literature. In flies, Cf2-II was consistently enriched, while in nematodes, Egr1, Sp1, Smad3, Klf4, Trl, Ebf1 and Zfx motifs were among the top-ranked motifs.

4.3.1 Cilia Genes

To assess the conservation of regulatory programs outside of muscle, we identified an additional collection of genes incorporating data from distant organisms. Ciliary genes are subject to control by Rfx transcription factors. Leroux et al. compiled known cilia genes from various metazoan species in a central repository named Ciliome DB. Three Rfx binding site profiles were obtained: (i) the nematode daf-19 profile (Li et al. 2004); (ii) vertebrate RFX1 profiles (Emery et al. 1996); and (iii) vertebrate RFX binding profiles from PBM studies (Badis et al. 2009). Cilia genes retrieved from Ciliome DB for each of the three species were analyzed with oPOSSUM. Another set of human cilia genes from Geremek et al. were analyzed as well. The summary violin plots are given in Figure 4-3, showing the distance to median distribution for the Rfx profiles in respective ortholog sets of each species and comparing them against the scores for randomly selected genes. For nematode and fly cilia genes, the Rfx profile distances are clearly separated from those for random genes. However, for the human gene set, the distance ranges of Rfx profiles are indistinguishable from the random gene set results.
(Figure 4-3a). As the gene lists for each species were created by merging multiple experimental data sets within the Ciliome DB, it is possible that this may reflect some aspect of the merging process, resulting in numerous paralogs under different regulatory programs being included in the human ortholog set. For the human Geremek gene set, Rfx profiles were better separated from other profiles than in the random gene set, although the separation was still lower compared to the orthologs for other species. It is possible that the human cilia genes may be subject to more complicated regulatory landscape in which the central importance of Rfx TFs has been diminished compared to other species.

**Figure 4-3. Score distributions for Rfx profiles from oPOSSUM TCA performed on cilia genes collected from the Ciliome DB and Geremek et al.**

a) Ciliome DB genes for human, fly and worm compared against randomly selected genes in each species.
b) Geremek et al. genes and their orthologs compared against randomly selected genes.

4.3.2 Nfe2l2 / Nrf2 Target Genes

A third system related to cellular detoxification was available in the three species. Nfe2l2 is a master regulator of the oxidative stress response. Nfe2l2 induces greater expression of many protective genes including oxidoreductases, glutathione S-transferases, and multidrug resistance-associated proteins. Malhotra et al. performed a genome-wide identification of Nrf2 binding sites using ChIP-Seq, and combined the data with gene expression profiles to specify two sets of potential Nrf2 target genes. The recent release of modENCODE data includes the ChIP-seq peaks for cnc in flies and skn-1 in nematodes, which are the orthologs of Nfe2l2 in the respective species. In addition to the TF-bound coordinates, the authors reported putative target genes based on their proximity to the identified regions.
The oPOSSUM TCA results for these gene lists are shown in Figure 4-4. When compared against results from randomly selected human genes, Nfe2l2 profiles are clearly separated from the median when analyzed on both of the mouse Nfe2l2 target genes, nematode orthologs, and skn-1 ChIP-Seq peak associated nematode genes. oPOSSUM TCA of the fly cnc target genes did not result in such separation for Nfe2l2. Review of the Fisher vs. Z-score plot confirms this observation; Nfe2l2 is not distinguished from other profiles (Figure 4-4b). Application of the oPOSSUM-3 SSA procedure to the same gene set, however, distinguishes the Nfe2l2 profiles (Figure 4-4c). Interestingly, oPOSSUM analysis of the nematode genes associated with skn1 ChIP-Seq peaks exhibited the opposite behaviour, with clear separation in TCA but not in SSA.
Figure 4-4. oPOSSUM TCA results on Nfe2l2, cnc and skn-1 target genes.

(a) Score distributions for Nfe2l2 gene sets. Mouse induced and mouse integrated gene sets are from Malhotra et al, and Fly and Worm gene sets are orthologs of mouse induced gene sets. Fly:cnc and Worm:skn1 are genes associated with ChIP-Seq peaks from modENCODE data.
b) oPOSSUM TCA results for cnc-associated fly genes. Nfe2l2 profiles are not clearly separated from other profiles included in the analysis. (Search region level: 3)

![Graph showing TCA results for cnc-associated fly genes.](image)

Z-score

---

c) oPOSSUM SSA results for cnc-associated fly genes. Nfe2l2 profiles received high Z- and Fisher scores, unlike in TCA results. (Search region level: 3)

![Graph showing SSA results for cnc-associated fly genes.](image)

Z-score
To examine in detail why the differences exist for cnc and skn-1 target genes, we performed motif discovery on cnc and skn-1 ChIP-seq peak sequences using MEME-ChIP (Figure 4-5). While the fly cnc motif resembles the vertebrate Nfe2l2 profile closely, none of the top three candidate motifs derived de novo resemble the Nfe2l2 profile.

Figure 4-5. Logos of Nfe2l2 orthologs in flies and nematodes

a) Fly cnc logo, generated from cnc ChIP-seq sequences using MEME-ChIP. Only the top scoring motif is shown. This motif is the reverse complement of the known mouse Nfe2l2 motif.

b) Nematode skn-1 logos generated from skn-1 ChIP-seq sequences, using MEME-ChIP. The logos are listed in the order of MEME scores. None of the profiles share similarity to the known mouse Nfe2l2 motif.
Further investigation into the differences among the gene sets from the three species was performed by looking into the gene ontology (GO) annotated properties of the genes. Table 4-4 lists the top 3 annotation clusters retrieved by DAVID GO-enrichment analysis for each species. While the Nfe2l2 is known to be involved in each of the functions listed, there is a clear emphasis in nematodes on early stages of development.

Table 4-4. Gene ontology analysis of target genes for Nfe2l2 and its orthologs using DAVID.

The top three annotation clusters are reported. For each cluster, select gene ontology terms that capture the main functions are reported to clarify presentation.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Mouse (Nrf2)</th>
<th>Fruit fly (cnc)</th>
<th>Nematode (skn-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Score</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.72</td>
<td>12.05</td>
<td>35.10</td>
</tr>
<tr>
<td>GO terms</td>
<td>Glutathione metabolism/transferase</td>
<td>Proteasome complex, peptidase activity</td>
<td>Embryonic development, larval development</td>
</tr>
<tr>
<td>2</td>
<td>Score</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.82</td>
<td>4.42</td>
<td>15.65</td>
</tr>
<tr>
<td>GO terms</td>
<td>Oxidoreductase</td>
<td>Proteasome regulatory particle, proteasome component region</td>
<td>Genitalia development, reproductive developmental process</td>
</tr>
<tr>
<td>3</td>
<td>Score</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.12</td>
<td>2.95</td>
<td>13.81</td>
</tr>
<tr>
<td>GO terms</td>
<td>Regulation of cell/nervous system/neurogenesis development and cell differentiation</td>
<td>Proteasome regulatory particle/subunit, ATPase</td>
<td>Non-membrane-bounded organelle, ribonucleoprotein</td>
</tr>
</tbody>
</table>
4.3.3 Muscle Genes

The analysis of regulatory program conservation for muscle gene sets is complicated by the challenge of defining the relationships between species-specific sub-types of muscle tissues. For instance, the nematode produces muscle fibres, but they differ in some characteristics from both human cardiac and skeletal muscle.

4.3.3.1 Observations on Enriched Motifs

In the following sections, I will provide qualitative observations regarding the distribution of the muscle regulatory TFs based on the oPOSSUM Fisher vs. Z-score plots, followed by discussion on the over-representation of specific profiles of interest when compared across all muscle data sets. The Fisher vs. Z-score plots used for the muscle gene sets are included in the supplemental data.

4.3.3.1.1 Human GNF Gene Sets

The first gene sets analyzed were the human muscle-enriched genes from GNF expression profiling studies (GNF-combined, GNF-skeletal, GNF-heart, GNF-smooth). Key regulatory TFs involved in regulation of vertebrate muscle differentiation are well documented, and were therefore assessed oPOSSUM reports enrichment for the TFs of the Myf family, and the MADS box TFs Mef2A and Srf.

1. For GNF-skeletal, the contribution of Mef2A is emphasized, with its score clearly separated from other TFs in both the Z and Fisher scores. Tbp is highly enriched, along with Srf and Nkx2-5/Lim-domain TFs.

2. For GNF-heart muscle, one observes a distinct set of high scoring TFs. Profiles for Klf4 score highly. The Myf family and Srf TFs again score highly, along with Sp1, Pparg, Rxra, Zeb1 and Egr1. Egr1 is important in vascular smooth muscle proliferation in vertebrates (Wada et al. 2003) and is linked to contractile machinery regulation in skeletal muscle, with increased Egr1 expression during skeletal muscle development (Irrcher and Hood 2004). Zfx, Nhlih1, Ebf1 and Klf7...
receive high scores as well (marked green), although no direct annotation of muscle regulation was reported for these TFs, Nhlh1 and Ebf1 both belong to the same family (bHLH) as Myf. The high scores received by Egr1, Zfx, Klf4, Klf7 and EBF are noteworthy, as they score highly in nematode muscle gene sets (section 1.3.1.2).

3. For smooth muscle, Mef2A is not strongly enriched. High Z-scores for Srf are reported, as well as Nfatc2, Tead1, Gata1, Nkx3-1, Elk1/4 and Tbp TFs. When ranked by Fisher scores, Klf4, Smad3 and Myf are found to be enriched. As summarized in the introduction, these TFs are known to regulate smooth muscle differentiation. oPOSSUM-3 is able to distinguish key mediating TFs for each subtype of muscle tissue.

Analysis of each of the GNF muscle gene sets with the oPOSSUM system identifies enrichment for motifs associated with key TFs in the corresponding tissue.

4.3.3.1.2 Meissner Nematode Gene Sets

To assess whether regulatory programs are conserved over long periods of evolution, a set of genes defined as co-expressed in worm muscle was identified from the literature. Two different gene sets were obtained from Meissner et al. The first is the Meissner/Stormo muscle set, the union of the RNAi-identified and previously annotated genes collected by Meissner et al., combined with the set of 41 annotated muscle genes from GuhaThakurta et al. The second is the Meissner-enriched gene set, which consists of genes enriched in nematode muscle compared to whole embryos using data derived by SAGE analysis. oPOSSUM TCA results on these gene sets reveal the following key observations:
1. In the Meissner/Stormo muscle (nematode) set analysis, the top enriched motifs include Klf4, Egr1, Sp1, Smad3 and Myf. Non-muscle (in terms of existing annotation) TFs include Tfap2A, Ebf1 and Trl. As stated earlier, many of the top enriched motifs from this gene set were also enriched in the GNF-heart (human) genes. The Tfap2A profile is unusual, almost entirely scoring for the presence of GCC trinucleotides. The prevalence of Tfap2A motifs might be an artefact particular to the nematode genome, which has lower GC content compared to human and fly genomes. In low GC regions, any short GC-rich motifs will be overemphasized. If a given set of sequences have over-represented motifs which include GC-rich motifs, Tfap2A is likely to receive high scores as well as it will predict binding sites overlapping such motifs. Trl is a GAGA factor, known as an activator of gene expression, is involved in chromatin remodelling and insulator activity (Ohtsuki and Levine 1998; Zhan et al. 2001; West et al. 2002; Shimojima et al. 2003). As the ensuing results demonstrate, this motif is enriched for most of the nematode muscle gene sets analyzed in this study. It may perform an important role selectively for nematode muscle.

2. In fly orthologs of Meissner/Stormo genes, Mef2A, Gata factors, Six6, Pax4 and Srf are among the top enriched motifs. Cf2_II rank highly in Z-scores, and this enrichment is often observed for other fly muscle gene sets tested in this study, suggesting that this TF may have a central role in fly muscle regulation. Atf-1 consistently ranked high in fly muscle gene sets. This protein belongs in the same class as Jun, which has known muscle regulatory roles (Blais et al. 2005). In the human orthologs, Srf, Pparg::Rxra, Gata factors, Myf and Tead1 are among the top-ranked by Z-scores. When ranked by Fisher scores, Klf4, Arnt::Ahr, Nfatc2, Myf and Egr1 are the most enriched motifs, along with non-muscle factors Ebf1 and Zfx. The human and fly results demonstrate that co-expressed genes from worms can be used to infer regulatory factors acting on orthologous genes.

3. In the Meissner-Enriched (nematode) set, the top enriched motifs include Myf6, Zeb1, Sp1, Trl, Egr1, Myf and Tfap2A when ranked by Z-scores. Many other known muscle TFs including Arnt::Ahr, Six4, Tead1, Ap1, Elk1/4, FoxF2, Pbx1
and Nkx3-1 rank highly in Fisher scores. In the human ortholog set, Mef2A is highly enriched, along with Nkx2-5, FoxF2 and Tbp. If ranked by Fisher, factors such as Myf6, Zfx, Ebf1, Elk1/4 and Nfatc2 are among the top motifs. The Sox family of TFs was ranked highly among non-muscle factors. For the fly ortholog set, Six6 is the top enriched muscle TF. While Srf and Mef2A are among the top ranked, their separation from the rest of the motifs is not as noticeable as in other analysis results. There is a clear shift in the enrichment of Six6 compared to the human orthologs, suggesting that this may be a fly-specific feature of the regulatory controls for this gene set. Also, forkhead TFs receive high rankings, in agreement with the observations made for nematodes and human ortholog sets.

**4.3.3.1.3 Schnorrer Fly Gene Sets**

Schnorrer et al. performed RNAi screens for muscle-related genes in flies, focusing on those that influence development of larval body wall muscle and adult indirect flight muscle (IFM). Adult IFM is thought to be most similar to vertebrate skeletal muscle. From the oPOSSUM TCA results, the following observations can be made:

1. **Schnorrer-Larval (fly) was unusual in that the Fisher scores obtained were lower than other gene set analyses. This difference suggests that only a small subset of the genes contained common motifs. In the Schnorrer-Larval (fly) set analysis, the top enriched motifs include Cf2, Gata factors, twist, Six6, Smad3, Srf, Egr1 and Mef2. Enriched TF motifs with no previous muscle annotation Irf, Arid5a, Jundm2 and P53. In the human orthologs, Srf, Pax4, Nkx, Myf and HNF4A motifs are enriched, while Osr, Plagl1, CTCF and Gabpa are the most enriched non-muscle TFs. Interestingly, the Fisher scores obtained from this analysis are lower than those observed from other analysis results. In the nematode orthologs, Smad3, Myf, Egr1, Elk, Nkx and FoxF2 motifs are enriched, and E2F, CTCF and che-1 are included in the enriched non-muscle motifs.**

2. **For the Schnorrer-Adult (fly) set, Cf2-II, FoxF2, Nkx factors, Klf4, Pax4 and PBX1 motifs are enriched, along with non-muscle HMG, Mafk and br TF motifs. In the human orthologs, Egr1, HNF4A, Pax4, Klf4, Srf and Smad3 were among the most**
enriched motifs. This differs from the distribution of the scores observed for GNF-skeletal (human). There are many non-muscle motifs that are enriched in the human orthologs, including E2F, Gmeb1 and Plag1. For Schnorrer-adult (nematode), no clear separation could be observed among the muscle TFs when ranked by Z-scores, although Smad3 and Myf were included in the top rankings. Runx1, Nkx, FoxF2, TEAD and Smad3 are some of the enriched muscle motifs. For the non-muscle motifs, Ewsr1-Fli1 is the most enriched, followed by others including E2F, Deaf1, HMG, CTCF and lin-14.

4.3.3.2 Comparison of Muscle Gene Sets and Corresponding Orthologs

As the complexity of muscle regulation is compounded by the presence of multiple tissue types, growth stages and species differences, we decided to focus on several key muscle TFs and compare their enrichment across multiple muscle gene sets and species. For this purpose, we generated violin plots based on the combined scores (distances to the median combined score, as explained in the Methods) for each of the key TFs including Mef2A, Myf, Srf, FoxF2 and Egr1. The scores for cilia, Nfe2l2/Nrf2 gene sets and random gene sets are included in each of the violin plots as non-muscle gene control set.
4.3.3.2.1 Mef2

Mef2 TFs are known for their central role in muscle development regulation in vertebrates and insects, but not in nematodes. In agreement with the literature, no meaningful combined scores were given to the Mef2A motif in Meissner nematode gene sets and nematode orthologs of other gene sets (Figure 4-6a,d). In human GNF gene sets, Mef2A is especially prominent in the skeletal muscle genes, while in smooth muscle genes, the score obtained by Mef2A is indistinguishable from the randomly selected genes. Mef2A is enriched in both of the Schnorrer fly genes, in agreement with the known importance of Mef2 in insect muscle development. Mef2A motif is also enriched in Nfe2l2/Nrf2 genes. Mef2A is also highly enriched in Meissner-enriched (human) and elevated in the fly ortholog sets (Figure 4-6b,c). However, Mef2A motif is enriched in the randomly selected fly genes as well; it will be necessary to further investigate whether this is a general trend in fly genes or just an artefact of the random selection process before we can make conclusive statements.
Figure 4-6. Enrichment of Mef2A motif in each gene set, as measured by the distance between the combined score of Mef2A motif and the median score of all the other motifs.

For control, results for cilia genes, Nfe2l2/Nrf2 genes, and randomly selected genes are also included.

a) Mef2A profile enrichment in each of the gene sets, in the original species where the genes were identified. Mef2A is strongly enriched in GNF-skeletal, Schnorrer fly genes, and Nfe2l2/Nrf2 mouse genes. (GNF genes: human; Meissner genes: nematode; Scnorrer: fly; Cilia: human; NRF2: mouse; Random: human)
b) Mef2A profile enrichment in human ortholog sets. In addition to the GNF-skeletal genes, human orthologs of Meissner-enriched genes also show high Mef2A enrichment.
c) Mef2A profile enrichment in fly ortholog sets.
d) Mef2A profile enrichment in nematode ortholog sets. Other than Nfe2l2/Nrf2 induced gene set orthologs, no Mef2 enrichment is evident.
4.3.3.2.2 Myogenic Regulatory Factors

The MRF family of TFs play central roles in vertebrate muscle development. In invertebrates, their orthologs have varying degrees of importance; the insect MRF ortholog is required for subsets of muscles only, while the nematode ortholog is found in body wall muscle. The analysis results across gene sets from the three species in general support the literature evidence, as MRF enrichment was strong in gene sets from human and nematode (especially GNF-Heart and Meissner-Enriched) but weaker in fly gene sets (Figure 4-7). The human orthologs of the Schnorrer gene sets also exhibited enrichment compared to the results for fly (the origin of the set).
**Figure 4-7.** Enrichment of Myf/Myf₆ motifs in each gene set, as measured by the distance between the combined score of Myf/Myf₆ and the median score of all the other motifs.

For control, results for cilia genes, Nfe2l2/Nrf2 genes, and randomly selected genes are also included.

a) Myf/Myf₆ profile enrichment in each of the gene sets, in the original species where the genes were identified. While Myf/Myf₆ is elevated in the GNF and Meissner gene sets, it is especially enriched in GNF-heart and Meissner-enriched genes. Enrichment is not evident in Schnorrer fly gene sets. (GNF genes: human; Meissner genes: nematode; Schnorrer: fly; Cilia: human; NRF2: mouse; Random: human)
b) Myf/Myf6 profile enrichment in human orthologs. Enrichment is elevated in Meissner-Stormo gene set compared to the results in the original species.
c) Myf/Myf6 profile enrichment in fly orthologs. While the plot for GNF-heart (fly) genes indicate that there is enrichment of Myf/Myf6, the low mean suggests that the analysis results were dependent on the search region distances.
d) Myf/Myf6 profile enrichment in nematode orthologs.
4.3.3.2.3 Serum Response Factor

Srf is another MADS domain class of TF that is important for muscle development in vertebrates and insects, but not in nematodes. In agreement, comparison of oPOSSUM TCA results across the gene sets tested reveals that Srf is strongly enriched in human (GNF-Skeletal and GNF-Smooth) and fly (Schnorrer-Larval) gene sets but not in nematode gene sets (Figure 4-8). Interestingly, Srf was enriched in human and fly ortholog sets of Meissner/Stormo (nematode) genes as well, further corroborating the conservation of regulatory control by Srf in both vertebrates and insects.
Figure 4-8. Enrichment of Srf motifs in each gene set, as measured by the distance between the combined score of Srf and the median score of all the other motifs. For control, results for cilia genes, Nfe2l2/Nrf2 genes, and randomly selected genes are also included.

a) Srf profile enrichment in each of the gene sets, in the original species where the genes were identified. Srf is enriched in the GNF human genes and Schnorrer-Larval fly genes. Srf enrichment is not evident in Meissner nematode genes. (GNF genes: human; Meissner genes: nematode; Scnorrer: fly; Cilia: human; NRF2: mouse; Random: human)
b) Srf profile enrichment in human orthologs. Most human ortholog sets show Srf enrichment. Strong enrichment in Schnorrer-larval genes in both humans and flies suggests conservation of regulatory controls for these genes.
c) Srf profile enrichment in fly orthologs. Srf shows strong enrichment in Meissner/Stormo and Schnorrer-Larval gene sets.
d) Srf profile enrichment in nematode orthologs. While there is some enrichment of Srf in GNF nematode orthologs, the low mean distances suggests that these results should be interpreted with caution.
4.3.3.2.4 Egr1

As stated previously, Egr1 is involved in vascular smooth muscle proliferation and skeletal muscle contractile machinery regulation. As summarized in Figure 4-9, oPOSSUM analyses of the muscle gene sets indicate that Egr1 is over-represented in GNF-Heart (human), GNF-Smooth (human), Meissner/Stormo (nematode) and Meissner-Enriched (nematode), with stronger enrichment observed in GNF-Heart (human) and Meissner/Stormo (nematode). Of the human orthologs of the muscle gene sets from other species, Schnorrer-Adult (human) genes also showed strong Egr1 enrichment, whereas in fly, it was GNF-Smooth that showed increased enrichment. These observations suggest that Egr1 may have significant roles in muscle development for all three species. It will be interesting to investigate the extent of the role that Egr1 has in nematode muscle development, where MADS TFs do not have muscle regulatory functions in contrast to vertebrates and insects.
Figure 4-9. Enrichment of Egr1 motifs in each gene set, as measured by the distance between the combined score of Egr1 and the median score of all the other motifs. For control, results for cilia genes, Nfe2l2/Nrf2 genes, and randomly selected genes are also included.

a) Egr1 profile enrichment in each of the gene sets, in the original species where the genes were identified. Egr1 is strongly enriched in the GNF-Heart (human) genes and Meissner/Stormo (nematode) genes. Egr1 also shows elevated enrichment in GNF-Smooth (human) and Meissner-Enriched (nematode) genes, as well as Nfe2l2l/Nrf2 induced mouse genes. (GNF genes: human; Meissner genes: nematode; Scnorrer: fly; Cilia: human; NRF2: mouse; Random: human)
b) Egr1 profile enrichment in human orthologs. While elevated in most of the muscle orthologs except GNF-Skeletal genes, it is especially enriched in GNF-Heart (human) and Schnorrer-Adult (human).
c) Egr1 profile enrichment in fly orthologs.
d) Egr1 profile enrichment in nematode orthologs. Egr1 shows strong enrichment in Meissner/Stormo genes.

4.3.3.3 Mef2 ChIP-Chip Gene Sets

As explained in the introduction, Mef2 is one of the central regulators of muscle development in vertebrates and insects, but no discernable muscle regulatory role has been observed in nematodes. Sandmann et al. performed large scale ChIP-chip with developing fly embryos targeting Mef2, from which they identified three temporal binding patterns for this TF. Figure 4-10 represents the Mef2A motif over-representation analysis results for the combined and the late stage fly Mef2 target genes in their orthologs. In fly, strong Mef2A enrichment could be observed in both Mef2 target gene sets as expected, while in human, only the Mef2 late stage genes were associated with such enrichment. In support of the literature evidence for the diverged
Mef2 regulatory role in nematodes, no strong enrichment could be observed in the nematode orthologs.

**Figure 4-10. oPOSSUM TCA results on Mef2 target genes.**

These genes were identified through combination of ChIP-chip analysis and expression profiling in fly. Mef2A is not enriched in nematode orthologs of the fly Mef2 target genes.
4.3.3.4 Biniou ChIP-chip Target Gene Sets

Biniou, the *Drosophila* ortholog of FoxF2, is important for the development of fly visceral muscle. Jakobsen *et al.* performed a large scale ChIP-chip study in developing fly embryos targeting Biniou, from which they identified three temporal binding patterns for this TF. Figure 4-11 represents the motif over-representation analysis results for the combined and the late stage Biniou target genes. Interestingly, even though the FoxF2 motif is strongly enriched in Biniou-All (human) and both of the nematode ortholog sets, it did not receive high enrichment scores in fly genes, despite the fact that these genes were originally selected based on Biniou ChIP-chip study in *Drosophila*. Foxa, another forkhead-box TF with a similar binding profile to FoxF2, received better scores than FoxF2. This was unexpected as the logo for Biniou TFBS profile generated by Jakobsen *et al.* is highly similar to that of FoxF2 profile in JASPAR. The Fisher vs. Z-score plot for Biniou-All (fly) genes indicate that it is Foxa that receives high Z-scores instead of FoxF2.
Figure 4-11 oPOSSUM TCA results on Biniou target genes.
These genes were identified by combination of ChIP-chip and expression profiling studies in fly.

a) Enrichment analysis of FoxF2 (biniou ortholog).
b) Fisher vs. Z-score plot of Biniou-All (fly). (Search Region Level: 1)
4.3.3.5 Gene Ontology Analysis

Gene ontology (GO) over-representation analysis using the DAVID system was performed on the tested muscle gene sets for each species in order to assess whether annotated gene functions were conserved. The DAVID result files are available in the accompanying Supplemental Materials. In this section, the GO analysis results for a subset of the gene lists will be discussed.

For the orthologs of nematode Meissner/Stormo-muscle genes, GO terms relating to striated muscle development and structure were the most over-represented terms (Table 4-5). This is consistent with the oPOSSUM analysis results, in which muscle TFs related to striated muscle were the most enriched.

**Table 4-5. Gene ontology analysis on Meissner/Stormo-muscle gene sets.**

<table>
<thead>
<tr>
<th>Human</th>
<th>Fly</th>
<th>Worm</th>
</tr>
</thead>
<tbody>
<tr>
<td>contractile fiber, myofibril, sarcomere, myosin complex</td>
<td>sodium transport, ion transport, transmembrane</td>
<td>actin-binding</td>
</tr>
<tr>
<td>muscle contraction, myosin complex, thick filament</td>
<td>glycolysis, glucose metabolic process, carbohydrate catabolic process</td>
<td>striated muscle development, myofibril assembly, actinomyosin structure reorganization</td>
</tr>
<tr>
<td>muscle organ development, striated muscle development, cardiac muscle development</td>
<td>cytoskeleton, actin</td>
<td>post-embryonic development, larval development</td>
</tr>
<tr>
<td>cell junction, anchoring junction</td>
<td>cytoskeleton, actin filament</td>
<td>cytoskeleton, myosin complex, motor activity, thick filament</td>
</tr>
<tr>
<td>cardiomyopathy, integrain beta subunit, EGF, cardiac muscle contraction, cell adhesion</td>
<td>glycosyl hydrolase</td>
<td>reproductive behaviour</td>
</tr>
</tbody>
</table>

For Schnorrer-larval genes, GO terms for myofibrils and sarcomeres are in the top rankings for all species, as well as the terms for ribonucleoprotein complexes (Table 4-6). For worms, terms related to embryonic and reproductive development are among the top ranked.
Table 4-6. Gene ontology analysis on Schnorrer-larval gene sets.

For worms, GO terms for striated muscle were in 6th place.

<table>
<thead>
<tr>
<th>Human</th>
<th>Fly</th>
<th>Worm</th>
</tr>
</thead>
<tbody>
<tr>
<td>ribonucleoprotein complex</td>
<td>cytoskeleton organization, mitotic cell cycle</td>
<td>larval development, post-embryonic development</td>
</tr>
<tr>
<td>organelle lumen, nuclear lumen</td>
<td>ribonucleoprotein complex</td>
<td>ribonucleoprotein complex</td>
</tr>
<tr>
<td>contractile fiber, myofibril,</td>
<td>mRNA processing, RNA splicing</td>
<td>positive regulation of growth</td>
</tr>
<tr>
<td>striated muscle thin filament,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sarcomere, muscle contraction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HTH, DNA-binding, centromere</td>
<td>SUMO binding, small conjugating protein binding</td>
<td>genitalia development, reproductive developmental process</td>
</tr>
<tr>
<td>protein Cenp-B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mRNA processing, RNA splicing</td>
<td>contractile fiber, sarcomere, myofibril, striated muscle thin filament</td>
<td>molting cycle</td>
</tr>
</tbody>
</table>

Table 4-7. Gene ontology analysis on Schnorrer-adult genes.

While Schnorrer-adult genes are striated muscle-specific genes, GO terms related to myofibrils and sarcomeres are absent from human and worm orthologs, other than the actin cytoskeleton for humans (Table 4-7). Further literature search is required to confirm if the over-represented GO terms for the other two species are related to muscle functions. It should be noted that the oPOSSUM results for human ortholog sets were not representative of striated muscle analysis results.
4.3.3.6 Overview of the Enriched Motifs in the Nematodes and Drosophila

4.3.3.6.1 Nematodes

Egr1, Klf4, Smad3 and Sp1 consistently score as highly enriched in nematode striated muscle genes, suggesting that these factors play central roles in striated muscle regulation. The known absence of Mef2 is apparent in the nematode analysis results. Srf is not enriched in the gene sets tested, suggesting the possibility that MADS family of TFs in general may not play central roles in nematode muscle regulation. Several potential regulatory TFs arose repeatedly in the analysis, including Trl, Tfp2A, Ebf1 and Zfx. GAGA binding sites have been found to be associated with muscle-specific TFs in nematodes previously. In their computational analysis of known muscle genes in nematodes, GuhaThakurta et al. identified three novel motifs that were over-represented. One of these motifs shares similarity with the Trl profile, a Drosophila GAGA factor. No direct reference of muscle development regulatory roles could be found for the other TFs.

4.3.3.6.2 Drosophila

Mef2A was reconfirmed as a central mediating regulatory factor in fly muscle development, as it was one of the highest enriched motifs in many of the tested gene sets. Other highly enriched TFs with muscle regulatory roles include Cf2_II and Six6.
4.4 Discussion

In order to investigate the extent of conservation in transcriptional regulatory mechanisms between evolutionarily divergent species, we performed TFBS over-representation analyses on selected groups of co-expressed genes using the oPOSSUM-3 system. The analysis of conserved regulatory programs allows for the detection of TFs serving as master regulators for fundamental biological systems.

Collections of muscle-enriched genes, ciliary genes, and Nfe2l2 target genes from vertebrates, fruit flies and nematodes were evaluated for similarities and differences in candidate regulatory TFs. In muscle, the main myogenic regulators such as Myf family, Srf and Mef2A were commonly detected in both vertebrates and flies, whereas in nematodes, the MADS domain TFs were absent, in agreement with the literature. In flies, Cf2-II was consistently detected as a likely regulator of muscle development. In nematodes, Egr1, Sp1, Smad3, Klf4, Trl, Ebf1 and Zfx motifs were among the top-ranked candidates. For ciliary genes, the importance of RFX family of TFs was affirmed. RFX motifs consistently received the top-ranks for enrichment in gene collections for all three species. As for Nfe2l2, nematodes differ from the other species. skn-1, the ortholog of Nfe2l2, does not share the same binding profile as its vertebrate and fly orthologs (based on the motif discovery software results on the available ChIP-seq data). Gene ontology analysis of the skn-1 target genes suggests different functional roles for these genes from the vertebrate or insect counterparts.

While the current system is able to detect the similarities and differences in the key regulatory TFs, improvements can be made. Despite the expansion of the JASPAR TFBS profile database, the nematode-specific profiles are limited in number, and the insect-specific profiles are heavily dominated by the homeodomain TF family. An increased number of profiles may improve the capacity to detect conserved regulatory programs. ModENCODE ChIP-Seq data sets for TFBSs in flies and nematodes offer a path for expansion. Also, large-scale gene knockout for all TFs in a given species, such as the data from the C. elegans Knockout Consortium, could be employed to examine whether
the over-represented motifs are associated with known phenotype change in the tissue or biomolecular structure of interest.

Further gene ontology analysis of the tested gene sets in each species indicates that the conservation of their functions may not be consistently true. Due to genome duplications, many vertebrate genes have paralogs that are not found in invertebrates. These genes may have diverged in their function, with different regulatory controls introduced. This may introduce difficulties when orthologs are identified and tested computationally, without affirming their conserved functions. A simple computational method of screening for functionally diverged genes would be to compare the associated gene ontology terms and confirm they remain consistent. If there is expression profiling data for the orthologous genes in the species being compared, their enrichment in the tissue of interest could be evaluated as well.

When comparing co-expressed genes from two different species, it may be advantageous to use smaller, more targeted gene lists that are under the control of fewer master regulatory TFs. Because of the dominance of the RFX TF family in the regulation of the ciliary genes, we hypothesized that the regulatory mechanism will be conserved and be easily elucidated. To this end, preparing subsets of genes according to each growth and development/differentiation stages would be useful. Comparing gene expression programs between stages of vastly different organisms is non-trivial. When determining target genes of a given TF based on ChIP-seq data, it would be preferable to obtain genetic data for TF-dependent gene expression to overlay with the ChIP-Seq peaks (rather than assuming peaks act upon the most proximal gene). Such approaches were performed by Malhotral et al. for Nfe2l2, Sandmann and Jakobsen et al. for fly ChIP-chip studies, and Meissner et al. for nematode RNAi studies.
With the increased availability of TFBS data, using TFBS over-representation analysis for the elucidation of the conserved regulatory programs among evolutionarily diverged species has great promise. The similarities and differences found among the analyzed regulatory programs should result in better understanding of the evolution of the regulatory mechanism and the genes under their control.
5 Discussion and Conclusions

Since the sequencing of the first whole genome of an organism, the field of genomics has been characterized by the continuous development of higher-throughput experimental technologies and the development of accompanying bioinformatics tools required for the analysis of the data generated from such technologies. These advances have made available an ever increasing amount of higher quality genomic data, empowering researchers to sequence genomes of multiple organisms and produce detailed annotations of their gene structures, transcriptome profiling, and epigenetic modifications. Despite the vast advances, elucidating the complete structures of gene regulatory networks remains as an unmet challenge.

To elucidate the properties of gene regulatory networks responsible for the correct development and functioning of a given tissue or biomolecular structures, researchers often resort to analyzing sets of genes sharing a common characteristic, such as co-expression. A major piece of the regulatory network puzzle is solved when the mediating TFs for the observed patterns are identified. In the genome, the binding sites for these TFs are often clustered together for cis-regulatory modules to assert combinatorial regulatory control on the regulatory target genes. If the set of key mediating TFs and their binding motifs are known, computational CRM prediction tools can be used to scan genomic sequences for clusters of these motifs.

However, the currently available tools are insufficient for direct annotation of a genome. If the key TFs are not known but the list of co-regulated genes are known, TFBS over-representation analysis can be used to determine the putative candidate TFs that are enriched in these genes compared to a control set of genes. The DNA binding domains of TFs and the sequence motifs they bind to tend to be well conserved even between species from different taxonomic groups. Furthermore, major tissues are shared among a large subset of metazoans, including the genes specialized for their development and function and the regulatory mechanisms controlling the expression of
these genes. TFBS over-representation analysis based on TFBS profiles pooled from multiple species can be used to identify the key mediating TFs that are conserved in the orthologs of these genes from species belonging to diverse taxonomic groups. This can facilitate the transfer of knowledge gained from experiments performed in model organisms to humans.

This thesis reports on the research efforts into the development and validation of the bioinformatics tools that are outlined above. A review of the developed tools and their experimental results are summarized in the following sections.

5.1 Identification of Tissue-Specific cis-Regulatory Modules Through Computational Prediction Programs

The main regulators of vertebrate myogenesis have been widely studied, with the expression stages and interactions among the regulator TFs identified. However, the number of muscle CRMs consisting of these TFs has remained relatively static despite the introduction of computational CRM prediction methods, which could be used perform a genome-wide search for the CRMs.

In Chapter 2, we reported on the efficacy of utilizing the prediction tools in the identification of novel skeletal muscle regulatory regions. We scanned the human genome for putative skeletal muscle regulatory regions with LRA, MSCAN and ClusterBuster and subjected a subset of the predictions to biological validation using reporter gene expression assay in C2C murine skeletal muscle cell culture, where we assessed the ability of the putative CRMs to drive reporter expression selectively in mature myotubes only. Of the 339 candidate sequences tested, 278 were successfully transfected into cells and the expression measurements of these constructs were taken. We identified 19 myotube-restricted promoter-enhancing sequences, and we established a new reference collection of skeletal muscle-specific regulatory regions by adding these regions to the known skeletal muscle reference CRM collection. Compared
to the non-functional predictions, the validated CRMs exhibit increased sequence conservation and G/C mononucleotide content. Such increased G/C content could be detected in brain-directing CRMs from the Pleiades Promoter Project, indicating the potential benefit of employing sequence composition as a component of CRM prediction methodology.

Given that only a small subset of the predicted regions was validated as being functional, it is apparent that the CRM prediction programs used in this study have inadequate predictive power for direct genome annotation. It is possible that the performance could be improved by extending the analysis to include all TFs known to be expressed in the target tissue, rather than restricting the analysis of a small core of characterized TFs. Such an extension, however, does not address the more fundamental missing information – the epigenetic data defining TF-accessible portions of the genome. The CRM search pipeline used in this study only relies on the primary sequence and its conservation. The predictive performance could be improved if such data was available for the tissue of interest and could be incorporated into the search pipeline. With the increasing availability of ChIP-chip and ChIP-seq data for individual TFs or co-activating proteins, another possible approach would be to make use of such data to determine the location of the CRMs. Again, in order for such an approach to be feasible, ChIP-based TF binding data needs to be obtained for each tissue type being analyzed. While there is insufficient large scale TF-chromatin binding data available to undertake such a strategy for most tissue types, the rapid developments in genomics will likely alleviate the limitations in the coming years. It should not escape the reader that this approach returns the burden onto laboratory analysis and minimizes the potential for computational annotation of regulatory modules in the human genome.
5.2 TFBS Over-Representation Analysis for the Identification of Main Regulatory TFs in Co-expressed Genes

Before the generation of ChIP-seq data for the TF mediating regulation of interest, the mediating TF itself must first be identified. In the skeletal muscle CRM prediction and validation study, a core subset of the main regulators of skeletal muscle differentiation was obtained from literature, as this tissue has been widely studied (a more comprehensive list is used for the work performed in later stages of research). However, such information is more often lacking; in many cases, researchers may wish to find the key regulatory controls responsible for differential expression of a select group of genes based on gene expression profiling studies for a specific tissue of interest. Even if some of the mediating TFs are known, the list is incomplete and there may be other TFs that also participate in the regulatory process. A key strategy for the identification of such TFs is to determine the sequence motifs that are over-represented in the cis-regulatory regions of gene set being co-regulated compared to some control gene set. With the available central repositories of TFBS profiles such as JASPAR, all known sequence motifs could be searched in a given co-regulated gene set and have the over-represented motif mapped back to the TF that binds to it.

In Chapter 3, we describe oPOSSUM-3, a significantly re-designed software system for the identification of statistically over-represented, predicted TFBSs in co-regulated gene sets. The oPOSSUM system uses two different scoring methods to measure over-representation: 1) Z-scores based on binomial distribution that measures the change in the number of TFBS motifs in the target gene set compared to the background set, and 2) Fisher p-values calculated using a one-tailed Fisher exact probability assessing the number of genes with the TFBS motifs in the target set versus the background set. The new oPOSSUM-3 system incorporates the phastCons multi-species conservation scores, removing the need for the restrictive pairwise sequence alignment phylogenetic footprinting procedures of previous versions. This change has the added benefit of making it possible to construct oPOSSUM services for any species represented in the
Ensembl genome annotation database and the UCSC phastCons track. Currently, oPOSSUM-3 services have been built for human, mouse, fruit fly (a species not available in the previous version), nematode and yeast. The system uses the updated JASPAR 2010 database, which features a significantly expanded set of annotated TFBS profiles (5-fold increase over the previous version) and new collections from PBM studies. The system also features new, overhauled analysis methods: 1) anchored combination site analysis (aCSA) that allows rapid identification of interacting pairs of TFBSs, and 2) TFBS cluster analysis (TCA), where highly similar TFBS profiles of the same family are clustered to remove redundancies and allow a streamlined reporting of results to users. Such clustering methods are especially important given the recent expansion of the available TFBS profiles in the JASPAR database. Most importantly, the oPOSSUM system contains new sequence-based methods, allowing the system to function as an analysis tool for large-scale ChIP-Seq sequence sets. The system was successfully validated with known co-regulated gene sets and published ChIP-Seq data sets.

5.3 Utility of Evolution in Identifying Conserved Regulation Programs

Metazoans from diverse branches of the taxonomic tree share common developmental stages and types of major tissues and organs, despite their evolutionary distances from one another. Specialized genes responsible for organ functions share similar expression patterns across different species, and the regulatory mechanisms for organ development tend to be maintained as well. Muscle is an example of such a tissue with conserved structure and regulatory controls. Muscle is one of the oldest tissues in metazoans, being present from jellyfish to human. The DNA binding domains of major tissue regulator TFs are often conserved as well. Such conservation of key regulatory components presents an opportunity to transfer the knowledge of the corresponding regulatory mechanisms from simpler model organisms to humans. Fruit fly and nematode are important model organisms that serve as popular vehicles for genetic research. If common regulatory mechanisms can be found among those species and
human, it would bring considerable benefits to the annotation efforts for human regulatory regions. A key to understanding the mechanisms responsible for co-regulation of genes is to identify the mediating TFs, which can be accomplished with TFBS over-representation analysis. Currently, oPOSSUM-3 has been constructed for human, mouse, fruit fly and nematode. New ChIP-Seq data for the model organisms have been released by the modENCODE consortium. The combination of the software tool set and the new data provide the required framework for the identification of conserved regulatory programs.

In Chapter 4, we performed a three-way comparison of over-represented TFBSs among human, fruit fly and nematode, in three co-regulated gene sets: muscle-enriched, cilia-components, and Nfe2l2 target genes. We collected from literature and expression studies the sets of candidate co-regulated genes and analyzed them using oPOSSUM-3 to evaluate the extent of conservation of key regulatory mechanisms. For muscle genes, enrichment of binding site motifs for key myogenic regulators with MADS domains, such as SRF and MEF2A, could be detected in both vertebrates and flies, but not in nematodes. The Cf2-II binding site motif was enriched in flies, whereas in nematodes, Egr1 and Zfx motifs were among the most enriched motifs. For ciliary genes, Rfx binding site motifs were consistently enriched across all three species. For the Nfe2l2 TFs, shared enriched motifs could be found for vertebrates and insects, while the nematode orthologous protein exhibited unique binding site properties from the other species. Gene ontology analysis of Nfe2l2 target genes suggests distinct functional properties for the proteins in each species. Overall, the analysis of conserved regulatory programs reveals complex evolutionary characteristics, with some systems conserved and others diverged for distinct functions.
5.4 Future Directions

With the bioinformatics framework for the identification of conserved regulatory programs, one can examine the species from major branching points of the metazoan evolutionary tree and investigate the conservation of the regulatory mechanisms. Comparing and contrasting the enriched TFBSs in orthologous gene lists between two adjacent taxonomic groups, and extending these comparisons to the entire metazoan tree, one may gain insights into how the gain or loss of regulatory controls accompanied the changes in the orthologous gene lists, reflecting the change in the phenotype regulated by the examined mechanism.

To analyze regulatory sequence properties for sets of orthologous genes across species, it is important that the initial set be subject to homogeneous regulatory mechanisms, such that the genes are under the control of a limited number of master regulatory TFs. As observed for the cilia genes, the dominant contribution of one TF allowed for the regulatory signature to be detected across evolution. Thus researchers should carefully consider in their experimental designs which orthologous sets of genes are compared, such that the temporal and spatial expression patterns of the genes selected are as restrictive as possible and that they share similar roles in developmental stages or structural functions between the compared species.

Over-representation analysis can be improved beyond the approaches introduced in this thesis. While the JASPAR database has been expanded in 2010, it is dominated by vertebrate profiles. Nematode profiles are severely under-represented. As seen with the analysis of Nfe2l2, the binding motifs for orthologous TFs are not always conserved across taxonomic groups. When binding profiles for the major TF families for multiple taxonomic groups are represented in the database, it will be possible to perform TFBS similarity clustering. Those orthologous TFs for which the binding site profiles fail to cluster may provide key insights into changing regulatory programs. A better understanding of the changes in their functions could be obtained by constructing an
evolutionary map of the binding profiles to determine if changes in binding motifs could be correlated with divergence of specific taxonomic groups.

The method for modeling the TFBS motif may be improved. Effort is ongoing to design models that resolve the restrictions of PSSMs. The existing method cannot represent variable spacers within motifs or interdependencies of the bases. While models such as HMM-based TFBS profiles have not supplanted the traditional PSSMs in popularity, this may change in the future as better models are now warranted in light of the binding site properties being revealed by the emergence of rich ChIP-Seq data. Currently, oPOSSUM can identify the enriched pairings of TFs given a user-specified anchoring TF. This is important as TFs often interact with each other and exert regulatory control as a unit, either by binding to the same CRMs or by binding strings of homogenous, multimeric sites. Utilizing pre-computed models of TFBS multimers may streamline the search process and present new opportunities for further analysis of the variations in the tissue-specific regulatory controls, as different multimers involving the same TF can have different regulatory effects.

One of the most significant developments in recent years has been next generation sequencing. The massively parallel sequencing method has driven down DNA sequencing costs. Taking advantage of the new sequencing methods, the ENCODE and modENCODE consortia are continuously generating new data sets appropriate for regulatory region analysis. With the availability of modENCODE ChIP-Seq data sets for TFBSs in flies and nematodes, it is expected that the collection of annotated TFBS profiles will grow rapidly. With the increased availability of TFBS data from multiple species, using TFBS over-representation analysis for the elucidation of the conserved regulatory programs among evolutionarily diverged species will achieve greater success. The similarities and differences found among the analyzed regulatory programs should result in better understanding of the evolution of the regulatory mechanism and the genes under their control.
High-throughput sequencing methods present researchers with the possibility of studying variations in the regulatory regions and their functional effects, if any. When variations that produce discernable phenotypic changes in the regulatory regions are identified at the nucleotide level, the altered TFBS can be identified and the importance of the modified base elucidated. Regulatory region variation data is important for biomedical reasons, as identification of regulatory region variations responsible for a genetic disease can improve our understanding of the disease and be utilized as diagnosis tools.

With the newly developed technologies, other sources of information can be incorporated into regulatory region identification. Peak regions from ChIP-Seq TF binding experiments can be filtered for tissue-specific histone modifications so that the binding regions are restricted to those likely to be functional. Chromosome conformation capture and its higher-throughput variations may allow one to observe whether the bound TFs and putative target genes are located proximal to one another, giving more confidence to the classification of the genes as the regulatory targets of the TFs. Such 3D maps could also be combined with functional insulator locations to obtain a more complete picture of the regulatory landscape (Ferraiuolo et al. 2010).

Moving forward, new bioinformatics approaches are required to identify clusters of bound TFs and proximal genes confined within epigenetic and/or insulator-delineated boundaries. Such integrative analysis will enable more accurate computational prediction of functional regulatory elements that drive specific expression patterns in a given condition, compared to the CRM prediction methods from Chapter 2 that were solely dependent on the primary genomic sequence. With the increased amount of TF-DNA binding data, more comprehensive and accurate models of TFBS profiles can be constructed, which will improve the performance of computational tools that rely on them. Motif over-representation analysis tools such as oPOSSUM will continue to serve important roles in regulatory region analysis, directing researchers to the candidate TFs that participate in regulating the gene co-expression pattern being studied. Ultimately, data generated from the new technologies will enable the construction of
spatial and temporal maps and simulation models of functional TF-DNA interaction for most of the TFs in a given species, providing important insights into the dynamics of transcription regulation throughout the developmental process of the species. When such maps become available for different organisms, the regulog strategy outlined in Chapter 4 can be applied to determine the extent of regulatory program conservation and characterize how the programs have evolved.

In summary, the technological advances continuously being made in the field of genomics are presenting researchers with exciting new data that could not be obtained even just a few years ago. This is especially important for those working on elucidating the structure and components of gene regulatory networks, as vital pieces of information have been lacking, frustrating the research efforts. With the increased availability of TFBS data and genome annotations for a diverse set of organisms, we will achieve a much better understanding of the evolution of the regulatory mechanism and the genes under their control in the near future. By transferring the new knowledge among different species, and ultimately to humans, we will be able to gain a clearer insight into the inner workings of human physiology and work towards determining the causes of debilitating diseases.
References


Evans K. J., 2010 Most transcription factor binding sites are in a few mosaic classes of the human genome. BMC genomics 11: 286.
Geremek M., Bruinenberg M., Ziętkiewicz E., Pogorzelski A., Witt M., et al., 2010  Gene expression studies in cells from primary ciliary dyskinesia patients identify 208 potential ciliary genes. Human genetics 129: 283-93.


Golovnin a, Gause M., Georgieva S., Gracheva E. and Georgiev P., 1999  The su(Hw) insulator can disrupt enhancer-promoter interactions when located more than 20 kilobases away from the Drosophila achaete-scute complex. Molecular and cellular biology 19: 3443-56.


Pan D. and Courcy A. J., 1992 The same dorsal binding site mediates both activation and repression in a context-dependent manner. The EMBO journal 11: 1837-42.


Shimizu S., Miyamoto Y. and Hayashi M., 2002 Cell-type dependency of two Foxa/HNF3 sites in the regulation of vitronectin promoter activity. Biochimica et biophysica acta 1574: 337-44.


