Genomic Analysis of Transcription and Alternative Splicing with Embryonic Stem Cell Differentiation and Myometrial Gestational Remodeling

by

Nathan G. Salomonis

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSPHY

In

Pharmaceutical Sciences and Pharmacogenomics

In the

GRADUATE DIVISION

Of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

Acknowledgments

Publication Reprints

The text in chapter 2 of this dissertation contains a reprint of materials as it appears in:
Salomonis N, Hanspers K, Zambon AC, Vranizan K, Lawlor SC, Dahlquist KD, Doniger SW, Stuart J, Conklin BR, Pico AR.
GenMAPP 2: new features and resources for pathway analysis.
BMC Bioinformatics. 2007 Jun 24;8:218.
The co-authors listed in this publication co-wrote the manuscript (AP and KH) and provided critical feedback (see detailed contributions at the end of chapter 2).

The text in chapter 3 of this dissertation contains a reprint of materials as it appears in:
Salomonis N, Cotte N, Zambon AC, Pollard KS, Vranizan K, Doniger SW, Dolganov G, Conklin BR.
Identifying genetic networks underlying myometrial transition to labor.
Genome Biol. 2005;6(2):R12. Epub 2005 Jan 28.
The co-authors listed in this publication developed the hierarchical clustering method (KP), co-designed the study (NC, AZ, BC), provided statistical guidance (KV), co-contributed to GenMAPP 2.0 (SD) and performed quantitative mRNA analyses (GD).

The text of this dissertation contains a reproduction of a figure from:
Yeo G, Holste D, Kreiman G, Burge CB.
Variation in alternative splicing across human tissues.
Genome Biol. 2004;5(10):R74. Epub 2004 Sep 13.
The reproduction was taken without permission (chapter 1), figure 1.3.

Personal Acknowledgments

The achievements of this doctoral degree are to a large degree possible due to the contribution, feedback and support of many individuals. To all of you that helped, I am extremely grateful for your support.

This dissertation would not have been possible without the guidance of my advisor, Dr. Bruce Conklin. In addition to contributing to the majority of the intellectual work provided herein, Bruce has been an extremely supportive mentor who has fostered the creativity of his lab members. Bruce has always remained confident in my abilities and has encouraged me to think outside of the box. I am grateful for his generous time and intellectual contribution to my projects. In addition to Bruce, I would like to thank the support of the Gladstone Institutes and in particular Gladstone president Dr. Robert Mahley for making Gladstone one of the best academic environments around to work in the world.

Much of the work presented herein was through the personal and professional support of my lab mates who I would like to specifically acknowledge.

Dr. Alex pico, a Conklin lab Postdoctoral fellow, is an extremely talented computational biologist who has provided ongoing support, help and encouragement with the bioinformatics of nearly all of the projects I've worked on. Alex has always been available to provide feedback, assistance and new insights into problems that I often thought were too daunting to achieve. Specifically, Alex developed several programs required to generate domain -evel predictions from splicing data to visualize splicing data at the level exons and introns in Cytoscape (SubgeneViewer). He developed tools to identify primers for validation of alternative splicing events in a high-throughput manner (AltPrimer) (Chapters 5-6).

Kristina Hanspers, a senior research associate in the lab has been a long-term colleague without her assistance, much of the work provided in this thesis would not be possible. Kristina is a multi-talented genomics researcher who is the heart and soul of the GenMAPP project. Kristina is always able to approach difficult problems and tackle complex projects with a level head and achieve great successes as a result. Kristina has been one of the main developers of the GenMAPP applications (chapters 2, 3, 4, 5 and 6) and contributed to the development of SubgeneViewer and BubbleRouter applications (Chapters 2, 5).

Alex Zambon, a former postdoctoral fellow in the Conklin lab, has been a long-term collaborator who has also contributed heavily to the work discussed in this thesis. Alex and I have written several papers together, working very successfully as a team. Alex has an excellent ability to see unique problems and design elegant solutions, both at the bench and at the computer. In particular, Alex did considerable experimental and analysis work for Chapters 2, 3, 4 and 6.

Karen Vranizan, a bioinformatics specialist and statistician in the Conklin lab, has provided wise advice and consul for all nearly all bioinformatics projects conducted during the course of this thesis and before. Karen has directed many of the specific analyses discussed in this dissertation including assistance with clustering (Chapter 3), algorithms for permutation based tests (Chapter 5-7), over-representation analysis (Chapters 2 and 4), splicing statistics (Chapter 5-7) and study design (Chapter 3).

Dr. David Williamson is a recent graduate of the bio-medical Informatics (BMI) program at UCSF who has provided valuable input into my projects. David is an excellent bioinformatics programmer with the ability to approach seemly impossible problems and identify quick and effective solutions.

Stan Gaj in the Chris Evelo group, a part of the BiGCAT bioinformatics consortium an early adopters of GO-Elite and provided considerable input and design suggestions that were incorporated into the final software (Chapter 4).

I would like to especially thank the members of my thesis and orals committee. Drs. Ru-Fang Yeh and Susan Fisher have served on both committees and provided excellent support, input and advice throughout this process. Drs. Patsy Babbitt (chair) and Andrej Sali, who served on my orals examination committee, made significant sacrifices to be there (Andrej’s first child was delivered the night before) and made critical contributions to the direction this work has taken.

Although most of my work has been done at the computer, I have relied heavily on the assistance and support of both academic and industrial collaborators on what have become fairly extensive and involved research projects. Chris Schlieve, a talented cell biologist, partnered up with me for the Tcf3 embryonic stem cell analyses (Chapter 6) and conducted a large portion of that work, resulting in the production of data that I would not have been able to produce on my own. Laura Pereira in the Brad Merrill lab, has significantly contributed to the Tcf3 isoform project and along with Brad has contributed greatly to the scientific content and descriptions in this analysis. In the Mark Mercola lab, the contributions by both Brandon Nelson and Mark have provided valuable reagents which have facilitated the large scale human exon array analyses conducted in this thesis (Chapter 5). Furthermore, the excellent work of the Gladstone Genomic's core, lead by Dr. Chris Barker and conducted by Linda Ta have been vital to obtaining high quality data.

In addition, industrial collaborators at Affymetrix and Agilent were also instrumental in this work. Without the assistance from John Blume, Melissa Cline, Hui Wang, and Allan Williams at Affymetrix, the technology and support needed to perform the mouse exon-exon junction array analyses would not be possible. At Agilent, work by Allan Kuchinsky, an expert Java programmer who is one of the primary contributors to Cytoscape development, worked closely with Kristina Hanspers and Alex Pico to develop the SubgeneViewer and BubbleRouter software, used extensively in these analyses.

To re-implement published algorithms (ASPIRE, linear regression analysis of splicing, perfect match DABG p-value calculation), heavily used in these projects, the authors of these algorithms graciously provided their assistance (Jernej Ule and Chuck Sugnet).

I am especially grateful to my friends, family and fellow students who have provided invaluable support and understanding throughout this process.   To my wife Heather, thank you so much for your love and support from day one of this process, I don’t know what I’d do with out you. You have patiently stood with me and helped throughout this process, from isolating EBs with me on your days off and editing the drafts of this thesis, to helping me cope with the ups and downs of graduate student life.

Finally, I would like to thank Dr. Frank Szoka and Debbie Acoba for their impressive administrative accomplishments in the Pharmaceutical Sciences and Pharmacogenomics graduate program. They have both been extremely helpful and supportive during my graduate studies, from day-to-day needs to funding and academic advice. I would like to further acknowledge the funding opportunities that allowed to me to accomplish this work, including the Achievement Rewards for College Scientists (ARCS) Scholarship 2005 – 2006, National Institutes of Health Training Grant (T32 GM07175) 2003-2004, and the kind awards provided by the PSPG graduate program (Ira Herskowitz award) and the Gladstone Institutes.

Abstract

Nathan G. Salomonis

The development of an organism from conception to adulthood requires the specification of cell types to distinct fates. In an adult organism, tissues can similarly undergo dramatic transformation, altering their structure, physiology, and overall biochemical properties. In both of these paradigms, the study of gene transcription and its contribution to protein content in the cells has been the primary focus. While clearly important, more recently, alternative splicing and microRNA regulation have been recognized as significant processes that can have crucial regulatory influences on the diversity of mRNAs produced and their over-all expression in the cell.

In this dissertation, I have set out to characterize the molecular changes that occur in two distinct cellular paradigms, muscle remodeling in the uterus throughout pregnancy and the differentiation of embryonic stem cells to distinct fates using DNA microarray technology. To achieve this goal, I developed several new software applications, designed specifically to assess the relevance of coordinated gene regulatory events along biological pathways (GenMAPP and GO-Elite) and characterize sequence level functional attributes of proteins and mRNAs regulated by alternative splicing (AltAnalyze).

Studies of the mouse uterus during gestation reveal novel coordinated transcriptional networks regulating quiescence, contraction, and involution when multiple time-points are considered. Analysis of alternative splicing in mouse and human embryonic stem cell differentiation uncovered novel mechanisms for the regulation of protein domain and microRNA binding site inclusion and at least for one gene, Tcf3, the requirement of splicing to properly promote the early steps of embryonic lineage commitment down multiple paths. In summary, I have used novel computational methods and genomic resources to uncover new regulatory networks and potential biological mechanisms from both differentiating and transitioning cells, that involve regulation at the level of transcription, alternative splicing, and translational regulation by micoRNAs.

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction: Utilizing Genome-Wide Methods to Gain Novel Biological Insights into Physiological and Developmental Programs

1.1 Genomic era opportunities and challenges

With the completion of the human genome sequencing project, researchers now have access to biological data on an unprecedented scale. This information has provided the means to design new assays and new technologies to efficiently and in an unbiased fashion, measure biological responses simultaneously on a molecular and genomic level. Access to such genome-level data provides both new opportunities and new challenges to integrate the wealth of complex information. I have focused on this area of research to help translate these data into findings that will hopefully impact human health and biological knowledge. To meet this challenge, I have built a series of software tools (e.g., GenMAPP, GO-Elite, and AltAnalyze) and have applied them to specific biological problems.

1.2 Interrogating mRNA content on a genome-wide level

An immediate application of genome sequencing data has been the development of biological assays to specifically measure the relative abundance of all known RNA transcripts at a cellular level. One of the primary tools to assess genomic responses in living cells and whole organisms is the DNA microarray. DNA microarrays are high throughput platforms for assaying the relative quantity of both DNA and RNA for specific targets with extremely high resolution. Microarrays are composed of either small (20-70mer) DNA probes or longer transcript sequences which are synthesized or chemically attached to glass or silicon substrate at picomolar levels, with only a few microns separating probes of distinct composition. Since microarray probes can be synthesized on a large scale and are composed of millions of unique, pre-designed oligomers onto an area that is a fraction of an inch, this technology is ideal for assessing highly complex gene level transcript variation. While typically utilized for the purposes of measuring gene expression, or transcriptional activity of a gene, current microarrays allow users to examine hundreds of thousands of distinct RNAs, including alternative splice variants, alternative promoter transcripts, and non-coding RNAs, such as microRNAs, with biological functions that are largely unknown (Figure 1.1). The role of these distinct biological entities is only recently coming to light, largely as a result of the development of new technologies such as whole-genome level transcript microarrays.

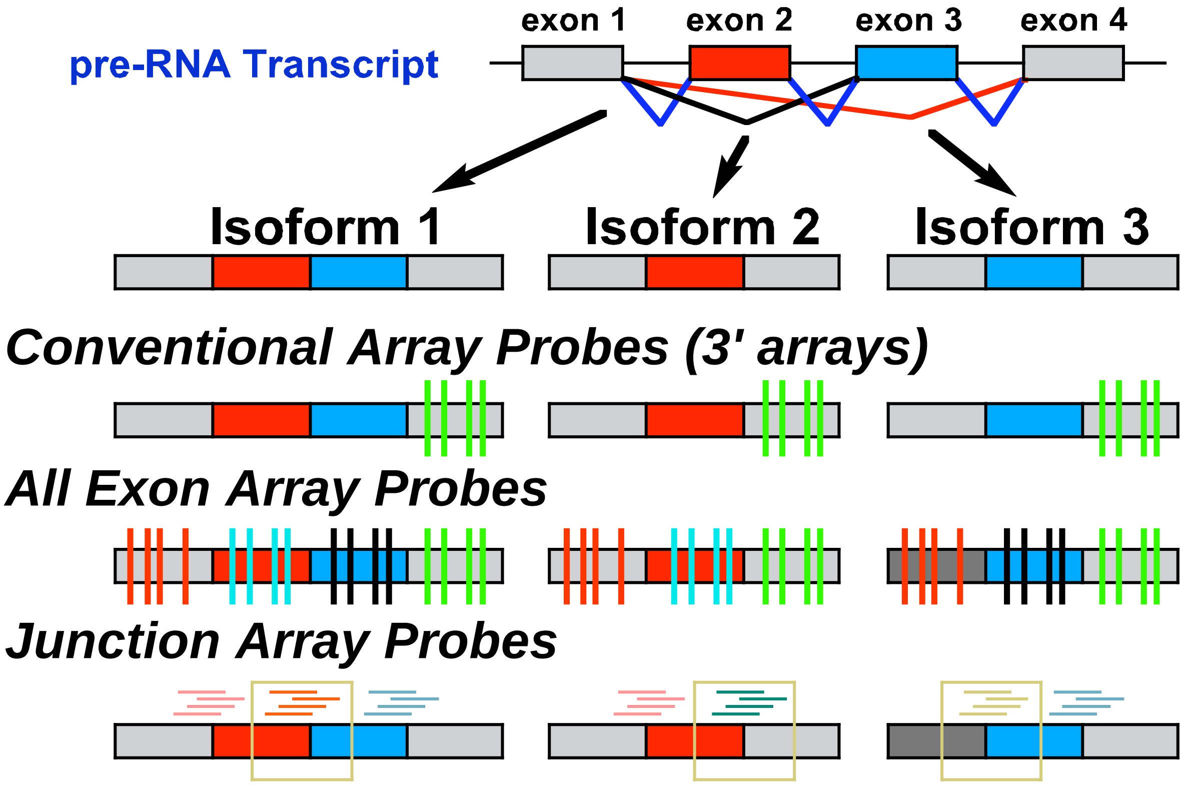


Figure 1.1. Interrogating multiple gene level features on a genomic scale. As microarray feature size has increased, so has the diversity of probe set selection and targets. Shown here are alternative exons produced through alternative splicing for a single pre-mRNA transcript, with different possible probe sets (colored lines). Underneath this depiction is a conventional microarray strategy, with a protocol that is biased towards amplification of the 3' end of the mRNA and thus, probe sets are typically associated with this region of the transcript. When multiple mRNA transcripts are produced, such a strategy may not yield optimal information on gene expression. An all-exon array or genomic tiling strategy has the advantage of assaying for all known regions of possible transcripts. However, this requires knowledge of which exons are informative for transcription and/or alternative modes of exon regulation. A junction array focuses specifically on the expression of alternative isoforms that can occur as a result of either alternative splicing

or alternative promoter selection. This strategy, however, is typically biased by existing mRNA information and has limited sequence space for optimal probe design.

1.3 Proteomic diversity through alternative splicing

Alternative splicing (AS) is the process by which the composition of a primary mRNA transcript is alternatively regulated to produce distinct processed mRNAs. These alternative mRNAs may produce different protein translations or be specifically targeted for degradation (non-sense mediated decay) (Cooper 2005). Splicing is an essential mechanism that cells utilize in order to excise long intronic sequences, at canonical splice sites, from the primary mRNA transcript since these sequences to do not contribute to translated protein products (Figure 1.2 A). This same mechanism can be used to alternatively include exons and/or introns in the processed mRNA sequences (Figure 1.2 B). Both exons and introns can contain binding sites for regulatory splicing factors, which can enhance or inhibit splice site selection and thus regulated splicing.

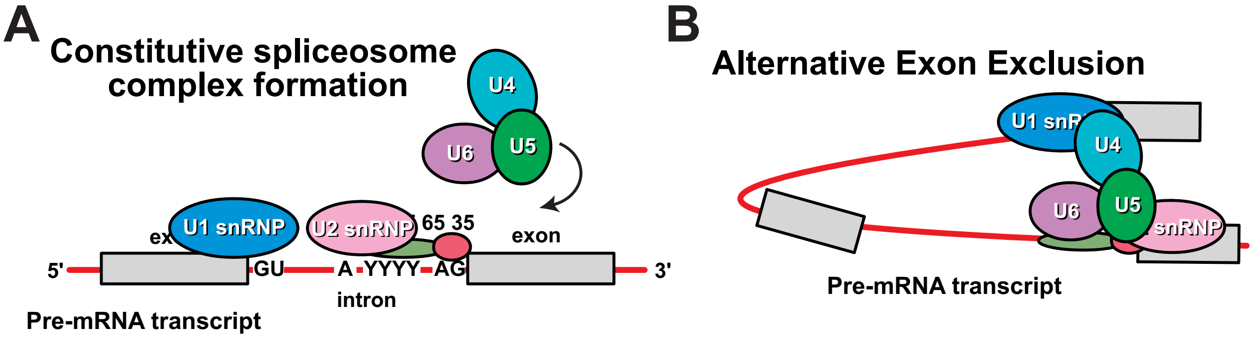


Figure 1.2. Regulation of mRNA composition by alternative splicing. (A) For constitutive forms of splicing (always present), factors not typically regulated in the nucleus bind to consensus splice site sequences to regulate exon inclusion and intron exclusion. This occurs through the formation of a lariat structure and subsequent cleavage. (B) With AS, factors not present at optimal concentrations fail to bind to the spliceosomal complex, resulting in both exon and intron exclusion. A lariat structure shown is shown with the red line indicating introns and the gray boxes indicating exons.

Splicing factors can consist of both proteins and RNAs that directly bind to primary mRNA transcripts. These factors can also associate with signaling components of the cell, such as regulatory kinases that can directly and indirectly effect expression, conformation, and localization of these factors. As a result, signaling within the cell can influence splicing. In humans 40-80% of all genes produce alternative transcripts, as compared to lower eukaryotes that have a similar number of genes but produce significantly fewer alternative transcripts (Ruzanov *et al.* 2007). This case is well-illustrated for the model organism Caenorhabditis elegans which has ~20,000 identified genes, of which only ~9% produce multiple mRNA transcripts (Ruzanov *et al.* 2007). AS therefore provides a potent means to produce a vast number of distinct mRNAs in different cell types and discrete developmental transitions, which are likely coordinated by the differential expression or activation of splicing factors in a developmentally controlled manner (Figure 1.3).

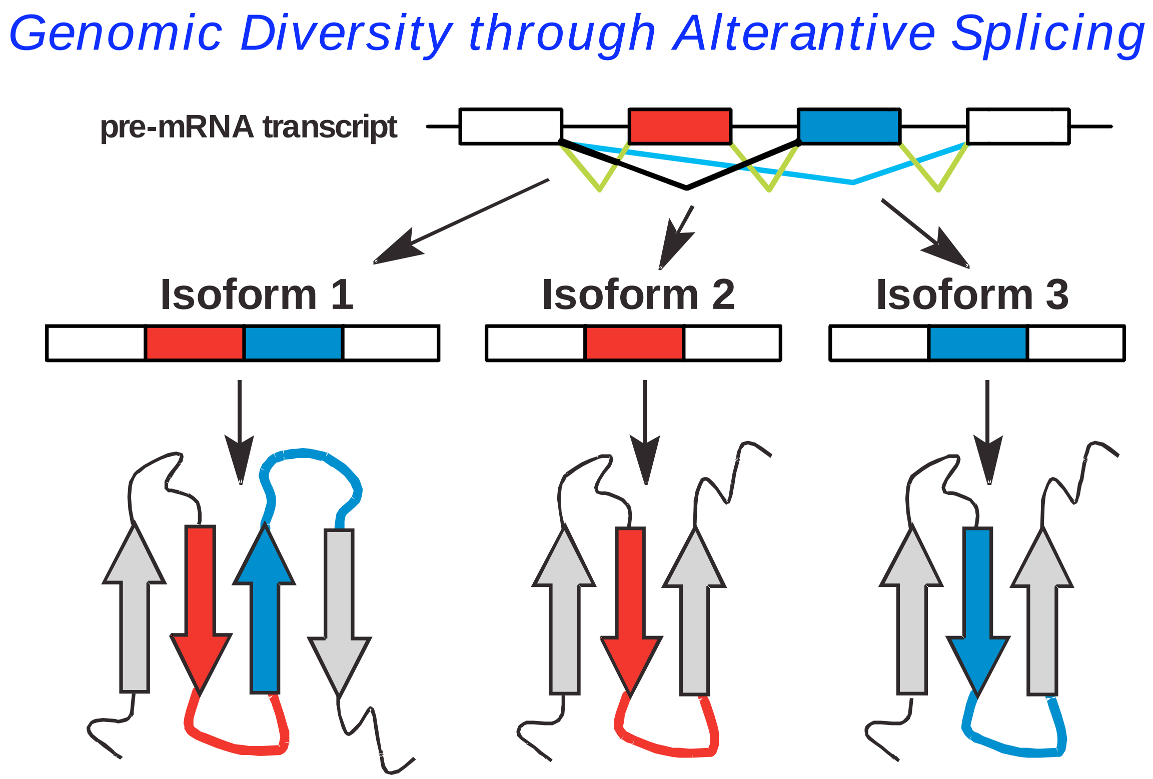
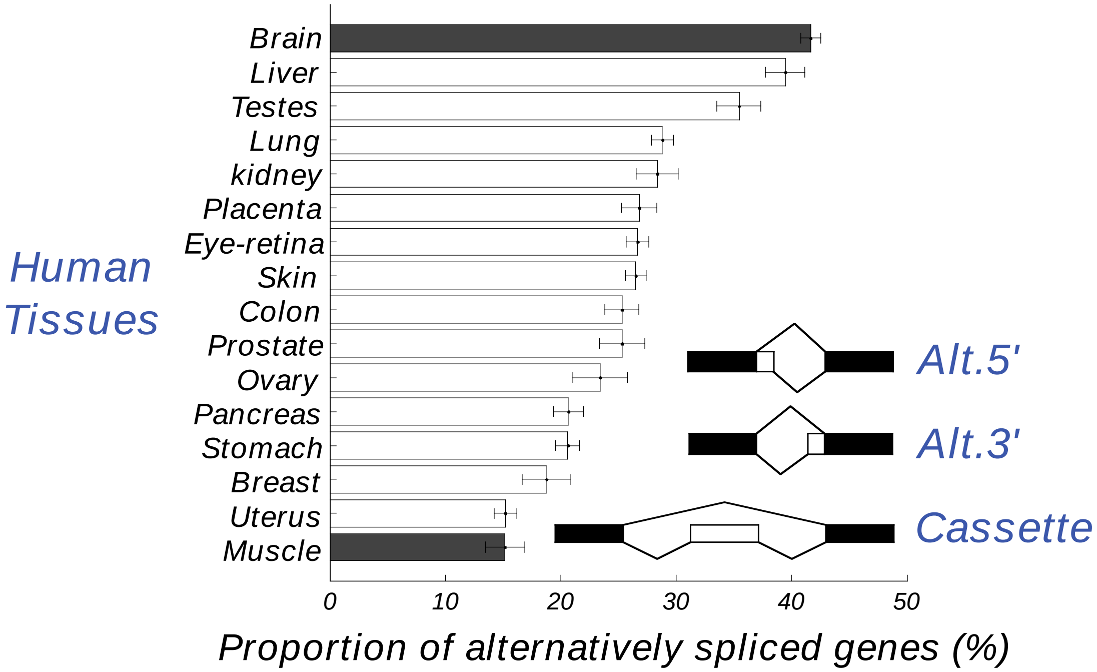


Figure 1.3. Protein diversity through alternative exon inclusion. A common outcome with AS is the inclusion or exclusion of sequences that are critical for protein function, localization, or expression. Thus, by regulating splice variant expression in distinct tissues, the cell can produce distinct isoforms that can differentially impact critical signaling cascades.

1.4 Genomic analysis of discrete cellular transitions

The majority of genome-wide AS analyses have been comparisons between distinct adult tissues. Such analyses have shown that there is a diverse range of transcripts expressed among tissues, with the largest differences often found between neural and muscle lineages (Figure 1.4) (Yeo *et al.* 2004). While these studies are useful as a means to assess diversity of transcripts for genes between tissues, they are convoluted by the fact that adult tissues possess highly distinct processes and are themselves composed of highly heterogeneous cell types. To better assess the functional contribution of specific splice variants to biological processes and development, we require comparisons between derivative cell types, such as the lineage commitment of cells as they differentiate or physiological transformation of cells as they undergo altered demand. Such comparisons decrease the number of variables assessed by the researcher and yield more biologically informative results. This is especially true when the conditions examined occur as a linear continuum of responses in the cell that can be correlated back to specific physiological differences.

With the recent availability of microarrays designed to assay for all known transcripts, a number of studies have begun to assess more discrete biological comparisons in order to identify splicing events that directly correlate with isoform-specific functions. These include the knock-out of the neuron-specific splicing factor, Nova2, in mouse (Ule *et al.* 2005) and AS of human embryonic stem cells to neural precursors (Yeo *et al.* 2007). In the case of Nova2 ablation, genome-wide AS analysis identified that Nova2 specifically regulates the splicing of genes that localize to the synapse. These changes correlate with a decrease in hippocampal synaptic plasticity as measured by electrophysiology recordings (Huang *et al.* 2005). Other studies have demonstrated that splicing variation is sufficient and necessary for regulating critical developmental transitions. These include mouse juvenile cardiac adaptation (Xu *et al.* 2005), sex-determination, and synaptogenesis (Burgess *et al.* 1999), all which result from a failure to alter the splice isoform distribution of genes during development.



Yeo GW et al. Genome Biol. 2004;5(10):R74

Figure 1.4. Distinct tissues possess varying degrees of transcript diversity. Data is shown for distinct EST sequences in public databases for various cell and tissue types. This data suggest that the brain encodes for a highly diverse set of transcripts compared to other tissues, while muscle encodes for a far less diverse set.

1.5 Primary research aims

To effectively assess the role of specific proteins, differentially expressed or alternatively spliced, that are critical for discrete cellular transitions, we must perform unbiased genome-wide analyses that integrate multiple genomic and informatics approaches. To achieve this goal I have focused my doctoral studies on the delineation of temporally regulated gene expression and splicing events restricted to two informative model systems: (1) the mouse myometrium as it transitions from virgin to term gestation and through to postpartum; and (2) the differentiation of mouse and human embryonic stem cells (ESC) to distinct cell fates.

Using these systems, I have been able to identify novel genetic programs that correspond to both unique and overlapping biological pathways that impact pluripotency and muscle remodeling. These studies highlight the important role of coordinated transcription and AS events in the regulation of cell physiology and cell development. By characterizing a single alternatively spliced microarray target, Tcf3, we found that specific splice isoforms had distinct roles in the regulation of ESC differentiation to distinct cell fates. These analyses required the development of new bioinformatics tools and methods, including pathway analysis (GenMAPP, MAPPFinder and GO-Elite), genomic expression clustering (GEMFinder), and AS/functional analysis software (AltAnalyze). In summary, this thesis outlines the development of a series of powerful new open-source bioinformatics tools and strategies to assess complex trends from large-scale genomic data. We believe such methodologies will become increasingly more necessary as genome-wide technologies and their applications evolve.

1.6 References

Burgess, R. W., Q. T. Nguyen, Y. J. Son, J. W. Lichtman and J. R. Sanes (1999). "Alternatively spliced isoforms of nerve- and muscle-derived agrin: their roles at the neuromuscular junction." Neuron **23**(1): 33-44.

Cooper, T. A. (2005). "Alternative splicing regulation impacts heart development." Cell **120**(1): 1-2.

Huang, C. S., S. H. Shi, J. Ule, M. Ruggiu, L. A. Barker, R. B. Darnell, Y. N. Jan and L. Y. Jan (2005). "Common molecular pathways mediate long-term potentiation of synaptic excitation and slow synaptic inhibition." Cell **123**(1): 105-18.

Ruzanov, P., S. J. Jones and D. L. Riddle (2007). "Discovery of novel alternatively spliced C. elegans transcripts by computational analysis of SAGE data." BMC Genomics **8**: 447.

Ule, J., A. Ule, J. Spencer, A. Williams, J. S. Hu, M. Cline, H. Wang, T. Clark, C. Fraser, M. Ruggiu, B. R. Zeeberg, D. Kane, J. N. Weinstein, J. Blume and R. B. Darnell (2005). "Nova regulates brain-specific splicing to shape the synapse." Nat Genet **37**(8): 844-52.

Xu, X., D. Yang, J. H. Ding, W. Wang, P. H. Chu, N. D. Dalton, H. Y. Wang, J. R. Bermingham, Jr., Z. Ye, F. Liu, M. G. Rosenfeld, J. L. Manley, J. Ross, Jr., J. Chen, R. P. Xiao, H. Cheng and X. D. Fu (2005). "ASF/SF2-regulated CaMKIIdelta alternative splicing temporally reprograms excitation-contraction coupling in cardiac muscle." Cell **120**(1): 59-72.

Yeo, G., D. Holste, G. Kreiman and C. B. Burge (2004). "Variation in alternative splicing across human tissues." Genome Biol **5**(10): R74.

Yeo, G. W., X. Xu, T. Y. Liang, A. R. Muotri, C. T. Carson, N. G. Coufal and F. H. Gage (2007). "Alternative splicing events identified in human embryonic stem cells and neural progenitors." PLoS Comput Biol **3**(10): 1951-67.

Chapter 2
GenMAPP 2: New Features and Resources for Pathway Analysis

2.1 Abstract

Background: Microarray technologies have evolved rapidly, enabling biologists to quantify genome-wide levels of gene expression, alternative splicing, and sequence variations for a variety of species. Analyzing and displaying these data present a significant challenge. Pathway-based approaches for analyzing microarray data have proven useful for presenting data and for generating testable hypotheses.

Results: To address the growing needs of the microarray community we have released version 2 of Gene Map Annotator and Pathway Profiler (GenMAPP), a new GenMAPP database schema, and integrated resources for pathway analysis. We have redesigned the GenMAPP database to support multiple gene annotations and species as well as custom species database creation for a potentially unlimited number of species. We have expanded our pathway resources by utilizing homology information to translate pathway content between species and extending existing pathways with data derived from conserved protein interactions and coexpression. We have implemented a new mode of data visualization to support analysis of complex data, including time-course, single nucleotide polymorphism (SNP), and splicing. GenMAPP version 2 also offers innovative ways to display and share data by incorporating HTML export of analyses for entire sets of pathways as organized web pages.

Conclusions: GenMAPP version 2 provides a means to rapidly interrogate complex experimental data for pathway-level changes in a diverse range of organisms.

2.2 Introduction

Advances in DNA microarrays, RNA interference, and genome-wide gene engineering have contributed a wealth of genomic data to the public domain. The average researcher is faced with the challenge of connecting these genome level results to specific biological processes. Therefore intuitive tools for integrating, analyzing, and displaying this data are welcomed by many biologists. One popular approach is pathway-oriented data analysis, which enables biologists to interpret genomic data in the framework of biological processes and systems, rather than in a traditional gene-centric manner.

We developed Gene Map Annotator and Pathway Profiler (GenMAPP) as a free, open-source, stand-alone computer program for organizing, analyzing, and sharing genome-scale data in the context of biological pathways (Dahlquist *et al.* 2002). GenMAPP was initially released in 2001 and has been widely used with over 15,000 unique user registrations and over 250 publications citing its use. GenMAPP allows users to view and analyze genome-scale data, such as microarray data, on biological pathways, Gene Ontology terms or any other desired grouping of genes. These groupings are represented and stored in

GenMAPP as “MAPPs”. GenMAPP automatically and dynamically colors genes on MAPPs according to data and criteria supplied by the user. In addition, GenMAPP allows investigators to easily access annotation for genes at major genomic databases, such as Ensembl (http://www.ensembl.org), Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez), and Gene Ontology (GO) (Ashburner *et al.* 2000). Using the integrated MAPPFinder tool, researchers can rapidly explore their data in the context of pathways and the GO hierarchy by over-representation analysis (Doniger *et al.* 2003).

GenMAPP was developed by biologists and remains focused on pathway visualization for bench biologists, our major user base as judged from publications citing GenMAPP. Unlike other computational systems biology tools (e.g., BioSPICE (Kumar *et al.* 2003), CellDesigner (Kitano *et al.* 2005), E-Cell (Tomita *et al.* 1997)), GenMAPP is not designed for cell/systems modeling. GenMAPP focuses on the immediate needs of bench biologists by enabling them to rapidly interpret genomic data with an intuitive, easy-to-use interface.

2.3 Implementation

GenMAPP is implemented in Visual Basic 6.0 and is available as a stand-alone application for Windows operating systems (Dahlquist *et al.* 2002). The program includes an automatic update feature that allows rapid and reliable updates to the program and documentation.

The three main data components in GenMAPP—experimental data (.gex), gene databases (.gdb), and pathways (.mapp)—are stored in separate files accessible by GenMAPP. All three file types are stored in Microsoft Jet format.

Experimental datasets store any data imported by the user, together with a set of custom coloring criteria (color sets). The gene databases contain species-specific gene annotation from a number of public resources. Databases are created through an ETL (Extract, Transform, and Load) process, by which information is collected from Ensembl, Entrez Gene, Affymetrix (http://www.affymetrix.com), and GOA (UniProt) (http://www.pir.uniprot.org) and reassembled. Annotations supported by GenMAPP include Ensembl gene IDs, UniProt IDs, Entrez Gene IDs, Gene Symbols, UniGene IDs, RefSeq protein IDs, HUGO IDs, GO terms, Affymetrix probe set IDs, RGD IDs (rat), MGI IDs (mouse), SGD IDs (yeast), FlyBase IDs (fruit fly), WormBase IDs (worm), ZFIN IDs (zebrafish), InterPro IDs, EMBL IDs, PDB IDs, OMIM disease associations, and Pfam IDs. MAPPs contain a set of gene or protein identifiers as well as optional graphical elements which are laid out manually.. It is up to the author of the MAPP to choose how to illustrate activation, inhibition, compartments, etc. There is no graph underlying MAPPs, there are no formal nodes and edges: the gene boxes are data-linked, but all lines, edges and sub-groupings are illustrations only. Each MAPP can also contain a record of the author and any relevant literature references. GenMAPP does not restrict users to particular semantics. A MAPP can represent any gene set whether it is a metabolic pathway, a signaling pathway, a disease process or an arbitrary set. The pathway archives GenMAPP distributes undergo general review and revision by the GenMAPP staff.

Databases and pathway archives are available through the Data Acquisition Tool in GenMAPP and from the GenMAPP website. The tools known as MAPPFinder 2 and MAPPBuilder 2 are bundled with and accessible from GenMAPP. MAPPBuilder creates .mapp files from imported lists of genes, and MAPPFinder (Doniger *et al.* 2003) computes permutation test P values for over-representation of differentially expressed genes in individual GO categories and MAPPs. Westfall-Young adjusted P values (Westfall *et al.* 1993) are included as a control for multiple testing.

2.4 Results and Discussion

GenMAPP version 2 provides 1) new built-in features to support user data import and mapping, 2) expanded pathway resources and 3) increased support for different high-throughput biological assays. These improvements substantially increase the usability and flexibility of this tool for pathway level genomic analysis.

2.4.1 GenMAPP version 2 new features

Several new features have been implemented in GenMAPP version 2. A new gene database schema supports a variety of gene and protein identifiers, annotations, and microarray probe set IDs, more thoroughly connecting user data to the archive of pathway MAPPs and Gene Ontology terms and to external gene annotation. A new visualization mode allows for simultaneous access to multiple data points, statistics or custom annotations. A new export option packages sets of pathways, including data, to a web-ready format for display and browsing.

2.4.2 Expanded gene and species support in GenMAPP version 2

A major shortcoming of GenMAPP 1.0 and other pathway analysis programs has been the limited number of species supported, permitting analysis of a few model organisms (human, mouse, rat, and yeast) and a few gene identifier or ID systems (GenBank, SGD, and UniProt). To solve this problem, the GenMAPP version 2 gene database schema has been redesigned to allow expanded gene content and greater species support. Support of many diverse gene and protein ID systems is essential to establish critical relationships between disparate sources of information, providing greater flexibility for users importing data associated with virtually any identifier. In addition to expanded gene and protein ID support, secondary annotation systems such as GO, OMIM, and PDB have been added into the GenMAPP gene databases. These IDs and annotations are provided on HTML “backpages” of MAPP gene objects, providing critical links to primary resources. As additional genomes are assembled and annotated, GenMAPP can readily integrate the information and support pathway analysis for these species.

Databases in GenMAPP version 2 are created through a semi-automated process, using information extracted from major public resources, primarily Ensembl, Entrez Gene, UniProt, and Affymetrix. The process of extracting gene information has been greatly simplified by populating our gene database with data from Ensembl's "mart" tables (Kasprzyk *et al.* 2004), which effectively integrates gene information for major sequenced genomes. GenMAPP.org currently distributes databases for eleven species: human, mouse, rat, yeast, worm, zebrafish, fruit fly, mosquito, chicken, dog, and cow. GenMAPP version 2 also supports user-defined additions to these databases as well as the creation of custom gene databases for any other species. The ability to create custom databases is of vital importance to research groups working with model organisms not supported by the major public databases. This feature is supported by only one other pathway analysis tool we are aware of (Hu *et al.* 2005). Creating a custom database is a collaborative effort where GenMAPP developers generate a template database containing relevant GO term associations for the species of interest. A user interface within GenMAPP version 2 allows users to add to the template database by importing additional gene and annotation information as a set of relational tables. The build process can be completed entirely using GenMAPP and common spreadsheet programs (e.g., Excel), without the need for specialized database software. The resulting database has full GenMAPP functionality, including the ability to display information on HTML backpages, link to external sources, and perform global GO queries using MAPPFinder. Custom GenMAPP version 2 databases are currently available for *Escherichia coli* K12 (KDD and John David N. Dionisio, personal communication) and *Saccharomyces pombe* (http://www.databases.niper.ac.in/Pombe/ S.pombe gene database for

GenMAPP). A detailed manual describing the process of creating a custom gene database is available at GenMAPP.org.

2.4.3 Visualizing complex genomic data

As microarray experimental designs grow increasingly complex researchers require tools to examine data across multiple time-points and conditions, and over multiple datasets. The types of biological entities measured have also increased. Various array platforms measure polymorphisms, splice variants, regulatory protein binding and genomic amplifications and deletions. Methods for visualizing the complex outputs from these technologies have not been well established and remain a critical challenge for researchers. With previous versions of GenMAPP, users could view multiple sets of criteria only serially. For example, genes up-regulated at different time-points of an experiment could be viewed by creating a custom set of coloring criteria (color set) for each time point. While informative, this method is not well suited to assess the temporal effects of gene level changes over an extended time or to examine multiple data simultaneously. To expedite the analysis of such datasets, GenMAPP version 2 now allows multiple color sets to be viewed simultaneously, depicted as vertical stripes within each gene box. In the case of multiple time points, the stripes could represent the criteria at each time point (Figure 2.5).

The ability to view multiple color sets concurrently can also be extended to datasets where different biological substrates are examined, such as transcription and mRNA splicing. Demand for this feature is increasing because current microarrays can assay distinct regions of mRNA transcripts, such as exons and exon junctions, thereby allowing assessment of both transcriptional changes and changes in splice isoform expression. While there are many possible ways to view such data, using multiple color sets in GenMAPP is now a powerful way to explore such complex data in a single view. Similar visualization options are only available in a few freely available (Yi *et al.* 2006) and commercial applications (Chu *et al.* 2001; Ekins *et al.* 2007).

2.4.4 Batch export of data to the web

In addition to visualizing data on pathway MAPPs, GenMAPP version 2 also exports pathways with data to various graphical formats and to the web. Because genome-scale data are difficult to share with a larger community, GenMAPP version 2 includes the option to export any number of MAPPs with their associated data to an organized web-ready format. This MAPP Set Export feature allows any or all established color sets to be exported with the pathway, including the striped view of multiple color sets. Instead of static images, each MAPP retains its interactive features, such as gene backpage information, including data display, gene annotations, and hyperlinks to external resources. The different criteria can be browsed through a pull-down tab on each exported MAPP. The MAPP Set can be navigated through an index of all MAPPs or through a gene index, which stores all gene-to-MAPP relationships for all related gene/protein IDs. MAPP Sets are stored in HTML format, ready for immediately posting on any web site, where collaborators can browse the data independently of the GenMAPP program. An example of how a GenMAPP MAPP Set can be used to display large-scale data is the International Gene Trap Consortium web site (http://www.genetrap.org), where thousands of publicly available gene trap ES cell lines can be viewed in the context of biological pathways (http://www.genetrap.org/dataaccess/pathways.html). This method of data presentation allows users to quickly share information over the Internet and perform efficient searches for gene pathway information. Batch export of fully interactive pathways and user data is not available in other pathway analysis tools we are aware of (http://cancer.cellmap.org/cellmap/; http://www.ingenuity.com/; Chu *et al.* 2001; Shannon *et al.* 2003; Hu *et al.* 2005; Mlecnik *et al.* 2005; Yi *et al.* 2006; Yuryev *et al.* 2006; Ekins *et al.* 2007; Mi *et al.* 2007).

2.4.5 New Pathway Resources

Integral to any pathway analysis tool is its access to pathway content. One of the goals of the GenMAPP project is to facilitate community curation of pathway content. GenMAPP's built-in drawing tool allows users to illustrate biology and associate gene objects with identifiers maintained in a given gene database. The ability to customize the layout and to annotate a pathway with basic graphics provides a powerful means of communication to the biological community. The expertise of the biological research community is the most important source of new pathway information, and GenMAPP's pathway content is primarily contributed by this community. We have added several new sources of MAPPs.

For example the NetPath project is a human pathway annotation project, initiated by the Pandey lab at Johns Hopkins University (http://pandeylab.igm.jhmi.edu; http://www.netpath.org) and the Institute of Bioinformatics (http://www.ibioinformatics.org). The NetPath group has produced 10 cancer and 10 immune pathways in GenMAPP, BioPAX (http://www.biopax.org), and PSI-MI (Hermjakob *et al.* 2004) formats, and are planning a substantial increase within the first year. Another ongoing pathway curation effort is being performed by undergraduate research students directed by Dr. Kam Dahlquist. These students have contributed 120 yeast pathways that were created by hand using the SGD BioCyc metabolic pathways (http://pathway.yeastgenome.org/biocyc/) as templates. The GenMAPP pathway archives also include selected content from KEGG (Kanehisa *et al.* 2000), Reactome (Joshi-Tope *et al.* 2005; Vastrik *et al.* 2007), The European Nutrigenomics Organization (http://www.bigcat.nl), Neurocrine Biosciences, PharmGKB (Hewett *et al.* 2002), and various academic laboratories. The content from these resources was manually migrated by the MAPP authors with the exception of the “KEGG Converted” archive, which is not updated or synchronized. The pathways from community resources are collected and organized at GenMAPP.org and automatically downloadable through the GenMAPP program.

We now also provide pathways that have been mapped through homology so that users with genomic data from relatively unsupported species can perform pathway analyses. These homology MAPPs represent a starting point for further curation, an interim solution until species-specific pathways are elucidated and contributed. Another means of increasing the biological content available to the user is the extension of existing pathways using interaction and coexpression data. Together, these methods only begin to address the paucity of pathway content available for the analysis of complex genomics data across the multitude of organisms.

2.4.6 Making homology MAPPs

Despite the relative ease with which we can gather gene information for many species, pathway information is generally not available for many of the newly supported species (Table 2.1). To address this problem, we implemented a strategy that utilizes the existing pathway content in our pathway archives. Using publicly available gene homology information (Kasprzyk *et al.* 2004; Wheeler *et al.* 2006) , we generated pathways for several species from our archive of existing human pathways (Figure 2.1).

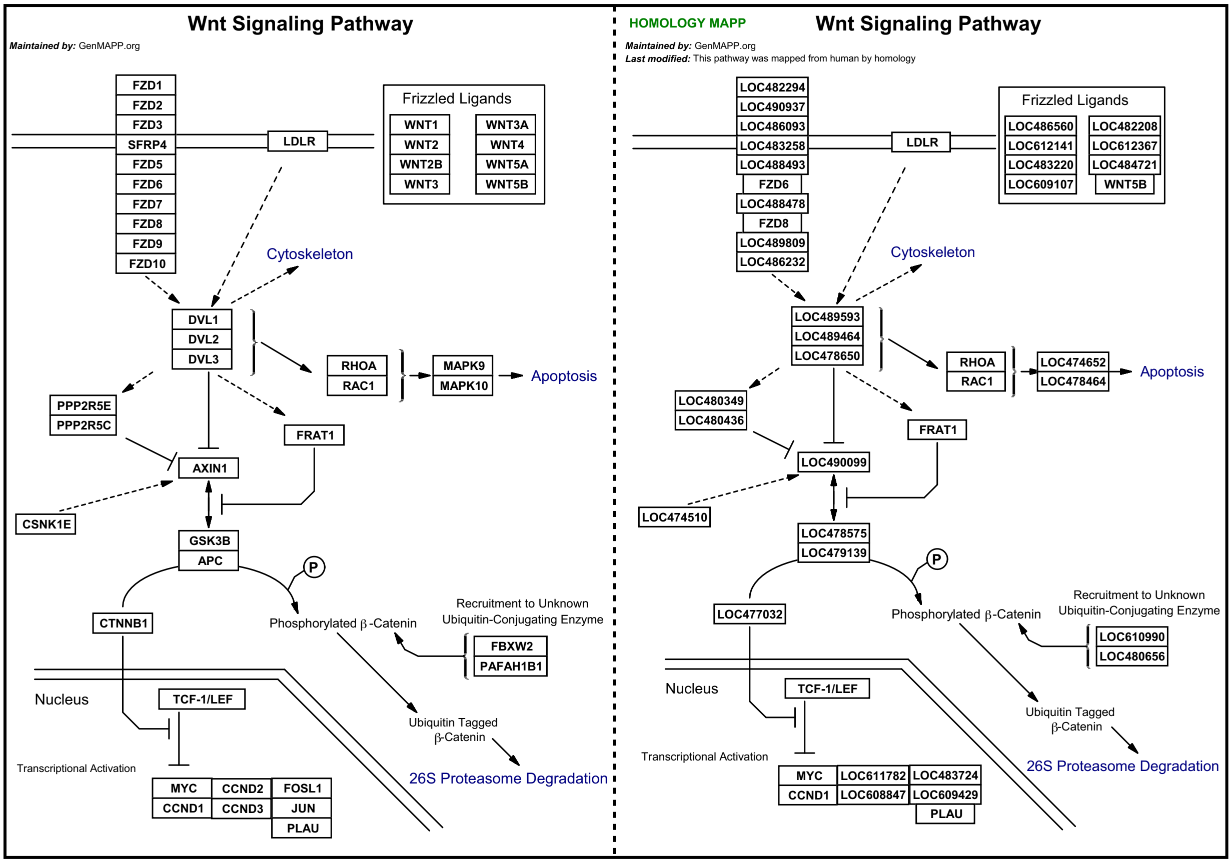


Figure 2.1. The WNT-signaling pathway is shown in GenMAPP for human and dog (left to right). The dog pathway MAPP was mapped from the original curated human pathway MAPP by using homology information from Homologene and Ensembl. Additional information in the top-left corner of the MAPP indicates the origin.

The process of rapidly mapping pathways between species relies on the Converter function in GenMAPP version 2, which allows for conversion of genes on MAPPs between gene ID systems in the database without altering the graphical layout of the MAPP. MAPP conversion is possible between any gene ID systems linked in the database; adding homology information to a GenMAPP database consequently enables conversion of MAPPs between species.

The GenMAPP human MAPPs were chosen as template MAPPs because they represent the highest quality of curation in our archive. Homology information between human and the applicable target species was obtained from Homologene (Wheeler *et al.* 2006) and form BioMart (Spudich *et al.* 2007)(for cow only)(Algorithm details in Supplemental Data). For simplicity we restricted the use of data from these resources to 1:1 gene relationships between template and target species. This restriction reduces clutter in the converted MAPPs and avoids potentially ambiguous homology relationships. Conversion rates (percentage of genes converted) were calculated for each pathway MAPP (Figure 2.2 and Supplemental Data). To maintain reasonable gene content on MAPPs, a cutoff of 50% for the conversion rate was set for inclusion in the MAPP archives. The cutoff of 50% was chosen based on the qualitative assessment of structure and pathway information retained following conversion (see supplemental data for MAPP conversion rates). Qualitatively, the conversion rates correlated with the expected conservation of biological processes across species. The MAPPs representing the central dogma of DNA replication, RNA transcription and translation, for example, were converted with high fidelity from human to each of the target species, whereas specialized signaling pathways failed to be translated beyond dog, cow and chicken. This strategy of utilizing public homology information, existing pathway information and the Converter function can be applied to any species with available homology information to a species for which pathway information exists. Instructions for translating MAPPs between species using the GenMAPP Converter is available at GenMAPP.org (http://www.genmapp.org/tutorials/Converting-MAPPs-between-species.pdf).

Species	Contributed	Homology	KEGG Converted
Human	109		
Mouse	109		
Rat	100		
Dog		94	
Cow		87	
Chicken		85	
Zebrafish		19	18
Fruit fly	2	25	89
Mosquito			
Worm		19	87
Yeast	122*	9	

Table 2.1. Number of GenMAPP MAPPs for GenMAPP supported species. Column Headers: Contributed: MAPPs contributed to the GenMAPP project. Homology: MAPPs mapped from human pathways using homology information. KEGG Converted: MAPPs automatically created from the KEGG resource. Note: *Includes 120 SGD metabolic MAPPs contributed by undergraduate research students mentored by Dr. Kam Dahlquist.

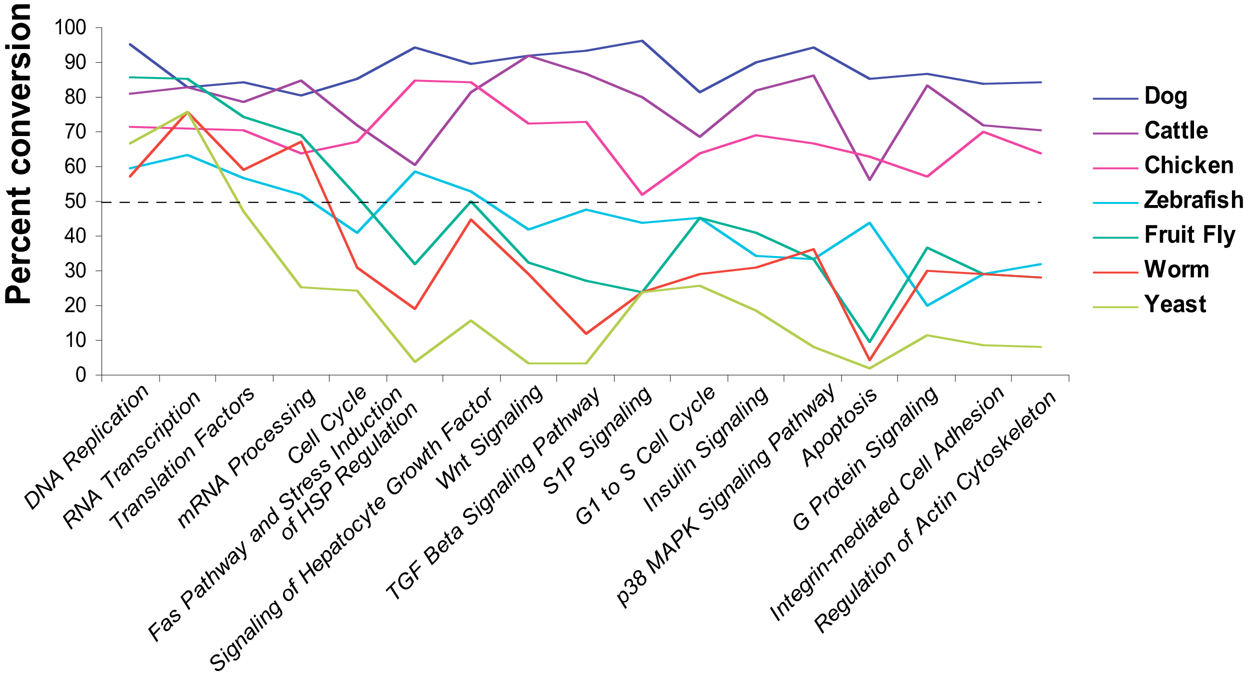


Figure 2.2. Conversion rate (percentage of genes converted) for MAPPs in the Cellular Process category of the Contributed archives. Colored lines indicate conversions per species; dashed line indicates the 50% cutoff for inclusion in the Homology MAPPs archive. As expected, highly conserved processes showed high conversion rates across species (far left), while more specialized processes were homologous only among more closely related species.

The development of homology MAPPs in GenMAPP builds upon similar efforts at other databases(Vastrik *et al.* 2007) and addresses the dearth of pathway content that can be queried computationally. However, it is important to note that these MAPPs are not genuine species-specific pathways, but rather translations of human pathways where target species genes have been mapped based on homology. This distinction is important since accurate pathway inference requires knowledge that the particular biological process and molecular interactions are conserved between organisms and that predicted homologues encode for gene products that perform the same biological function. Another current limitation is that, unlike several other resources (Mao *et al.* 2006; Wu *et al.* 2006; Vastrik *et al.* 2007), the reactions in a GenMAPP pathway are illustrations rather than computable networks that allow for identification of conserved interactions. Furthermore, pathways for non-mammalian species are mapped from human rather than the most closely related organism. As such, these homology MAPPs are by no means equal to the quality of manually curated MAPPs. For that reason homology MAPPs are distributed as a separate archive, accompanied by a README file explaining the nature of these MAPPs. They nonetheless offer an immediate and concrete solution for many researchers studying organisms with minimally annotated genomes not supported by other analysis programs. It is our hope that these pathways will serve to nucleate additional curated pathways. Furthermore, the information provided by pathway representations of known biology, especially for minimally annotated genomes, is crucial not only for analyzing large-scale datasets, but also for assigning gene function.

2.4.7 Extending pathways

The use of homology mapping addresses the critical need to extend biological representations across species. Yet it is also necessary to expand the pathway content within a given species. In the case of mammalian model organisms, such as mouse, only ~14% of annotated genes in the genome are represented in curated pathways (from the combined archives of GenMAPP, KEGG, BioCarta (http://www.biocarta.com) and Reactome). Figure 2.3 illustrates the collection of curated pathways in the GenMAPP archive over time, which, in terms of gene content is >90% redundant with BioCarta and Reactome. The collection of curated pathways from the scientific community is a slow, iterative process that requires the synthesis of a variety of evidence. Such evidence is being cataloged in numerous databases as protein-protein interactions, genetic interactions, and coexpression patterns, which are rapidly expanding with the advent of large-scale, high-throughput assays. But it remains a challenge to form meaningful networks from this data and grow our understanding of pathways.

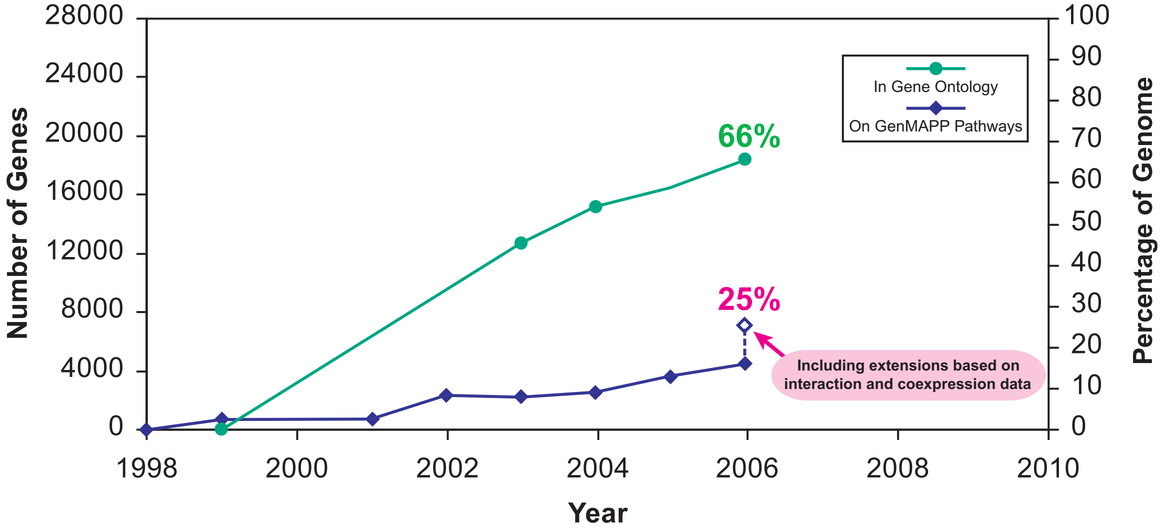


Figure 2.3. Number of mouse genes represented on GenMAPP Pathways and in Gene Ontology. The unique gene content in the GenMAPP pathway archive is traced over time (blue) for the mouse genome in terms of the number of genes (left axis) and corresponding percentage of the genome (right axis). For comparison, the unique gene content annotated by Gene Ontology is shown (green). Significant gains in absolute gene content were made by collecting new pathways targeting new biology (e.g., NetPath) and by extending pathways with orthogonal datasets from coexpression and protein-protein interaction networks (latest GenMAPP count at 7095 genes, or ~25% of the genome).

To address this challenge, we created a new pathway resource, which incorporates additional genes into our existing set of pathways using prior evidence. This method of pathway extension has been previously used to include new genes predicted to expand and enhance the content of existing pathways and gene sets (Novak *et al.* 2006). The method can work with any type of data that can be modeled as pairs of linked genes. The most obvious example is protein-protein interactions, where the link between genes represents the physical association of two proteins. The link could also represent coexpression, transcriptional regulation, or literature search results. The extension method is currently implemented as a set of in-house Perl scripts used as an accessory to GenMAPP to expand a given pathway. Each pathway MAPP is processed individually. First, the gene IDs are extracted from the pathway and converted to a uniform ID system (e.g., Entrez Gene). The resulting gene list is used to query one or more specified databases (e.g., protein-protein interactions). A threshold is set for including new genes (e.g., one or more links to original gene list). Finally, the new genes are added to a side panel of the original MAPP, separate from the curated pathway, and the interaction partners are noted and stored in the remarks field of each involved gene (Figure 2.4).

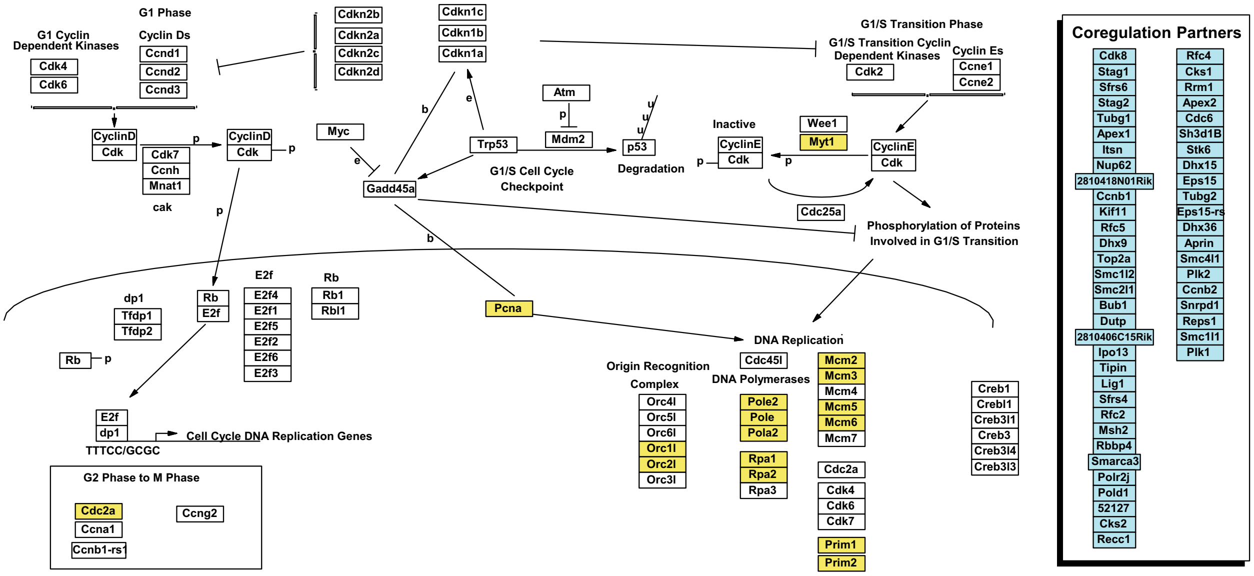


Figure 2.4. G1 to S cell-cycle control pathway extended with genes from a coexpression network. All genes assigned to the original pathway were queried against the coexpression network. Yellow designates the genes found in the coexpression network and blue designates their coexpression partners that were extracted from the network and added to the pathway.

Using this approach, we extended the GenMAPP curated pathway archives for mouse with two types of data: protein-protein interactions and coexpression data (Stuart *et al.* 2003) (see supplemental data). The coexpression links were derived from a network analysis of correlated gene expression across multiple species networks (Stuart *et al.* 2003) under the premise that genes that maintain an evolutionary conservation of coregulation often participate in a related biological process (van Noort *et al.* 2003; Bergmann *et al.* 2004). With the additional genes added from these datasets, we have significantly increased the coverage (~25%) per genome (Figure 2.3). It is important to distinguish the added genes from those originally in the pathway since the added genes are not necessarily *involved* in the pathway; rather, they are *related* to the pathway by a particular type of evidence. Having access to this related information in the same view as the pathway allows for simultaneous data visualization and statistical analysis using MAPPFinder. These extended pathways may also serve as launching points for improved pathway curation by the community and as a predictive method for identifying new pathway interactions.

2.4.8 Examples of pathway analysis

Here we explore three of the many examples of how GenMAPP version 2 can be used to analyze data from complex genomic experiments and the types of biological insights potentially gained.

2.4.8.1 Gene expression time course analysis

In figure 2.5, we display gene expression data from multiple time-point comparisons for the myometrium during gestation(Salomonis *et al.* 2005). There are two baselines in this analysis: virgin non-pregnant (NP) myometrium and mid-pregnancy myometrium. The comparison allows the user to simultaneously examine the effects of pregnancy as compared to non-pregnant animals and the specific temporal effects leading up to labor through postpartum.

The prostaglandin synthesis and regulation pathway contains molecular interactions that are critical in the transition of the myometrium from a relatively quiescent tissue throughout pregnancy to a highly contractile tissue at term. By viewing multiple time-point comparisons in this pathway, one can easily see which genes are differentially expressed just prior to the onset of labor (18 days of pregnancy) compared to mid-pregnancy (14 days of pregnancy) (e.g. Ptgs2, Edn1 and Hsd11b1) alongside the relative expression of these genes at mid-pregnancy versus the virgin state (first stripe). Making such comparisons in the new version of GenMAPP is relatively straightforward and flexible, supporting not only multiple data points, but also multiple types of data (see SNP example, figure 6c). In GenMAPP version 2, the user can also select any combination of color sets to view on a given MAPP simply by selecting them from the “Choose Color Set” pull-down.

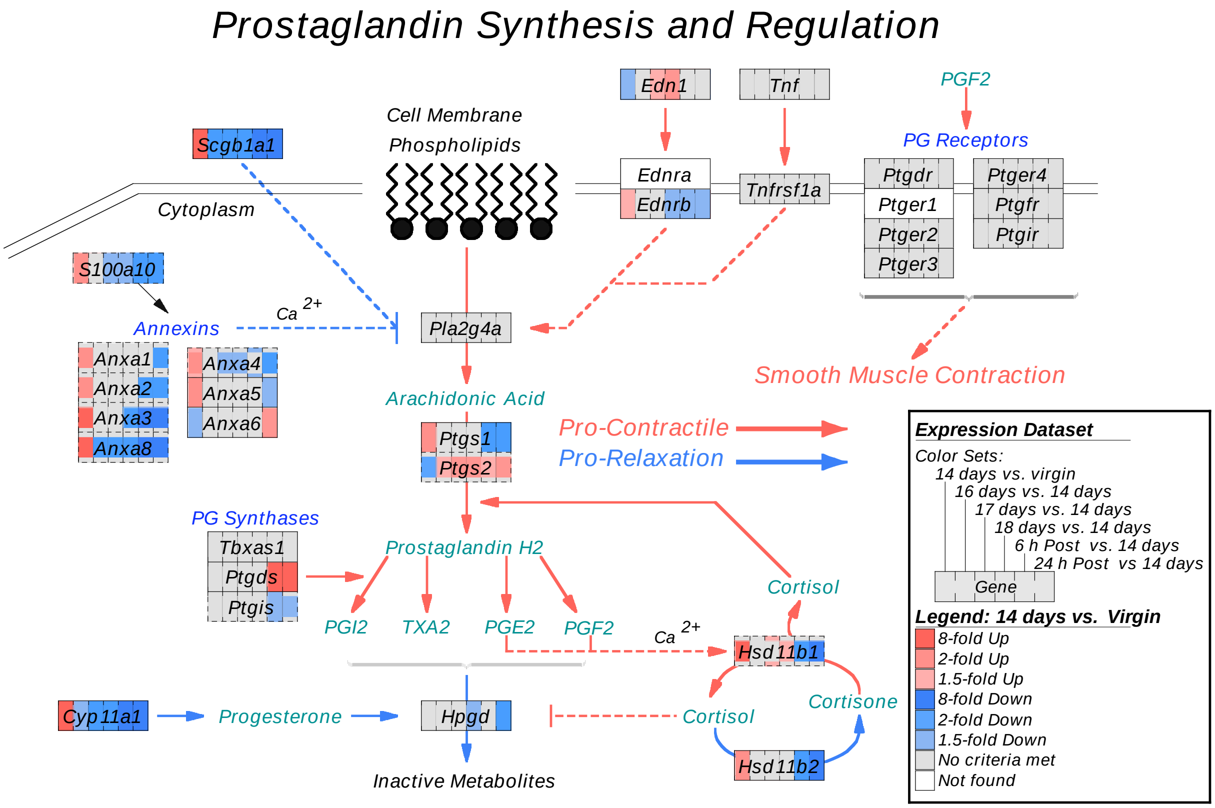


Figure 2.5. Striped view of multiple time-point comparisons. Gene expression data from mouse uterine smooth muscle from mid-to-late pregnancy through postpartum are shown as striped color sets on a pathway for prostaglandin synthesis and regulation. For each comparison, a separate color set was generated, with eight colors designating different fold thresholds. The order of the criteria dictates the priority for how a gene box is colored. For this dataset, the striped view allows comparison of expression changes that are predicted to promote versus block contraction during distinct phases of pregnancy. Multiple probe sets from the microarray linking to a gene are indicated by a dashed edge around the gene box. The central color of the gene box corresponds to the predominant criterion met (mode) and the rim colors represent the second most prevalent criterion met.

2.4.8.2 Analysis of whole-genome exon array data

As the feature size of DNA microarrays have decreased, the number of probes hybridizing to specific targets has increased by well over an order of magnitude. In the example shown in Figure 2.6 A, we examined a publicly available microarray dataset that measured the expression of all known and predicted exons from 11 different adult human tissues (Figure 2.6 A) (http://www.affymetrix.com/support/technical/sample_data/exon_array_data.affx). From these data, both gene expression changes between tissues and splicing scores can be calculated for all genes (see supplemental methods). GenMAPP version 2 can display this information in each gene box, with the central color stripes indicating relative expression change for each tissue (red or blue) and the rim color designating a threshold for the significance of an alternative splicing call (green, gray, or white). This strategy takes advantage of how GenMAPP prioritizes assignment of central and rim colors of a gene box based on the order of the underlying data. Viewing related identifiers to a given gene as a secondary rim criterion can provide critical information to the analysis and is a unique feature of GenMAPP. When viewed in the context of Monoamine G-protein coupled receptors, we can clearly identify in which tissues a gene is most highly expressed (bright red center color) and which genes have a significant alternative splicing call (green rim color). By creating a color set for each of the 11 tissues and selecting “all” for visualization, both the tissue specific regulation of gene expression and the likelihood of splicing can be assessed in a single view. The results from this dataset can be exported for any given set of pathways with web-ready images and HTML backpages for each and every gene. The web export function allows researchers to navigate and effectively communicate the impact of both gene expression and splicing on specific pathways and genes (see the GenMAPP website (http://www.genmapp.org/multiple_cs.html) for this example and others).

2.4.8.3 Combining proteomic and gene expression data

In another example, gene expression and proteomic data (Griffin *et al.* 2002) is viewed concurrently as two adjacent stripes of color (Figure 2.6 B). The example displays data from an experiment measuring both mRNA and protein levels in yeast in response to changes in carbon source. Simultaneously visualizing changes at the transcript and protein level in the context of pathways represents a more informative depiction of the system-level changes occurring in the organism than if either data was analyzed alone. The flexibility of combining any number of disparate data types in a single view is a relatively uncommon feature in pathway analysis tools. To view two data types side by side, datasets are combined into a single spreadsheet before import into GenMAPP. There are no restrictions on the nature of data that can be viewed as independent, adjacent color sets, provided that the data links to the GenMAPP gene database.

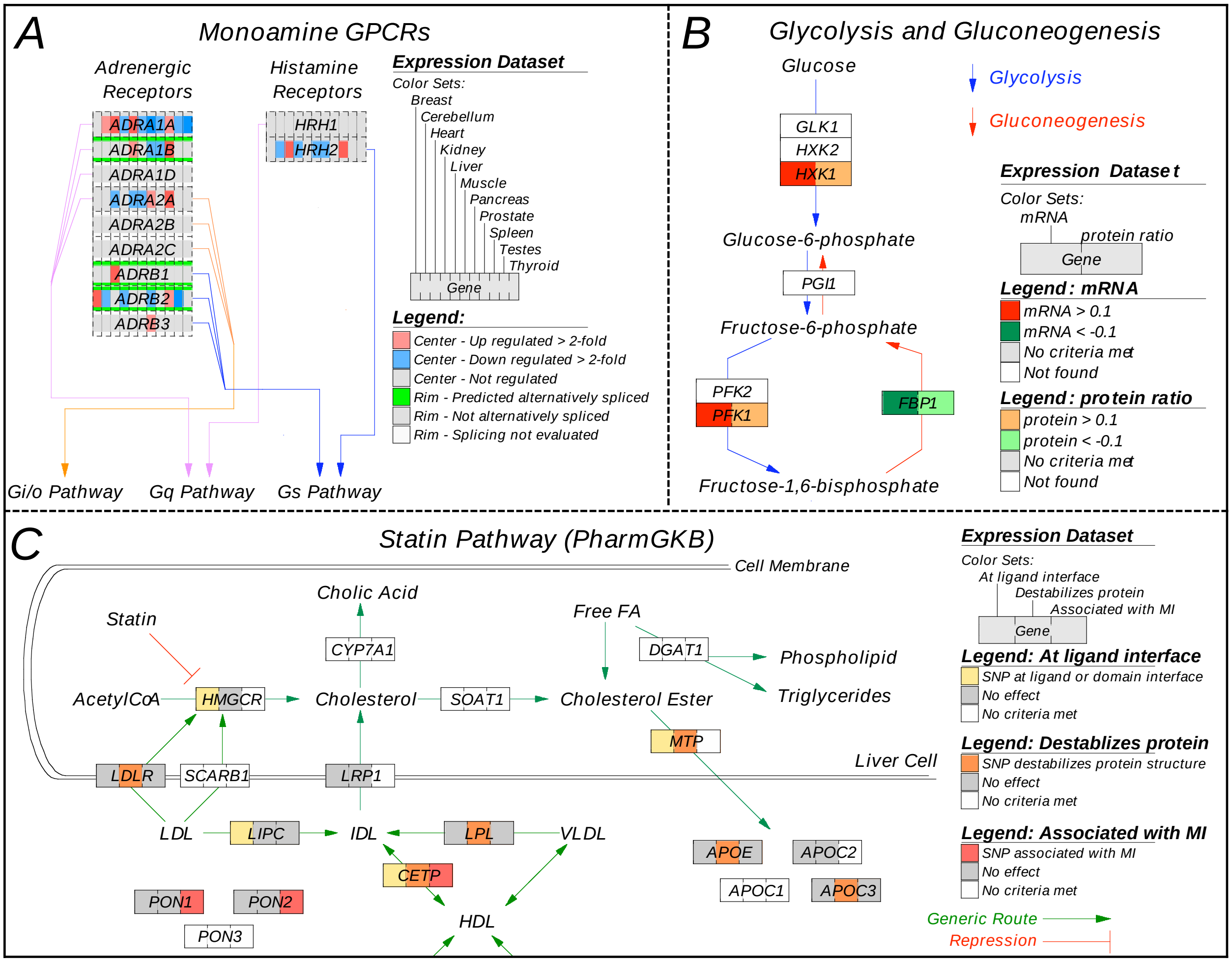


Figure 2.6. Striped view of multiple data types. (A) Transcription and splicing data, collected on whole-genome exon tiling microarrays (see supplemental methods), are represented by stripes of color on a functionally organized list of monoamine G-protein-coupled receptors. Transcriptional changes for 11 different human tissues are displayed as the center color of the gene box, and splicing for the gene across all tissues is displayed as the rim color. (B) In the context of glycolysis and gluconeogenesis, mRNA and protein levels change in response to carbon source perturbation in Saccharomyces cerevisiae growing on galactose or ethanol. The color on the left side of the gene box illustrates mRNA changes; the color on the right indicates corresponding protein-level changes. (C) A variety of SNP parameters can be viewed simultaneously using the striped view. SNP distribution (dbSNP www.ncbi.nih.gov/SNP), structure-based functional

predictions (LS-SNP alto.compbio.ucsf.edu/LS-SNP/), and myocardial infarction (MI) association data are combined to asses the coverage of the SNP panel over a pathway depicting the role of statin drugs. To view the complete versions of these MAPPs with live backpages see Supplemental Data.

2.4.8.4 Integrating genomic, phenotypic and structural information for polymorphism data

One of the key principles of pathway analysis is the integration of multiple pieces of information in order to assess new data in the context of known biology. In studying polymorphic, or SNP, differences that may contribute to disease, the ability to compare the distribution of polymorphisms in the population along with phenotypic and protein product effects in the context of biological pathways provides both a birds-eye view and detailed dissection of how specific changes might impact larger biological systems. An example of how these different types of biological data can be combined is shown in Figure 2.6 C using data from a whole-genome myocardial infarction SNP array experiment (Tobin *et al.* 2004). Displaying data in this format highlights genes evidenced by association, experimental and bioinformatics predictions (e.g. CETP, MTP) as well as their relationship to each other and with other genes upstream and downstream of these components. Display formats such as this allow access to multiple modes of gene regulation from a single display.

Although these examples illustrate three possible methods for displaying complex results, users can customize such views and apply them to any combination of data types that have been merged and ordered before import to

GenMAPP. This feature provides a means to assess multiple modes of gene regulation and thus new avenues of insight into complex biological relationships.

2.4.8.5 **Ongoing development of GenMAPP**

GenMAPP version 2 provides new tools for analyzing complex data in the context of biological pathways for a variety of genomes. Although the new features of GenMAPP version 2 are a useful starting point for the analysis of complex microarray data, there are still a number of obstacles to overcome. These obstacles include providing cross-platform tools for integrating pathway resources, representing gene features (such as SNPs and splicing variation), and supporting structured pathway vocabularies for more efficient pathway migration, update, curation and exchange.

To accelerate development and take full advantage of the growing base of open source pathway tools we are actively working with the Cytoscape Consortium (www.cytoscape.org)(Shannon *et al.* 2003) and BioPAX (www.biopax.org) developers to implement GenMAPP-style visualization and analysis methods in a new software framework. The primary aims are (1) to transition to a platform-independent Java code base that is readily integrated with online resources, (2) to support dynamically generated gene databases that not only organize identifiers and aliases, but also sub-gene entities such as transcripts, exons, and polymorphisms, and (3) to provide innovative analysis tools to preprocesses high-throughput datasets preparing them for integration with gene databases and statistical analyses, as well as for abstracted visualization at multiple levels of resolution. We are also working on an XML-based pathway data format that captures relationships, coordinates, and annotations, as well as a Web tool that facilitates pathway content migration, and curation from the community. We anticipate that open source bioinformatics tools such as GenMAPP and Cytoscape will provide researchers with a new view of biology that integrates genomic data with our growing knowledgebase of pathways.

2.5 Conclusions

GenMAPP version 2 represents a step towards fostering the critical link between the biologist and their data, providing powerful analyses and intuitive representations of increasingly large and complex high-throughput datasets.

2.6 Availability and requirements

Project Name: GenMAPP

Project Home Page: http://www.genmapp.org

Operating System: Windows

Programming Language: Visual Basic

Requirements: Species-specific databases and pathway file collections distributed by GenMAPP.org

License: Open-source (Apache)

Any Restrictions to Use by Non-academics: None

2.7 Authors' contributions

GenMAPP version 2 features were conceived by BC, KD, AP, NS, SD, KH, KV, AZ and SL. Computer code for the GenMAPP version 2 application was written by SL and SD. AP, NS and KH drafted the manuscript. The pathway extension method was performed by AP and JS; and homology mapping was performed by KH. All authors read and approved the final version of the manuscript.

2.8 Acknowledgments

We thank the GenMAPP user community for valuable suggestions, pathway submissions, and the inspiration to keep improving GenMAPP. We thank Jeff C Lawlor for his work on GenMAPP installer and downloader and Stephen Ordway and Gary Howard for editorial assistance. This work was supported by the J. David Gladstone Institutes, American Heart Association grant 0425278Y (to ACZ), National Human Genome Research Institute grant HG003053 (to BRC), and National Heart, Lung, and Blood Institute grant HL66621 (to BRC). ARP was supported by the David R. and Mary Phillips Distinguished Postdoctoral Fellowship.

2.9 References

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

Bergmann, S., J. Ihmels and N. Barkai (2004). "Similarities and differences in genome-wide expression data of six organisms." PLoS Biol **2**(1): E9.

Chu, L., E. Scharf and T. Kondo (2001). "GeneSpring: Tools for Analyzing Microarray Expression Data." Genome Informatics **12**: 227-229.

Dahlquist, K. D., N. Salomonis, K. Vranizan, S. C. Lawlor and B. R. Conklin (2002). "GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways." Nat. Genet. **31**: 19–20.

Doniger, S. W., N. Salomonis, K. D. Dahlquist, K. Vranizan, S. C. Lawlor and B. R. Conklin (2003). "MAPPFinder: Using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data." Genome Biol. **4**: R7–R7.12.

Ekins, S., Y. Nikolsky, A. Bugrim, E. Kirillov and T. Nikolskaya (2007). "Pathway mapping tools for analysis of high content data." Methods Mol Biol **356**: 319-50.

Griffin, T. J., S. P. Gygi, T. Ideker, B. Rist, J. Eng, L. Hood and R. Aebersold (2002). "Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae." Mol Cell Proteomics **1**(4): 323-33.

Hermjakob, H., L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering, B. Roechert, S. Poux, E. Jung, H. Mersch, P. Kersey, M. Lappe, Y. Li, R. Zeng, D. Rana, M. Nikolski, H. Husi, C. Brun, K. Shanker, S. G. Grant, C. Sander, P. Bork, W. Zhu, A. Pandey, A. Brazma, B. Jacq, M. Vidal, D. Sherman, P. Legrain, G. Cesareni, I. Xenarios, D. Eisenberg, B. Steipe, C. Hogue and R. Apweiler (2004). "The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data." Nat Biotechnol **22**(2): 177-83.

Hewett, M., D. E. Oliver, D. L. Rubin, K. L. Easton, J. M. Stuart, R. B. Altman and T. E. Klein (2002). "PharmGKB: the Pharmacogenetics Knowledge Base." Nucleic Acids Res **30**(1): 163-5.

http://cancer.cellmap.org/cellmap/. "CellMap." from http://cancer.cellmap.org/cellmap/.

http://pandeylab.igm.jhmi.edu. "Pandey Lab." from http://pandeylab.igm.jhmi.edu.

http://pathway.yeastgenome.org/biocyc/. from http://pathway.yeastgenome.org/biocyc/.

http://www.affymetrix.com. "Affymetrix." from http://www.affymetrix.com.

http://www.affymetrix.com/support/technical/sample_data/exon_array_data.affx. "Affymetrix-Exon Array Dataset." from http://www.affymetrix.com/support/technical/sample_data/exon_array_data.affx.

http://www.bigcat.nl. "BiGCaT Bioinformatics." from http://www.bigcat.nl.

http://www.biocarta.com. "Biocarta - Charting Pathways of Life." from http://www.biocarta.com.

http://www.biopax.org. "BioPAX Home." from http://www.biopax.org.

http://www.databases.niper.ac.in/Pombe/ (S.pombe gene database for GenMAPP).

http://www.ensembl.org. "Ensembl Genome Browser." from http://www.ensembl.org.

http://www.genetrap.org. "IGTC, International Gene Trap Consortium." from http://www.genetrap.org.

http://www.genetrap.org/dataaccess/pathways.html. "IGTC, International Gene Trap Consortium." from http://www.genetrap.org/dataaccess/pathways.html.

http://www.genmapp.org/multiple_cs.html. "Visualizing Multiple Color Sets." from http://www.genmapp.org/multiple_cs.html.

http://www.genmapp.org/tutorials/Converting-MAPPs-between-species.pdf. "Converting GenMAPP MAPPs between species using homology." from http://www.genmapp.org/tutorials/Converting-MAPPs-between-species.pdf.

http://www.ibioinformatics.org. "Institute of Bioinformatics." from http://www.ibioinformatics.org.

http://www.ingenuity.com/. "Ingenuity Systems." from http://www.ingenuity.com/.

http://www.ncbi.nlm.nih.gov/entrez. "Entrez PubMed." from http://www.ncbi.nlm.nih.gov/entrez.

http://www.netpath.org. "NetPath - Signal Transduction Pathways." from http://www.netpath.org.

http://www.pir.uniprot.org. "Welcome to UniProt-UniProt [the Universal Protein Resource]." from http://www.pir.uniprot.org.

Hu, Z., J. Mellor, J. Wu, T. Yamada, D. Holloway and C. Delisi (2005). "VisANT: data-integrating visual framework for biological networks and modules." Nucleic Acids Res **33**(Web Server issue): W352-7.

Joshi-Tope, G., M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney and L. Stein (2005). "Reactome: a knowledgebase of biological pathways." Nucleic Acids Res **33**(Database issue): D428-32.

Kanehisa, M. and S. Goto (2000). "KEGG: Kyoto encyclopedia of genes and genomes." Nucleic Acids Res. **28**: 27–30.

Kasprzyk, A., D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox and E. Birney (2004). "EnsMart: a generic system for fast and flexible access to biological data." Genome Res **14**(1): 160-9.

Kitano, H., A. Funahashi, Y. Matsuoka and K. Oda (2005). "Using process diagrams for the graphical representation of biological networks." Nat Biotechnol **23**(8): 961-6.

Kumar, S. P. and J. C. Feidler (2003). "BioSPICE: a computational infrastructure for integrative biology." Omics **7**(3): 225.

Mao, F., Z. Su, V. Olman, P. Dam, Z. Liu and Y. Xu (2006). "Mapping of orthologous genes in the context of biological pathways: An application of integer programming." Proc Natl Acad Sci U S A **103**(1): 129-34.

Mi, H., N. Guo, A. Kejariwal and P. D. Thomas (2007). "PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways." Nucleic Acids Res **35**(Database issue): D247-52.

Mlecnik, B., M. Scheideler, H. Hackl, J. Hartler, F. Sanchez-Cabo and Z. Trajanoski (2005). "PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways." Nucleic Acids Res **33**(Web Server issue): W633-7.

Novak, B. A. and A. N. Jain (2006). "Pathway recognition and augmentation by computational analysis of microarray expression data." Bioinformatics **22**(2): 233-41.

Salomonis, N., N. Cotte, A. C. Zambon, K. S. Pollard, K. Vranizan, S. W. Doniger, G. Dolganov and B. R. Conklin (2005). "Identifying genetic networks underlying myometrial transition to labor." Genome Biol **6**(2): R12.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome Res **13**(11): 2498-504.

Spudich, G., X. M. Fernandez-Suarez and E. Birney (2007). "Genome browsing with Ensembl: a practical overview." Brief Funct Genomic Proteomic **6**(3): 202-19.

Stuart, J. M., E. Segal, D. Koller and S. K. Kim (2003). "A gene-coexpression network for global discovery of conserved genetic modules." Science **302**(5643): 249-55.

Tobin, M. D., P. S. Braund, P. R. Burton, J. R. Thompson, R. Steeds, K. Channer, S. Cheng, K. Lindpaintner and N. J. Samani (2004). "Genotypes and haplotypes predisposing to myocardial infarction: a multilocus case-control study." Eur Heart J **25**(6): 459-67.

Tomita, M., K. Hashimoto, K. Takahashi, T. Shimizu, Y. Matsuzaki, F. Miyoshi, K. Saito, S. Tanida, K. Yugi, J. C. Venter and C. A. Hutchison (1997). "E-CELL: Software Environment for Whole Cell Simulation." Genome Inform Ser Workshop Genome Inform **8**: 147-155.

van Noort, V., B. Snel and M. A. Huynen (2003). "Predicting gene function by conserved co-expression." Trends Genet **19**(5): 238-42.

Vastrik, I., P. D'Eustachio, E. Schmidt, G. Joshi-Tope, G. Gopinath, D. Croft, B. de Bono, M. Gillespie, B. Jassal, S. Lewis, L. Matthews, G. Wu, E. Birney and L. Stein (2007). "Reactome: a knowledgebase of biological pathways and processes." Genome Biol **8**(3): R39.

Westfall, P. H. and S. S. Young (1993). Resampling-based multiple testing: examples and methods for p-value adjustment. New York, Wiley.

Wheeler, D. L., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L. Y. Geer, W. Helmberg, Y. Kapustin, D. L. Kenton, O. Khovayko, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, K. D. Pruitt, G. D. Schuler, L. M. Schriml, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, T. O. Suzek, R. Tatusov, T. A. Tatusova, L. Wagner and E. Yaschenko (2006). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res **34**(Database issue): D173-80.

Wu, J., X. Mao, T. Cai, J. Luo and L. Wei (2006). "KOBAS server: a web-based platform for automated annotation and pathway identification." Nucleic Acids Res **34**(Web Server issue): W720-4.

Yi, M., J. D. Horton, J. C. Cohen, H. H. Hobbs and R. M. Stephens (2006). "WholePathwayScope: a comprehensive pathway-based analysis tool for high-throughput data." BMC Bioinformatics **7**: 30.

Yuryev, A., Z. Mulyukov, E. Kotelnikova, S. Maslov, S. Egorov, A. Nikitin, N. Daraselia and I. Mazo (2006). "Automatic pathway building in biological association networks." BMC Bioinformatics **7**: 171.

Chapter 3
Identifying Genetic Networks Underlying Myometrial Transition to Labor

3.1 Abstract

Background: Early transition to labor remains a major cause of infant mortality, yet the causes are largely unknown. Although several marker genes have been identified, little is known about the underlying global gene expression patterns and pathways that orchestrate these striking changes.

Results: We performed a detailed time-course study of over 9,000 genes in mouse myometrium at defined physiological states: nonpregnant, mid-gestation, late gestation, and postpartum. This dataset allowed us to identify distinct patterns of gene expression that correspond to phases of myometrial "quiescence," "term activation," and "postpartum involution." Using recently developed functional mapping tools (e.g., HOPACH, GenMAPP 2.0), we have identified new potential transcriptional regulatory gene networks mediating the transition from quiescence to term activation.

Conclusions: These results implicate the myometrium as an essential regulator of endocrine hormone (cortisol and progesterone synthesis) and signaling pathways (cAMP and cGMP stimulation) that direct quiescence via the transcripitional up-regulation of both novel and previously associated regulators. With term activation, we observe the up-regulation of cytoskeletal remodeling mediators (intermediate filaments), cell junctions, transcriptional regulators, and

the coordinate down-regulation of negative control checkpoints of smooth muscle contractile signaling. This analysis provides new evidence of multiple parallel mechanisms of uterine contractile regulation and presents new putative targets for regulating myometrial transformation and contraction.

3.2 Introduction

The initiation of mammalian labor is a complex physiological process that requires the expression and secretion of many factors, both maternal and fetal (Norwitz *et al.* 1999; Challis *et al.* 2000). The majority of these factors exert their effect on the myometrium, the smooth muscle responsible for expelling the fetus from the uterus. While species differences in labor regulation have been observed, several common signaling pathways and factors have been implicated as key regulators across species. During mid to late gestation, myometrial quiescence is maintained by several contractile inhibitors, such as relaxin, adrenomedullin, nitric oxide, prostacyclin, and progesterone (Norwitz *et al.* 1999; Challis *et al.* 2000). A number of these regulators stimulate cAMP- and cGMP-mediated signaling pathways. Smooth muscle contraction is inhibited by the phosphorylation of myosin light chain kinase by the cAMP-dependent protein kinase. This inhibition is believed to promote quiescence. In addition, the myometrium undergoes major structural changes throughout pregnancy that are required to generate the necessary contractile force for labor, including hypertrophy and hyperplasia of smooth muscle, connective tissue, focal adhesion, and cytoskeletal remodeling (Lopez *et al.* 2000).

The transition to labor results in synchronous contractions of high amplitude and high frequency by the myometrium. Factors previously associated with the regulation of myometrial activation include the oxytocin receptor, gap junction protein connexin-43, voltage-gated calcium channels, prostaglandin receptor subtypes, estrogen, cortisol, and transcription factors c-jun and c-fos. Most of these proteins participate in pathways that stimulate calcium release (e.g., calcium-calmodulin G protein signaling) and the formation of intracellular junctions, leading to stimulation of contractions. Although several important components that regulate the initiation of labor have been identified, the mechanisms that guide this transition are poorly understood.

A difficult challenge in identifying the regulatory events that control the switch from myometrial quiescence to activation is developing tools for examining whole genome expression profiles in the context of known biology. Recent efforts to identify transcriptional changes from laboring and non-laboring human myometrium have proven to be valuable in identifying putative physiological regulators (Aguan *et al.* 2000; Bethin *et al.* 2003; Charpigny *et al.* 2003; Rehman *et al.* 2003; Havelock *et al.* 2004); however, the lack of gestational time points examined have limited these approaches to interrogating only those genes with large fold changes at term activation without exploring the global patterns of gene expression over the time-course of myometrial transformation. While gene profiling of the rodent uterus during gestation has proved fruitful in revealing some of the large scale patterns of gene expression throughput pregnancy (Bethin *et al.* 2003; Girotti *et al.* 2003), there is still a critical need to improve the global view of myometrial gene expression with greater temporal resolution using newly developed bioinformatics tools.

To identify molecular mechanisms involved in the transition from myometrial quiescence to labor, we analyzed gene expression changes in mouse myometrium at midgestation, throughout late gestation, and during the postpartum period. Our results reveal several novel patterns of expression occurring along the phases of myometrial quiescence to term activation and postpartum involution. Analysis of putative quiescence and term activation regulators in the context of well-defined biological pathways revealed new putative functional roles for several previously unassociated genes in the suppression of contraction throughout gestation and activation of phase-dependent contractions at labor. This analysis further implicates the regulation of several novel pathways, including smooth muscle/extracellular matrix interactions throughout late gestation and cell junction/cytoskeletal interactions immediately prior to the onset of labor.

3.3 Results

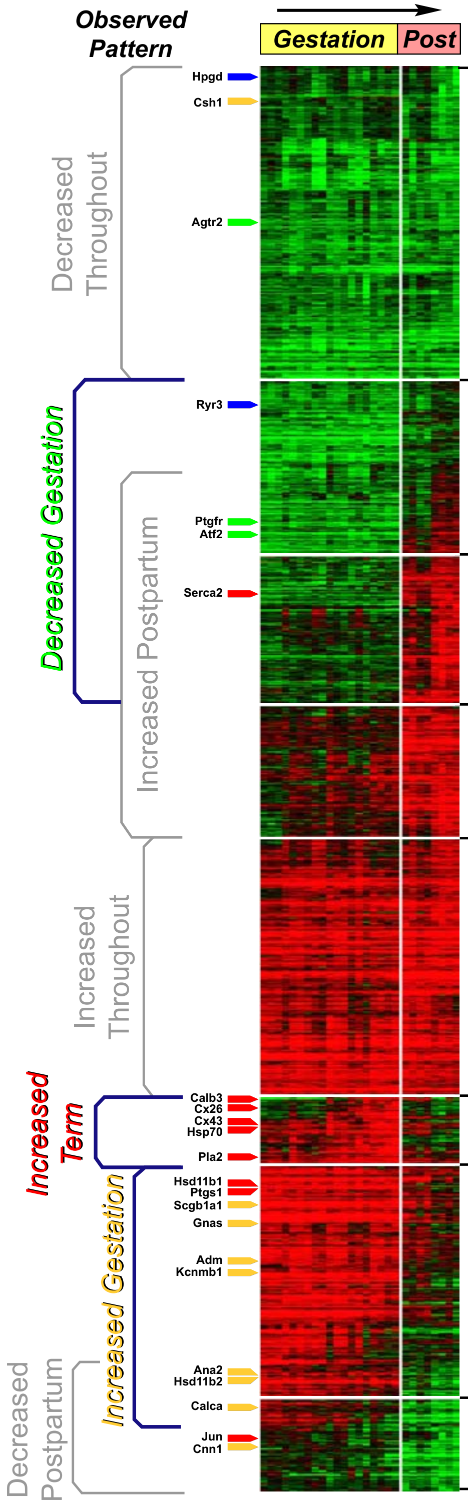
3.3.1 Clustering of expression changes in gestational myometrium

mRNA transcript levels were measured from isolated myometrium of 35 time-mated mice at four time-points of late gestation (14.5–18.5 days), at postpartum (6 and 24 h after labor), and from a nonpregnant control group. In all, ~13,000 probe sets corresponding to ~9,000 unique cDNAs and ESTs were probed with oligonucleotide microarrays. About 35% these transcripts (p<0.05 and >20% fold change) were regulated throughout gestation and postpartum (14.5 days through 24 h postpartum).

Analysis of these probe sets with HOPACH (http://www.bioconductor.org; Pollard *et al.* 2002; van der Laan *et al.* 2003) revealed eight primary cluster groups and 133 subclusters. The majority of these clusters showed a clear association with known physiological phases of uterine gestation: quiescence (clusters 2, 3, 7,and 8), term activation (cluster 6), and postpartum involution (clusters 3, 4, and 7). In addition to these clusters, we observed two cluster groups with genes down-regulated or up-regulated throughout the analyzed time-course (clusters 1 and 5) (Figure 3.1).

3.3.2 MAPPFinder Analysis

To characterize the major biological processes, molecular functions, and cellular components associated with the HOPACH pattern groups, we used MAPPFinder (a component of GenMAPP version 2.0) (http://www.genmapp.org; Ashburner *et al.* 2000; Dahlquist *et al.* 2002; Doniger *et al.* 2003). MAPPFinder produced a statistically ranked list (based on p value) of Gene Ontology (GO) biological categories associated with each cluster, from which the most significant nonsynonymous groups are listed (Figure 3.1, GO Categories). In each cluster, several highly significant biological associations were identified (adjusted permutation p<0.05).



Cluster	Gene Ontology Category	Changed/ Total
1	cytosolic ribosome (sensu Eukarya) *	33/39
	structural constituent of ribosome *	56/105
	cation transporter activity	24/97
	muscle contraction	12/38
	heat shock protein activity	8/24
	antigen binding	7/21
	nucleolus	10/41
	NADH dehydrogenase (ubiquinone) activity	5/11
	proton transport	9/33
	muscle development	11/49
	eukaryotic translation elongation factor 1 complex	3/4
2	chromatin assembly/disassembly	9/32
	nucleus	106/1279
	primary active transporter activity	17/120
	viral nucleocapsid	5/15
	RNA processing	12/75
	RNA binding	23/201
3	metalloendopeptidase activity	9/43
	ubiquitin-dependent protein catabolism	10/53
	transcriptional repressor activity	4/14
	apoptosis	15/122
	guanyl-nucleotide exchange factor activity	6/26
4	proteasome complex *	12/24
	ubiquitin-dependent protein catabolism *	19/53
	cytoplasmic microtubule	4/7
	motor activity	8/53
	protein binding	52/683
5	lysosome *	25/52
	MHC class I receptor activity *	6/8
	hydrolase activity, acting on glycosyl bonds *	16/51
	Golgi apparatus	28/136
	catalytic activity	220/1963
	coated pit	7/20
	peptidase activity	43/276
	glucose catabolism	9/31
6	cell junction	8/35
	endopeptidase inhibitor activity	6/63
	protein targeting	6/65
	amine biosynthesis	3/24
	structural constituent of cytoskeleton	6/74
7	cell growth *	13/32
	extracellular matrix structural constituent *	15/42
	calcium-dependent phospholipid binding *	6/10
	protein-lysine 6-oxidase activity	3/3
	phospholipase A2 inhibitor activity	3/3
	calcium ion binding	41/233
8	muscle development *	12/49
	muscle contraction *	10/38
	collagen	6/26
	structural constituent of cytoskeleton	10/74
	defense/immunity protein activity	8/68
	isomerase activity	7/61

previously associated regulators

increased term	🟥
increased quiescence	🟨
decreased quiescence	🟩
decreased term	🟦

* Adjusted permute p<0.05

Figure 3.1. Clustering of Myometrial Expression Profiles with HOPACH. Gene expression profiles for 27 microarrays (vertical axis) and 4,510 probe sets (horizontal axis) are shown in the context of the HOPACH cluster map (nonpregnant data excluded). The array groups correspond to mid-to-late gestation (14.5, 16.5, 17.5, and 18.5 days), and postpartum (6 and 24 h). Eight clusters of genes are arranged vertically. Physiological phase groups are assigned based on visual observation and association with previously associated regulators. MAPPFinder results are shown for the top ranking distinct biological process, molecular function, and cellular component groups based on a permuted p value. Previous associated regulators of uterine quiescence and activation are indicated by a colored line next to the location of the corresponding gene probe set in the cluster map.

3.3.3 Association of Expression Clusters with Previously Associated Uterine Quiescence and Activation Genes

Gene expression groups associated with the maintenance of pregnancy (quiescence) or induction of labor (activation) were confirmed by mapping lists of previously identified regulators of uterine quiescence and activation onto our HOPACH cluster map. Extensive literature searches for such regulators identified 66 genes, of which 23 were regulated in our dataset (Figure 3.1, Previously Associated Regulators). Genes hypothesized to regulate quiescence by transcriptional up-regulation or secretion were largely associated with clusters 7 and 8 (increased “quiescence”), while putative activators of uterine activation were largely associated with cluster 6 (increased “term activation”). Although only three down-regulated quiescence regulators were associated with HOPACH clusters, two of them mapped to cluster 2 (decreased “quiescence”), as predicted.

3.3.4 Functional Analysis of “Quiescence” and “Term Activation” Pattern Groups

To further elucidate specific genes and pathways linked to the regulation of uterine quiescence and the initiation of labor, we examined pattern groups linked to quiescence and term activation, in the context of GO categories, GenMAPP pathway maps, and literature associations. While low magnitude fold changes have been included within these functional analyses to broaden our survey of biological groups, we have largely restricted our discussion to transcripts with fold changes greater than 2.

3.3.5 Up-regulation of Pathways of Relaxation and Remodeling during Quiescence

Analysis of genes up-regulated throughout gestation (“increased quiescence”) revealed a number of biological categories associated with uterine quiescence. These categories contain a large number of highly regulated genes coupled to the inhibition of prostaglandin and cortisol synthesis, stimulation of cAMP and cGMP signaling pathways, extracellular matrix remodeling, cytolysis, and regulation of cell growth (Figure 3.2 and Table 3.1). To explore the potential relationships between the products of these transcriptionally regulated genes, we mapped the data onto respective metabolic and signaling pathways (Figure 3.3 A, B).

Besides well-established quiescence regulators (Adm, Cgrp, Hsd11b2, Gnas, Cnn1, and Utg)(see Tables 1-3 for complete gene names), several genes previously unassociated with the maintenance of quiescence were identified along the same or related biological pathways. The most highly regulated of these genes were those implicated in the induction of cGMP and cAMP signaling pathways (Guca2b and Cmkor1), calcium-dependent phospholipid binding genes (Anxa1, Anxa2, Anxa3 and Anxa8), and the Anxa2 dimerization partner S100A10 (Figure 3.3 A). Other changes in expression from this pattern group were observed among cytolysis-inducing proteases (granzymes B–G), regulators of cell growth (Igfbp2 and Il1r2), and transcriptional regulation (Sfrp4 and Klf4). Several of these and other genes were found to have highly reproducible patterns of expression using quantitative real-time PCR (TaqMan), with typically larger fold changes produced by TaqMan than by GeneChip (consistent with the more conservative folds typically produced after RMA normalization) (Supplemental Figure 1).

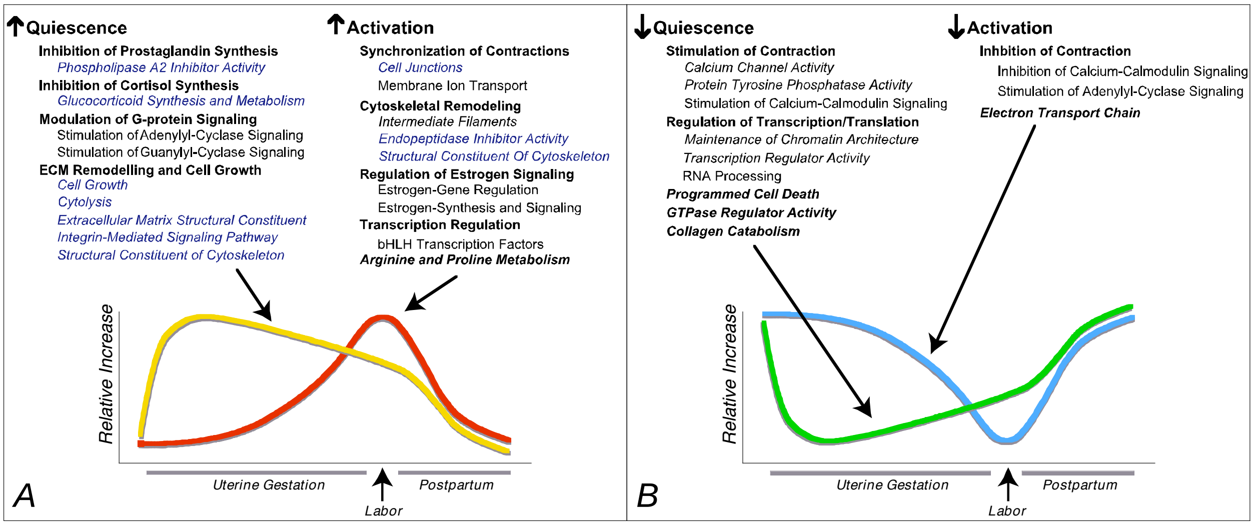


Figure 3.2. Association of Quiescence and Term Activation Pattern Groups with Biological Pathways. Significant associations to GO classification groups and GenMAPP pathways were determined for each of the four examined expression pattern groups, (A) increased quiescence, increased activation, (B) decreased quiescence and decreased activation. GO terms and GenMAPP pathways highlighted by analysis with the program MAPPFinder are indicated by italicized blue text. Biological processes identified by literature association are indicated in black text or bold black text for general contraction associated biological groups.

Up Quiescence Expression Group > 2 Fold

Increased Gestation Pattern Group	Gene Symbol	14 days fold
Prostaglandin and Cortisol Synthesis		
hydroxysteroid 11-beta dehydrogenase 1	Hsd11b1	10.6
decidual/trophoblast prolactin-related protein	Dtprp	6.2
hydroxysteroid 11-beta dehydrogenase 2	Hsd11b2	3.6
cytochrome P450, 11a	Cyp11a1	2.3
prostaglandin-endoperoxide synthase 1	Ptgs1	2.0
Phospholipase Inhibition		
annexin A8	Anxa8	4.4
annexin A3	Anxa3	3.1
uteroglobin	Utg	2.7
calpactin	S100a10	2.5
annexin A1	Anxa1	2.4
annexin A2	Anxa2	2.1
Proteolysis and Peptidolysis		
kidney-derived aspartic protease-like protein	Kdap	8.2
CTLA-2-beta	Ctla2b	8.1
cathepsin Z	Ctsz	3.1
dipeptidase 1	Dpep1	3.1
procollagen C-proteinase enhancer protein	Pcolce	2.6
lipocalin 7	Lcn7	2.6
Serine-Type Endopeptidases		
granzyme G	Gzmg	71.4
granzyme D	Gzmd	45.7
granzyme F	Gzmf	40.2
granzyme E	Gzme	19.8
granzyme C	Gzmc	10.7
RIKEN cDNA 2210021K23 gene	2210021K23Rik	2.9
cathepsin G	Ctsg	2.2
protease, serine, 11 (Igf binding)	Prss11	2.2
granzyme B	Gzmb	2.1
Protease inhibitors		
tissue factor pathway inhibitor 2	Tfpi2	4.2
serine protease inhibitor 14	Serpinb9e	3.3
plasma protease C1 inhibitor	Serping1	2.7
ECM Remodelling and Cell Growth		
Regulation of Cell Growth		
insulin-like growth factor binding protein 2	Igfbp2	12.4
interleukin 1 receptor, type II	Il1r2	5.0
glucocorticoid-induced leucine zipper	Gilz	3.8
tumor necrosis factor, alpha-induced protein 2	Tnfaip2	3.3
c-fos induced growth factor	Figf	3.2
related RAS viral (r-ras) oncogene homolog 2	Rras2	3.0
cysteine rich protein 2	Crip2	2.9
MORF-related gene X	Morf4l2	2.6
epithelial membrane protein 1	Emp1	2.5
four and a half LIM domains 1	Fhl1	2.3
S100 calcium binding protein A6 (calcyclin)	S100a6	2.3
insulin-like growth factor binding protein 6	Igfbp6	2.1
transforming growth factor, beta 2	Tgfb2	2.0
Integrin-Mediated Signaling Pathway		
secreted phosphoprotein 1	Spp1	17.3
connective tissue growth factor	Ctgf	2.8
caveolin, caveolae protein	Cav	2.5
ras homolog gene family, member A2	Arha	2.4
Structural Constituent of Cytoskeleton		
gelsolin	Gsn	2.4
tropomyosin 4	Tpm4	3.1
tubulin, beta 2	Tubb2	2.2
Extracellular Matrix Structural Constituent		
microfibrillar associated protein 5	Mfap5-pending	6.9
elastin	Eln	3.1
procollagen, type XI, alpha 1	Col11a1	3.0
fibromodulin	Fmod	2.4
fibrillin 1	Fbn1	2.3
procollagen, type V, alpha 2	Col5a2	2.2
laminin, gamma 1	Lamc1	2.2
procollagen, type I, alpha 2	Col1a2	2.2
G protein Signaling		
guanylate cyclase activator 2b	Guca2b	15.2
chemokine orphan receptor 1	Cmkor1	5.0

adrenomedullin	Adm	2.0
guanine nucleotide binding protein, gamma 11	Gng11	2.0
Transcriptional Regulation		
secreted frizzled-related sequence protein 4	Sfrp4	4.2
Kruppel-like factor 4	Klf4	3.0
C/EBP delta	Cebpd	2.3
inhibitor of DNA binding 1	Idb1	2.1
X-box binding protein 1	Xbp1	2.0
Kruppel-like factor 2	Klf2	2.0

Table 3.1. Genes Up-regulated with Quiescence. Only up-regulated genes with a relative fold change versus nonpregnant mice ≥2 at 14.5 days gestation and linked to biological categories highlighted by the expression analysis are shown. Full gene lists can be obtained online (Supplemental Table 2).

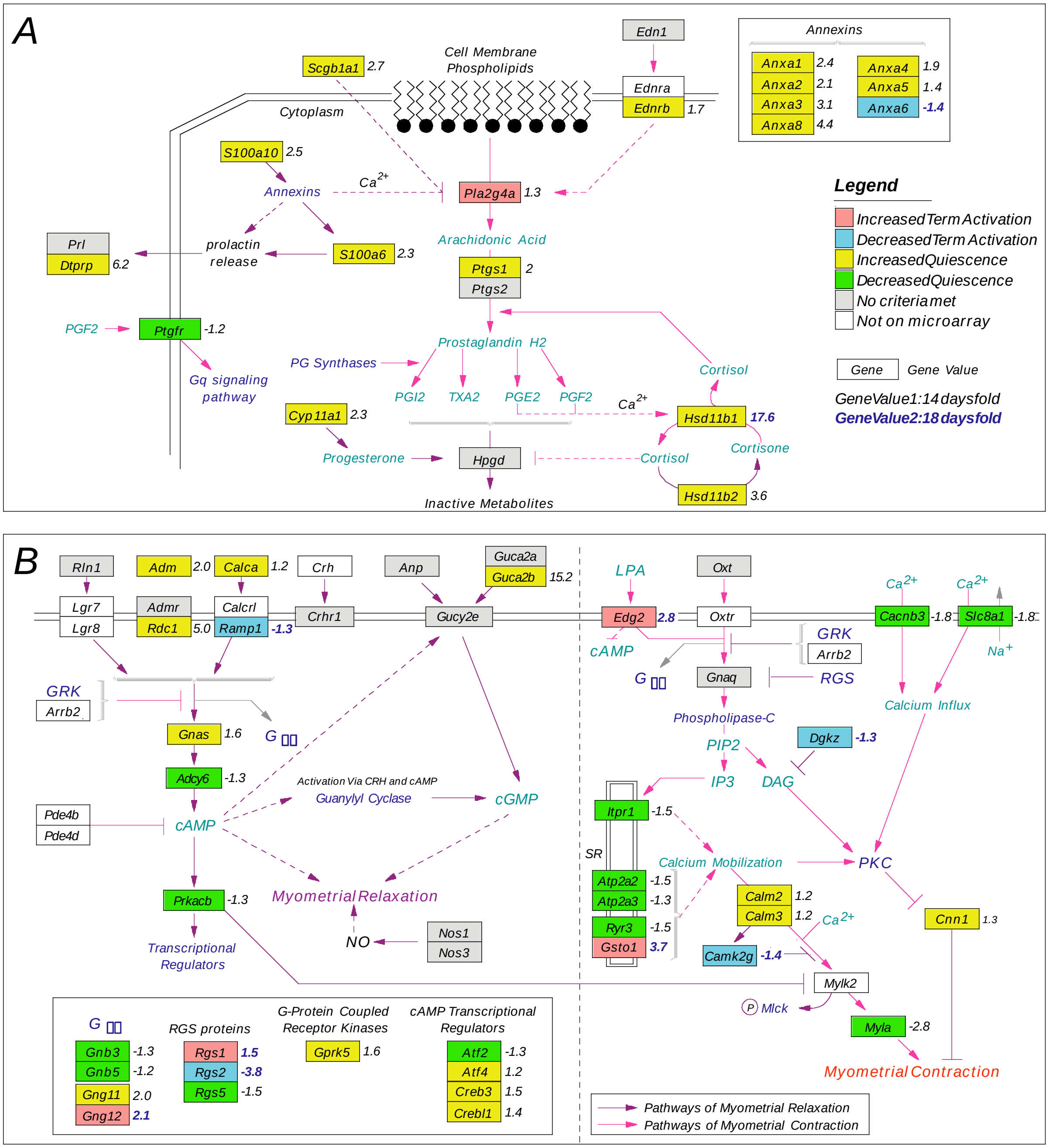


Figure 3.3. Analysis of Pathways of Uterine Smooth Muscle Contraction. Prostaglandin synthesis (A) and G protein signaling (B) pathways in the myometrium are overlaid with gene expression color criterion and fold changes from the program GenMAPP. Interactions suggested by results of this microarray analysis are included in these figures. Some of the genes in these pathways that are not significant in this analysis are indicated by blue

text. Detailed gene-expression data, statistics and full gene annotations are available on the GenMAPP interactive version of these pathways online.

Several cAMP response element transcription factors were also found within the “increased quiescence” group (Atf4, Crebl1, and Creb3, see Figure 3.3 B). These genes are all members of a larger group of basic leucine zipper (bZip) transcription factors not previously associated with quiescence, which also includes the CCAAT/enhancer binding protein Cebpd, the Maf protein Mafk, the nuclear factor, interleukin 3, regulated Nfil3, and the X-box binding protein Xbp1, also up-regulated with quiescence.

3.3.6 Down-regulation of mRNA Processing and Contraction-Associated Signaling during Quiescence

MAPPFinder analysis of genes in the “decreased quiescence” group identified a wide variety of cell maintenance, transcription, and cell signaling biological processes. Many of these GO categories were associated with the onset of labor (calcium ion transport and protein tyrosine phosphatase activity) or myometrial postpartum involution (programmed cell death, collagen catabolism, and ubiquitin conjugating enzyme activity). These results are in accordance with the inhibition of contraction and suppression of cell death in late gestation. Unlike term-related biological processes, categories shared between the “decreased quiescence” and “increased postpartum involution” group appear to be largely the result of a common transcript expression profile (Figure 3.1, cluster 3; Figure 3.2).

Although similar numbers of genes were down-regulated or up-regulated with “quiescence” (~480–520 genes), very few genes were down-regulated more than two fold at 14.5 days of gestation (Table 3.2). One of the most down-regulated transcripts was the myosin light chain gene Myl4, the primary target for oxytocin-induced phosphorylation leading to uterine contraction at term. Several additional putative components of the oxytocin contractile signaling pathway (calcium-calmodulin signaling pathway) were also present in this expression group (Iptr1, Ryr3, Plcg1, and Atp2a2) (Figure 3.3 B). Another large set of coordinately down-regulated genes include factors involved in RNA processing. Alternative splicing of putative quiescence and term activation regulators has been proposed to be a critical mechanism of the physiological switch to labor (Benkusky *et al.* 2000; Pollard *et al.* 2000).

Down Quiescence Expression Group > 2 Fold

Decreased Gestation Pattern Group	Gene Symbol	14days fold
Regulation of Cell Growth		
myosin light chain, alkali, cardiac atria	Myl4	-2.8
N-myc downstream regulated 2	Ndr2	-2.7
actin, beta, cytoplasmic	Actb	-2.2
Calmodulin-Signlaing		
MARCKS-like protein	Mlp	-2.2
Proteolysis		
matrix metalloproteinase 3	Mmp3	-2.2
Ion Channels		
expressed sequence AW538430	Kctd12	-2.9
Transcriptional Regulation		
SRY-box containing gene 4	Sox4	-2.9
homeobox protein Meis2	Mrg1	-2.5
special AT-rich sequence binding protein 1	Satb1	-2.1
D site albumin promoter binding protein	Dbp	-2.1
RIKEN cDNA 1110033A15 gene	1110033A15Rik	-2.1
myeloid ecotropic viral integration site 1	Meis1	-2.0
Regulation of Alternative Splicing		
CDC-like kinase	Clk	-2.1

Table 3.2. Genes Down-regulated with Quiescence. Only down-regulated genes with a relative fold change versus nonpregnant mice ≥2 at 14.5 days gestation and linked to biological categories highlighted by the expression analysis are shown. Full gene lists can be obtained online (Supplemental Table 3.2).

3.3.7 Transition from Remodeling and Relaxation to Cell-Cell Signaling and Transcriptional Regulation with Activation of the Myometrium at Term

A large percentage of genes regulated with "quiescence" continued to be highly regulated at term. This result emphasizes the importance of expression changes immediately before labor to counteract the effects of quiescence. Consistent with the number of up-regulated genes, MAPPFinder analysis of the "increased term activation" group identified a smaller set of GO terms and pathways. Prominent among these were genes associated with the formation of cell junctions, kinesin complexes, and endopeptidase inhibitors. In addition, functionally related transcription factors (basic helix-loop-helix members or BHLH), ion transport proteins and ion transport regulators were coordinately up-regulated at term.

Within these biological categories, several contractile regulators, both associated and unassociated with parturition, were highly up-regulated. These genes include cell junction molecules (Cx43, Cx26, Ocln, and Dsp), the pulmonary smooth muscle contractile regulator and complement component C3, the estrogen signaling regulator Hsp70, the chloride conductance regulator Fxyd3, and the ryanodine receptor regulator Gsto1 (Table 3.3). These changes occurred in concert with the up-regulation of signaling molecules, such as growth factors (Inhba, Inhbb), G protein signaling components (Edg2, Gng12) (Figure 3.3 B), and collagen catabolism proteins (Pep4, Mmp7). On the whole, however, this pattern group was predominated by the up-regulation of genes that encode for proteins that are largely epithelial cell–specific. Most prominent among these are the cytokeratin intermediate filaments, Krt2-7, Krt2-8, Krt1-18, and Krt1-19, and the cytokeratin transcriptional regulator Elf3, which are among the most highly up-regulated genes at term.

3.3.8 Down-regulation of Pathways of Calcium Mobilization and G Protein Signaling in Term Myometrium

HOPACH analysis with a metric that disregarded the direction of fold change (Supplemental Figure 2) revealed a small number of down-regulated genes at term that mirror the “increased term activation” group. Among these, we observed two highly down-regulated genes, regulator of G-protein signaling 2 (Rgs2), a potent inactivator of Gαq-GTP bound activity and inhibitor of DNA binding 2 (Idb2), a bHLH factor that heterodimerizes with other HLH proteins to inhibit their function. Rgs2 is one of the most down-regulated genes throughout the gestation-postpartum time-course, in addition to being highly expressed in nonpregnant myometrium and throughout gestation. Additional term down-regulated G protein signaling proteins that act to antagonize calcium-calmodulin signaling are illustrated in Figure 3.3 B.

Up Term Activation Expression Group > 2 Fold

Title	Gene Symbol	18days fold
Regulation of Cell Growth		
inhibin beta-B	Inhbb	3.1
inhibin beta-A	Inhba	2.2
Cell Death		
growth arrest and DNA-damage-inducible 45 γ	Gadd45g	3.2
baculoviral IAP repeat-containing 1a	Birc1a	2.1
clusterin	Clu	2.0
Cell Junctions		
occludin	Ocln	2.8
gap junction membrane channel protein α 1	Gja1	2.8
desmoplakin	Dsp	2.8
G Protein Signaling		
lysophosphatidic acid receptor Edg-2	Edg2	2.8
guanine nucleotide binding protein, γ 12	Gng12	2.1
Structural Constituent of Cytoskeleton		
villin 2	Vil2	3.1
Kinesin Complex		
keratin complex 1, acidic, gene 19	Krt1-19	7.8
keratin complex 2, basic, gene 7	Krt2-7	4.6
keratin complex 2, basic, gene 8	Krt2-8	4.6
keratin complex 1, acidic, gene 18	Krt1-18	4.5
surfactant associated protein D	Sftpd	3.4
Metabolism and Biosynthic Reactions		
lipoprotein lipase	Lpl	4.5
aldehyde dehydrogenase family 1, subfamily A2	Aldh1a2	3.9
glutathione S-transferase omega 1	Gsto1	3.7
branched chain aminotransferase 1, cytosolic	Bcat1	3.4
protein phosphatase 1, regulatory subunit 3C	Ppp1r3c	2.2
carbonic anhydrase 2	Car2	2.1
Proteolysis and Peptidolysis		
cytosolic nonspecific dipeptidase	0610010E05Rik	3.2
transmembrane protease, serine 2	Tmprss2	2.1
kallikrein 5	Klk5	2.1
Collagen Catabolism		
peptidase 4	Pep4	2.3
matrix metalloproteinase 7	Mmp7	2.2
Proteolysis inhibitors		
complement component 3	C3	4.3
RIKEN cDNA 1600023A02 gene	1600023A02Rik	2.9
extracellular proteinase inhibitor	Expi	2.8
Transcriptional Reglation		
Transcription Factors		
myeloblastosis oncogene	Myb	2.5
hairy and enhancer of split 1	Hes1	2.3
E74-like factor 3	Elf3	2.1
Androgen Regulation		
kidney androgen regulated protein	Kap	33.9
heat shock protein 4	Hspa4	3.1
alpha fetoprotein	Afp	3.1
Transport		
FXYD domain-containing ion transport regulator 3	Fxyd3	2.8
lipocalin 2	Lcn2	2.5
lactotransferrin	Ltf	2.2
solute carrier family 16, member 1	Slc16a1	2.1
fatty acid binding protein 5, epidermal	Fabp5	2.0

Table 3.3. Genes Up-regulated with Term Activation. Only up-regulated genes with a relative fold change versus nonpregnant mice ≥2 at 18.5 days gestation and linked to biological categories highlighted by the expression analysis are shown. Full gene lists can be obtained online

(Supplemental Table 3.2).

3.3.9 Global Mechanisms of Transcriptional Regulation

One of the most prominent observations in this dataset is the highly significant correlation in the expression and genomic position of eight serine-type endopeptidases (Gzmb–Gzmg, Mcpt8, and Ctsg) during the phase of quiescence. Genes within this multi-gene cluster undergo tight coordinate regulation in response to cell stimulus (Pham *et al.* 1996; Allen *et al.* 1998). Examination of this expression cluster group in the context of genomic position reveals a novel pattern of positional gene regulation, where relative fold increases from the peripheral members in the cluster to the center of the gene cluster (Figure 3.4 A).

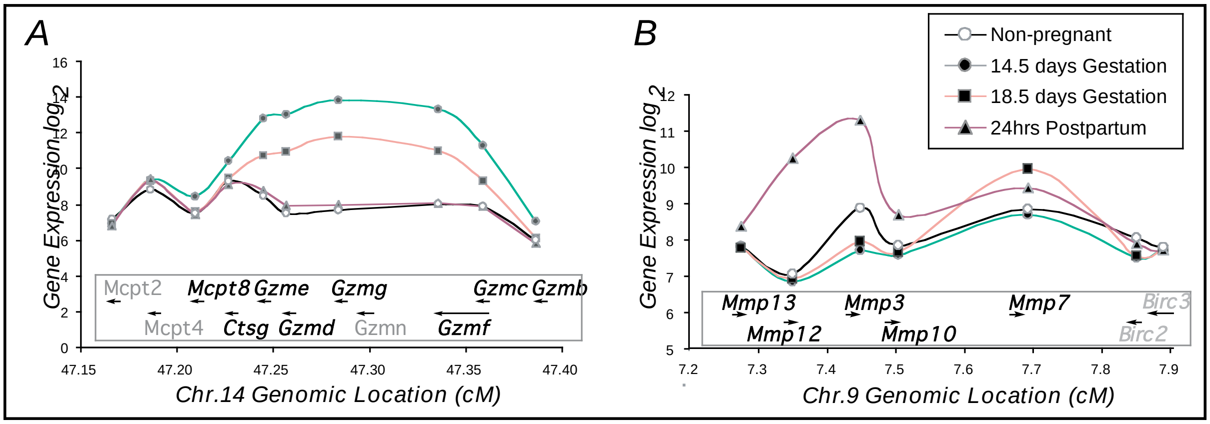


Figure 3.4. Association of Genomic Localization with Expression Coregulation. (A) and (B), Chromosomal gene clusters contain highly correlated expression changes among multiple members. Global patterns of gene expression within these genomic intervals are visualized by representing mean log expression for four of the myometrium time-point groups (nonpregnant, 14.5 and 18.5 days gestation, and 24 h postpartum), versus relative gene position on the chromosome. Gene strand orientation and position is designated by the orientation of arrows. Gene symbols above and below arrows are shown, where italicized black text indicates coregulated genes (same HOPACH cluster) and italicized grey genes noncoregulated for increased quiescence (A) and increased postpartum involution (B). Non-italicized grey text indicates genes not probed by the arrays.

To determine whether other gene clusters exhibit a similar form of positional coregulation, we developed a program to identify genomic intervals containing several co-expressed genes. Searching for regions with three or more members in a broad genomic interval (500 kb), we identified 11 clusters of genes that are co-localized and coregulated (same HOPACH cluster)(Supplemental Figure 3). Among these, we were able to identify at least one other gene cluster that possessed a genomic pattern of gene expression similar to that of the granzyme cluster, with genes maximally up-regulated postpartum (Figure 3.4 B). These genes, which encode several of the collagen catabolism matrix metalloproteinases Mmp3, Mmp10, Mmp12, and Mmp13, are among the most highly up-regulated genes postpartum. Since we do not have data from full genome arrays, it is difficult to determine if these coregulated clusters of genes occur more frequently. However, these coregulated gene clusters suggest coordinated gene regulation by an unknown mechanism.

3.4 Discussion

This time-course analysis provides the first global view of gene-expression changes in mouse myometrium from uterine quiescence through the activation of the myometrium before labor and to its postpartum involution. Examination of multiple time points, the use of replicates, robust array normalization, and powerful clustering tools enabled us to delineate and characterize unique patterns of gene expression throughout this physiological process. In addition to partitioning clusters of genes, analysis with the program HOPACH also provides us with a continuum of expression changes that reveals an overall transition in the expression of genes from one cluster group to another (Figure 3.1). Annotation of these clusters with GO terms provides a bird's eye view of the major processes regulating each of these pattern groups. These results support the hypothesis that mid-to-late gestation is predominated by changes in the expression of genes related to cell growth and extracellular matrix remodeling (cluster 7), term gestation by changes in the content of cell junctions (cluster 6), and postpartum by targeted protein degradation, collagen digestion, and apoptosis (clusters 3 and 4). Furthermore, results from genes up-regulated throughout gestation through postpartum suggest a continual local uterine immune response throughout this process (cluster 5). To help visualize the large scale gene expression changes in the context of myometrial physiology, we have depicted the data in an animation (see supplementary data Flash movie) that summarizes our major findings.

A number of studies emphasize the importance of fetal regulation of the switch from quiescence to term activation, particularly increased cortisol and estrogen output from the fetal adrenal gland (Norwitz *et al.* 1999; Challis *et al.* 2000). Interestingly, our studies provide evidence of a dynamic interplay between the myometrium and the fetus, particularly at the level of cortisol and progesterone synthesis (Figure 3.3 A). Genes highly up-regulated with quiescence include Hsd11b, which converts cortisol to the inactive cortisone, and Cyp11a1, which promotes the synthesis of progesterone. Conversely, Hsd11a, which catalyzes the synthesis of cortisol, increased expression from 11- to 18-

fold throughout gestation, suggesting that local regulation of cortisol levels are important for myometrial activation. While we observed the up-regulation of the estrogen signaling regulator, Hsp70, with "term activation," downstream markers of estrogen action are among the most highly up-regulated genes with term activation, supporting the role of the fetus in myometrial activation.

Examination of highly up-regulated putative quiescence and term activation genes revealed several novel changes within important assoicated pathways for quiescence and activation (cAMP and cGMP signaling, calcium and calmodulin signaling and prostaglandin synthesis). Such factors include Guca2b (uroguanylin), Anxa3, and Anxa8 with quiescence and C3, Edg2, Gsto1, and Fxyd3 during activation (see Figure 3.3). These factors may represent novel targets for controlling gestational length. This is evidenced by the parallel observed up-regulation of Guca2b from a recent microarray analysis of rat uterine gestation, where this factor has also been proposed to be a crucial regulator of cGMP mediated smooth muscle relaxation throughout late pregnancy(Girotti *et al.* 2003; Buxton 2004). We have validated the expression patterns of a number of these genes using quantitative real-time PCR (Supplemental Figure 3.1). In addition to these mentioned candidates, a number of other highly up-regulated genes, whose functions have not been elucidated are also found in these two expression groups (Supplemental Table Set 2).

Although a number of genes up-regulated with quiescence or with term activation can be clearly implicated in the regulation of contractile pathways or uterine growth, several more groups of genes with little known functional connection to these processes were coordinately expressed. Highlighted among these groups are serine endopeptidases (granzymes) and bZip transcription factors, up-regulated during quiescence, and endopeptidase inhibitors and bHLH factors, up-regulated with term activation. In addition to cytolysis, granzyme expression has been associated with the breakdown of extracelluar matrix proteins in the uterus during pregnancy by secretion from T-lymphocytes (Garcia-Sanz *et al.* 1990; Croy *et al.* 1997; Benkusky *et al.* 2000). Interestingly, the up-regulation of serine endopeptidases appears to be antagonized prior to the onset of labor by the up-regulation of several serine endopeptidase inhibitors with term activation. A similar antagonistic relationship may also exist for bHLH factors up-regulated at term with inhibitors of HLH function that are up-regulated with quiescence and become down-regulated at term.

Although the myometrium is considered to be relatively homogenous, many of the largest changes in gene expression at term occurred in genes that are not normally associated with muscle such as the keratins, tight junction and desmosome junction proteins. Indeed, altered gene expression due to changes in cell type distribution or the invasion of the myometrium by the decidua and endometrium would not be distinguished if those changes occur consistently between gestational myometrium preparations. Further inspection of the literature reveals that the cytokeratins, which compose the bulk of this group, are expressed within smooth muscle and likely function as intermediate filaments of the cytoskeleton (Brown *et al.* 1987; Gown *et al.* 1988; Stiemer *et al.* 1995; Yu *et al.* 1998). Furthermore, several components of desmosome spot junctions and hemidesmosomes, which interact with keratin intermediate filaments and the extracellular matrix to impart tensile strength between cells, are also up-regulated with term activation (see Supplemental Animation). These data suggest that an increase in rigidity imparting cell junctions and remodeling of the cytoskeleton immediately before labor may promote coordinate contractions. However, further studies are needed to determine if cytokeratin expression at term occurs within resident or infiltrating cells.

In addition to the capability to group and annotate clusters of genes, pattern analysis with HOPACH can be used to interrogate gene clusters in the context of genomic location. For this analysis, we developed a program to isolate gene clusters that are likely to be coregulated based on genomic location, similar to other reported methods (Gabrielsson *et al.* 2000; Caron *et al.* 2001; Megy *et al.* 2003; Trinklein *et al.* 2004). Using this program, we identified genomic regions that undergo correlated changes in gene expression associated with specific phases of the myometrial time-course. These groups highlight novel forms of gene regulation during quiescence and postpartum to coordinate cell responses (serine-protease activation and collagen catabolism). The prominent coregulation among members of these two gene clusters further suggests that immune cell trafficking and activation also play important roles in the progression towards labor and recovery from pregnancy.

3.5 Conclusions

We have identified several highly regulated genes not previously associated with myometrial quiescence or activation, in addition to families of genes coregulated at different phases of the myometrial time-course. In addition to providing new hypotheses about how the switch from quiescence to term activation may be facilitated (see Figure 3.5), these data highlighted several proteins which may serve as new candidate pharmacological targets for regulating myometrial contraction and thus the onset of labor. Such analyses will also be useful in predicting and correlating gene expression changes in human pregnancy, where several time-points are often difficult to obtain(Aguan *et al.* 2000; Bethin *et al.* 2003; Charpigny *et al.* 2003; Rehman *et al.* 2003; Havelock *et al.* 2004). Similar studies in other species using complementary methods of transcript measurement will also be necessary to validate these changes and understand the species-specific and regional myometrium transcriptional differences that likely occur. A detailed examination of the precise physiological roles of these regulators and mechanisms of regulation will be essential for developing a more detailed view of the regulation of labor.

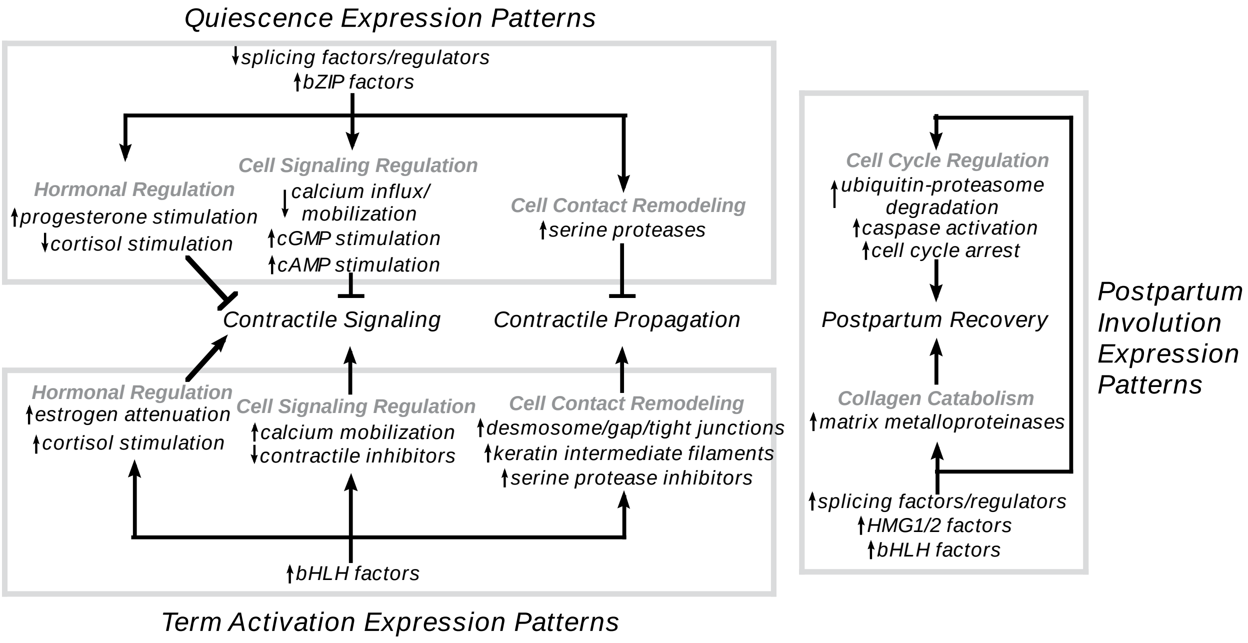


Figure 3.5. Proposed Maternal Model of Uterine-Directed Contractile Regulation. Theoretical model based on the major gene expression pattern groups for quiescence, term activation, and postpartum involution (light grey box outline). Arrows next to gene processes and functional groups indicate the predominant direction of fold change as indicated by HOPACH analysis. This model proposes new roles for transcriptional regulators, regulators of mRNA processing, local hormone regulation, protease activity, and cell junction formation in the control of both contractile signaling and contraction propagation in the myometrium during pregnancy. A model of postpartum involution is also presented based on additional results (Supplemental Table Sets 1 and 2).

3.6 Materials and Methods

3.6.1 Tissue Harvesting

FVB/N mice (Jackson Laboratory) were sacrificed in the morning (10 to noon) at 14.5 (n = 3), 16.5 (n = 4), 17.5 (n = 5), or 18.5 days (n = 7) after timed mating, and 6 (n = 4) or 24 hours (n = 4) after delivery. Control myometrium was harvested from nonpregnant littermate females (n = 8) 1 day after timed mating with a vasectomized male. After dissection of both uterine horns, the tissue closest to the cervix was removed. Each horn was washed with PBS and opened longitudinally. Pups and placenta were discarded, and the decidua was removed by blunt dissection. The myometrium from each horn was then immediately frozen in liquid nitrogen and stored at –80 ºC.

3.6.2 Sample Preparation and Microarray Data Normalization

For each sample, labeled cRNA was prepared from 20 μg of purified total RNA and hybridized to Affymetrix Mu11k A and B arrays according to the manufacturer's instructions. Tissue from each mouse was hybridized individually to one array set. Microarrays were scanned at a photomultiplier tube (PMT) setting of 100%. Resulting .cel files were generated with Affymetrix Microarray Suite 5.0 and analyzed with robust multi-array average (RMA) (Irizarry *et al.* 2003).

3.6.3 Statistical Analysis

To identify transcripts differing in mean expression across the seven experimental groups, p values were calculated from a permutation test with the F-statistic function from the mult test package of Bioconductor (Dudoit *et al.* 2003). Fold changes in transcript levels were calculated from the mean $\log_2$ expression values of each time-point group versus the mean of nonpregnant controls. For cluster analysis, the dataset was filtered for probe sets with a $p<0.05$ across the full expression time-course and fold change of >20% (positive or negative) for at least one time-point group versus nonpregnant controls. Additional filters were used downstream of clustering for genes related to uterine quiescence and term activation. For clusters related to “quiescence” and “term activation,” a fold change >20% was required for the midgestation (14.5 days) and term (18.5 days) time-points, respectively, versus non-pregnant controls.

3.6.4 Clustering and Pattern Analysis

Gene expression clustering for 4,510 significant probe sets was performed using the program HOPACH (hierarchical ordered partitioning and collapsing hybrid), with uncentered correlation distance (http://www.bioconductor.org; Pollard *et al.* 2002; van der Laan *et al.* 2003). HOPACH produced a tree with six levels of clusters (eight primary level clusters and 133 main clusters). To examine expression patterns independently of the direction of the fold change, HOPACH was re-run with absolute uncentered correlation distance. Associations with GO biological process, molecular function, cellular component groups, and GenMAPP biological pathways were obtained with MAPPFinder 2.0, a part of the

GenMAPP 2.0 application package(Ashburner *et al.* 2000; Dahlquist *et al.* 2002; Doniger *et al.* 2003). A permuted p value was calculated by MAPPFinder 2.0 to adjust for multiple hypothesis testing (supplemental Methods). Due to the highly redundant nature of the oligonucleotide arrays used, redundant probe sets corresponding to a single gene were identified from the Affymetrix NetAFFX website (Liu *et al.* 2003).

3.6.5 Real Time PCR validation of Microarray Data

Real-time (RT) PCR was used to validate the expression patterns of several highly regulated genes associated with specific phases of myometrium gestation. Gene-specific primers for multiplex real time RT-PCR were designed for each gene of interest (n=18) using “Primer Express” software (Perkin Elmer, Foster City, CA) based on sequencing data from NCBI databases and purchased from Biosearch Technologies, Inc. (Novato, CA). Sequence data for all oligos are available online. Total RNA concentration and quality was assessed using the Agilent Bioanalyzer 2001 (Agilent Technologies, Palo Alto, CA). First strand cDNA synthesis was performed using total cellular RNA (BD Biosciences Clontech, Palo Alto, CA), Powerscript™ reverse transcriptase (BD Biosciences Clontech), and random hexamer primers. Finally, an equivalent of 10 ng of total RNA from the first strand cDNA synthesis reaction was used in 10 μl of each TaqMan gene quantification in 384-well format. Universal Master Mix for real time PCR was purchased from InvitrogenTM life technologies (Carlsbad, CA). Raw data from an ABI Prizm7900 were processed into Excel spreadsheets and conversion of raw Ct values to relative gene copy numbers (GCN) were done as described previously (Dolganov *et al.* 2001). Gene-expression analysis requires proper internal control genes for normalization. By using an endogenous control as an active reference, quantification of an mRNA target can be normalized for differences in the amount of total RNA added to each reaction. For this purpose, we used four mouse housekeeping genes—PPIA, GAPDH, PGK1 and S9. Moreover, using GeNorm (Vandesompele *et al.* 2002), we selected PGK and GAPDH as the two most stable housekeeping genes across all 12 specimens and used their geomeans for normalization. Normalized data were graphed and compared to the data generated on similar specimens via microarrays. Genes could be broken down into the following groups: a) 13 genes with concordant microarray-TaqMan patterns, b) 1 false negative result by microarray (Acta2), c) 3 genes with high TaqMan variability (Mmp9, Krt19, Id1) and d) 1 gene with evidence of alternative splicing (Csb) (supplemental Figure 1). It should be noted that Acta2 baseline expression was relatively high for both microarray and TaqMan results. Since both of these techniques probed different regions of the Acta2 gene, we can not exclude the possibility of alternatively splicing.

3.6.6 Chromosomal Localization Analysis

We constructed a program to link HOPACH expression data to chromosome start site location and strand orientation, obtained from the Ensembl database. Co-localized clusters of genes were identified as those genes clustered within a 500-kb genomic interval, belonging to the same HOPACH cluster, with a z-score

>1.96, and an average pair-wise Pearson correlation among cluster members of r >0.65. See Supplemental Methods for calculation details and the full Supplemental Chromosome Cluster Lists online.

3.7 Supplemental Methods

Filename: supplemental_methods.doc

see: http://www.genmapp.org/myometrium.html or Genome Biology website.

3.8 Supplemental Figures/Movie/Datasets: Complete Expression Dataset. Myometrium gene expression time-course expression dataset with statistics is supplied as a MS-Excel spreadsheet in addition to a GenMAPP format GEX file for use with GenMAPP format pathway maps (MAPP files). MAPP files can be downloaded from http://www.genmapp.org/.

File name (XLS): full_dataset.zip

File name (GEX – GenMAPP format): GenMAPP_myometrium_timecourse.zip

For all additional supplemental files, see:

http://www.genmapp.org/myometrium.html

3.9 Acknowledgments

We would like to thank Chris Barker, Kristina Hanspers, Yanxia Hao, and Anita Chow from the Gladstone Genomics Core and Michael McMaster for his assistance with uterine dissections. We would like to thank Susan Fisher, Janet A. Warrington, Gary Howard, Bethany Taylor, and members of the Conklin lab for her helpful discussions and editorial assistance. This work is supported by the J.

David Gladstone Institutes and grants from the National Institutes of Health: NHLBI, R01-HL61689 (B.R.C.) NHGRI R01-HG002766 (B.R.C) and SBIR 1R44DK53325-01 (Janet A. Warrington and B.R.C.), T32 GM07175 (N.S.).

3.10 References

Aguan, K., J. A. Carvajal, L. P. Thompson and C. P. Weiner (2000). "Application of a functional genomics approach to identify differentially expressed genes in human myometrium during pregnancy and labour." Mol. Hum. Reprod. **6**(12): 1141–1145.

Allen, M. P. and M. Nilsen-Hamilton (1998). "Granzymes D, E, F, and G are regulated through pregnancy and by IL-2 and IL-15 in granulated metrial gland cells." J. Immunol. **161**(6): 2772–2779.

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

Benkusky, N. A., D. J. Fergus, T. M. Zucchero and S. K. England (2000). "Regulation of the Ca2+-sensitive domains of the maxi-K channel in the mouse myometrium during gestation." J. Biol. Chem. **275**(36): 27712-27719.

Bethin, K. E., Y. Nagai, R. Sladek, M. Asada, Y. Sadovsky, T. J. Hudson and L. J. Muglia (2003). "Microarray analysis of uterine gene expression in mouse and human pregnancy." Mol. Endocrinol. **17**(8): 1454–1469.

Brown, D. C., J. M. Theaker, P. M. Banks, K. C. Gatter and D. Y. Mason (1987). "Cytokeratin expression in smooth muscle and smooth muscle tumours." Histopathology **11**(5): 477–486.

Buxton, I. L. (2004). "Regulation of uterine function: a biochemical conundrum in the regulation of smooth muscle relaxation." Mol Pharmacol **65**(5): 1051-1059.

Caron, H., B. van Schaik, M. van der Mee, F. Baas, G. Riggins, P. van Sluis, M. C. Hermus, R. van Asperen, K. Boon, P. A. Voute, S. Heisterkamp, A. van Kampen and R. Versteeg (2001). "The human transcriptome map: clustering of highly expressed genes in chromosomal domains." Science **291**(5507): 1289–1292.

Challis, J. R. G., S. G. Matthews, W. Gibb and S. J. Lye (2000). "Endocrine and paracrine regulation of birth at term and preterm." Endocr. Rev. **21**(5): 514–550.

Charpigny, G., M. J. Leroy, M. Breuiller-Fouche, Z. Tanfin, S. Mhaouty-Kodja, P. Robin, D. Leiber, J. Cohen-Tannoudji, D. Cabrol, C. Barberis and G. Germain (2003). "A functional genomic study to identify differential gene expression in the preterm and term human myometrium." Biol. Reprod. **68**(6): 2289–2296.

Croy, B. A., B. A. McBey, L. A. Villeneuve, K. Kusakabe, Y. Kiso and M. van den Heuvel (1997). "Characterization of the cells that migrate from metrial glands of the pregnant mouse uterus during explant culture." J Reprod Immunol **32**(3): 241-63.

Dahlquist, K. D., N. Salomonis, K. Vranizan, S. C. Lawlor and B. R. Conklin (2002). "GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways." Nat. Genet. **31**: 19–20.

Dolganov, G. M., P. G. Woodruff, A. A. Novikov, Y. Zhang, R. E. Ferrando, R. Szubin and J. V. Fahy (2001). "A novel method of gene transcript profiling in airway biopsy homogenates reveals increased expression of a Na+-K+-Cl- cotransporter (NKCC1) in asthmatic subjects." Genome Res **11**(9): 1473-1483.

Doniger, S. W., N. Salomonis, K. D. Dahlquist, K. Vranizan, S. C. Lawlor and B. R. Conklin (2003). "MAPPFinder: Using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data." Genome Biol. **4**: R7–R7.12.

Dudoit, S., R. C. Gentleman and J. Quackenbush (2003). "Open source software for the analysis of microarray data." Biotechniques **Suppl**: 45–51.

Gabrielsson, B. L., B. Carlsson and L. M. Carlsson (2000). "Partial genome scale analysis of gene expression in human adipose tissue using DNA array." Obes. Res. **8**(5): 374–384.

Garcia-Sanz, J. A., H. R. MacDonald, D. E. Jenne, J. Tschopp and M. Nabholz (1990). "Cell specificity of granzyme gene expression." J Immunol **145**(9): 3111-3118.

Girotti, M. and H. H. Zingg (2003). "Gene expression profiling of rat uterus at different stages of parturition." Endocrinology **144**(6): 2254–2265.

Gown, A. M., H. C. Boyd, Y. Chang, M. Ferguson, B. Reichler and D. Tippens (1988). "Smooth muscle cells can express cytokeratins of "simple" epithelium. Immunocytochemical and biochemical studies in vitro and in vivo." Am. J. Pathol. **132**(2): 223–232.

Havelock, J. C., P. Keller, N. Muleba, B. A. Mayhew, B. M. Casey, W. E. Rainey and R. A. Word (2004). "Human Myometrial Gene Expression Before and During Parturition." Biol Reprod: Epub ahead of print.

http://asthmagenomics.ucsf.edu/pubs/publication/Myometrium.htm. from http://asthmagenomics.ucsf.edu/pubs/publication/Myometrium.htm.

http://www.bioconductor.org. from http://www.bioconductor.org.

http://www.ensembl.org/Multi/martview. from http://www.ensembl.org/Multi/martview.

http://www.genmapp.org. from http://www.genmapp.org.

http://www.genmapp.org/supplemental/MAPPs/supp_fig3.html. from http://www.genmapp.org/supplemental/MAPPs/supp_fig3.html.

Irizarry, R. A., B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs and T. P. Speed (2003). "Summaries of Affymetrix GeneChip probe level data." Nucleic Acids Res. **31**(4): e15.

Liu, G., A. E. Loraine, R. Shigeta, M. Cline, J. Cheng, V. Valmeekam, S. Sun, D. Kulp and M. A. Siani-Rose (2003). "NetAffx: Affymetrix probesets and annotations." Nucleic Acids Res. **31**(1): 82–86.

Lopez, B. A. and R. L. Tamby-Raja (2000). "Preterm labour." Baillieres Best Pract. Res. Clin. Obstet. Gynaecol. **14**: 133–153.

Megy, K., S. Audic and J. M. Claverie (2003). "Positional clustering of differentially expressed genes on human chromosomes 20, 21 and 22." Genome Biol. **4**(2): P1.

Norwitz, E. R., J. N. Robinson and J. R. Challis (1999). "The control of labor." N. Engl. J. Med. **341**(9): 660–666.

Pham, C. T., D. M. MacIvor, B. A. Hug, J. W. Heusel and T. J. Ley (1996). "Long-range disruption of gene expression by a selectable marker cassette." Proc. Natl. Acad. Sci. USA **93**(23): 13090–13095.

Pollard, A. J., C. Sparey, S. C. Robson, A. R. Krainer and G. N. Europe-Finner (2000). "Spatio-temporal expression of the trans-acting splicing factors SF2/ASF and heterogeneous ribonuclear proteins A1/A1B in the myometrium of the pregnant human uterus: a molecular mechanism for regulating regional protein isoform expression in vivo." J. Clin. Endocrinol. Metab. **85**(5): 1928-1936.

Pollard, K. S. and M. J. van der Laan (2002). "A method to identify significant clusters in gene expression data." Proceedings of 6th World Multiconference on Systemics, Cybernetics and Informatics (SCI2002) **II**: 318–325.

Rehman, K. S., S. Yin, B. A. Mayhew, R. A. Word and W. E. Rainey (2003). "Human myometrial adaptation to pregnancy: cDNA microarray gene expression profiling of myometrium from non-pregnant and pregnant women." Mol Hum Reprod **9**(11): 681-700.

Stiemer, B., R. Graf, H. Neudeck, R. Hildebrandt, H. Hopp and H. K. Weitzel (1995). "Antibodies to cytokeratins bind to epitopes in human uterine smooth muscle cells in normal and pathological pregnancies." Histopathology **27**(5): 407–414.

Trinklein, N. D., S. F. Aldred, S. J. Hartman, D. I. Schroeder, R. P. Otillar and R. M. Myers (2004). "An abundance of bidirectional promoters in the human genome." Genome Res. **14**(1): 62–66.

van der Laan, M. J. and K. S. Pollard (2003). "A new algorithm for hybrid clustering with visualization and the bootstrap." J. Stat. Planning Infer. **117**: 275–303.

Vandesompele, J., K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, A. De Paepe and F. Speleman (2002). "Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes." Genome Biol **3**(7): research0034.1-0034.11.

Yu, J. T. and A. Lopez Bernal (1998). "The cytoskeleton of human myometrial cells." J. Reprod. Fertil. **112**(1): 185–198.

Chapter 4
Optimized Selection of Pathways for Over-representation analysis from Gene Expression and Alternative Splicing Data

Nathan Salomonis[1,2], Stan Gaj[3,4], Alexander C. Zambon[1,5], Karen Vranizan[1,6], Chris Evelo[3] and Bruce R. Conklin[1,2]

[1]Gladstone Institute of Cardiovascular Disease, University of California, San Francisco, CA, [2]Pharmaceutical Sciences and Pharmacogenomics Graduate Program, University of California, San Francisco, CA, [3]Department of Bioinformatics, BiGCaT Bioinformatics, Maastricht University, Maastricht, The Netherlands, [4]Nutrigenomics Consortium, Top Institute Food and Nutrition, Wageningen, The Netherlands, [5]Department of Pharmacology, University of California at San Diego, La Jolla, CA, [6]Functional Genomics Laboratory, University of California, Berkeley, CA

4.1 Abstract

Background: Microarray experiments provide a powerful means to elucidate the genetic and expression networks that regulate cellular functions. Numerous tools are available to annotate microarray data with Gene Ontology (GO) and pathway annotations to help illuminate higher-level biological processes related to such data. GO has a hierarchical structure that provides a means to assess gene changes at specific levels of the GO tree as well as nested relationships that include gene associations from children terms. While nesting such relationships provides richer annotations at every level within the GO hierarchy, this approach is disadvantaged in its ability to report a minimal set of significant GO associations, often with many highly related terms reported.

Implementation: To improve on this technique, we have written a stand-alone program, named GO-Elite (http://www.genmapp.org/go_elite), with an efficient algorithm to derive and prune GO and pathway results to provide a minimal set of

non-redundant terms to describe a set of input genes. By considering both the over-representation score of a GO term and its relative position along each trunk and branch of the GO hierarchy, this approach identifies a minimal set of descriptive terms for the original GO categories. GO-Elite can prune up to 90% of redundant GO results, while retaining as much as a 100% of the original associated genes and biological trends. When combined with gene redundancy pruning, GO-Elite can further compress GO and pathway results. In addition to redundancy filtering, GO-Elite provides multiple levels of gene annotation along with GO and pathway-level gene data averaging. To facilitate community contribution and update, this software has built-in tools for the addition of new gene and species relationships as well as a simple underlying data structure.

4.2 Introduction

In recent years, the use of pathway/ontology over-representation analysis (ORA) has become the gold standard for obtaining biological insights into data from genome-scale experiments. This method has wide-ranging applications, from understanding the basis of phenotypic differences in cell and animal models from whole-genome mRNA expression data, to identifying DNA polymorphisms which co-occur within biologically pathways (Wang *et al.* 2007). With an ever-increasing amount of data produced from microarray and high-throughput sequencing technologies, efficient and informative pathway-level analyses are in great demand.

We previously described a freely available tool called MAPPFinder (Doniger *et al.* 2003) for linking genomic datasets to curated biological pathways and hierarchically organized GO terms (Ashburner *et al.* 2000) for ORA. Unlike many other tools, MAPPFinder, which is a component of the GenMAPP application (Salomonis *et al.* 2007), uses user-defined criteria to perform ORA on curated and custom pathways and nested GO terms. Results are displayed in a hierarchically ordered tree, and gene-level changes for any biological term can be visualized in GenMAPP. More recent versions of MAPPFinder perform a permutation analysis of z-scores (normal approximation to the hypergeometric distribution) to determine an overall and multiple hypothesis corrected likelihood that over-representation of each biological category is due to chance. MAPPFinder’s ease of use has made it a highly popular application among computational biologists and typical bench biologists.

Both GO and pathway-level perspectives can provide distinct insights into user data. The GO has a hierarchical structure and thus allows for distinct annotations and ORA at each level of the hierarchy. This structure allows for the user to assess changes for each parent, child and sibling GO term. The genomic coverage of GO terms is typically far greater than that observed for pathways (~70% versus ~ 25%) (Salomonis *et al.* 2007) and can include electronically inferred content including biological interactions that have not been confirmed. Nonetheless, pathway information can provide detailed interactions, additional cellular context and annotation information as well as highlighting critical rate limiting steps.

In addition to the approach used by MAPPFinder, several other methods have been developed that exploit additional relationships either from the user data or from pathway interactions to additionally weight their ORA statistics. One approach, Gene Set Enrichment Analysis (GSEA) (Subramanian *et al.* 2005), uses a prior ranking of the gene data (e.g., by expression clustering) to identify more highly related sets of genes (based on their distance by ranking) that are specifically enriched in certain pathways, GO categories and genomic loci. An advantage of this approach is that stringent filtering of the data may not be necessary since ranking can be based on the pattern and robustness of the change. However, the user must appropriately rank their genes, and that ranking must be biologically significant for ORA. Another intriguing strategy is used by the tool Pathway-Express: (Draghici *et al.* 2007), a component of Onto-Tools (Khatri *et al.* 2007). Pathway-Express weights gene changes based on magnitude and also considers the position of the protein within the context of known interaction networks to determine the overall impact of that change and others on the entire network.

A principle challenge for all of these approaches is to identify a minimal set of non-redundant terms to describe genome-level results. Such methods are necessary to efficiently summarize ORA results in a publication-ready manner. The most popular GO ORA tools, including MAPPFinder, include child gene associations with those from the parent often producing many related GO terms (sharing overlapping gene content) as being over-represented. Methods such as GO-Slim (http://www.geneontology.org), provide a slimmed down version of GO hierarchy to address this, but can result in lost associations to highly specific terms. Although some algorithms provide methods to cluster (DAVID) (Dennis *et al.* 2003) and identify the most statistically enriched GO terms among a set of ORA results (TopGO) (Alexa *et al.* 2006), these tools provide little control over how to best identify an optimal set of descriptive terms based on different enrichment options and to obtain gene associations for these terms. Redundancy is also an issue for pathway/network ORA, where multiple highly related pathways are often present within a single pathway archive.

To address these challenges we developed a new approach, called GO-Elite, to eliminate or highlight redundant GO terms or pathways from ORA results for genomic datasets. This method can be applied directly to genomic data or to existing ORA results from other sources to reduce these results to a manageable, optimal set of descriptive terms. Along with GO-Elite terms, this approach includes maximally informative gene-level annotation and data summarization results, which allow users to easily view associated data at glance. In addition to GO terms, GO-Elite includes pathway-level ORA along with the ability to assess gene redundancy between pathways. This software was written with the bench biologist in mind and thus requires no prior expertise with bioinformatics applications. With a modular structure and simple built-in tools for database file creation, GO-Elite will allow any user to analyze data for any number of custom species or gene relationship systems.

4.3 Methods

4.3.1 Over-representation analysis

There are three critical steps in the GO-Elite analysis: (1) building ORA files, (2) establishing criterion for filtering and (3) GO-Elite pruning and gene/data summarization. Input ORA files for the GO-Elite analysis can be created in one of three ways: (1) by directly using the GO-Elite-ORA in GO-Elite from gene-level data, (2) by using existing MAPPFinder 2.0 (a component of the GenMAPP 2.0 application (http://www.GenMAPP.org, Dahlquist et. al 2002) results and by using (3) output from other GO analysis programs re-formatted into the MAPPFinder format. When building ORA from scratch in GO-Elite, users must begin with at least two input gene lists: a set of regulated genes (numerator) and a list with all gene identifiers (IDs) initially examined (denominator), stored in an existing species-specific directory. Multiple gene ID systems are supported in GO-Elite, including Affymetrix, Ensembl, and EntrezGene, however, additional systems can easily be added by modifying or adding database text files (see Building New Relationship). By default, the user has the option of using either Ensembl or EntrezGene as the primary gene system to link to GO or supplied pathways, where GO relationships are supplied by the respective resource. For non-primary gene systems, such Affymetrix, relationships to Ensembl and EntrezGene are stored in database file folders, which can be augmented by users using built-in tools.

GO-Elite uses the locally stored versions of the latest OBO files (supplied with the program or updated by users from files posted at GeneOntology.org) along with gene to GO relationship files to build a nested tree of gene to GO relationships, similar to MAPPFinder. Unique nested gene to GO relationships are identified for each GO term, a permuted non-adjusted p-value from the GO-Elite-ORA z-score and a Benjamini-Hochberg false-discovery rate p-value are exported to output files. Z-scores are calculated using a normal approximation to the hypergeometric distribution as previously described (Doniger *et al.* 2003), where a particular gene or array ID is counted only once per GO term independent of the number of times it is present in the regulated gene list or redundant gene associations present to IDs in the GO term. To calculate a permuted p-value for each GO term, the GO-Elite-ORA function randomly selects the same number of input genes or array IDs in the user's input gene list from the denominator list, 2000 times (or user-defined) to determine the likelihood of obtaining a z-score greater than or equal to the empirically derived z-score. This p-value is adjusted for multiple hypothesis testing, using the Benjamini-Hochberg (Benjamini *et al.* 1995) correction method and saved as a second set of p-values in the output file. Run-time for ORA in GO-Elite is typically one to dozens of minutes per gene list, depending on the number of genes in the regulated list and number permutations selected.

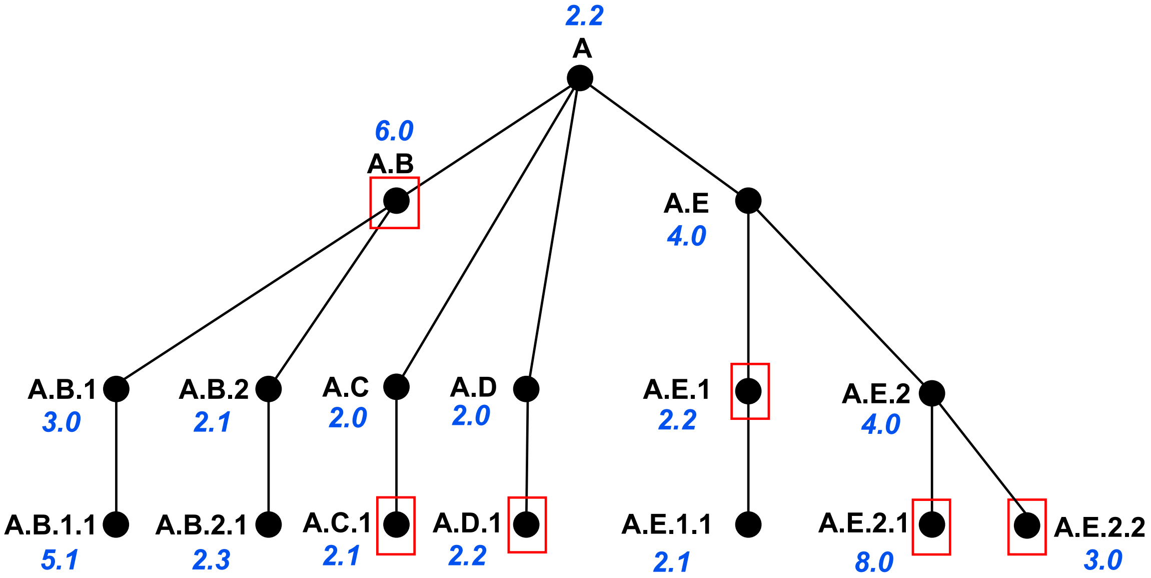
As an alternative to running ORA directly in GO-Elite, such results can also be supplied by MAPPFinder or from other methods supplied in the MAPPFinder format. The MAPPFinder analysis method in GenMAPP is similar to that employed by GO-Elite-ORA function, except that the multiple hypothesis correction method used is based on the Westfall-Young method (Westfall *et al.* 1993). GO-Elite can use the text files automatically produced by MAPPFinder along with corresponding input gene lists to build a minimal non-redundant set of GO terms or pathways matching the user's original GenMAPP criterion along with gene/data summary information.

4.3.2 Filtering statistics

The over-representation z-score, number of genes changed, and non-adjusted permutation p-value generated by either the GO-Elite-ORA function or MAPPFinder are the default statistics used to prune GO terms and pathways in GO-Elite. When ORA data from other GO and pathway analysis programs are used as input for GO-Elite filtering, analogous statistics are recommended. Upon import of ORA data, only those GO terms and pathways that meet the user-defined minimum filters (by default, permuted p-value < 0.05, z-score > 1.96 and number of genes change > 2) are processed for redundancy. Once imported, GO-Elite will compare related GO terms based one of three possible options: (1) the z-score, (2) the number of genes changed or (3) the z-score weighted by genes changed (i.e. combination). The z-score option ranks GO terms only based on the z-score, ranked from highest to lowest. The gene number option allows ranking based on the number of genes changed in the GO hierarchy, again from highest to lowest. The combination option is a weighted metric based on both number of genes changed and the z-score, generated by multiplying the z-score by the log base 2 of the number of genes changed for a given GO term. These scores are used to select GO terms to report by GO-Elite.

4.3.3 GO-Elite Algorithm

GO-Elite can process different types of ORA files, corresponding to GO results (file suffix "-GO.txt") or pathway results (file suffix "-local.txt"). For GO results, once GO terms are initially filtered based on user-defined statistics, all possible parent-child relationships are built and stored for these GO terms, where each parent is the key in the database (Python dictionary object) and all of its children are the values. This full database is stored for later queries, while the full parent-child paths (agglomerated path relationships) for all entries are generated by iterating this process. The program then searches these relationships in a hierarchical manner to identify the highest scoring GO term that either has a higher score (see Algorithm Statistics) than all of its children (along that branch of the tree) or sibling terms (children of a single parent, each representing distinct branches), where at least one of the sibling terms on a branch scores greater than the parent. For these sibling terms, if one sibling branch scores higher than the parent and another branch does not, the highest scoring term from the latter sibling branch is also selected for the GO-Elite output, but the parent term is not. A visual representation of this pruning strategy is shown for a theoretical set of parent-child relationships with corresponding z-scores (Figure 4.1).



Step 1) Build all possible parent-child relationships.
Step 2) Find parents from this list more significant (see score options) than all of their children
Step 3) Find the most significant child terms (downstream of the last bifurcation)
Step 4) Eliminate terms from step 3 that are children of any other term from step 3
Step 5) Report the most signficant parent OR child terms

Figure 4.1. Gene GO-Elite Node Selection Strategy. A theoretical GO tree with GO paths (black text) is displayed along with associated z-scores (blue text). Here, a single GO term can be represented by multiple paths, since the GO is represented as a directed acyclic graph. Red boxes indicate the "reported significant term" selected by GO-Elite. Since multiple paths can exist for each GO term, if the GO paths A.B.2 and A.C.1 correspond to the same GO term (e.g., *apoptosis*), only A.B would be reported since the GO term for both A.B.2 and A.C.1 is a child of A.B. As a result, 16 distinct GO paths would be reduced to five GO terms. The order of GO-Elite operations is listed below this network.

This process allows the user to view the highest scoring term(s) for a particular branch of GO terms and eliminates redundancy of GO terms within the same global category (e.g., *biological process*, *molecular function* and *cellular component*) without needing to consider associated gene content. Since some terms and branches are replicated within the GO hierarchy (redundant), already eliminated or selected GO terms are removed from the results from other branches. Finally, GO-Elite reports the list of pruned GO terms along with a summary of gene symbols associated with input regulated genes, user data, and other reported GO-Elite terms that have gene content that is redundant with that term (if gene lists are supplied by user). These data are stored in a results file for each input gene list, a combined file with all lists run in the batch analysis and detailed gene association files with gene ID, symbol, description, array IDs and data provided in the user input files.

4.3.4 Building New Relationship Files

GO-Elite was built with a modular design, that allows for the addition of new gene ID systems (primary and secondary) and species support, with minimal work and no specific expertise. Gene ID system and species support are controlled by two configuration files in the main database directory (“Databases”) of GO-Elite (“species.txt” and “source_data.txt”). Support for additional species requires the addition of the species code (e.g., Hs) and species name (e.g., Homo sapiens) to the species configuration file (via text or spreadsheet editor), creation of new species directories in the “input” and “Databases” directories (see existing structure), and creation of relationship text files. The latter can be done in an automated fashion using either Affymetrix microarray annotation files for the species (relationships extracted directly by GO-Elite), the cross-species GO-EntrezGene relationship file provided by NCBI (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz) or direct downloads for Ensembl relationships from BioMart (http://www.ensembl.org). The resulting tables, along with GO hierarchy files supplied at http://www.geneontology.org, are sufficient to analyze data for a new species. To add additional gene ID systems, the user must add a new gene ID system name, system code (optionally included in the gene input file to ensure proper gene ID system selection), and an indicator if the system is a possible primary ID (which links directly to GO). For new primary ID systems, the user must supply a gene annotation table, gene-GO relationship table and any array ID or unique ID to primary system (e.g., Ensembl) tables in the corresponding folders. GO-Elite will automatically build intermediate tables from this information (e.g., nested gene-GO relationships). For more information, please see the online or packaged GO-Elite Help file.

4.3.5 Compatibility and Installation

GO-Elite is a Python (Python 2.3.4) program that is provided as a stand-alone Windows executable application and as cross-platform source-code, compressed as a ZIP file. Program files and additional documentation can be found at http://www.genmapp.org/go_elite/ and at http://sourceforge.net. No additional files are required, beyond those typically packaged with the operating system (Python for Mac OS X and Linux, but not required for PCs when using the stand-alone executable).

4.4 Results and Discussion

4.4.1 GO-Elite Increases Gene Ontology ORA Specificity for Example Queries

To estimate the specificity which GO-Elite summarizes biological associations for a given set of genes without including redundant or nonspecific terms, we compared this method to different ORA strategies for three examples. As a test case, we presumed that given a list of genes matching a single GO category, ORA would report only the most descriptive GO term, the input GO term. Although we are only analyzing genes belonging to one GO category, with typical ORA, we would expect an increasing number of additional nonspecific and child GO terms to be reported with an increasing number of genes associated with the input parent category. Therefore, we tested the ability of GO-Elite to report a minimal, non-redundant set of GO terms for input gene lists containing all genes for a GO category with a large (*apoptosis*), mid-range (*response to unfolded protein*), and small number (*stem cell division*) of gene associations.

For each of three input GO category gene sets, GO-Elite was compared to typical ORA, non-nested ORA, and GO-Slim association, all derived using the

GO-Elite-ORA function using different ontology tables or nesting options for more accurate comparison (Table 4.1 A). Although in each case, genes for only one GO term were provided, the smallest input gene set, *stem cell division*, had an additional 57 categories over-represented by conventional ORA filters, although, this method accurately assigned the highest z-score for each query to the same input GO category supplied (e.g., *apoptosis* was the highest scoring term for the *apoptosis* input list, with 100% of genes 'regulated'). For non-nested and GO-Slim ORA, only 6 and 5 additional GO categories were associated with the *stem cell division* test list, respectively; however, neither method included the category *stem cell division* (although the child term, *somatic stem cell division* was selected by non-nested ORA) (Table 4.1 B), reflecting the lack of specificity for reporting nested results for specific terms. GO-Elite analysis for this test case retained *stem cell division* as the top scoring GO term in addition to 13 additional nonspecific terms, eliminating 44 related redundant terms. For the two additional test cases, GO-Elite reduced the outputs by a similar magnitude. While the overall number of categories reported by GO-Slim was also relatively small, this method produced associations to less specific or unrelated terms with much lower overall z-scores. In the case of *response to unfolded protein*, GO-Elite did not report the test GO category but rather its parent term *response to protein stimulus*. Since both the parent and the child terms contained identical possible gene content, GO-Elite selected the parent term, because additional specificity could not be found in the child. Furthermore, when gene content redundancy is considered, only a single GO-Elite term remained for each test case (the positive control), since all other categories contained a subset of genes from this ‘regulated’ input GO category. Thus for the three test cases examined, GO-Elite selected a representative set of high-scoring non-redundant GO terms that reduced the number of results produced by typical approaches.

4.4.2 GO-Elite Eliminates up to 90% of Redundant GO terms without Decreasing Associated Gene Content from Experimental Data

To observe how GO-Elite performs with real data, we examined published microarray datasets for two biological processes highlighted by large phenotypic transitions: (1) differentiation of human embryonic stem cells (hESCs) in to cardiomyocytes (Kita-Matsuo, Barcova et al. submitted) and (2) mouse uterine gestation (Salomonis *et al.* 2005). The first of these datasets, hESC differentiation, possess over 1,500 up-regulated genes (fold>2 and t-test p<0.05). For the second dataset, we looked at two criterion: (1) up or down-regulation across 7 time-points (fold>2 as compared to non-pregnant and f-test p<0.05) corresponding to ~2,500 genes and (2) a smaller subset of these genes (~150) specifically expressed in the myometrium just before to the onset of labor (HOPACH clustering) (Salomonis *et al.* 2005). For each of these datasets, the highlighted pathways and GO terms can be correlated to physiologically relevant functions, in this case regulation of contraction and structural myocyte changes.

A.

GO terms changed	non-nested	GO-Slim	nested	GO-Elite	GCR GO-Elite
Response to unfolded protein	39	16	160	38	1
stem cell division	6	5	58	14	1
apoptosis	571	48	1639	254	1

GCR: Gene Content Redundancy filtering

B.

Top scoring association	non-nested	GO-Slim	nested	GO-Elite	
response to unfolded protein	same	protein complex	same	response to protein stimulus	top-term
	120	15	134	134	z-score
stem cell division	same	multicellular organismal development	same	same	top-term
	110	5	134	134	z-score
apoptosis	same	protein binding	same	same	top-term
	91	14	134	134	z-score

Table 4.1. Comparison of GO-Elite to Alternative ORA Methods Using Simulated Data. A summary of ORA and GO-Elite analyses using different strategies to determine which methods produce the least and most specific set of GO associations given all genes associated with 3 different GO categories (*response to unfolded protein*, *stem cell division* and *apoptosis*). ORA methods include GO-Elite-ORA selecting non-nested GO terms (genes specifically associated with each term), nested (nesting child associated terms to parents), GO-Slim, GO-Elite filtering, and GO-Elite filtering including gene content redundancy filtering (GCR). (A) The number of GO terms that were produced (after common default filtering) for each input GO-category list. (B) The top ranked GO term (based on z-score) for each method along with the GO term’s z-score value are shown, where “same” indicates that the top-ranked GO term is the same as the GO-category input gene list.

To determine if sensitivity is affected by GO-Elite pruning, we compared excluded GO terms to the retained GO-Elite terms and corresponding gene content to see what information may be lost upon filtering (Tables 4.2 A-B). For this analysis, we compared the three built-in filtering strategies that can be employed by GO-Elite: (1) ordering by z-score, (2) number of genes changed and (3) z-score weighted by genes changed (combination). Traditional ORA for each of the datasets and default filtering resulted in 520 GO terms being reported for hESC regulated, 845 for uterine gestational and 56 term-regulated criterion. GO-Elite pruning, ranking by z-score alone, resulted in a 61- 88% reduction in the number of GO categories reported, where the percentage of GO terms excluded is inversely proportional to the number of originally filtered ORA categories. This effect is due to the improved ability of GO-Elite to filter terms given more related sets of terms. Examination of eliminated versus retained GO terms revealed that the eliminated content was largely redundant with the retained set and preserved the ORA reported biological trends (Table 4.3 A-B).

A.

		GO-Elite Pruning Options		
GO terms changed	GO-Elite-ORA	z-score	gene count	combination
hESC cardiac differentiation	520	146 (47)	110 (42)	109 (37)
uterine gestation time-course	845	103 (50)	20 (11)	69 (32)
Uterine term pregnancy	56	22 (6)	16 (7)	19 (6)

B.

		GO-Elite Pruning Options		
Associated Genes	GO-Elite-ORA	z-score	gene count	combination
hESC cardiac differentiation	872	753	872	847
uterine gestation time-course	2553	2362	2553	2438
uterine term pregnancy	111	98	111	106

Table 4.2. GO-Elite Filtering For Muscle Differentiation and Remodeling Paradigms. Different GO-Elite filtering options are compared with the GO-Elite-ORA. The three pruning methods shown are z-score, gene-count, and combination (gene-count-weighted z-scores), obtained from the input ORA files. (A) The number of GO terms changed using the different strategies are shown, with the number of additional GO terms that are redundant based on gene content in parentheses (e.g., for hESC cardiac differentiation, of 146 highlighted GO terms with the GO-Elite z-score method, 47 have gene content redundant with other GO-Elite terms). (B) Number of genes associated with each set results produced by the different GO-Elite filtering methods.

A.

GO Name	GO Type	Number Changed	Z-score	Permute P	Adjusted P	redundant with terms	inverse redundant
Intercellular junction	C	8	6.35	0	0		intercalated disc
B cell differentiation	P	4	5.38	0.0005	0.352		
intermediate filament	C	6	5.03	0.0005	0.352		
Intercalated disc	C	4	4.98	0	0	intercellular junction	
basolateral plasma membrane	C	5	4.54	0.0015	0.5025		
structural constituent of cytoskeleton	F	6	4.28	0.0055	1		
cytokine and chemokine mediated signaling pathway	P	4	4.16	0.0015	0.5025	cell development, death	
positive regulation of growth	P	3	3.98	0.0125	1		
extracellular region	C	43	3.89	0	0		endopeptidase inhibitor activity, serine-type endopeptidase activity
erythrocyte homeostasis	P	3	3.20	0.016	1	cell development	
endopeptidase inhibitor activity	F	5	3.17	0.011	1	extracellular region	
apical plasma membrane	C	3	3.13	0.016	1		
cellular structure morphogenesis	P	12	3.00	0.0055	1		
regulation of cell adhesion	P	3	2.91	0.0175	1		
amine biosynthetic process	P	3	2.85	0.018	1		
morphogenesis of an epithelium	P	5	2.71	0.011	1		
intrinsic to membrane	C	45	2.70	0.016	1		
carbohydrate biosynthetic process	P	3	2.45	0.041	1		
cell development	P	21	2.26	0.0185	1		cytokine and chemokine mediated signaling pathway, death, erythrocyte homeostasis
serine-type endopeptidase activity	F	5	2.26	0.029	1	extracellular region	
epidermis development	P	3	2.17	0.0445	1		
death	P	13	2.14	0.0265	1	cell development	cytokine and chemokine mediated signaling pathway

B.

GO Name	GO Type	Z-score
apical junction complex	C	6.25
apicolateral plasma membrane	C	6.11
intermediate filament cytoskeleton	C	4.97
tight junction	C	4.70
plasma membrane	C	3.75
cell junction	C	3.60
protease inhibitor activity	F	3.13

GO Name	GO Type	Z-score
biological adhesion	P	2.57
cell adhesion	P	2.57
structural molecule activity	F	2.54
integral to membrane	C	2.53
lymphocyte differentiation	P	2.27
enzyme regulator activity	F	2.26
hemopoietic or lymphoid organ development	P	2.24

apical part of cell	C	3.11	developmental process	P	2.23
serine-type endopeptidase inhibitor activity	F	3.11	growth	P	2.22
membrane	C	3.07	regulation of growth	P	2.18
cell morphogenesis	P	3.00	nitrogen compound biosynthetic process	P	2.17
B cell activation	P	2.92	cell death	P	2.14
enzyme inhibitor activity	F	2.88	serine hydrolase activity	F	2.11
anatomical structure morphogenesis	P	2.83	serine-type peptidase activity	F	2.11
plasma membrane part	C	2.80	endopeptidase activity	F	2.08
anatomical structure development	P	2.71	immune system development	P	2.05
membrane part	C	2.60	apoptosis	P	1.97

Table 4.3. Comparison of Uterine Term Gestation Elite and Non-Elite GO terms. (A) 22 GO-Elite terms for a cluster of genes, maximally up-regulated just prior to the onset of labor from a mouse uterine microarray analysis. (B) 34 GO terms that were filtered out of the original set of 56 ORA highlighted terms.

When comparing total gene content before and after pruning, we observe a loss of 11-14% of genes associated with input GO categories as compared to GO-Elite terms. Examining these lost genes revealed that they typically align to more general GO categories with relatively high gene content. These data therefore suggests that GO-Elite pruning dramatically reduces the amount of redundant content without compromising biological sensitivity.

While a 14% loss in associated gene content is largely acceptable, given the large decrease in redundant GO terms, we wanted to assess the cost versus benefit of alternate GO ranking strategies in more detail. For both gene number and combination methods we observed an increase in the associated gene content, relative to using the z-score method alone. In fact, ranking by gene number yielded no loss in gene content while further decreasing the number of reported GO categories to as much as 13% the number of original ORA terms (hESC differentiation)(Table 4.2 A). While this observation suggests that gene number is the preferred method for GO-Elite analysis, this method can result in agglomeration of highly descriptive child terms, into large, less descriptive parents, which may prevent highlighting important biology.

Combination filtering, on the other hand, largely preserved gene content and decreased the number of reported terms, but did not largely effect the biological description of these datasets, compared to z-score alone or the published reports. This is likely because gene number ranking will always favor the GO term that is highest up on the tree and thus contains more genes, while the z-score ranking can commonly favor more specific child terms with sibling terms that will be retained by GO-Elite, because they are on distinct branches. The combination of gene number and z-score however, will tend to favor parents that can occur before a bifurcation as opposed to child terms, since the scores are close but lower in the parent because they have a larger denominator. In the case of uterine term pattern genes, 15 GO terms were shared between the combination and z-score methods, whereas five GO terms were unique to combination and eight for z-score. Comparison of these distinct results reveals that the GO terms *B-cell differentiation*, *epidermis development*, *cellular structure morphogenesis*, and *morphogenesis of an epithelium*, found using the z-score method, are represented by their common upstream parent term, *anatomical structure development,* which is a general development category. Therefore, we conclude that the combination z-score approach is suitable for retaining associated gene content from ORA analyses, though it produces a loss of some biologically descriptive content.

4.4.3 Gene Content Redundancy Additionally Increases Specificity for Both GO and Pathways

We have shown that pruning of the hierarchy is a useful means of selecting a set of fairly non-redundant GO terms. However, this method does not take into account terms that cannot be directly related to each other within the GO hierarchy but contain overlapping gene content. In addition to the primary GO-Elite analysis strategy, considering gene redundancy, can be useful in further pruning of GO terms and even non-GO pathways with redundant gene content.

In GO-Elite, two additional columns of information are included with the summary results, terms redundant with a given category and vice versa (Table 4.3 A). As an example, two terms reported from the uterine gestation GO-Elite analysis, *gas transport* (biological process) and *oxygen transport activity* (molecular function) have overlapping gene content, where *oxygen transport activity transport* genes (n=4) are a subset of *gas transpor*t (n=5). As a result, in the first redundancy column, *gas transport* will be reported as having a super-set of genes for *oxygen transport activity*. In the second column, GO-Elite reports the reciprocal relationship, that *oxygen transport activity* is redundant with *gas transport*. These annotations allow the user to filter for relationships that are maximally descriptive for their dataset, eliminating either categories containing redundant gene content with another or the inverse, which can highlight or more specialized, less generic terms. Considering the three experimental datasets analyzed here, after removing terms annotated as being redundant with other GO terms, we observe as much as an additional 38% decrease in redundant terms (Table 4.2 A). Since these terms have entirely redundant gene content, there is no decrease in the total associated gene content. As a result, however, more descriptive GO categories with fewer genes tend to be 'absorbed' into larger, less descriptive categories (e.g., *creatine kinase activity* is redundant with the term *cytoplasm*). Therefore, this annotation, in the result summary file, can be used as a guide to decide which general categories (encompassing the redundant terms and genes) or more specific categories do not introduce additional novel biological information, prior to removal.

While the GO-Elite method can only be applied to data from hierarchically organized ontology data, the gene redundancy annotations can be applied to non-ontology data, such as GenMAPP pathway associations, also computed by GO-Elite-ORA. This approach can be useful when a set of genes are regulated that correspond to core biological processes (e.g., *apoptosis*, *cell cycle*, *integrin-mediated signaling*), which are often described among in several cell-type or cell response specific versions of those pathways (see http:/www.wikipathways.org). In addition to providing pre-packaged GenMAPP pathway gene association data, GO-Elite has a simple format for adding or replacing the existing non-GO pathway content for ORA and gene redundancy analysis. Thus, this method is a general multipurpose approach for assessing GO term or pathway gene-content redundancy.

4.4.4 Extended Annotation and Data Summarization of GO-Elite Level Terms

A common limitation of pathway/ontology analysis tools is efficient gene annotation and interpretation of user data after ORA. For example, MAPPFinder allows users to export lists of associated user input IDs associated with a particular biological term or view color criterion for associated genes in the context of a GO list or GenMAPP pathway. However, to access the set of associated genes (as opposed to an array ID) along with annotations and associated data (e.g., gene expression statistics and annotations), complex queries are required against the ORA results and the input data. A second challenge is how to summarize these gene annotations and associated quantitative data in an efficient format for direct publication or downstream analyses (e.g., clustering of mean pathway expression values at different time-points).

After automatic GO-Elite pruning, three methods of gene summarization are provided with the GO-Elite output: (1) associated gene symbol column in the GO and pathway summary files, (2) specific gene associations, annotations and data for each GO term and pathway (gene-association file), and (3) gene ranking for over-representation among GO terms and pathways (gene-ranking file). Gene symbols provided in the annotation file provide a simple means to summarize associated data. Detailed gene associations allow the user to not only see which input IDs link to primary gene IDs for each biological category, but also allows access to detailed gene annotations (symbol and description) in addition to any user data provided in the input ID file. Since these are pruned list of terms, there should be minimal overlap in gene content between distinct terms. Nonetheless, to better assess if certain genes are over-represented among certain GO-Elite categories, the user can view the gene-ranking files which show which genes tend to be most represented among GO-Elite terms, providing a low-level means of eliminating GO terms and pathways with redundant content. This method not only highlights promiscuous genes that tend to be associated with several GO-Elite biological categories, but also provides a gene-centric view of associated GO-Elite terms.

In addition to summarizing gene annotations, GO-Elite can summarize numerical data for pruned GO terms and pathways. This method is similar to GO-Quant (Yu *et al.* 2006), which links prior MAPPFinder ORA biological terms to gene IDs in a GenMAPP data file to calculate mean expression values for each biological term, for any given number of time-points/conditions. The analogous option in GO-Elite allows any column of numerical data present with the input gene IDs (e.g., array IDs) to be averaged first at the level of primary gene IDs (e.g., Ensembl, EntrezGene) and next at the level of GO-Elite terms for inclusion in the GO-Elite summary results file. An example of this process is illustrated with log2 fold changes for all differentially regulated genes for the different time-points of uterine gestation compared to baseline, non-pregnant mouse uterus (Figure 4.2). Here, GenMAPP pathway MAPP results generated by GO-Elite's basic ORA algorithm are clustered for all uterine time-point mean pathway fold changes relative to baseline non-pregnant uterine mRNA profiles. This method of analysis allows for a global view of gene expression changes directly linked to biological processes. While we have used the mean of fold changes for summarization, this could easily be applied to any other values linked to the input gene ID data.

4.5 Conclusions

A critical challenge in the analysis of large-scale genomic datasets has been proper description and summarization of gene-associated changes in the context of known biology. Given the rate at which such data are produced and commonly utilized in publications, researchers require tools capable of highlighting the most relevant biological associations. In addition to GO categories, this includes over-representation of pathways, efficient summarization of gene content and associated gene data, and customizable tools that can be easily updated and optimized for new species/gene analyses.

Here, we describe a new stand-alone software package called GO-Elite for the analysis of user genomic data that quickly identifies a minimal set of non-redundant GO terms, pathways, and associated gene data. While the method itself is not entirely novel (Barriot *et al.* 2007), it is the only available application we are aware of that is currently capable of performing such integrated analyses. GO-Elite was specifically written with the needs of the genomics community in mind, providing flexibility in the type of input gene IDs, species and types of applications required. In addition to pruning of existing ontology and pathway data, GO-Elite can perform batch ORA analyses of user gene data, with superior speed and performance than our previous application, MAPPFinder.

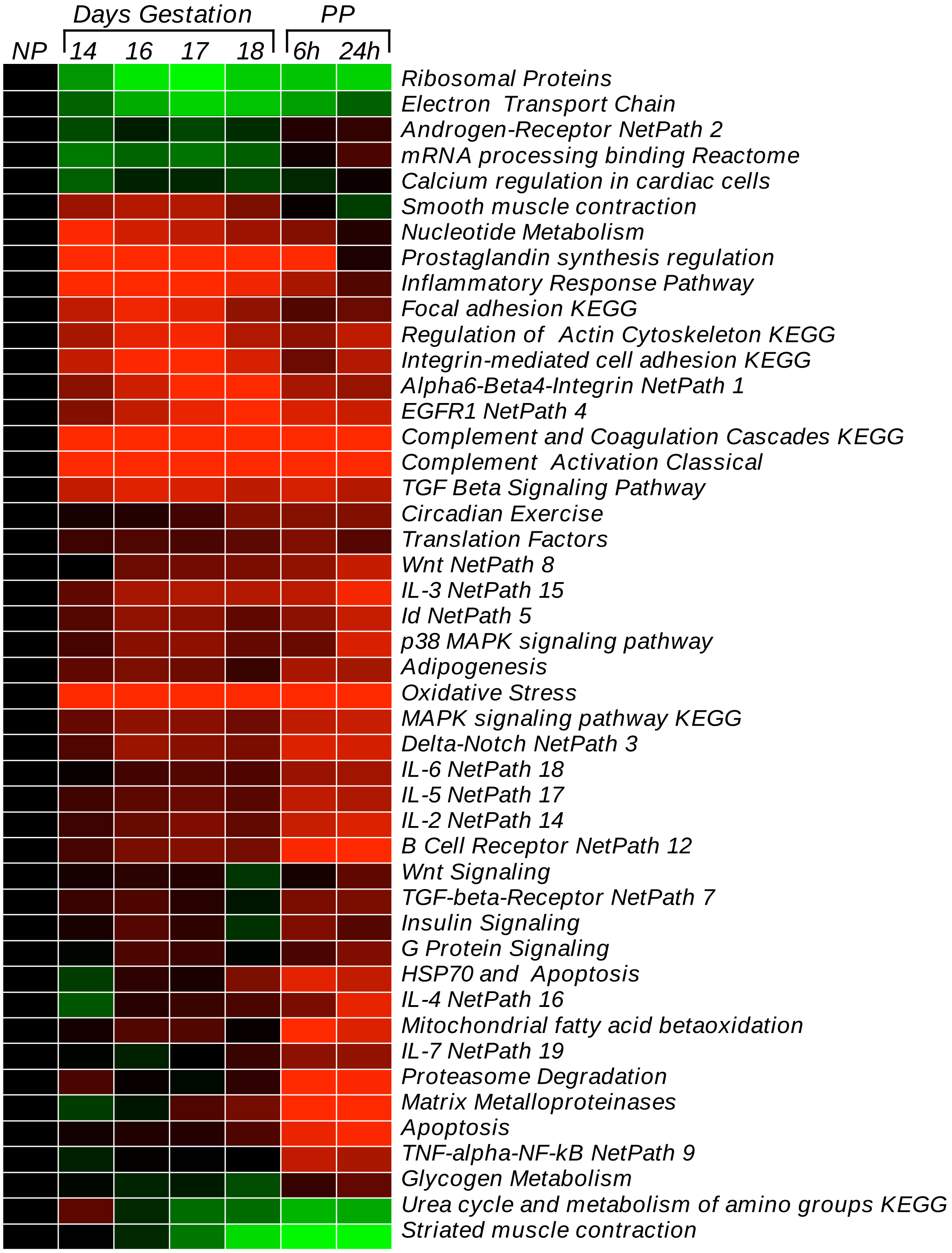


Figure 4.2. Pathway-level Analysis of Numerical Gene Data. The mean pathway fold change (log2) at multiple time-points of mouse uterine gestations are shown for Affymetrix probe sets linked to GO-Elite selected

GenMAPP MAPPs. These fold changes are relative to non-pregnant (NP) mouse uterus, for days post fertilization (14.5, 16.5, 17.5 and 18.5 days) in addition to postpartum (PP) time-points (6 and 24hrs). This data was clustered using the program HOPACH, where red indicates up-regulation and green, down-regulated mean expression changes for all regulated gene linked probe sets, for a given pathway.

Application of this method to both mouse and human expression data dramatically reduced the results produced by conventional ORA, with a reduction of up to 90% of the original GO categories (when including gene content redundancy pruning), while preserving both descriptive categories and associated gene content. While this method of ORA is relatively basic as compared to more recent methods, such as GSEA (Khatri *et al.* 2007) or pathway topology analysis (Pathway-Express(Draghici *et al.* 2007), it is compatible with nearly any ORA data, given the proper input (see associated documentation). However, this method does not attempt to select the most statistically significant term relative to other related terms, but rather provides different options for users to rank terms using different ORA statistics.

In addition to these methods, GO-Elite includes multiple levels of gene annotation GO-Elite terms and support for pathway-level gene data summarization. We anticipate this approach will be an important addition to the tool kit used by biologists for large-scale, genome-level ORA in the years to come.

4.6 Acknowledgments

We would like to the thank Stephen Ordway, Benjamin Phillips, Kristina Hanspers and Alexander R. Pico for editorial help with manuscript and the GO-Elite user community for the helpful feedback and suggestions. This work was supported by the National Heart, Lung, and Blood Institute grant HL66621.

4.7 References

Alexa, A., J. Rahnenfuhrer and T. Lengauer (2006). "Improved scoring of functional groups from gene expression data by decorrelating GO graph structure." Bioinformatics **22**(13): 1600-7.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

Barriot, R., D. J. Sherman and I. Dutour (2007). "How to decide which are the most pertinent overly-represented features during gene set enrichment analysis." BMC Bioinformatics **8**: 332.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate—a new and powerful approach to multiple testing." Journal of the Royal Statistical Society B **57**: 289-300.

Dennis, G., Jr., B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane and R. A. Lempicki (2003). "DAVID: Database for Annotation, Visualization, and Integrated Discovery." Genome Biol **4**(5): P3.

Doniger, S. W., N. Salomonis, K. D. Dahlquist, K. Vranizan, S. C. Lawlor and B. R. Conklin (2003). "MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data." Genome Biol **4**(1): R7.

Draghici, S., P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu and R. Romero (2007). "A systems biology approach for pathway level analysis." Genome Res **17**(10): 1537-45.

Khatri, P., C. Voichita, K. Kattan, N. Ansari, A. Khatri, C. Georgescu, A. L. Tarca and S. Draghici (2007). "Onto-Tools: new additions and improvements in 2006." Nucleic Acids Res **35**(Web Server issue): W206-11.

Salomonis, N., N. Cotte, A. C. Zambon, K. S. Pollard, K. Vranizan, S. W. Doniger, G. Dolganov and B. R. Conklin (2005). "Identifying genetic networks underlying myometrial transition to labor." Genome Biol **6**(2): R12.

Salomonis, N., K. Hanspers, A. C. Zambon, K. Vranizan, S. C. Lawlor, K. D. Dahlquist, S. W. Doniger, J. Stuart, B. R. Conklin and A. R. Pico (2007).

"GenMAPP 2: new features and resources for pathway analysis." BMC Bioinformatics **8**: 217.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proc Natl Acad Sci U S A **102**(43): 15545-50.

Wang, K., M. Li and M. Bucan (2007). "Pathway-Based Approaches for Analysis of Genomewide Association Studies." Am J Hum Genet **81**(6).

Westfall, P. H. and S. S. Young (1993). "Resampling-based multiple testing: Examples and methods for p-value adjustment." John Wiley & Sons.

Yu, X., W. C. Griffith, K. Hanspers, J. F. Dillman, 3rd, H. Ong, M. A. Vredevoogd and E. M. Faustman (2006). "A system-based approach to interpret dose- and time-dependent microarray data: quantitative integration of gene ontology analysis for risk assessment." Toxicol Sci **92**(2): 560-77.

Chapter 5

Analyzing alternative splicing along multiple lineage commitment pathways in the context of protein function and regulation by non-coding RNAs

Nathan Salomonis[1, 2], Brandon Nelson[3], Karen Vranizan[4], Alexander R. Pico[1], Kristina Hanspers[1], Allan Kuchinsky[5], Linda Ta[1], Mark Mercola[3] and Bruce R. Conklin[1, 2]

[1]Gladstone Institute of Cardiovascular Disease, San Francisco, CA, [2]Pharmaceutical Sciences and Pharmacogenomics Graduate Program, University of California, San Francisco, CA, [3]Burnham Institute for Medical Research, La Jolla, CA, [4]Functional Genomics Laboratory, University of California, Berkeley, CA, [5]Agilent Technologies, Santa Clara, CA

5.1 Abstract

Background: Although several of the essential core transcriptional control elements in human and mouse embryonic stem cells (ESCs) have been identified, the specific protein isoforms that enable ESCs to maintain self-renewal and pluripotency or promote tissue lineage specification are still largely unknown. To better define these crucial regulatory cues, we require new tools to interrogate ESCs and lineage-restricted cells as homogenous populations at both the level of transcription and alternative splicing (AS).

Results: To assess the transcriptional and splicing profile of human ESCs and ESC derived cardiac cells, we modified the H9 ESC line to allow for drug selection of mouse pluriptotent ESCs and alpha myosin heavy chain expressing cardiac spheroids (CSs). Exon-level microarray expression data from undifferentiated ESCs, day 40 CSs, and other lineage-restricted cells and tissues were used to identify splice isoforms with cardiac-restricted or common differentiation expression patterns. A new, open-source application called AltAnalyze was developed to identify hundreds of splice events for each of these two pattern groups corresponding to the pathways of cell death, serine-threonine kinase activity, muscle specification, and cytoskeletal-remodeling. Integration of these data with protein level annotations and predicted microRNA binding sites highlighted novel changes in domain and binding site architecture that have profound implications for the biology of AS proteins.

Summary: By combining robust, genome-wide AS predictions with new functional annotations, we have uncovered potential mechanisms hypothesized to influence lineage commitment and ESC maintenance at the level of specific splice isoforms and microRNA regulation.

5.2 Introduction

Embryonic stem cell (ESC) differentiation is a powerful system for dissecting out developmental cues required for lineage commitment *in vitro*. Similar to their *in vivo* counterpart, the cells of the inner cell mass of the blastocyst, ESCs self-renew and direct differentiation to all three adult germ layers. The maintenance of pluripotency and self-renewal are dependent on the expression of a core set of transcription factors, including Oct4, Sox2, and Nanog. Whole-genome expression (Ivanova *et al.* 2006), microRNA (miRNA) (Mitschischek 1991) and epigenetic analyses (Boyer *et al.* 2005; Loh *et al.* 2006) of ESC differentiation have led to the discovery of additional factors with similar expression patterns that interact with this core set of transcription factors to regulate pluripotency.

While these studies are informative, there is still a large gap in our understanding of the mechanisms that regulate ESC maintenance up-stream and down-stream of these core regulatory components and the steps required for proper cell fate commitment. These challenges exist, to a large extent, due to a difficulty in obtaining pure populations of fully differentiated cells and lack of detailed transcript expression profiles that allow for the prediction of alternative splicing (AS).

As many as 80% of all human genes undergo AS to produce multiple mRNA transcripts with the differential inclusion of exons and introns (Lee *et al.* 2005). This mechanism results in considerable variation in transcripts and proteins among distinct cell types. This variation often results in unique proteins with biologically distinct composition and function. The functional impact of splicing includes altered domain composition and cellular localization, both of which can lead to distinct signaling properties of the resulting protein, while AS in untranslated mRNA regions can impact RNA stability and localization (Cooper 2005). Disruption of AS for a single gene can have profound effects on cellular development, ranging from improper neonatal cardiac adaptation (Xu *et al.* 2005) to sex-determination (Hammes *et al.* 2001) and synaptogenesis (Burgess *et al.* 1999).

Since ESCs can differentiate into all lineages of cells, characterizing isoform expression along specific lineage paths requires efficient methods to obtain pure populations of cells. A step toward this goal was recently made with the profiling of multiple human ESC (hESC) lines differentiated to neural precursors, isolated using an effective neural differentiation protocol, with whole-genome exon-arrays (Yeo *et al.* 2007). This analysis highlights coordinate AS of serine/threonine kinases and helicases, suggesting that coordinated programs may exist in ESCs to direct both cell-type-specific and general differentiation programs.

To identify AS occurring with the differentiation to cardiac progenitors, we have performed exon-level genome profiling of homogenous populations of human undifferentiated ESCs and cardiac spheroids (CSs) obtained using a new selectable marker strategy (Kita-Matsuo, Barcova et al. submitted). To further identify AS events that are enriched in cardiac progenitor differentiation or common to multiple lineages, we have used an analysis of variance method (ANOVA) to compare these profiles to existing differentiation datasets and adult tissues. Alignment of alternatively regulated exons that match to two patterns (common to neural and cardiac differentiation or enriched just in cardiac), identifies unique proteins and functional elements with novel domain-level changes that could significantly alter protein function. Analysis of mRNA sequences corresponding to known and predicted miRNA binding sites revealed the gain and loss of such sequences as a direct result of alternative exon inclusion, suggesting an additional mechanism for regulation of translation inhibition.

5.3 Results and Discussion

5.3.1 Characterization of hESC-derived cardiac spheroids

To isolate a homogenous population of both undifferentiated hESCs and derived cardiac progenitors, the H9 ESC line was modified to stably express two drug selection markers driven by the pluripotent-specific REX-1 and cardiac-specific myosin heavy chain a (MHCα or MHY6) promoters. This strategy allowed for the selection for REX-1 positive (Rex+) pluripotent ESCs. To isolate cardiac progenitors, embryoid bodies were selected for using the cardiac marker MHCα and cultured for an additional 27 days (day 40 of differentiation). Day 40 CSs possessed action potentials and axial force measurements equivocal to normal fetal cardiomyocytes (Kita-Matsuo, Barcova et al. submitted). RNA harvested from selected hESCs and CSs were processed and hybridized to Affymetrix human exon 1.0 arrays. The resulting data were combined with a dataset of neural progenitor (NP) differentiation previously described (Yeo *et al.* 2007) and a dataset of 11 adult human tissues, analyzed using the same microarray platform (see materials and methods). Un-biased comparison of gene expression profiles derived from exon-level probe sets using hierarchical clustering (Eisen *et al.* 1998) demonstrated that Rex+ hESCs and two independent hESC lines, Cythera and HUES6 hESCs, co-segregate from differentiated hESC-derived cells (CSs and NP) (Figure 5.1 A). These data demonstrate that distinct hESC lines are more similar to each other than respectively derived differentiated lineage-restricted cells. Not surprisingly, all cell culture conditions had greater overall expression similarity to each other than to adult tissues, with the exception of cerebellum. Similar results were also obtained using an independent clustering method, HOPACH (Salomonis *et al.* 2005) (data not shown). Although the CSs were not closely correlated with samples from adult heart, both of these conditions possessed highly similar gene expression levels for all established cardiac markers examined (Figure 5.1 B).

5.3.2 Segregation of putative pluripotent and cardiac-specific gene expression changes occurs along predicted pathways

The primary aim of our analysis was to identify differential transcription and AS events that specifically correspond to either cardiac-specification or inhibition/promotion of differentiation. To accomplish this, we set out to design a similar strategy for analysis of both biological paradigms that utilizes multiple ESC differentiation datasets.

Using conventional gene expression filters (>2 fold, t-test p<0.05), 3,044 genes were found to be differentially expressed out of the 29,151 Ensembl genes examined in the day 40 CSs relative to Rex+ hESCs. Since this comparison only considers two conditions (hESCs and CSs), the differentially expressed genes will include transcripts that are (1) selectively regulated in the transition to cardiac precursors, (2) specifically associated with an ESC or differentiated cells, and (3) common to multiple but not all cell lineages. By directly comparing our cardiac differentiation profiles to non-cardiac differentiation profiles (NP differentiation), we can begin to segregate these gene expression changes into these more discrete categories and gain insight into the molecular mechanisms that contribute to these different programs.

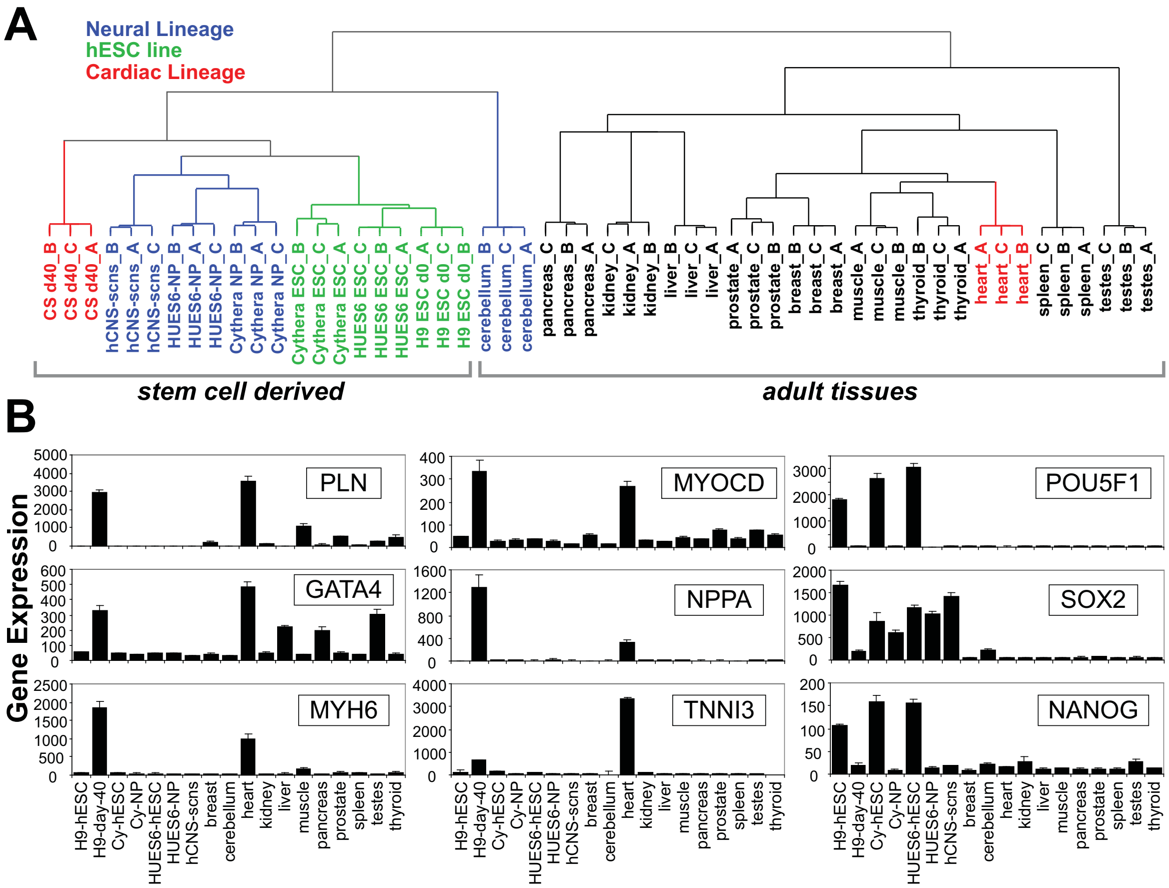
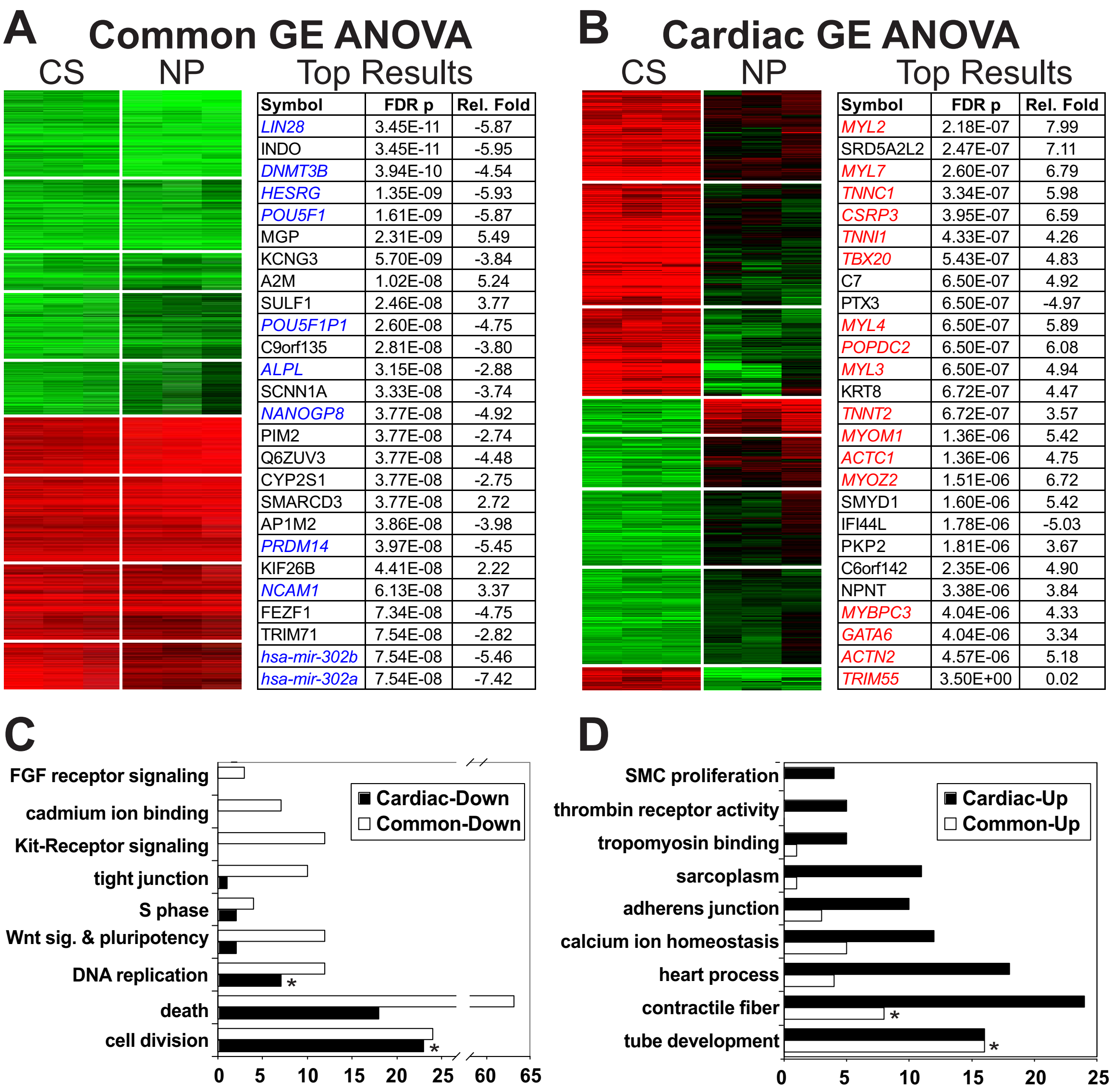


Figure 5.1. Cardiac precursors and hESCs have consistent expression profiles with in vitro and in vivo analogues. (A) Human Affymetrix exon array data are compared for Rex+ hESCs and derived CSs, Cythera, and HUES6 lines differentiated to NPs and 11 adult tissues, normalized together and clustered by array. All stem cells or stem cell-derived data, clusters into a distinct node of the dendrogram, with a high degree of correlation between hESCs from distinct cell lines. (B) Gene expression profiles for this combined dataset, for specific markers of cardiac-specification and pluripotency.

To implement this comparison, we used a two-way ANOVA strategy, comparing the day 40 CS samples to NP samples along with their respective pluripotent ESC controls using the LIMMA package in Bioconductor (Dudoit *et al.* 2003). Since the in vitro differentiation data clusters into distinct sub-groups, all statistical comparisons were performed only using this data (separate normalization and background correction) and the combined *in vitro* and *in vivo* data used only for gene expression clustering and down-stream comparisons. For these comparisons, the Cythera hESC line data was used to examine NP differentiation, since this dataset had small sample-to-sample variability than the HUES6 hESC line data, when analyzed with RMA (data not shown). Each of the differentiated conditions was normalized to the mean of its appropriate hESC reference set for this comparison and used to calculate a p-value to assess whether both the neural precursor and cardiac cell profiles have a common or opposite pattern (differentiation or interaction effect, respectively). Of 3,030 differentially expressed day 40 CS genes, 1,962 had a common expression pattern between cardiac and neural differentiation (differentiation $p<0.05$) and 951 of these were preferentially regulated in the differentiation to CSs (interaction $p<0.05$).

Among the genes with the lowest ANOVA differentiation p-value were the pluripotency inducing factors LIN28 (p=3.45E-11) and OCT3/4 (p=1.61E-09). Clustering of these genes across the examined conditions reveals that the majority of the genes are consistently up- or down-regulated relative to hESCs (Figure 5.2 A), leading to the hypothesis that these are hESC- or differentiation specific-transcripts. Both up- and down-regulated genes were significantly associated with the regulation of Wnt signaling by pathway over-representation analysis with the program GO-Elite (http://www.genmapp.org/go_elite). Down-regulated genes were specifically enriched in pathways of DNA-replication, cell cycle control, and regulation of pluripotency, whereas up-regulated genes were over-represented among pathways of stem cell differentiation, organ system development (bone, brain, muscle, immune, fat and circulatory), TGF-β signaling, and focal adhesion formation (Figure 5.2 C).

The top-ranked genes expressed with an ANOVA defined cardiac-specific expression pattern, largely consisted of well-described cardiac markers (TNNC1, TNNI1, TNNI3, MYH6, MYH7, PLN, GATA4, GATA6, NPPA) and signaling (CHRNA1, CHRM2) and developmental cardiac regulators (TBX5, TBX20) (Figure 5.2 B). The up-regulated cardiac-specific genes were highly enriched in early cardiac developmental pathways, muscle proliferation, cardiac muscle contraction, adherens junction, and blood vessel and tube development, while down-regulated genes were associated with G1-to-S cell cycle control, chromatin remodeling, mRNA processing, and androgen receptor signaling (Figure 5.2 D). These results demonstrate that direct comparison of independent differentiation datasets using an ANOVA strategy is sufficient to segregate regulated genes into tissue-restricted categories that conform to the expect outcome. Therefore, this method was deemed sufficient to identify alternative exons, either in common among differentiation paradigms or that are specifically regulated in CSs.



Symbol	FDR p	Rel. Fold
LIN28	3.45E-11	-5.87
INDO	3.45E-11	-5.95
DNMT3B	3.94E-10	-4.54
HESRG	1.35E-09	-5.93
POU5F1	1.61E-09	-5.87
MGP	2.31E-09	5.49
KCNG3	5.70E-09	-3.84
A2M	1.02E-08	5.24
SULF1	2.46E-08	3.77
POU5F1P1	2.60E-08	-4.75
C9orf135	2.81E-08	-3.80
ALPL	3.15E-08	-2.88
SCNN1A	3.33E-08	-3.74
NANOGP8	3.77E-08	-4.92
PIM2	3.77E-08	-2.74
Q6ZUV3	3.77E-08	-4.48
CYP2S1	3.77E-08	-2.75
SMARCD3	3.77E-08	2.72
AP1M2	3.86E-08	-3.98
PRDM14	3.97E-08	-5.45
KIF26B	4.41E-08	2.22
NCAM1	6.13E-08	3.37
FEZF1	7.34E-08	-4.75
TRIM71	7.54E-08	-2.82
hsa-mir-302b	7.54E-08	-5.46
hsa-mir-302a	7.54E-08	-7.42

Symbol	FDR p	Rel. Fold
MYL2	2.18E-07	7.99
SRD5A2L2	2.47E-07	7.11
MYL7	2.60E-07	6.79
TNNC1	3.34E-07	5.98
CSRP3	3.95E-07	6.59
TNNI1	4.33E-07	4.26
TBX20	5.43E-07	4.83
C7	6.50E-07	4.92
PTX3	6.50E-07	-4.97
MYL4	6.50E-07	5.89
POPDC2	6.50E-07	6.08
MYL3	6.50E-07	4.94
KRT8	6.72E-07	4.47
TNNT2	6.72E-07	3.57
MYOM1	1.36E-06	5.42
ACTC1	1.36E-06	4.75
MYOZ2	1.51E-06	6.72
SMYD1	1.60E-06	5.42
IFI44L	1.78E-06	-5.03
PKP2	1.81E-06	3.67
C6orf142	2.35E-06	4.90
NPNT	3.38E-06	3.84
MYBPC3	4.04E-06	4.33
GATA6	4.04E-06	3.34
ACTN2	4.57E-06	5.18
TRIM55	3.50E+00	0.02

Figure 5.2. Segregation of transcriptional profiles using comparison of neural and cardiac differentiation. Patterns of gene expression are shown for the extracted pattern groups, (A) common to neural and cardiac differentiation or (B) specific to CSs. Adjacent to each heatmap are the top-ranked genes based on the ANOVA score for each specific pattern; genes highlighted in blue are associated with ESCs or self-renewal, and genes in red with cardiac-specification. Gene Ontology (GO) terms and pathways enriched in the (C) common or (D) CS pattern groups are

displayed as compared to the number of associated gene changes in the alternate pattern group. Asterisks indicate significant GO-Elite scores (permute p<0.05) in the alternate pattern group.

5.3.3 Alternative splicing significantly contributes to transcript variation in hESC-derived cardiac cells

A custom application, called AltAnalyze, was created to identify alternative exons for day 40 CSs compared to Rex+ hESCs and link to these results to predicted functional outcomes (see details in methods and supplemental data). For exon array analysis, AltAnalyze uses the previously described splicing index approach to calculate a gene expression corrected probe set fold change and t-test. For this analysis, AltAnalyze was parameterized to include only probe sets with a relative fold change > 2, t-test p < 0.05, and to exclude probe sets with a MiDAS p > 0.05 and constitutive fold change > 3. Only regulated probe sets linked to exons or introns previously observed in mRNAs (Ensembl or UCSC) were used for further analyses. Of the 13,576 genes with evidence of expression, 15.1% (2,045) were predicted to have at least one alternative exon or intron regulated in the day 40 CSs (Table 5.1 A). Of these alternatively regulated genes, 58.6% (1,198) were connected to splicing events, intron retention, or alternative promoters (supported by mRNA evidence); whereas the remainder were probe sets linking to constitutive regions of the mRNAs. The majority of these exon-level changes (57.8%) can be attributed to AS (cassette-exon inclusion or exclusion or alternative 5’ or 3’ splice site selection), 26.9% to alternative promoter use (alternative N-terminal exon), 18.8% to intron retention, and the remainder to other splicing events classified by either UCSC (e.g., exon bleeding) or our algorithm (alternative C-terminally spliced exons). We therefore estimate that ~18% of all genes regulated in the differentiation of hESCs to CSs can be attributed to AS.

As a next step in AltAnalyze, probe sets aligned to mRNAs and proteins were analyzed for the gain or absence of known sequence elements (protein domains, modified residues, and miRNA binding sites). This method is conceptually similar to several described approaches (Xu *et al.* 2002; Taneri *et al.* 2004), but can easily be applied to any exon-level dataset with AltAnalyze. Since most human AS events produce large variations in mRNA and protein sequences or absence of translation, such analysis has the potential to provide greater insight into the functional consequences of altered exon expression. For all alternatively regulated probe sets, 93.9% aligned to at least one mRNA and specifically did not align to at least one other mRNA for that gene. For protein sequences aligning to these mRNAs, 83.1% of these comparisons yielded modification or an absence of one or more predicted functional elements (e.g., protein domains). In our analysis of all alternatively regulated probe sets, the typical exon-inclusion event produced a 530-residue increase or decrease in overall predicted protein sequence length (including predictions that would severely truncate protein sequence). Considering probe sets that only aligned to annotated protein sequences (as opposed to predicted based the mRNA sequences) produced a similar result (514 residues). Dozens of protein domains and functional residues were enriched among alternatively regulated genes, using an over-representation z-score, chief among them were spectrin and plectin repeats, asymmetric dimethylarginine, phosphoserine- and phosphothreonine-modified residues, spectrin-actin, DNA binding and START lipid binding domains, and SH, PH, CH, RRM, FERM, laminin, collagen, kinesin, RhoGEF, and protein kinase domains. When compared to the analysis of Cythera hESC neural precursor differentiation, several of the same enriched functional protein sequences were shared with the CS comparison, including spectrin, SH, PH, RhoGEF, and protein kinase domains (supplemental datasets).

In addition to these protein level changes, 13.1% of alternatively regulated genes (272 of 2,045) resulted in the inclusion or exclusion of at least one predicted miRNA binding site (supplemental datasets). Among genes with regulation of these binding sites, one-third occurred in an exon with evidence of AS or an alternative C-terminus, whereas the remainder occurred within a constitutive exon. To determine whether miRNA binding site inclusion occurred preferentially in hESCs as opposed to CSs, we compared the percentage of genes containing up- or down-regulated exons with these predicted binding sites. Interestingly, ~75% of genes with alternative inclusion of these binding sites were down-regulated in hESCs or up-regulated in CSs, compared to 62% of genes with probe sets down-regulated relative to the gene’s constitutive expression levels. This data suggests that miRNA binding site inclusion is substantially decreased in self-renewing hESCs.

A

	gene count	*out of*
Differentially Expressed Genes	**3,030**	30,473
Alternative Exons	**2,045**	13,576
▪ mRNA evidence	1,198	
– *alternative splicing*	*693*	
– *alternative promoter*	*322*	
– *intron retention*	*225*	

Alternative Splicing Events	*gene count*
– *alternative 5' splice sites*	*116*
– *alternative 3' splice sites*	*115*
– *alternative cassette exons*	*575*

B

	gene count	*out of*
Differentially Expressed miRNAs	**26**	210
Alternatively Regulated miRNA Binding Sites	**272**	11,079
▪ Evidence of AS	90	1,009
– *up-regulated in hESCs*	*18*	*313*
– *down-regulated in hESCs*	*73*	*729*
▪ No Evidence of AS	188	1,400
– *up-regulated in hESCs*	*57*	*666*
– *down-regulated in hESCs*	*135*	*860*

Table 5.1. Alternative Gene Regulation with hESC to CS Differentiation. (A) Transcriptional regulated genes (mean of constitutive gene features) and genes linked to alternative regulated features highlighted by AltAnalyze analysis. Gene expression values were calculated for 30,473 Ensembl gene identifiers, of which only 13,576 contained features expressed in both undifferentiated H9 ESCs and derived CSs. Alternative splicing, intron retention, and alternative promoter predicted events are shown for unique genes linked to the respective annotations. (B) Transcriptional regulated miRNAs and unique genes associated with alternatively regulated probe sets that contain predicted miRNA binding sites were calculated in AltAnalyze. Of the 13,576 expressed genes, 11,079 had features containing predicted

miRNA binding sites. The pattern of probe set regulation is indicated for hESC relative to CSs.

5.3.4 Confirmation and novel predictions for splice variants with previously established functional differences

Several of the identified splicing events in this experiment have been previously verified during hESC differentiation. These included SLK, SORBS1 (Yeo *et al.* 2007) and NFYA (Grskovic *et al.* 2007), all observed in differentiation to NPs. Splice events observed in the specification to cardiac/muscle lineage were also observed in our dataset (ATP2A2 (Misquitta *et al.* 2002; Periasamy *et al.* 2007), NF1 (Gutman *et al.* 1993), PKM2 (Imamura *et al.* 1986) and ANXA7 (Magendzo *et al.* 1991)) all with the predicted pattern of expression. Interestingly, AS of exons for CALD1, VCL, and ACTN1 in the CS comparison were also observed and verified in a large-scale colon cancer analysis using the same microarray platform, where the pattern of exon inclusion in the Rex+ hESCs is mimicked in proliferating tumor cells as opposed to normal colon. In our analysis, SLK and ANXA7 had two of the largest splicing index scores, for exon inclusion in hESCs and inclusion in CSs respectively. Analysis of six of these splice variants (ANXA7, ATP2A2, NF1, PKM2, SLK, and VCL) by RT-PCR verified clear shifts in isoform expression for each (Figure 5.3 A, B).

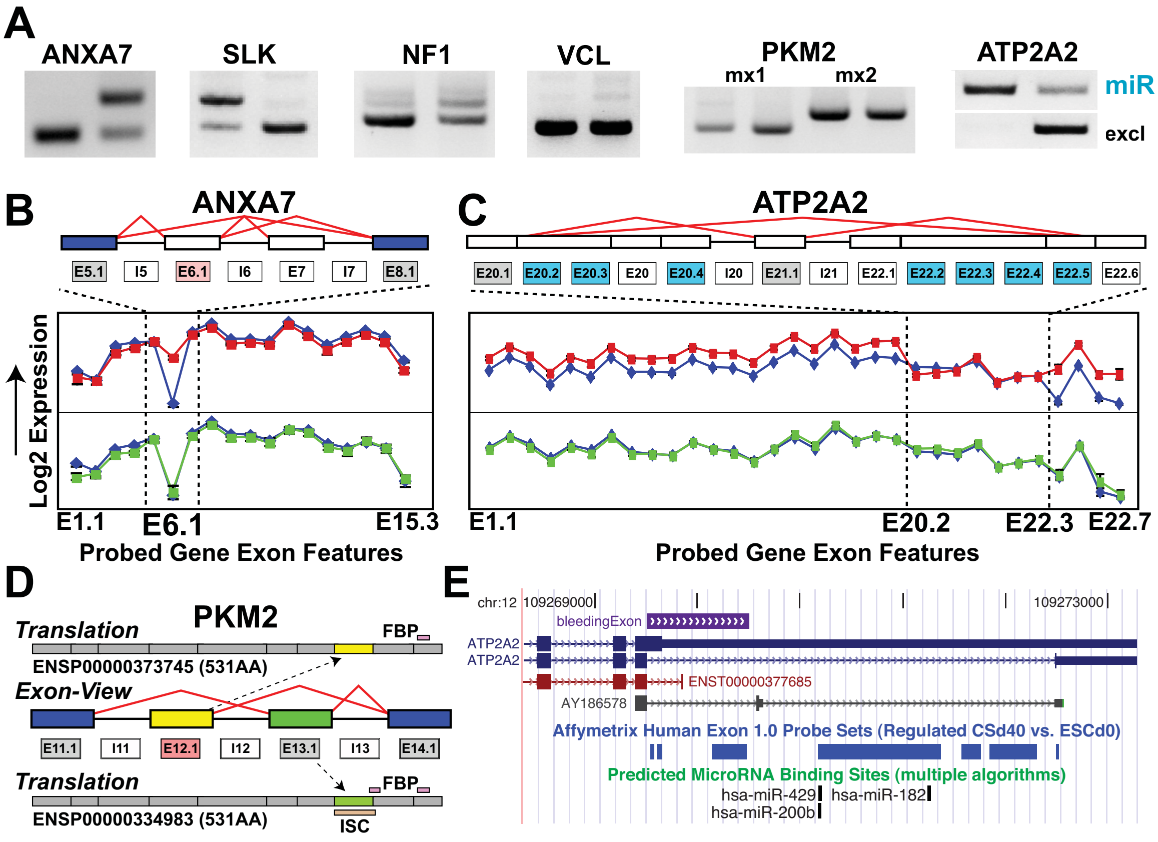


Figure 5.3. Analysis of verified splicing changes identifies novel functional associations. (A) Expression of splice isoforms validated by RT-PCR analysis of genes with prior evidence of AS, identified by AltAnalyze. These include ANXA7, SLK, NF1, and VCL validated using a flanking primers, and PKM2 and ATP2A2, validated using isoform-specific primers. DNA agarose gel images, with Rex+ hESCs RNA on the left side of the gel and CSs on the right. The notation miR indicates the presence of putative miRNA binding sites in the isoform, while excl indicates the exon-exclusion isoform. (B) Exon structure and expression profiles for two previously verified AS events in the genes ANXA7 and ATP2A2. Probe set exon-level expression data (log2) is displayed for both Rex+ hESCs and CSs (top graphs) and NP differentiation (bottom graph), both output from MS-Excel, with probe sets ranked in order of genomic position on the X-axis. Changes in probe set expression (relative to gene expression) are shown for probe sets aligning to exons and introns in the Cytoscape plugin

SubgeneViewer. Red boxes indicate up-regulation, blue down-regulation and gray not significant. (C) Protein functional regions aligning to only a single predicted variant are shown for the gene PKM2, predicted to encode for two 531 amino acid proteins with distinct, mutually exclusive exons. The two mutually exclusive isoforms produce proteins differing in the predicted inclusion of an FBP (fructose-1,6-bisphosphate) binding region and intersubunit contact (ISC) sequence as defined by UniProt. Yellow and green mutually exclusive exons are shown according their relative translated positions in resulting proteins. (D) AS of the ATP2A2 gene in the most distal 3' exon (inverse of intron retention), yields two isoforms with and without coding and UTR sequence, overlapping with miRNA binding sites predicted by at least two independent algorithms. Exons are displayed 5' to 3' (forward strand) along with aligning probe sets, down-regulated in this dataset (blue boxes).

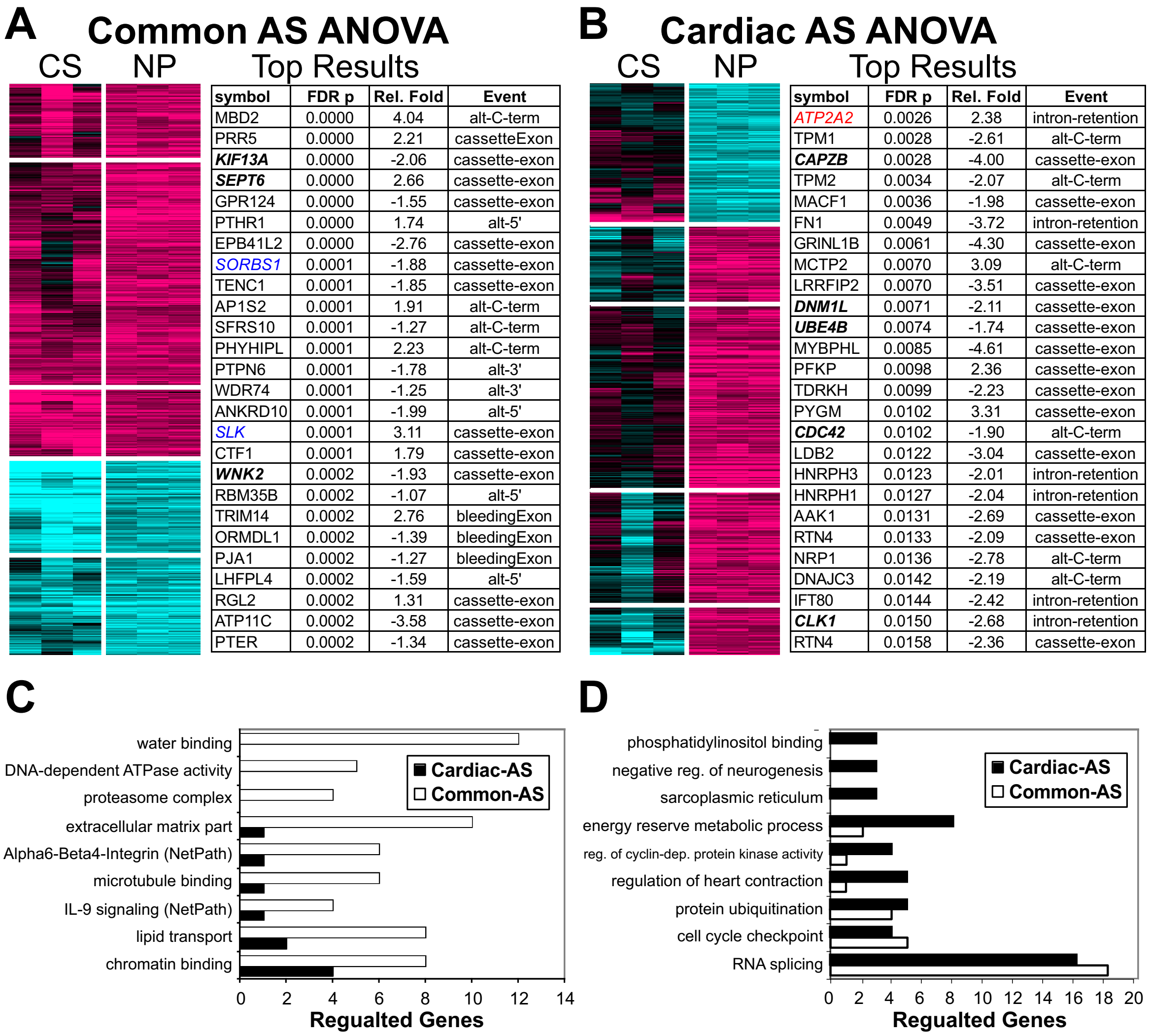
For at least three of the previously verified events, AS modifies the functional properties of the resulting proteins, producing differences in cell metabolism (PKM2), signaling (VCL), or mRNA stability (ATP2A2). The PKM2 or pyruvate kinase gene can encode two isoforms M1 and M2 through mutually-exclusive splicing of two 167 base pair (bp) exons (Imamura *et al.* 1986). Although the alternatively spliced exons are the same length and have 60% protein sequence identify to each other, they differ in their tissue developmental expression patterns, domain composition, and *in vivo* functions. In particular, the M1 isoform is largely present in normal adult heart, skeletal muscle, and brain and is not allosterically regulated by fructose-1,6-bisphosphate (FBP), whereas the M2 isoform is present only during embryonic development and in tumors, is regulated by FBP, and promotes proliferation (Dombrauckas *et al.* 2005; Lee *et* *al.* 2008). Isoform expression levels and protein level predictions from the present study confirm the existing data and in addition suggest negative alteration of the FBP binding region and intersubunit contact (as described by UniProt) with up-regulation in CSs of the M1 exon by our software (Figure 5.3 C). In the case of vinculin (VCL), expression of a 204bp exon in the C-terminal region produces a 68 amino acids (aa) insert that is enriched in muscle. Compared to the shorter variant, this longer isoform, metavinculin that is increased in CSs and appropriately predicted by AltAnalyze, has altered ligand binding properties (Witt *et al.* 2004), which correspond to the gain of a vinculin/alpha-catenin sequence (InterProt). For ATP2A2 (cardiac sarco/endoplasmic reticulum calcium ATPase), the alternative exclusion of mRNA sequence in the 3' terminal exon (4068bp) is predicted by AltAnalyze to preferentially remove a section of the cytoplasmic topological domain (45aa) and 3' UTR. Expression of the long C-terminal form of ATP2A2 (hESC-enriched) results in increased mRNA degradation of this transcript *in vitro* (Misquitta *et al.* 2002). In addition to the protein prediction, our tool also reported the loss of several predicted miRNA binding sites in this 3'UTR (hsa-miRNA-429, 200b and 182), each one supported by evidence from multiple miRNA binding site prediction algorithms (Figure 5.3 D). Interestingly, miRNA-microarray analysis has shown two of these miRNAs, miRNA-200b and miRNA-182 are highly enriched in cardiac cells derived from mouse ESCs (Ivey *et al.* 2008). These predictions therefore provide a new possible mechanism for increased degradation of the long 3'UTR form. Thus, for each of these splicing events, AltAnalyze sequence and domain/motif-level prediction complements the *in vitro* functional data and provides additional predictive insight into functional differences between isoforms.

5.3.5 Regulation of distinct pathways for cardiac and differentiation associated splicing events

To identify AS events in common to cardiac and neural differentiation or specific to cardiac differentiation, we applied our segregation ANOVA strategy to alternatively regulated probe sets. This analysis identified 565 alternatively regulated genes with a common splicing pattern during hESC differentiation to either CSs or neural precursors and 414 genes with a distinct pattern of alternative exon regulation in CSs (Figure 5.4 A, B). In both groups, we considered only probe sets for which there was previous evidence of AS.

Similar to previous results for NP differentiation, pathway analysis of all probe sets with evidence of AS showed that serine/threonine protein kinases (e.g., SLK, FER, FYN, MARK3, CDC42BPA, CLK1, WNK2) were highly enriched with the differentiation to CSs. When applied to genes with a common differentiation-splicing pattern, the most enriched ontology categories/pathways included water binding, RNA and chromatin binding, integrin-mediated signaling, microtubule binding, extracellular matrix and lipid transport (Figure 5.4 C). In contrast, alternatively spliced genes with a CS-specific pattern were enriched in pathways for phosphatidylinositol binding, sarcoplasmic reticulum, negative regulation of neurogenesis, regulation of heart contraction, ubiquination, Wnt receptor signaling, and regulation of cyclin-dependent protein kinase activity. Both sets were enriched in actin cytoskeletal, cell-matrix adhesion, RNA splicing, and cell cycle arrest genes (Figure 5.4 D). These results imply that the loss of pluripotency corresponds to AS of genes that regulate cell-cell contact formation and signaling, while cardiac-enriched events favor contractile pathways, inhibition of neurogenesis, and extracellular matrix signaling in addition to regulation of distinct metabolic pathways (full results provided as supplemental data). Likewise, over-representation analysis of protein domain-level annotations in these two pattern groups highlight distinct functional sequences present among alternatively spliced genes (supplemental Figure 5.1). Thus, both sets are largely distinct but complementary from those biological processes regulated at the level of gene expression in the analogous pattern groups.

When the same data are viewed in the context of adult tissue exon expression by clustering (supplemental data), we find that the common differentiation group largely had consistent splicing changes in all samples (relative to Rex+ hESCs), while the NP exon-level folds were largely in disagreement with the *in vivo* neural data (hCNS stem cells and cerebellum). Similar disagreements were also seen in comparison of CS regulated exons with adult heart for the cardiac-enriched group. These results suggest that the NPs derived by Yeo and colleagues (Yeo *et al.* 2007) and the CSs analyzed in this study have distinct features from their *in vivo* analogues. Thus, these precursor cells may be at a distinct developmental stage, but for our studies, suitable for examining early differences between early cardiac and neural differentiation.



symbol	FDR p	Rel. Fold	Event
MBD2	0.0000	4.04	alt-C-term
PRR5	0.0000	2.21	cassetteExon
KIF13A	0.0000	-2.06	cassette-exon
SEPT6	0.0000	2.66	cassette-exon
GPR124	0.0000	-1.55	cassette-exon
PTHR1	0.0000	1.74	alt-5'
EPB41L2	0.0000	-2.76	cassette-exon
SORBS1	0.0001	-1.88	cassette-exon
TENC1	0.0001	-1.85	cassette-exon
AP1S2	0.0001	1.91	alt-C-term
SFRS10	0.0001	-1.27	alt-C-term
PHYHIPL	0.0001	2.23	alt-C-term
PTPN6	0.0001	-1.78	alt-3'
WDR74	0.0001	-1.25	alt-3'
ANKRD10	0.0001	-1.99	alt-5'
SLK	0.0001	3.11	cassette-exon
CTF1	0.0001	1.79	cassette-exon
WNK2	0.0002	-1.93	cassette-exon
RBM35B	0.0002	-1.07	alt-5'
TRIM14	0.0002	2.76	bleedingExon
ORMDL1	0.0002	-1.39	bleedingExon
PJA1	0.0002	-1.27	bleedingExon
LHFPL4	0.0002	-1.59	alt-5'
RGL2	0.0002	1.31	cassette-exon
ATP11C	0.0002	-3.58	cassette-exon
PTER	0.0002	-1.34	cassette-exon

symbol	FDR p	Rel. Fold	Event
ATP2A2	0.0026	2.38	intron-retention
TPM1	0.0028	-2.61	alt-C-term
CAPZB	0.0028	-4.00	cassette-exon
TPM2	0.0034	-2.07	alt-C-term
MACF1	0.0036	-1.98	cassette-exon
FN1	0.0049	-3.72	intron-retention
GRINL1B	0.0061	-4.30	cassette-exon
MCTP2	0.0070	3.09	alt-C-term
LRRFIP2	0.0070	-3.51	cassette-exon
DNM1L	0.0071	-2.11	cassette-exon
UBE4B	0.0074	-1.74	cassette-exon
MYBPHL	0.0085	-4.61	cassette-exon
PFKP	0.0098	2.36	cassette-exon
TDRKH	0.0099	-2.23	cassette-exon
PYGM	0.0102	3.31	cassette-exon
CDC42	0.0102	-1.90	alt-C-term
LDB2	0.0122	-3.04	cassette-exon
HNRPH3	0.0123	-2.01	intron-retention
HNRPH1	0.0127	-2.04	intron-retention
AAK1	0.0131	-2.69	cassette-exon
RTN4	0.0133	-2.09	cassette-exon
NRP1	0.0136	-2.78	alt-C-term
DNAJC3	0.0142	-2.19	alt-C-term
IFT80	0.0144	-2.42	intron-retention
CLK1	0.0150	-2.68	intron-retention
RTN4	0.0158	-2.36	cassette-exon

Figure 5.4. AS genes associate with novel pathways for common or cardiac-enriched patterns. AS AltAnalyze predictions with evidence of either (A) a common neural/cardiac or (B) a CS-specific expression pattern, relative to undifferentiated hESCs. Adjacent to each heatmap are the top ANOVA scoring genes, similar to Figure 5.2. Gene names in blue have prior AS evidence with non-cardiac differentiation and genes in red have prior AS evidence with cardiac differentiation. Genes associated with GO terms and pathways are graphed that are over-represented in the AS (C) common or (D) CS-specific pattern group.

5.3.6 RT-PCR analysis of predicted AS show robust changes in isoform expression

In order to reliably focus in on a set of splicing predictions from this analysis, we conducted RT-PCR to examine shifts in the expression of alternate isoforms linked to regulated probe sets in the two pattern groups. As a pre-selection method, we largely restricted confirmation to events with the following criteria: (1) prior evidence of AS or intron retention OR presence of predicted miRNA binding sites, and (2) readily observable splicing patterns when data is viewed in the context of exon structures, and/or (3) predicted changes in protein domain structure or other functional sequence elements. To assess the second criterion, we viewed the raw log2 intensities of individual probe sets for Rex+ ESCs and CSc as a graph (Figure 5.3 A) and visualized AltAnalyze scores within a custom program called SubgeneViewer. SubgeneViewer is implemented as a plug-in for the network visualization software Cytoscape (Cline *et al.* 2007), that allows color criteria to be mapped onto exon and splicing structures. These results can be dynamically viewed from any gene/protein level interaction network/pathway (examples shown in Figure 5.3 C).

Using the described selection criteria, 53 predicted alternative events were selected for confirmation with both a differentiation and CS-specific expression pattern. Upon confirmation, a significant shift in isoform expression was observed for 37 of these target events, in line with our microarray analysis; an additional 10 events were verified to a much smaller extent (Figure 5.5 A). Only 6 of the 53 primer sets produced either inconclusive results or missing PCR products. Genes in the differentiation group that had large isoform expression changes were associated with a diverse range of biological categories, including serine/threonine kinases (SLK, FER, FYN, WNK2, MARK3), spectrin-actin binding (SPTBN1, ADD3), and cell-cell communication (TJP1). For CS-restricted splicing events, changes of similar magnitude were observed for proteins involved in calcium signaling (ASPH, ANXA7, ATP2A2) and cell metabolism (PKM2, OGDH); genes associated with ubiquitin protein degradation (UBE4B, NEDD4), double-stranded RNA binding (LRRFIP1, STAU1) and development (NUMB, TCF3, NAV2) were associated with both patterns. Exon-level array data are shown for two verified exons in the genes CAPZB, exon 12 and KIF13A, exon 41 (Figure 5.5 B), along with cross-tissue exon expression levels as compared to gene expression levels for the two respective exons. Overall, we conclude that the large majority of examined alternatively spliced and regulated exons/introns examined have consistent patterns with those predicted computationally with the AltAnalyze software.

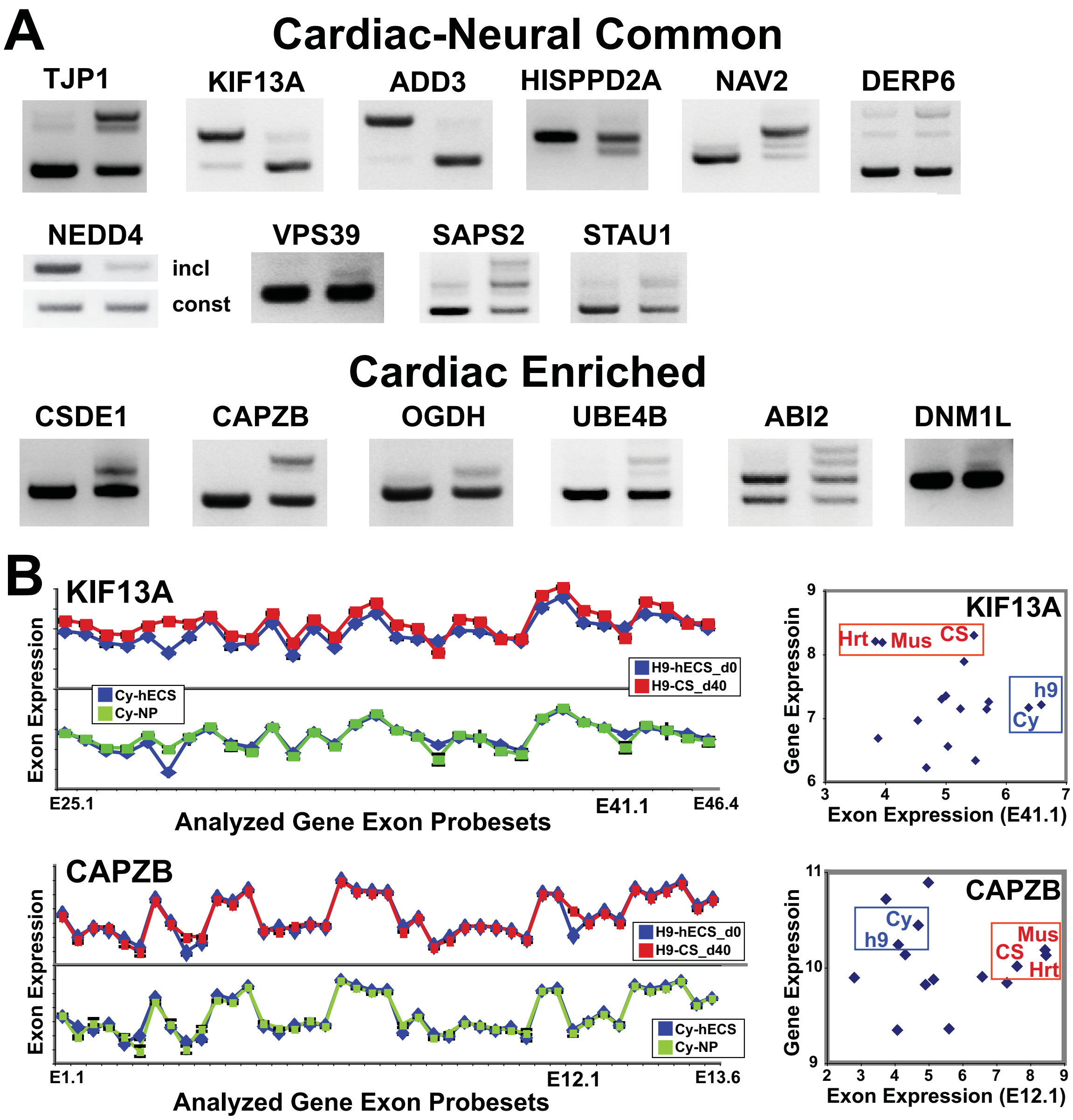


Figure 5.5. Validation of AS with distinct lineage commitment patterns. (A) Highlighted RT-PCR results for a panel of AltAnalyze predictions with multiple lines of evidence (overall exon expression patterns, specificity of splicing event, AltAnalyze score), with both a common and cardiac-enriched ANOVA pattern. Isoform-specific amplicons (incl) or constitutive probed exon-exon junction flanking amplicons (const) are indicated. (B) Log2 expression values for exon aligning probe sets for the genes KIF13A and CAPZB; probe sets ranked

in order of genomic position on the x-axis. For each gene, the mean gene expression value is plotted against the expression of the interrogated AS exon, for all tissues examined. This exon for CAPZB is E12 and E41 for KIF13A. h9 = Rex+ hESC, Cy = Cythera hESCs, Mus = muscle, Hrt = heart, and CS = cardiac spheroid.

5.3.7 Specificity of Domain-Level Protein Predictions

Examination of confirmed splicing events reveals distinct domain-level changes due to alternative exon expression. For such predictions, AltAnalyze links regulated probe sets to the longest transcripts containing the probe set sequence and missing that sequence (alternate isoform sequence). The majority of verified splicing events (35 of 47) had predicted changes that corresponded to altered protein sequence and annotated functional elements. However, several alternative exons and introns that did not align to known functional protein elements significantly altered protein sequence (NUMB, SAPS2, MADD, CSDE1, DERP6, TRAF6, and SEPT6).

To determine the validity of these domain-level changes, we performed a detailed analysis of the protein predictions using manual curation. This analysis verified predicted functional changes for 30 of the 35 splicing events; the remaining 5 events were specifically the result of secondary protein differences that were not directly associated with the splicing event (TJP1, OGDH, HISPPD2A, CDC42BPA, and NAV2). Of these 30 alternative exon and intron events, 17 were specifically associated with the splicing event (no additional regions of the protein affected); for the remaining 13, the regulation of an internal or C-terminal exon also segregated with a predicted change in N-terminal sequence (HIF3A, EWSR1, NEDD4, SPTBN1, NF1, CLK1, LRRFIP1, HDAC9, WNK2, VCL, CAPZB, ASPH, DNM1L)(UCSC genome browser). While five of these splicing events consistently were linked to an alternative promoter (EWSR1, SPTBN1, WNK2, VCL, LRRFIP1), the remainder did not. Although these alternative N-terminal predictions typically occurred because our algorithm chose the longest associated proteins, as opposed to those with the least overall differences, they none-the-less reflected possible outcomes.

5.3.8 Alteration of Kinase and DNA-binding Domains During Cardiac Differentiation

For validation, we preferentially selected genes with domain-level changes predicted to alter the function of the resulting proteins, chiefly kinase and DNA-binding domains. Such alterations could significantly alter the activity of transcriptional and signaling networks in the cell and contribute to altered cell physiology. Furthermore, both domain classes were highly over-represented among regulated domain predictions by AltAnalyze. Among the splicing events verified, six corresponded to genes with AS impacting kinase (FYN, FER, CLK1, WNK2) or kinase-like (SLK, MARK3) and four to predicted DNA-binding domains (TCF3, HIF3A, EWSR1, LRRFP1). These domain-level differences are the result of either the direct (HIF3A, FER, WNK2) or indirect (EWSR1, LRRFIP1, CLK1) introduction of a premature stop codon, the mutual exchange of exons with alternative domain sequences (TCF3, FYN), or cassette exon-exclusion (SLK, MARK3) (Figure 5.6 A, B). In each of these cases, the predictions by AltAnalyze were confirmed to accurately reflect the protein level changes that correspond to the regulated exons or introns when compared to reference protein sequences (UniProt/Ensembl).

To identify potential splice variants with the introduction of a premature stop codons, we searched for protein predictions that suggest either an absence of translation (e.g., EWSR1) or a markedly shorter protein product. In the instance of hypoxia-inducible factor-3α (HIF3A), selection of an alternative 5' splice site (CS-restricted pattern) is predicted by AltAnalyze to significantly alter protein domain composition in hESCs, including predicted loss/disruption of its DNA-binding, PAS and oxygen-dependent degradation domain (ODD) as well as its helix-loop-helix motif. This prediction matches the previously described domain composition differences of this variant (HIF3α6) (Maynard *et al.* 2003) compared to the full-length proteins. While the precise function of this variant is unclear, it has splicing and domain features similar to a mouse variant of this gene known inhibitory PAS domain protein (IPAS), which functions as a dominant-negative regulator HIF transcription factors induced under hypoxic conditions (Makino *et al.* 2002).

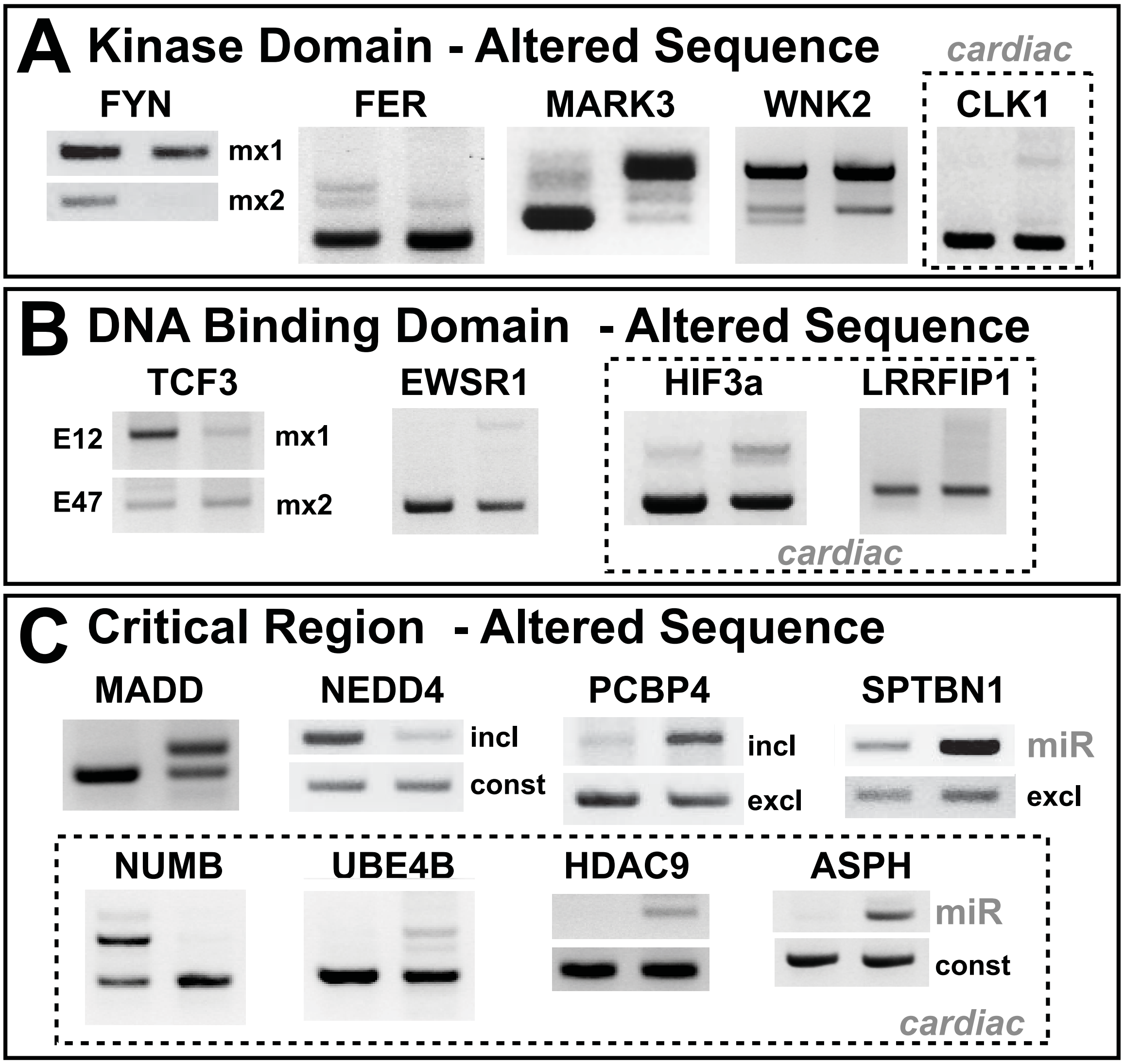


Figure 5.6. Altered composition of critical protein sequences by validated AS genes. RT-PCR of splice variants for genes with predicted disruption or alteration of (A) kinase or (B) DNA-binding domains or (C) other regions critical for protein function. Dashed boxes indicates genes with CS-restricted expression. All other genes have a common expression pattern.

Unlike the gain or loss of a critical protein domain (e.g., by protein truncation), assessing the precise functional impact on a domain with altered sequence is less clear. This was the case for the E2A immunoglobulin enhancer-binding factor TCF3 and for the serine/threonine and protein-tyrosine kinase FYN, in which a DNA-binding or kinase domain is specifically altered by the mutual-exclusive exchange of a cassette exon of similar lengths. For both of these splicing events, Ensembl annotates the respective InterProt domain as present in both isoforms, with 76% and 46% protein sequence identity between the mutually exclusive TCF3 and FYN exons (pairwise BLAST). Interestingly, both TCF3 and the FYN mutually exclusive isoforms have different biochemical properties (Davidson *et al.* 1994; Vitola *et al.* 1996), suggesting the domain level alterations predicted by AltAnalyze correlate with function. Our analysis of FYN detects the described T-lympohocyte isoform (FynT) as specifically expressed in undifferentiated hESCs while the brain form (FynB) appears to be expressed in both differentiated and undifferentiated cells (array data). Functional comparison of these isoforms showned that FynT has greater oncogenic transformation activity when activated by point mutations than the FynB isoform activated by the same mutations. Similarly, comparison of TCF3 isoforms has shown that the hESC-enriched isoform (E12) has less DNA-binding affinity than the differentiation-enriched form (E47). In addition to a basic helix-loop-helix motif with altered sequence, our study and previous studies revealed the absence of an inhibitory domain in the E12 form, which has been linked to this functional difference. Such forms of AS provide a potent means to modify specific residues within a sequence block without significantly changing overall protein length. These data provide further evidence that sequence changes and or deletions in critical protein domains can significantly alter the function of associated proteins in differentiated versus undifferentiated cells.

5.3.9 Altered Domain Expression Among Regulators of Proliferation and Cardiac Development

In addition to DNA-binding and kinase domains, AS was predicted to alter the composition of several critical domains for a functionally diverse set of proteins. These included large-scale changes, such as the removal and critical disruption of entire putative protein domains, missing or inserted sequences into such domains, or differential inclusion of small functionally significant protein residues.

Our AltAnalyze results predict significant changes to the domain composition of at least six genes with readily detected isoform expression changes (Figure 5.6 C). Two genes had known differences in isoform function (ASPH, SPTBN1), but the remaining AS predictions appear to be relatively novel. For common differentiation splice events, novel observations include the removal of the C2 calcium-dependent membrane targeting domain in the NEDD4 protein with exclusion of a 72aa block of exons in CSs; intron retention in the PCBP4 gene, which results in a shorter alternate N-terminus and disrupts a KH domain preferentially in hESCs and the exclusion of a 61aa encoding exon resulting in missing neutrophil cytosol factor domain and a proline-rich sequence in undifferentiated hESCs. For CS-restricted splicing events, novel findings include truncation of HDAC9 from a 1070 to 21aa protein specifically in CSs and the disruption of a phosphopantetheine attachment site in the UBE4B protein with the insertion of a 129AA encoding cassette exon. Since these domains play crucial roles in the annotated functions of these genes, the predicted loss or disruption of these sequences could considerably affect their function. An example is PCBP4, an RNA-binding protein characterized by presence of the KH domain. While the characterized form of this protein can suppress cell proliferation by inducing apoptosis, the isoform containing the complete KH domain sequence is expressed at lower levels than alternative form in hESCs, possibly hindering its apoptotic effects in undifferentiated hESCs. In cardiomyocytes, truncation or non-sense mediated decay (NMD) of the histone deacetylase HDAC9 should alleviate its repressive action on the expression of myocyte enhancer factor MEF2 transcription factors (Zhang *et al.* 2002). Therefore, de-repression of this protein through AS could present an important means for promoting cardiac development.

The genes aspartyl beta-hydroxylase (ASPH) and spectrin, beta, non-erythrocytic 1 (SPTBN1) both had similar changes and had prior evidence of functionally distinct splice variants, linked in this case to the regulation of cardiac physiology. ASPH encodes multiple splice variants, including a cardiac/skeletal-muscle specific form, junctin, which has a shorter N- and C-terminis and lacks enzymatic activity, a second short form with an alternate C-terminus (junctate); and the reference long form, which is ubiquitously expressed, has a distinct cellular localization and possesses hydroxylase enzymatic activity (Hong *et al.*

2007). Our splicing analysis identified both up-regulated junctate and junctin exons in CSs and down-regulated of the alternative N-terminal long-form of ASPH. RT-PCR verifies the up-regulation of junctin, which specifically complexes with cardiac contractile components (calsequenstrin, triadin, and the ryanodine receptor) (Fan *et al.* 2008) in the release of sarcoplasmic calcium, in CSs. In addition, we found there was no change in the expression of a region common to the different long forms of ASPH, which do not appear to participate in cardiac contractile control but rather regulate growth factor activity and are highly expressed in neoplastic cells (de la Monte *et al.* 2006). Domain-level analysis predicts the loss of several functional elements with up-regulation of junctin, including the loss of an N-terminal cytoplasmic and C-terminal luminal topological domain (UniProt).

Like ASPH, SPTBN1 was found to have both down-regulation of N-terminal and up-regulation of an alternate C-terminal exon, with confirmation from multiple probe sets per exon (supplemental dataset). Proteins for this gene can be found in the sarcomere along the muscle Z-line and likely contribute to structural stability (Hayes *et al.* 2000). We verified the up-regulation of a bleeding (overlapping with intron sequences in other transcripts and missing a 5' splice-site) C-terminal exon, linked to a shorter alternative N-terminus than the alternative isoform. Loss of the pleckstrin homology domain in the shorter bleeding exon variant correlates with the loss of inositol-1,4,5 triphosphate binding (Chen *et al.* 2001), in addition to the presence of an additional kinase phosphorylation site (Bignone *et al.* 2007), both of which were consistent with our predictions. The CS up-regulated short form is present in the sarcomere M-line and tetramers comprised of this short form are more stable than the long when associated with spectrin alpha 2 (Baines *et al.* 2005). For both ASPH and SPTBN1, the expression patterns of the isoforms fit with a model that would promote contractile signaling in CS and oppose it in undifferentiated hESCs.

In addition to domain-level changes predicted by our method, at least two other genes display functional isoform differences, both affecting pathways of proliferation and apoptosis that were not predicted (Figure 5.6 C). These were splicing of the Drosophila orthologue NUMB, which is involved in early cell-fate decisions (Yan *et al.* 2008) and the MAP-kinase activating death domain protein (MADD), which is a membrane-bound cytoplasmic adaptor protein that interacts with the TNF-α receptor 1 to transduce apoptotic signals (Mulherkar *et al.* 2007)(Figure 5.6 C). While the CS-enriched isoform of NUMB is anti-proliferative, the hESC-enriched form (p71), with a longer proline-rich region (PRR), retains its proliferative properties (Verdi *et al.* 1999; Toriya *et al.* 2006). Likewise, while the expression of the CS-enriched MADD isoform (IG20) can promote apoptosis, the hESC-enriched isoform (DENN) is anti-apoptotic and typically over-expressed in tumors. Since the protein databases used to derive the domain-level annotations for NUMB had a PRR with a shorter sequence, the alternatively spliced variants had no predicted difference. In the case of MADD, no predicted functional elements overlapped with the CS-enriched included exon. Nonetheless, both cases provide tantalizing evidence for splicing events that could regulate the proliferative or apoptotic properties of proteins in a developmentally specific manner.

5.3.10 Subtle Changes to the Composition of Essential Functional Elements

In addition to verified splicing events with domain prediction differences, several of the exons with the most apparent changes in isoform expression aligned to domains that were annotated in both isoforms (InterPro Ensembl associations). These include the removal of 32aa in the C-terminal aldehyde ferredoxin oxidoreductase domain of the ADD3 protein, insertion of 13aa into the dynamin GTPase region of DNM1L and a change in C-terminal sequence at the end of the F-actin capping protein, beta subunit region of the CAPZB protein and removal of 11aa from the N-terminal Citron homology domain (CNH) of VPS39. In each case, except VSP39, altering the sequence has unknown consequences on protein function. VPS39 is a putative adaptor protein that has decreased inclusion of a cassette exon in hESCs. The CNH domain in this protein is required for the clustering and fusion of late endosomes and lysosomes (Caplan *et al.* 2001). Interestingly, the isoform lacking this exon, called the TRAP-1 homologue, does not mediate lysosomal clustering but rather it specifically associates with the TGF-beta signaling pathway, suggesting modification of the CNH domain is sufficient to alter is properties. Other observed splicing events had more subtle function predictions, such as the microtubule-dependent motor protein KIF13A, in which removal of a 35aa coding exon results in the loss of one of three phosphoserine sites indicated by UniProt. If modulated directly by a protein kinase, however, such a change on its own could substantially alter the regulation of the resulting protein.

5.3.11 Developmental Regulation of MicroRNA Binding Site inclusion

A number of recent studies have demonstrated a critical connection between miRNA expression and the maintenance of pluripotency or the differentiation of cardiac cells from ESCs. In our gene expression analysis we observed up- and down-regulation of 26 miRNAs during differentiation to cardiac and neural lineages (supplemental dataset). Among these we find a number of previously implicated pluripotency (mir-302a, 302b) (Lakshmipathy *et al.* 2007) and cardiac (mir-133, 23b, 26a) (Ivey *et al.* 2008) regulated miRNAs, all appropriately segregated by the ANOVA pattern analysis strategy (Figure 5.7 A).

Although much effort has been devoted to defining the expression patterns and novel targets of miRNAs, little is known about the potential role of alternative splicing in miRNA binding site inclusion in processed mRNA transcripts. Traditional gene expression microarrays focus on the coding regions of transcripts and ignore the non-coding exons that can be alternatively spliced to produce different C-terminal exons or 3'UTRs of a gene. In contrast, exon-tiling arrays provide data on non-coding exons and provide a means to assess these distinct mRNA features in tandem with existing predictions for miRNA binding site position on a global basis.

Our analysis strategy highlighted 287 putative miRNA binding sites overlapping with exon-array probe sets that were alternatively expressed, including probe sets that did not align to alternatively spliced regions. We tested and validated 9 out of 10 of these alternatively included mRNA regions by RT-PCR, counting the SPTBN1 and ASPH variants described earlier. Putative miRNA binding sites were alternatively included as a result of the regulation of an alternative cassette (ASPH, SEPT6) or C-terminal exon (CDC42, C6orf134), bleeding-exon (SPTBN1), exon-region-exclusion (opposite of intron-retention) (ATP2A2), or 3'UTR with a longer or shorter sequence (LEFTY1, MAFB, CDK6) (Figure 5.7 B-D). At least one of these alternatively regulated genes (MAFB), with predicted regulation of a mir-130a binding site, is a known target of this miRNA (Garzon *et al.* 2006) (Figure 5.7 C). We were unable to find any other validated interactions from the literature. However, several of the predictions from our algorithm also contain overlapping predictions from multiple miRNA binding site algorithms (ATP2A2, C6orf134, CDC42, CDK6, LEFTY1, and MAFB), although some overlapping predictions were not originally found due to different miRNA names given by the different resources (e.g., MAFB).

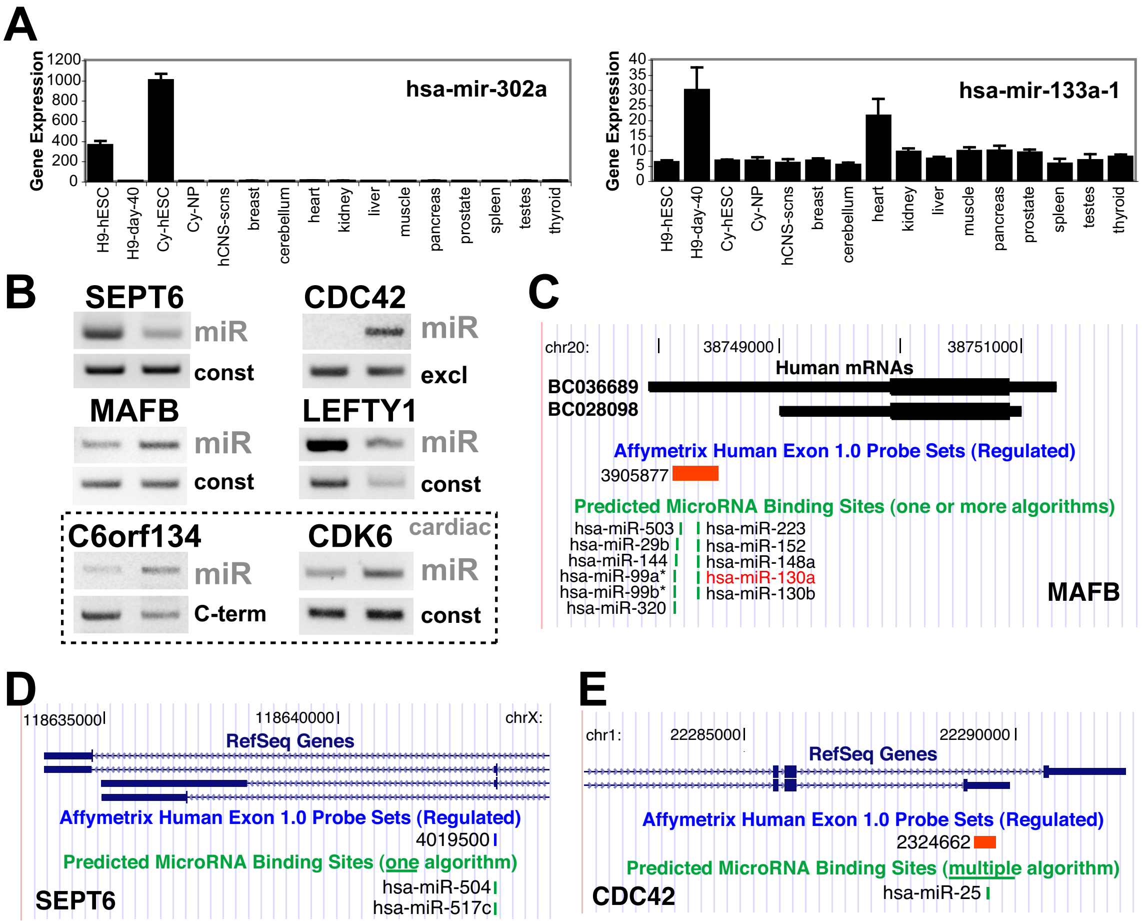


Figure 5.7. Regulation of miRNAs and miRNA binding sites within alterative exons. (A) The expression profile of two previously characterized microRNAs, mir-302a and mir-133-1, from combined tissue gene expression data. (B) RT-PCR isoform expression of genes with putative miRNA binding sites within the regulated probe set. The presence of one or more putative miRNAs is indicated by notation miR. (C-E) The 3' region of genes corresponding to three genes are shown, where the regulated isoforms are displayed from the UCSC genome browser along with regulated probe sets and putative microRNA binding site locations. UTR regions are indicated by thinner lines than coding regions. Each gene (MAFB, SEPT6, and CDC42) represents distinct possible modes of exon regulation leading to altered microRNA binding

site inclusion: shorter 3'UTR, alternate cassette-exon inclusion, and alternate C-terminal exon, respectively. Both MAFB and SETP6 are on the reverse genomic strand, where orientation is 3' to 5'.

Examination of miRNAs with previously established ESC or cardiac differentiation expression patterns highlighted the presence of mir-302a, 302c (ESC) and mir-26a (cardiac) binding sites, in the alternative bleeding exon of SPTBN1, in addition to the presence of mir-1 (cardiac) binding sites in CDK6. One of the most interesting cases is the presence of putative miRNA binding sites in the 3'UTR of the ATP2A2 gene, where this region promotes mRNA degradation. These data suggests a tantalizing new mechanism for miRNA regulation of such genes, largely dependent on AS. Since miRNAs can promote and inhibit the translation of targets dependent on cell-cycle stage (Vasudevan *et al.* 2008), there is the opportunity for complex modes of regulation by these predicted targets *in vivo*.

5.4 Conclusions

A necessary step in understanding the control of ESC pluripotency and lineage specification is to elucidate not only transcriptional events that occur with these transitions, but the secondary processing steps that can lead to fundamental changes in the amount and composition of proteins in these cells. We have implemented a new strategy for segregating whole-genome mRNA tiling data into distinct lineage expression pattern groups. Application of this method to gene expression changes results in the successful classification of known markers of both pluripotency and cardiac-specification. When applied to alternative splicing profiles for hESC differentiation in to cardiac and neural precursors, AltAnalyze highlights previously documented as well as novel predictions with large shifts in isoform expression that were readily confirmed by RT-PCR. Functional predictions from AltAnalyze, based on both protein sequence annotations and predicted miRNA binding sites for alternatively spliced genes, identifies clear functional changes along cardiac, differentiation, and pluripotency pathways for specific splice isoforms. These data provide new hypotheses that can readily be tested.

Among the most prominent predictions produced by AltAnalyze were the modification or disruption of DNA-binding and protein kinase domains, each enriched among all annotated protein regions regulated. Several of our predictions have been previously validated, suggesting this method is a useful prediction tool for identifying novel functional differences. Some of the most interesting genes tested were involved in apoptosis and proliferation pathways, enriched among AS events common to neural and cardiac differentiation. Isoforms encoded by the apoptosis genes PCBP4 and MADD both produce forms that do not activate apoptosis in undifferentiated hESCs. Conversely, the gene NUMB encodes an isoform in hESCs that activates proliferation and switches entirely to an isoform which inhibits proliferation in cardiomyocytes. These results suggest the intriguing possibility that splicing may act to coordinately alter the functional repertoire of distinct members of the same pathway to elicit a biological effect. We also observed AS for the apoptotic regulators CSDE1 and UBE4B along with previously demonstrated tumor suppressor genes ANXA7, EWSR1, and PKM2. Since both PKM2 and the proto-oncogene EWSR1 directly interact with the pluripotency transcription factor OCT3/4 to promote OCT3/4 activity (Lee *et al.* 2005) (Lee *et al.* 2008), specific isoforms of these genes may be critical in the regulation of ESC maintenance.

For AS events selectively enriched with differentiation to CSs, we observe the splicing of the cardiac contractile regulator ASPH, where the cardiac-enriched isoform-specifically functions to promote contraction. Likewise, AS of the cardiac inhibitor HDAC9 produced a highly truncated form specifically in CSs. This data further supports a role for AS in the direct specification of cardiac precursors.

Finally, exploration of the overlap between predicted miRNA binding sites and alternatively regulated probe sets has revealed a new potential mechanism by which specific cell types may regulate miRNA activity independently of miRNA expression. Such events were observed with both AS exons as well as the differential expression of distal terminal exons, where the mechanism regulating exon length is unclear. Two recent analyses have further demonstrated the interaction between miRNAs and alternatively spliced isoforms (Duursma *et al.* 2008) or UTRs of different length (Sandberg *et al.* 2008). Given that miRNA expression is thought to fine-tune protein expression, downstream of transcription, alternative exon inclusion may be a parallel means of regulating miRNA binding site selection, while still retaining full-length protein expression.

Future application and refinement of these analyses to additional cell lineages and time-points may yield greater resolution of AS events that will likely provide new insights into important mechanisms for cell fate commitment and maintenance of ESC pluripotency.

5.5 Materials and Methods

5.5.1 Isolation of hESCs and Cardiac Spheroids

The engineered ESC lines and cardiac electrophysiology are described in detail in an accompanying report (Kita-Matsuo, Barcova et al. submitted). In brief, H9 ESCs were infected by lentivirus with two drug selection cassettes, a neomycin resistance gene under the control of the REX-1 promoter (Rex-Neo^r) and a puromycin resistance under the control of the αMHC promoter (aMHC-Puro^r). Clonal stable lines were selected to allow for drug selection of a homogenous population of undifferentiated hESCs and CSs. Total RNA for biological triplicates of the Rex-Neo^r hESCs and Rex-Neo^r, aMHC-Puro^r day 40 CSs were extracted and prepared for hybridization to human 1.0 ST GeneChip arrays as previously described (Yeo *et al.* 2007). As starting material, ~1ug of total RNA was purified with the RiboMinus human Transcriptome Isolation it (Invitrogen), cDNA for hybridization generated using the GeneChip® WT cDNA Synthesis and WT Terminal Labeling kits (Affymetrix), by the Gladstone Institutes Genomics Core. The resulting fragmented and labeled cDNA was hybridized to individual GeneChip arrays and scanned as per the manufacturer’s instructions. Human exon array CEL files for the Cythera neuronal precursor differentiation datasets (Cy-ESCs and Cy-NPs), HUES6 cell line experiment (HUES6-ESCs and HUES6-NPs), and fetal human CNS stem cells (hCNS-SCs) were provided by the Gage laboratory (http://www.snl.salk.edu/~geneyeo/stuff/papers/supplementary/ES-NP) (Yeo *et al.* 2007). Eleven different adult human tissues, corresponding to 33 array CEL files, were obtained from the Affymetrix website (http://www.affymetrix.com/support/technical/sample_data/exon_array_data.affx).

5.5.2 Gene expression analysis

Probe set RMA (Irizarry *et al.* 2003) expression values and detection p-values were obtained for all probe sets using the Affymetrix program ExpressionConsole (http://www.affymetrix.com/products/software/specific/expression_console_software.affx). To calculate gene expression values from this exon-level data, we wrote a Python program (ExpressionBuilder - see following section for additional functions) that uses existing probe set to transcript associations from the Affymetrix probe set annotation file (HuEx-1_0-st-v2.na23.hg18.probe set.csv) and predicted splicing event information to identify exons most common (constitutive) to all transcripts. Predicted splicing event information was obtained directly from the knownAlt table provided by the UCSC genome bioinformatics website (http://www.genome.ucsc.edu) or predicted by comparing the structure of Ensembl transcripts (BioMART) (Spudich *et al.* 2007) and GenBank mRNA transcripts (UCSC) for all Ensembl annotated genes. Constitutive probe sets for each gene that were not associated with alternatively regulated transcripts are used by this program to calculate constitutive gene expression values using the mean of the probe set log2 intensity values. If no constitutive probe sets are present, gene expression is calculated by using the mean of all exon-associated gene linked probe set intensities.

5.5.3 Alternative exon analysis with AltAnalyze

To perform our alternative exon analyses, we created a custom application called AltAnalyze, composed of multiple modules written in Python. A detailed description of this software is available as supplemental materials. Briefly, AltAnalyze consists of three main modules: (1) LinkEST, (2) ExpressionBuilder, and (3) AltAnalyze. The program LinkEST builds relationship files used to connect probe sets to aligning proteins and microRNA binding sites (see following section), used by the AltAnalyze program. The LinkEST program is run only once with each new release of Ensembl. The ExpressionBuilder program is used to both assemble splicing annotations for exon-level microarrays and then process array expression datasets prior to splicing analyses. Like LinkEST, the annotation process (see previous section) is run only once with each build of Ensembl. The program AltAnalyze, imports ExpressionBuilder processed array datasets, calculates splicing scores, links this data to protein and microRNA binding site annotations and performs an over-representation analysis on both protein functional elements or miRNA binding sites, gained or lost among unique genes.

The ExpressionBuilder program, in addition to calculating gene expression values, organizes exon-level data for each experimental sample into biological groups, and then filters these probe sets based on detection above background

(DABG) p-values (obtained from a separate file generated by ExpressionConsole). The relationships between samples and groups and which groups in the expression dataset should be compared, can be indicated by the user, within two input text files. For non-constitutive probe sets, if either biological group has a mean DABG p <0.05 that probe set is retained. For constitutive probe sets, however, if either biological group has a mean DABG p >0.05, then the probe set is filtered out of the resulting expression file. This expression file is stored as input for the AltAnalyze module.

In the AltAnalyze program, the likelihood and extent of alternative splicing are calculated using the splicing index method (Srinivasan *et al.* 2005) (Gardina *et al.* 2006) for all Ensembl genes with one or more constitutive probe sets. Probe sets considered for this analysis consist of the Affymetrix 'core' set, probe sets associated with an alternatively regulated exon (associated with an alternative promoter, N-terminus, C-terminus, cassette exon, alternative 3' or 5' splice site, retained intron or exon-region exclusion region), and probe sets aligning to an analyzed mRNAs (data from ExpressionBuilder). The splicing index is a constitutive corrected exon-level fold change. Constitutive gene expression values are calculated as described in the previous section, for probe sets contained within the ExpressionBuilder filtered dataset files. Two probability estimates for alternative exon regulation are calculated based on a 1-way analysis of variance model, MiDAS (Gardina *et al.* 2006), using the Affymetrix Power Tools software (version 1.4.0) (http://www.affymetrix.com/support/developer/powertools/index.affx) software, and t-test of constitutive adjusted exon expression values for both comparison groups (AltAnalyze). These constitutive adjusted exon expression values are then used to calculate a splicing index fold value for each probe set (Gardina *et al.* 2006). Exon or intron representing probe sets with a MiDAS and adjusted expression t-test p<0.05 and splicing index value > 1 were reported as alternatively regulated. Splicing events confirmed from the literature were obtained by manually comparing the sequence of the alternatively spliced exons from the array and the previous report or matching up Affymetrix probe set IDs and verifying that the directionality of the splicing event is in common. Such events were initially identified by both manual and automated literature searches (LitSearch- http://www.agilent.com/labs/research/litsearch.html).

5.5.4 Identifying functional sequence elements with AltAnalyze

For all alternative exon and intron linked probe sets, multiple types of functional consequences were assessed using the AltAnalyze application (supplemental methods). In short, AltAnalyze uses two tables built using custom modules (LinkEST), relating microarray probe sets to matching and non-matching protein sequences or predicted miRNA binding sites. These scripts identified perfect or partial probe set consensus sequence matches to mRNAs and expressed sequence tags (ESTs) from Unigene and Ensembl. For the longest, high-quality matches and non-matches, corresponding protein sequences were identified or derived via *in silico* translation and stored for further analysis. Likewise, putative miRNA binding sites (PicTar (Krek *et al.* 2005), miRanda (http://www.microrna.org), miRbase (Griffiths-Jones *et al.* 2008) and TargetScan (http://www.targetscan.org)) contained within probe set consensus sequences were stored in a second table for import by AltAnalyze. In AltAnalyze, the predicted consequences of splicing on protein domains and functional regions was assessed by identifying the gain or loss of annotated protein regions and domains in the UniProt (http://www.pir.uniprot.org/) and Ensembl protein databases, by comparing the 'best' matching and non-matching proteins linked to an individual probe set. Over-representation of functional regions, domains, and miRNA binding sites was further evaluated with an over-representation z-score. Open-source code for AltAnalyze is provided under the Apache open source license along with an executable version compatible with multiple operating systems (see: http://www.genmapp.org/AltAnalyze).

5.5.5 Validation of Alternative Exon Expression

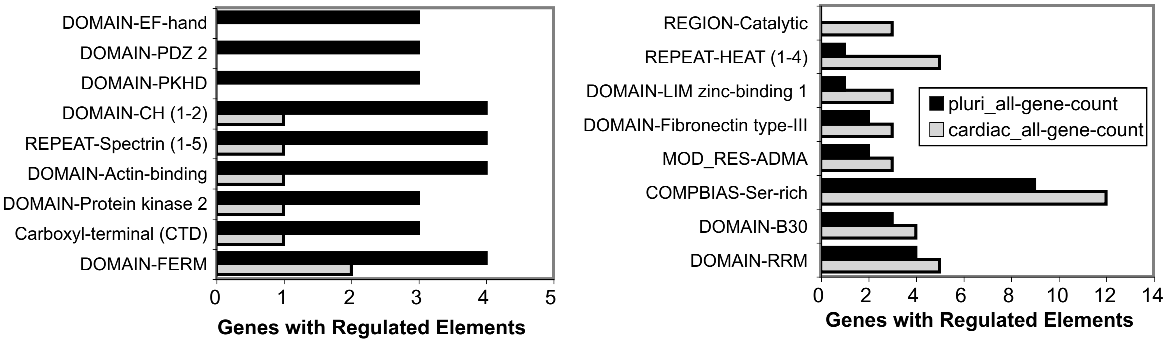
Alternatively regulated genes were selected for validation after bioinformatics filtering of genes and probe sets using AltAnalyze, SubgeneViewer, and visualization of probe set genomic location in the UCSC genome browser. Probe sets were largely selected for validation based on the prediction that the associated exon or intron occurred in a previously annotated full-length mRNA that was predicted to undergo alternative regulation, and also contained functionally informative sequence level changes. Based on these filters, we selected 52 alternative exon/intron sequences for validation. Exons for RT-PCR were manually selected by examination of exon structure at the UCSC and

Ensembl genome browsers and optimal flanking, isoform-specific, or constitutive primers designed using a custom implementation of primer 3 called AltPrimer (http://conklinwolf.ucsf.edu/tools/picoprimer.html). For RT-PCR, total RNA was diluted to ~10ng/μl and RT-PCR was amplified with the OneStep Superscript III RT-PCR kit (Invitrogen) for 28, 35, or 40 cycles with annealing temperatures of 55 or 58°C using isoform-specific or constitutive flanking primers and analyzed on a 2-2.5% DNA-agarose gel using ethidium bromide staining.

5.6 Supplemental Data

Supplemental datasets and figures are available at:

http://www.genmapp.org/supplemental/Salomonis_2008/hESC_exon/



Supplemental Figure 5.1. Distinct functional protein sequences regulated by alternative splicing for cardiac and neural or cardiac expression patterns. Unique genes that have a predicted gain or loss of a particular annotated protein segment (e.g., domain or site) are shown for protein elements enriched with a common differentiation pattern (left) or cardiac pattern (right) as determined by an over-representation z-score.

5.7 References

Baines, A. J. and J. C. Pinder (2005). "The spectrin-associated cytoskeleton in mammalian heart." Front Biosci **10**: 3020-33.

Bignone, P. A., M. D. King, J. C. Pinder and A. J. Baines (2007). "Phosphorylation of a threonine unique to the short C-terminal isoform of betaII-spectrin links regulation of alpha-beta spectrin interaction to neuritogenesis." J Biol Chem **282**(2): 888-96.

Boyer, L. A., T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, D. K. Gifford, D. A. Melton, R. Jaenisch and R. A. Young (2005). "Core transcriptional regulatory circuitry in human embryonic stem cells." Cell **122**(6): 947-56.

Burgess, R. W., Q. T. Nguyen, Y. J. Son, J. W. Lichtman and J. R. Sanes (1999). "Alternatively spliced isoforms of nerve- and muscle-derived agrin: their roles at the neuromuscular junction." Neuron **23**(1): 33-44.

Caplan, S., L. M. Hartnell, R. C. Aguilar, N. Naslavsky and J. S. Bonifacino (2001). "Human Vam6p promotes lysosome clustering and fusion in vivo." J Cell Biol **154**(1): 109-22.

Chen, Y., P. Yu, D. Lu, D. A. Tagle and T. Cai (2001). "A novel isoform of beta-spectrin II localizes to cerebellar Purkinje-cell bodies and interacts with neurofibromatosis type 2 gene product schwannomin." J Mol Neurosci **17**(1): 59-70.

Cline, M. S., M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. R. Pico, A. Vailaya, P. L. Wang, A. Adler, B. R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker and G. D. Bader (2007). "Integration of biological networks and gene expression data using Cytoscape." Nat Protoc **2**(10): 2366-82.

Cooper, T. A. (2005). "Alternative splicing regulation impacts heart development." Cell **120**(1): 1-2.

Davidson, D., M. Fournel and A. Veillette (1994). "Oncogenic activation of p59fyn tyrosine protein kinase by mutation of its carboxyl-terminal site of tyrosine phosphorylation, tyrosine 528." J Biol Chem **269**(14): 10956-63.

de la Monte, S. M., S. Tamaki, M. C. Cantarini, N. Ince, M. Wiedmann, J. J. Carter, S. A. Lahousse, S. Califano, T. Maeda, T. Ueno, A. D'Errico, F. Trevisani and J. R. Wands (2006). "Aspartyl-(asparaginyl)-beta-hydroxylase regulates hepatocellular carcinoma invasiveness." J Hepatol **44**(5): 971-83.

Dombrauckas, J. D., B. D. Santarsiero and A. D. Mesecar (2005). "Structural basis for tumor pyruvate kinase M2 allosteric regulation and catalysis." Biochemistry **44**(27): 9417-29.

Dudoit, S., R. C. Gentleman and J. Quackenbush (2003). "Open source software for the analysis of microarray data." Biotechniques **Suppl**: 45-51.

Duursma, A. M., M. Kedde, M. Schrier, C. le Sage and R. Agami (2008). "miR-148 targets human DNMT3b protein coding region." Rna **14**(5): 872-7.

Eisen, M. B., P. T. Spellman, P. O. Brown and D. Botstein (1998). "Cluster analysis and display of genome-wide expression patterns." Proc Natl Acad Sci U S A **95**(25): 14863-8.

Fan, G. C., Q. Yuan and E. G. Kranias (2008). "Regulatory roles of junctin in sarcoplasmic reticulum calcium cycling and myocardial function." Trends Cardiovasc Med **18**(1): 1-5.

Gardina, P. J., T. A. Clark, B. Shimada, M. K. Staples, Q. Yang, J. Veitch, A. Schweitzer, T. Awad, C. Sugnet, S. Dee, C. Davies, A. Williams and Y. Turpaz (2006). "Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array." BMC Genomics **7**: 325.

Garzon, R., F. Pichiorri, T. Palumbo, R. Iuliano, A. Cimmino, R. Aqeilan, S. Volinia, D. Bhatt, H. Alder, G. Marcucci, G. A. Calin, C. G. Liu, C. D. Bloomfield, M. Andreeff and C. M. Croce (2006). "MicroRNA fingerprints during human megakaryocytopoiesis." Proc Natl Acad Sci U S A **103**(13): 5078-83.

Griffiths-Jones, S., H. K. Saini, S. van Dongen and A. J. Enright (2008). "miRBase: tools for microRNA genomics." Nucleic Acids Res **36**(Database issue): D154-8.

Grskovic, M., C. Chaivorapol, A. Gaspar-Maia, H. Li and M. Ramalho-Santos (2007). "Systematic identification of cis-regulatory sequences active in mouse and human embryonic stem cells." PLoS Genet **3**(8): e145.

Gutman, D. H., L. B. Andersen, J. L. Cole, M. Swaroop and F. S. Collins (1993). "An alternatively-spliced mRNA in the carboxy terminus of the neurofibromatosis type 1 (NF1) gene is expressed in muscle." Hum Mol Genet **2**(7): 989-92.

Hammes, A., J. K. Guo, G. Lutsch, J. R. Leheste, D. Landrock, U. Ziegler, M. C. Gubler and A. Schedl (2001). "Two splice variants of the Wilms' tumor 1 gene have distinct functions during sex determination and nephron formation." Cell **106**(3): 319-29.

Hayes, N. V., C. Scott, E. Heerkens, V. Ohanian, A. M. Maggs, J. C. Pinder, E. Kordeli and A. J. Baines (2000). "Identification of a novel C-terminal variant of beta II spectrin: two isoforms of beta II spectrin have distinct intracellular locations and activities." J Cell Sci **113 (Pt 11)**: 2023-34.

Hong, C. S., S. J. Kwon and H. Kim do (2007). "Multiple functions of junctin and junctate, two distinct isoforms of aspartyl beta-hydroxylase." Biochem Biophys Res Commun **362**(1): 1-4.

Imamura, K., T. Noguchi and T. Tanaka (1986). "Regulation of isozyme patterns of pyruvate kinase in normal and neoplastic tissues." Markers of Human Neuroectodermal Tumors: 191-222.

Irizarry, R. A., B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs and T. P. Speed (2003). "Summaries of Affymetrix GeneChip probe level data." Nucleic Acids Res **31**(4): e15.

Ivanova, N., R. Dobrin, R. Lu, I. Kotenko, J. Levorse, C. DeCoste, X. Schafer, Y. Lun and I. R. Lemischka (2006). "Dissecting self-renewal in stem cells with RNA interference." Nature **442**(7102): 533-8.

Ivey, K. N., A. Muth, J. Arnold, F. W. King, R. F. Yeh, J. E. Fish, E. C. Hsiao, R. J. Schwartz, B. R. Conklin, H. S. Bernstein and D. Srivastava (2008). "MicroRNA regulation of cell lineages in mouse and human embryonic stem cells." Cell Stem Cell **2**(3): 219-29.

Krek, A., D. Grun, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel and N. Rajewsky (2005). "Combinatorial microRNA target predictions." Nat Genet **37**(5): 495-500.

Lakshmipathy, U., B. Love, L. A. Goff, R. Jornsten, R. Graichen, R. P. Hart and J. D. Chesnut (2007). "MicroRNA expression pattern of undifferentiated and differentiated human embryonic stem cells." Stem Cells Dev **16**(6): 1003-16.

Lee, C. and Q. Wang (2005). "Bioinformatics analysis of alternative splicing." Brief Bioinform **6**(1): 23-33.

Lee, J., H. K. Kim, Y. M. Han and J. Kim (2008). "Pyruvate kinase isozyme type M2 (PKM2) interacts and cooperates with Oct-4 in regulating transcription." Int J Biochem Cell Biol **40**(5): 1043-54.

Lee, J., B. K. Rhee, G. Y. Bae, Y. M. Han and J. Kim (2005). "Stimulation of Oct-4 activity by Ewing's sarcoma protein." Stem Cells **23**(6): 738-51.

Loh, Y. H., Q. Wu, J. L. Chew, V. B. Vega, W. Zhang, X. Chen, G. Bourque, J. George, B. Leong, J. Liu, K. Y. Wong, K. W. Sung, C. W. Lee, X. D. Zhao, K. P. Chiu, L. Lipovich, V. A. Kuznetsov, P. Robson, L. W. Stanton, C. L. Wei, Y. Ruan, B. Lim and H. H. Ng (2006). "The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells." Nat Genet **38**(4): 431-40.

Magendzo, K., A. Shirvan, C. Cultraro, M. Srivastava, H. B. Pollard and A. L. Burns (1991). "Alternative splicing of human synexin mRNA in brain, cardiac, and skeletal muscle alters the unique N-terminal domain." J Biol Chem **266**(5): 3228-32.

Makino, Y., A. Kanopka, W. J. Wilson, H. Tanaka and L. Poellinger (2002). "Inhibitory PAS domain protein (IPAS) is a hypoxia-inducible splicing variant of the hypoxia-inducible factor-3alpha locus." J Biol Chem **277**(36): 32405-8.

Maynard, M. A., H. Qi, J. Chung, E. H. Lee, Y. Kondo, S. Hara, R. C. Conaway, J. W. Conaway and M. Ohh (2003). "Multiple splice variants of the human HIF-3 alpha locus are targets of the von Hippel-Lindau E3 ubiquitin ligase complex." J Biol Chem **278**(13): 11032-40.

Misquitta, C. M., J. Mwanjewe, L. Nie and A. K. Grover (2002). "Sarcoplasmic reticulum Ca(2+) pump mRNA stability in cardiac and smooth muscle: role of the 3'-untranslated region." Am J Physiol Cell Physiol **283**(2): C560-8.

Mitschischek, E. (1991). "[Diagonal incision in capsulotomy for extracapsular cataract extraction]." Klin Monatsbl Augenheilkd **199**(6): 406-8.

Mulherkar, N., K. V. Prasad and B. S. Prabhakar (2007). "MADD/DENN splice variant of the IG20 gene is a negative regulator of caspase-8 activation. Knockdown enhances TRAIL-induced apoptosis of cancer cells." J Biol Chem **282**(16): 11715-21.

Periasamy, M. and A. Kalyanasundaram (2007). "SERCA pump isoforms: their role in calcium transport and disease." Muscle Nerve **35**(4): 430-42.

Salomonis, N., N. Cotte, A. C. Zambon, K. S. Pollard, K. Vranizan, S. W. Doniger, G. Dolganov and B. R. Conklin (2005). "Identifying genetic networks underlying myometrial transition to labor." Genome Biol **6**(2): R12.

Sandberg, R., J. R. Neilson, A. Sarma, P. A. Sharp and C. B. Burge (2008). "Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites." Science **320**(5883): 1643-7.

Spudich, G., X. M. Fernandez-Suarez and E. Birney (2007). "Genome browsing with Ensembl: a practical overview." Brief Funct Genomic Proteomic **6**(3): 202-19.

Srinivasan, K., L. Shiue, J. D. Hayes, R. Centers, S. Fitzwater, R. Loewen, L. R. Edmondson, J. Bryant, M. Smith, C. Rommelfanger, V. Welch, T. A. Clark, C. W. Sugnet, K. J. Howe, Y. Mandel-Gutfreund and M. Ares, Jr. (2005). "Detection and measurement of alternative splicing using splicing-sensitive microarrays." Methods **37**(4): 345-59.

Taneri, B., B. Snyder, A. Novoradovsky and T. Gaasterland (2004). "Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific." Genome Biol **5**(10): R75.

Toriya, M., A. Tokunaga, K. Sawamoto, K. Nakao and H. Okano (2006). "Distinct functions of human numb isoforms revealed by misexpression in the neural stem cell lineage in the Drosophila larval brain." Dev Neurosci **28**(1-2): 142-55.

Vasudevan, S., Y. Tong and J. A. Steitz (2008). "Cell-cycle control of microRNA-mediated translation regulation." Cell Cycle **7**(11).

Verdi, J. M., A. Bashirullah, D. E. Goldhawk, C. J. Kubu, M. Jamali, S. O. Meakin and H. D. Lipshitz (1999). "Distinct human NUMB isoforms regulate differentiation vs. proliferation in the neuronal lineage." Proc Natl Acad Sci U S A **96**(18): 10472-6.

Vitola, S. J., A. Wang and X. H. Sun (1996). "Substitution of basic amino acids in the basic region stabilizes DNA binding by E12 homodimers." Nucleic Acids Res **24**(10): 1921-7.

Witt, S., A. Zieseniss, U. Fock, B. M. Jockusch and S. Illenberger (2004). "Comparative biochemical analysis suggests that vinculin and metavinculin cooperate in muscular adhesion sites." J Biol Chem **279**(30): 31533-43.

Xu, Q., B. Modrek and C. Lee (2002). "Genome-wide detection of tissue-specific alternative splicing in the human transcriptome." Nucleic Acids Res **30**(17): 3754-66.

Xu, X., D. Yang, J. H. Ding, W. Wang, P. H. Chu, N. D. Dalton, H. Y. Wang, J. R. Bermingham, Jr., Z. Ye, F. Liu, M. G. Rosenfeld, J. L. Manley, J. Ross, Jr.,

J. Chen, R. P. Xiao, H. Cheng and X. D. Fu (2005). "ASF/SF2-regulated CaMKIIdelta alternative splicing temporally reprograms excitation-contraction coupling in cardiac muscle." Cell **120**(1): 59-72.

Yan, B., F. M. Omar, K. Das, W. H. Ng, C. Lim, K. Shiuan, C. T. Yap and M. Salto-Tellez (2008). "Characterization of Numb expression in astrocytomas." Neuropathology.

Yeo, G. W., X. Xu, T. Y. Liang, A. R. Muotri, C. T. Carson, N. G. Coufal and F. H. Gage (2007). "Alternative splicing events identified in human embryonic stem cells and neural progenitors." PLoS Comput Biol **3**(10): 1951-67.

Zhang, C. L., T. A. McKinsey, S. Chang, C. L. Antos, J. A. Hill and E. N. Olson (2002). "Class II histone deacetylases act as signal-responsive repressors of cardiac hypertrophy." Cell **110**(4): 479-88.

Chapter 6
Defining the functional repertoire of alternative transcripts in embryonic stem cell differentiation and myometrial gestation

Nathan Salomonis[1,2], Christopher R. Schlieve[1], Laura Pereira[3], Alexander C. Zambon[4], Karen Vranizan[5], Matthew J. Spindler[1,2], Alexander R. Pico[1], Melissa S. Cline[6], Alan Williams[6], John E. Blume[6], Bradley J. Merrill[3], Bruce R. Conklin[1,2]

[1]Gladstone Institute of Cardiovascular Disease, San Francisco, CA, [2]Pharmaceutical Sciences and Pharmacogenomics Graduate Program, University of California, San Francisco, CA, [3]Department of Biochemistry and Molecular Genetics, University of Illinois at Chicago, [4]Department of Pharmacology, University of California at San Diego, La Jolla, CA, [5]Functional Genomics Laboratory, University of California, Berkeley, CA, [6]Affymetrix, Inc., Santa Clara, CA

6.1 Abstract

Background: Two major goals for regenerative medicine are to reproducibly transform adult somatic cells to a pluripotent state and control their differentiation into specific cell-fates. These goals could be furthered by obtaining a complete picture of the RNA isoforms produced by these cells due to alternative splicing (AS) and alternative promoter selection (APS).

Results: To investigate the role of AS and APS, reciprocal exon-exon junctions were interrogated on a genome-wide scale in differentiating mouse embryonic stem cells (ESCs) with a prototype Affymetrix microarray. Using a custom analysis package name AltAnalyze, we identified 171 putative isoform variants for 143 genes, the majority of which were predicted to alter protein sequence and domain composition. Among the most robust verified isoform changes was a novel ESC enriched isoform of the pluripotency transcription factor Tcf3, encoding a protein with a gain of 14 amino acids. This longer form of Tcf3 was able to repress the transcription of Nanog and β-catenin reporters similar to the shorter reference isoform. Knockdown (KD) with short-hairpin RNAs (shRNAs) directed against Tcf3-short, long, or all isoforms had delayed differentiation upon removal of LIF. With differentiation to embryoid bodies (EBs), Tcf3-short and Tcf3-all KD blocked induction of early and late lineage markers, while Tcf3-long KD specifically blocked induction of late lineage markers. Although teratomas derived from wild-type and Tcf3-long KD ESCs produced all three primordial germ layers, Tcf3-short and Tcf3-all did not.

Conclusions: Analysis of exon-exon junction microarray data revealed AS of Tcf3 isoforms, which have distinct functions in the differentiation of pluripotent stem cells.

6.2 Introduction

ESCs are a vital tool for studying the events that regulate early embryonic propagation and cell-fate decisions. Research in this area has lead to the development of new technologies for adult somatic cell reprogramming and insights into the steps required for lineage commitment (Yamanaka 2008). Despite these significant advances, considerable challenges remain in elucidating the precise mechanisms that mediate these biological transitions.

Several factors critical for maintaining pluripotency have been identified using both conventional biochemical screens and whole-genome gene expression studies of ESCs. These include the transcription factors Oct4, Sox2, and Nanog, which interact with a common set of promoters to promote self-renewal and pluripotency (Boyer *et al.* 2005). Recently Tcf3, a β-catenin responsive transcription factor, was implicated in this core transcriptional network as a direct transcriptional repressor of both Oct4 and Nanog (Pereira *et al.* 2006; Tam *et al.* 2008) and is itself a target of these factors (Cole *et al.* 2008). While ESCs with little or no Tcf3 expression can be maintained in LIF-independent conditions for extended periods, these cells have delayed or hindered differentiation (Tam *et al.* 2008; Yi *et al.* 2008). Inhibition of other Wnt signaling components, including the protein phosphatase PPP3R2 (Miyabayashi *et al.* 2007) and GSK3β (Sato *et al.* 2006), also result in increased propensity of ESCs to self-renew in the absence of LIF. Similar effects can be elicited by exogenous administration of Wnt3a or expression of an activated form of β-catenin (Takao *et al.* 2007), further demonstrating a role for Wnt signaling in ESC maintenance.

AS and APS are potentially potent ways to regulate transcript diversity in undifferentiated ESCs and differentiation to distinct cell lineages (Pritsker *et al.* 2005; Yeo *et al.* 2007; Kunarso *et al.* 2008). In higher eukaryotes, AS and APS are prominent features that contribute to proteomic diversity by increasing the number of compositionally distinct mRNAs from a single primary transcript. In distinct tissues and cellular states, transcript variation can alter protein interaction networks by removing or inserting protein domains, alter localization in the cell or regulate gene expression (Cooper 2005). In at least one documented case, AS can also alter the inclusion of binding sites for translational repression by microRNAs (Duursma *et al.* 2008). A better understanding of how AS and APS regulate proteomic diversity and translational repression during ESCs differentiation may provide critical insights into this process.

To determine the extent of AS and APS, we measured the expression of competitive exon-exon junctions for over 7,500 genes upon mouse ESC differentiation. This analysis led to the identification transcripts with profound predicted functional changes, including differences in protein domain and microRNA binding site architecture. Among the most intriguing verified transcript changes was an unexplored variant of the transcription factor Tcf3, specifically enriched in ESCs (Tcf3-long). To study the contribution of Tcf3-long and short isoforms in the regulation of pluripotency and differentiation, we have selectively expressed or inhibited expression of these isoforms in mouse ESCs using parallel approaches. Both long and short isoforms caused similar effects on known target gene activity and self renewal in undifferentiated ESC, however, the short isoform was uniquely required for differentiated cell types after induction of lineage commitment in embryoid body and teratoma induction assays. These results reveal multiple functions of Tcf3 isoforms for self-renewal and differentiation processes and suggest that Tcf3 isoform-specific properties/activities differentially control cell fate decisions

6.3 Results

6.3.1 AS and ASP are prominent features with ESC differentiation

ESCs and derived EBs were profiled with a prototype Affymetrix exon-exon junction microarray, which interrogates ~7,500 genes and over 40,000 putative mRNA transcripts. As in previous studies using this array platform, we identified competing exon-exon junctions that indicate the alternative inclusion of cassette-exons, alternative 3’ or 5’ splice sites or alternative promoters using a linear regression based method (Sugnet *et al.* 2006). This algorithm has been incorporated into a free open source stand-alone analysis package named AltAnalyze, specifically designed to analyze exon-exon junction or exon level high-throughput expression data. In addition to scoring for alternative exon events (AEEs), this software can assess the likelihood of a regulated AEE score occurring by chance, assign protein associations to regulated probe sets, and identify functional sequence elements differing between aligning alternate mRNAs and proteins (Figure 6.1 A).

Analysis of ESC differentiation using AltAnalyze identified 144 genes corresponding to 171 unique AEE (see methods) out of 4,269 genes with evidence of isoform expression. Pathway analysis of these genes show enrichment in Wnt and TGF-beta receptor signaling pathways, actin-binding, lipid transport, muscle contraction, chromatin remodeling, and embryonic development among others (Table 6.1). These data suggest that AS and APS regulated genes aligning to processes commonly associated with the regulation of ESC pluripotency and differentiation.

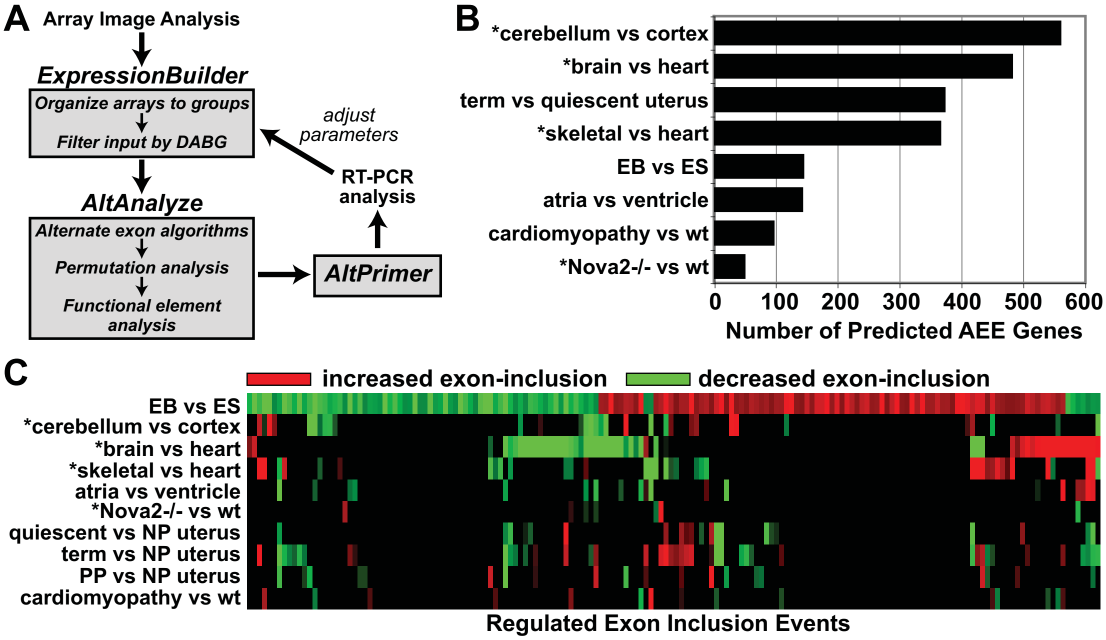
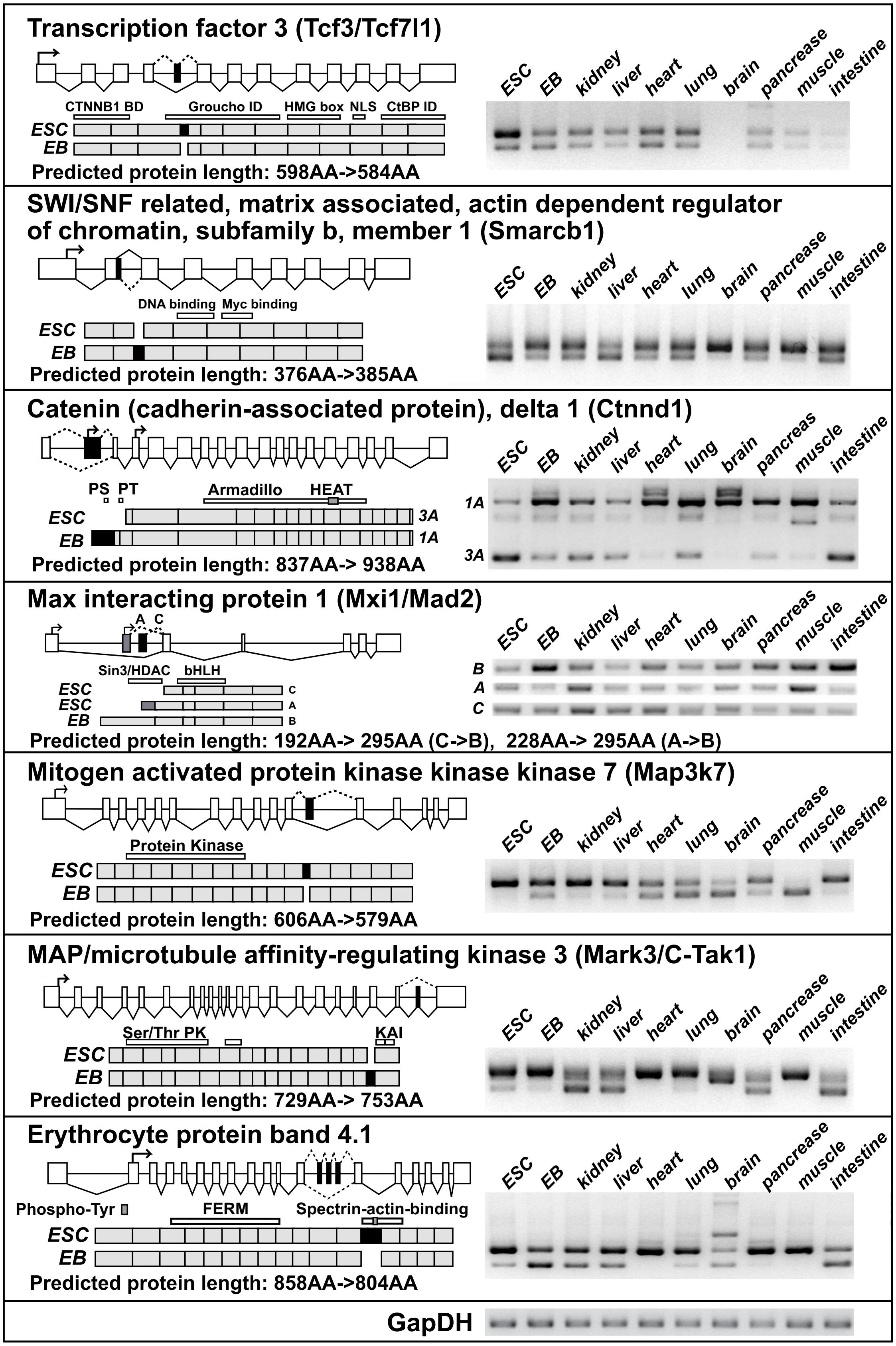


Figure 6.1. Unique alternative exon profiles with ESC differentiation. (A) AltAnalyze was used to process array intensities after normalization (ExpressionBuilder program), calculate alternative exon scores for reciprocal exon-junction probe sets, annotate these events based on existing mRNA structure annotations (e.g., alternative splicing events), and align probe sets to mRNA, protein, and function annotations (AltAnalyze program). These predictions were used for primer design (AltPrimer), RT-PCR validation, and subsequently to re-optimize filtering parameters. (B) Resulting data for unique genes associated with AEEs were compared for datasets indicative of AS. These comparisons include adult tissue conditions from previous studies (asterisks) or collected in parallel with array data for ESC differentiation. (C) Comparison of these profiles for AEEs unique to ESC differentiation. Black indicates the absence of the predicted AEE event for a comparison (no scoring thresholds), red indicates exon inclusion and green indicates exon exclusion for the numerator of the comparison (e.g., EB/ESCs).

Of these AEEs, 94 had clear evidence of AS and 17 had evidence of APS, with the majority aligning to predicted cassette-exons (83%). When APS and AS events were linked to protein sequences (through alignment of competitive junctions to mRNAs), 77 of the 109 AEEs showed predicted differences in protein domain or functional region composition between the alternatively regulated isoforms (InterPro/UniProt). For four AEEs with evidence of AS, multiple putative microRNA binding sites were identified within retained or excluded exons (Atp2a2, Pdlim7, Rbm35a, Eomes) (supplemental dataset), indicating that AS may selectively alter the ability of microRNAs to regulate these proteins. Thus an analysis of AEEs in the context of protein sequence and putative microRNA binding sites highlights a diverse set of putative functional differences.

Name	Type	Changed/ Measured	Z Score	Permute P
positive regulation of myeloid cell differentiation	P	3/6	8.06	0
TGF-β receptor signaling pathway	P	4/24	4.90	0.0005
actin binding	F	10/122	4.63	0
lipid transport	P	5/39	4.59	0.001
in utero embryonic development	P	6/55	4.48	0.002
bone remodeling	P	5/47	4.01	0.0025
chromatin remodeling	P	3/21	3.83	0.0085
smooth muscle contraction	WP	7/76	3.70	0.0005
protein amino acid phosphorylation	P	15/286	3.67	0
cell matrix adhesion	PP	3/26	3.29	0.0165
calcium regulation in cardiac cells	WP	6/68	3.29	0.0015
protein kinase activity	F	13/265	3.14	0.001
Mesoderm development	PP	8/126	2.73	0.0035
Wnt receptor signaling pathway	P	4/57	2.53	0.021

Table 6.1. Pathway analysis of AEEs during ESC differentiation. Highest-scoring pathways and Gene Ontology (GO) terms identified using GO-Elite analysis for genes with AEEs in mouse EBs versus to ESCs. The number of genes associated with AEEs relative to the number of genes measured in each category is reported. The permutation p-value (Permute P) is calculated from 2,000 random permutations of the input gene identifiers. Non-GO term pathways were obtained from http://www.wikipathways.org (WP) or www.pantherdb.org/pathway (PP).



Transcription factor 3 (Tcf3/Tcf7l1)

Predicted protein length: 598AA->584AA

SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily b, member 1 (Smarcb1)

Predicted protein length: 376AA->385AA

Catenin (cadherin-associated protein), delta 1 (Ctnnd1)

Predicted protein length: 837AA-> 938AA

Max interacting protein 1 (Mxi1/Mad2)

Predicted protein length: 192AA-> 295AA (C->B), 228AA-> 295AA (A->B)

Mitogen activated protein kinase kinase kinase 7 (Map3k7)

Predicted protein length: 606AA->579AA

MAP/microtubule affinity-regulating kinase 3 (Mark3/C-Tak1)

Predicted protein length: 729AA-> 753AA

Erythrocyte protein band 4.1

Predicted protein length: 858AA->804AA

Figure 6.2. Validated alternative exons show diverse expression and protein structure variation. Exon and protein structures are shown for genes with large differences in isoform expression between ESCs and EBs. For each gene, the detected alternative exon is indicated as a black box in the exon structure graph of each panel. Surrounding dashed and solid lines indicate alternative events (AS or APS). Below each graph are protein segments corresponding to each exon (not to scale) for both aligning protein isoforms. Functional protein sequences (e.g., domains) are annotated above the two protein representations. Predicted changes in protein length are indicated for the down-regulated (ESC enriched) and up-regulated (EB enriched) isoforms (ES->EB). ESC, EB, and cross-tissue expression profiles from select adult mouse tissues are shown adjacent to the gene structures for each of the probed mRNA isoforms. PS, phosphoserine; PT, phosphotyrosine (UniProt).

To determine the relative extent and overall diversity in alternative exon usage, AEEs from ESCs during differentiation were compared to a panel of tissue and cell remodeling paradigms, from public and in-house collected data. Datasets with previous evidence of large-scale AEEs include cardiac versus brain or skeletal muscle (Sugnet *et al.* 2006), atria versus ventricle (Sato *et al.* 2003; Chu *et al.* 2004) and Nova2-/- versus wild-type brain (Ule *et al.* 2005). Since AS and APS have also been implicated in distinct cell remodeling paradigms we also analyzed a time-course of myometrial gestational and cardiomyopathy remodeling (Biesiadecki *et al.* 2002; Curley *et al.* 2004; Dabertrand *et al.* 2007; Tyson-Capper 2007). Combined analysis of these datasets using the same AltAnalyze analysis parameters shows a wide range in the number of alternatively regulated genes predicted (Figure 6.1 B and supplemental data).

Although only a modest number of predicted AEEs were found with ESC differentiation as compared to other datasets, these events were mainly restricted to this dataset (Figure 6.1 C).

6.3.2 Verified alternative exons align to pathways of Wnt signaling and cell cycle control

To verify AEE predictions from AltAnalyze, we analyzed 24 AS and 4 APS predicted ESC differentiation events using RT-PCR. Twenty-two RT-PCR reactions produced amplicons from which 15 AS and 3 APS events had the predicted pattern of isoform expression (supplemental data). To further characterize these verified isoform changes, we examined protein-associated functional changes along with isoform expression across multiple adult tissues (Figure 6.2). Several of these verified AEEs were specifically found to alter the domain/functional residue composition of proteins (e.g., Tcf3, Mxi1, Epb4.1), suggesting that alternative exon inclusion may alter protein function. Although most isoforms enriched in ESCs were also expressed in other tissues, expression was highest in ESCs for Tcf3, Smarcb1, Map3k7, and Cttnd1 for one of the two regulated isoforms. Analysis of these genes in the context of previously demonstrated interactions reveals a putative signaling network that connects many of these genes to the regulation of self-renewal and differentiation pathways (Figure 6.3) (Ishitani *et al.* 1999; Bachmann *et al.* 2004; Huang *et al.* 2005; Imbalzano *et al.* 2005; Spring *et al.* 2005; Pereira *et al.* 2006; Dugast-Darzacq *et al.* 2007; Katoh *et al.* 2007). These include the regulation of the Wnt signaling components Tcf3, Ctnnd1, and Map3k7 and the cell cycle regulators Mark3, Smarcb1, Mxi1, and Epb4.1.

6.3.3 Tcf3-long is enriched in ESCs and retains Tcf3-short transcriptional repression activity

One of the most compelling findings from this dataset was the AS of the Wnt signaling transcription factor Tcf3. Although Tcf3 regulates pluripotency through interactions with Nanog and Oct4, the characterized isoform is the non-ESC enriched form (short form). Tcf3 or transcription factor 3 (TCF7L1 in human) is a member of the Tcf/Lef family of transcription factors that mainly represses transcription of downstream target genes. Genome-wide analyses suggest that Tcf3 regulates the transcription of many crucial pluripotency and developmental regulators (Cole *et al.* 2008; Tam *et al.* 2008; Yi *et al.* 2008). Specifically, repression of Nanog and Oct4 transcription by Tcf3-short has been hypothesized to be essential for ESCs to balance lineage-commitment and pluripotency (Pereira *et al.* 2006).

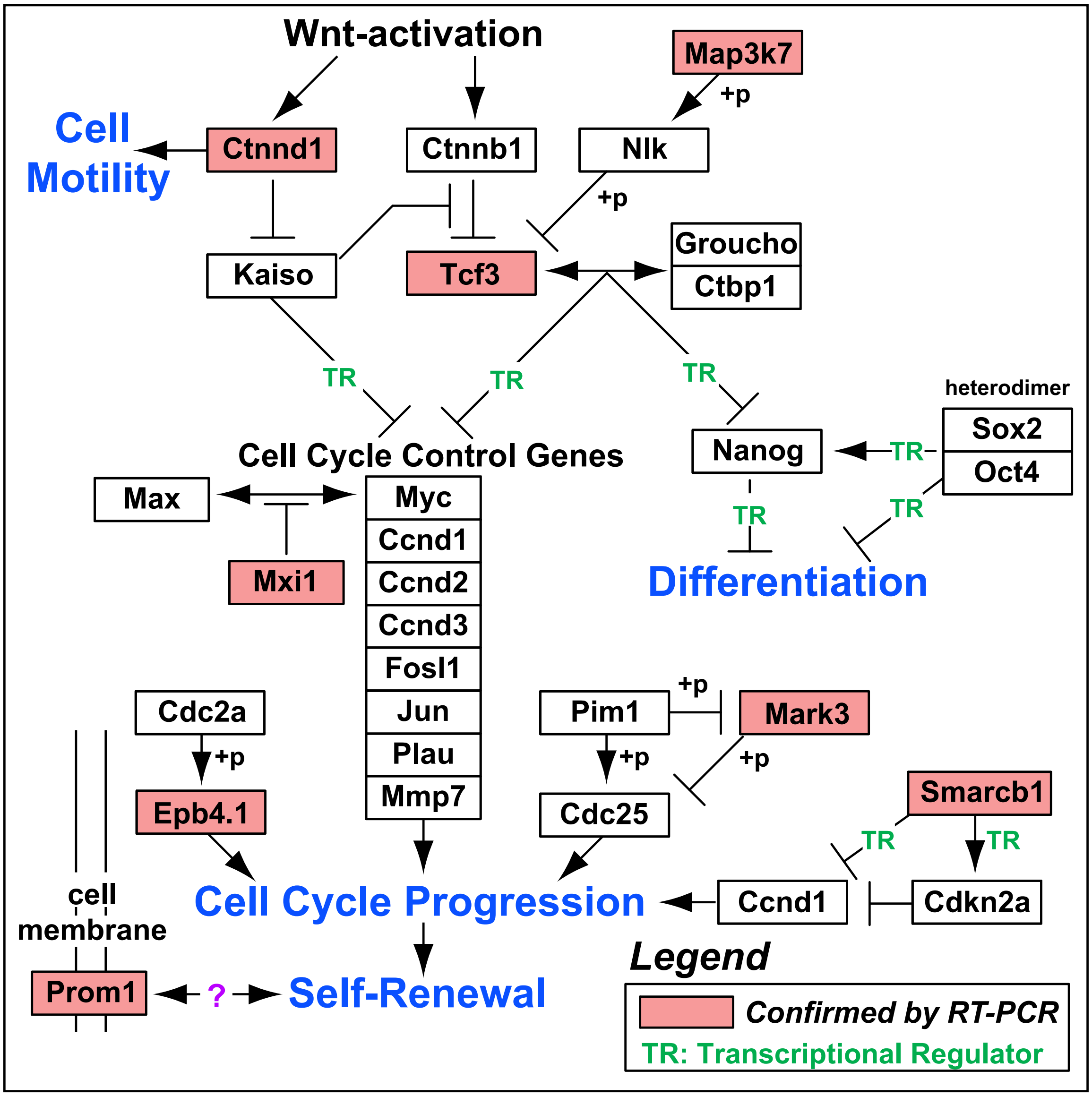


Figure 6.3. AS and APS proteins intersect with pathways of pluripotency and differentiation. Proposed model of interaction between Wnt signaling, cell cycle control, and the regulation of differentiation and pluripotency for validated AS and APS genes (red boxes). Arrows indicate promotion and T-bars indicate inhibition.

The two Tcf3 isoforms detected by our microarray analysis differ in the inclusion of a 42 base pair (bp) cassette exon, which encodes an additional 14 amino acids that overlap with the described Groucho binding domain (Cavallo *et al.* 1998; Roose *et al.* 1998). Tcf3-long is up-regulated 2 fold in mouse ESCs relative to EBs while the short form is expressed at roughly equivalent levels, as shown by quantitative PCR (qPCR) and RT-PCR (supplemental data). To determine if both isoforms encode for translated proteins, we expressed cDNAs encoding each Tcf3 isoform in a Tcf3-null ESC line (G4). Using an antibody directed against the carboxy terminal region by both isoforms, we found that both cDNAs encode proteins of similar size, with wild-type ESCs expressing both isoforms (Figure 6.4 A).

To assess the ability of Tcf3 isoforms to repress the transcription of known targets, both isoforms were transiently expressed in Tcf3-/- ESCs and assayed for expression of β-catenin transcriptional (TOPFlash) (Veeman *et al.* 2003) or Nanog promoter luciferase reporters (Pereira *et al.* 2006). Both isoforms produced equivalent repression of the TOPFlash and Nanog promoter reporters with transfection of increasing amounts of cDNA (Figure 6.4 C, D). Therefore, the ESC-enriched long isoform of Tcf3 retains transcription repression activity for β-catenin targets and Nanog. Although this data suggests both Tcf3 isoforms have identical transcriptional activities, these in vitro effects may not be recapitulated on endogenous promoters.

6.3.4 Delayed differentiation of Tcf3 knockdown lines

Given that both Tcf3 isoforms are expressed in ESCs and EBs and show functional activity, we utilized RNA interference (RNAi) to explore the specific relationships of each isoform in pluripotent ESCs and upon differentiation to multiple cell lineages. Stable clonal mouse E14 ESCs expressing a Tcf3-short shRNA construct targeting the exon 3-4 junction, a Tcf3-long shRNA construct targeting the 42bp inclusion exon, or both isoforms were obtained through lentiviral infection and puromyocin antibody resistance selection (Figure 6.5 A). This strategy yielded up to 90% KD of the targeted isoforms, with minimal or no reduction in the non-targeted isoform in ESCs and EBs (Figure 6.5 B). Tcf3 protein could not be detected with Tcf3-all KD by Western blot in undifferentiated ESCs (data not shown).

Isoform-specific Tcf3 KD lines displayed unique cell colony morphology in the presence and absence of LIF. In the presence of LIF, constitutive Tcf3 KD (Tcf3-all KD) ESCs had a highly clustered, round colony morphology, whereas wild-type E14 ESCs were predominantly distributed in a monolayer on gelatinized culture plates (typical for this cell line). Tcf3-short KD had a similar clustered appearance to Tcf3-all KD, whereas Tcf3-long KD formed very small clusters. By day 3 of LIF removal, wild-type ESCs had a differentiated morphology and could not be passaged further, while Tcf3-all and Tcf3-short KD lines flattened out by day 3 or day 6 with a more E14-like morphology. All Tcf3-KD lines could be further passaged to day 6 and did not have a predominantly differentiated morphology. These results agree well with recent reports that complete Tcf3 knock-out and KD cells can be maintained under LIF-independent conditions for extended periods in culture (Tam *et al.* 2008; Yi *et al.* 2008).

6.3.5 Preferential regulation of Nanog and Oct4 by Tcf3-long isoform knockdown

Since Nanog transcriptional repression by Tcf3 isoforms was assessed using an artificial Nanog reporter and cDNA was expressed at non physiological levels, we examined the expression of both Nanog and Oct4 using the isoform-specific shRNAs. Comparison of Nanog expression levels in ESCs by qPCR for each Tcf3 shRNA revealed maximal up-regulation with Tcf3-all KD (4.0-fold), followed by Tcf3-long (3.5-fold), and Tcf3-short (1.8-fold) KD. Interestingly, while Oct4 was up-regulated with Tcf3-long KD (2.0-fold), Oct4 expression levels were not significantly changed with either Tcf3-short or Tcf3-all KD relative to wild-type ESCs (t-test $p<0.05$). Although Tcf3-long KD produced preferential up-regulation of both Oct4 and Nanog compared to Tcf3-short KD, this effect could be due to the higher expression of Tcf3-long in undifferentiated ESCs.

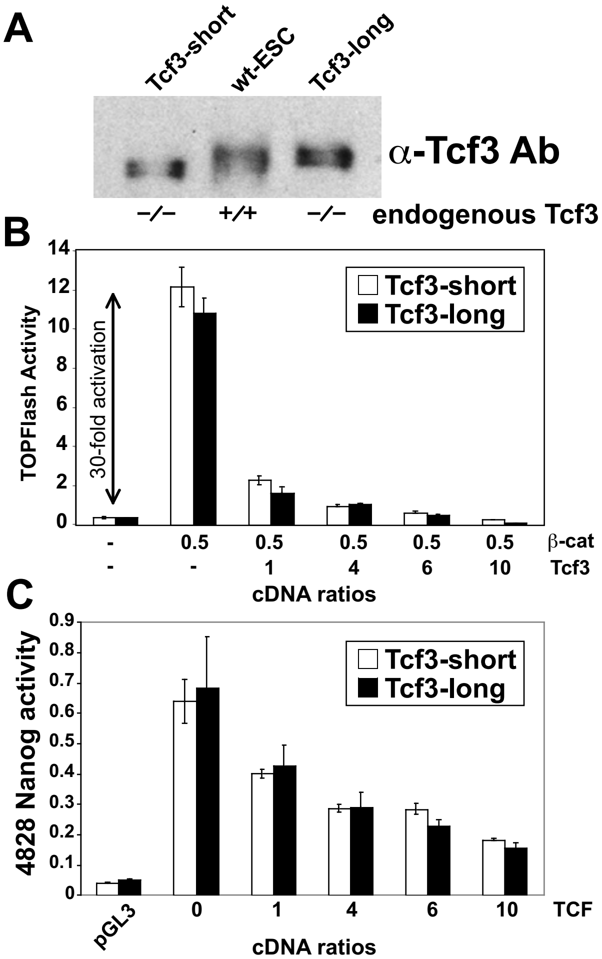


Figure 6.4. Both Tcf3 isoforms are expressed in mouse ESCs and repress reporter transcription. (A) Tcf3-/- ESCs were transfected with full-length cDNAs for both isoforms expressed under the control of a cytomegalovirus promoter and compared to wild-type ESCs. Expression of Tcf3-short and long isoforms resulted in a shift in the detection of Tcf3 protein on a polyacrylamide gel using a Tcf3 antibody common to both isoforms. To assess activity of these isoforms, Tcf3-/- ESCs were transfected with (B) a β-catenin transcriptional reporter construct (TOPFlash) or (C) Nanog promoter reporter, driving luciferase expression, in the presence or absence of increasing amounts of transfected cDNA for the Tcf3 isoforms. To assay for transcriptional repression in the TOPFlash reporter assay, cells were co-transfected with a stable form of β-catenin (β-cat).

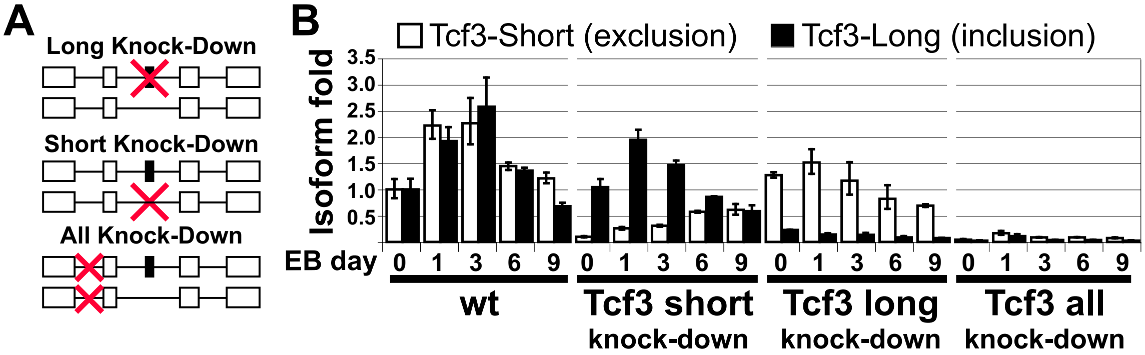


Figure 6.5. Selective knockdown of Tcf3 isoforms in ESCs and differentiating EBs. (A) To achieve isoform-specific KD, Tcf3 regions unique to or in common to each isoform were targeted using short-hairpin RNAs directed against either exon-junctions (E3-E4 for all or E4-E5 for short) or exons (exon E4a for long). (B) KD efficiency for Tcf3-short, Tct3-long, and Tcf3-all KD clonal lines by isoform-specific qPCR with ESC differentiation to EBs for 1, 3, 6 and 9 days.

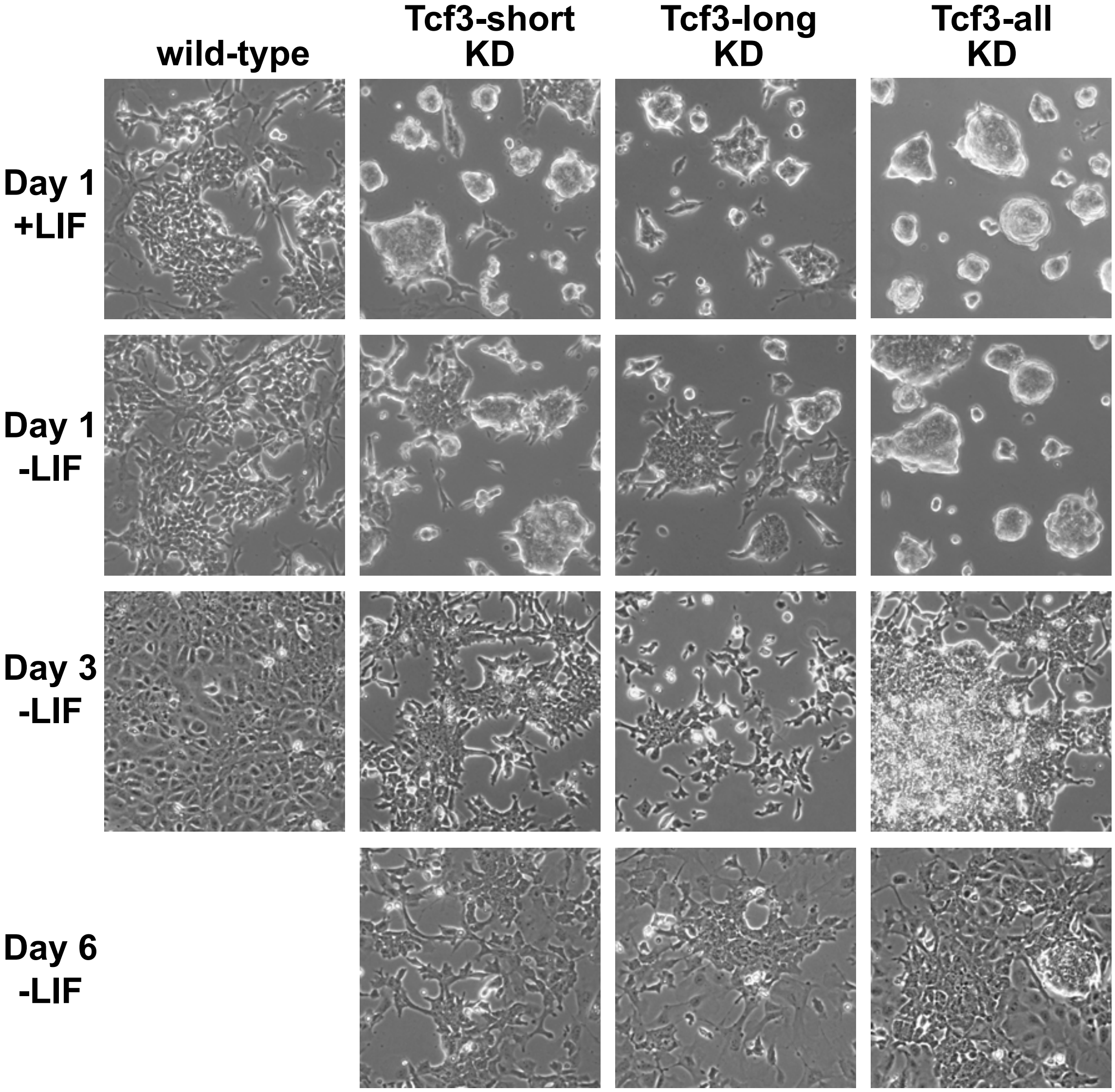


Figure 6.6. Tcf3 knockdown cells fail to differentiate upon LIF removal. Cell morphology, clustering and differentiation were observed for wild-type, Tcf3-short, Tcf3-long, and Tcf3-all KD ESCs after removal of LIF (-LIF).

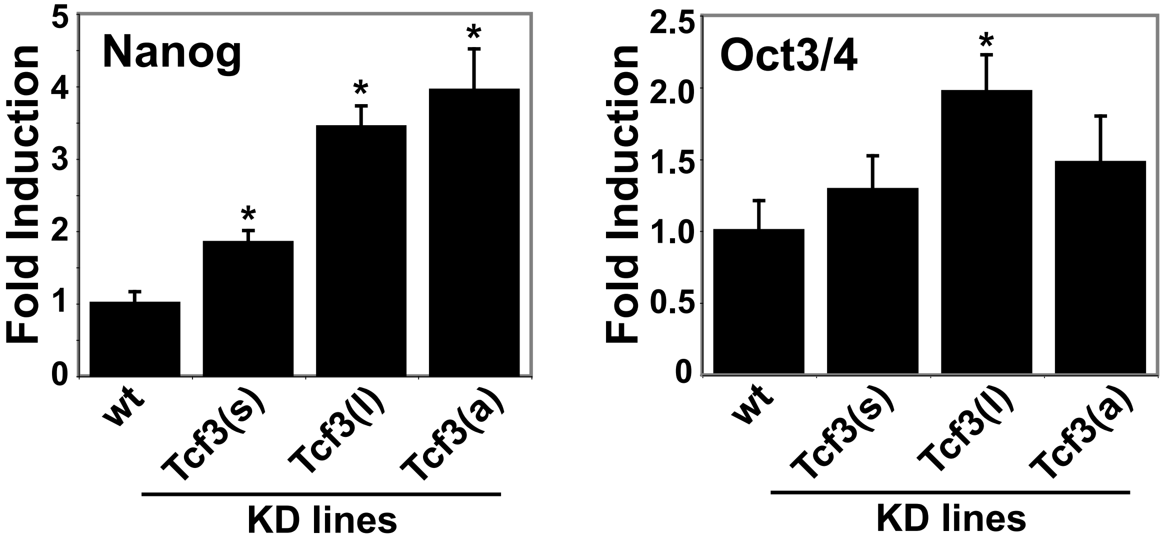


Figure 6.7. Oct4 and Nanog are preferentially up-regulated with knockdown of Tcf3-long. Quantitative expression levels for Oct4 and Nanog relative to the endogenous controls Ppia and Rpl7 are reported for undifferentiated wild-type (wt) and isoform specific Tcf3 KD lines cultured in the presence of LIF. Fold induction is relative to wild-type ESCs. Tcf3(a), Tcf3-all; Tcf3(s), Tcf3-short; Tcf3(l), Tcf3-long. Values are mean ± SEM of biological triplicates. An asterisk indicates t-test $p<0.05$.

6.3.6 Tcf3 isoform knockdowns differentially block differentiation pathways

For wild-type and Tcf3 KD lines, EBs differentiated for 1, 3, 6, and 9 days were analyzed by qPCR for expression of tissue lineage markers for all three early germ layers (mesoderm, endoderm and ectoderm) and late adult cell markers (cardiac and neural). Normal induction of gene expression was observed for all lineage markers examined in wild-type E14 ESCs; induction of early ectoderm markers (Fgf5 and Nestin) was maximal at 6 days while induction of all other markers was maximal at 9 days (Figure 6.6). Surprisingly, Tcf3-short KD blocked the expression of nearly all markers examined, with the greatest repression of markers for early mesoderm (Vegfr2, Gata4 and Gata6), followed by endoderm (FoxA2, Afp), and to a far lesser extent ectoderm (Fgf5 and Nestin). The only marker that was induced at levels similar to or higher than wild-type was the early mesoderm marker Brachyury (Bry), which is only induced ~2-fold by day 6 in wild-type cells, with similar induction in Tcf3-short KD at day 3. Since Bry is transiently induced at day 4 of E14 ESC differentiation, it was likely missed by our analysis (Ivey *et al.* 2008). The least repressed marker, other than Bry, was Sox1, a late neural marker, followed by Nestin and Fgf5, both markers of ectoderm. This effect was not due to a global decrease in gene induction, since expression of cell cycle and self-renewal regulators show similar or increased expression relative to wild-type ESCs throughout EB differentiation (supplemental data). Thus, expression of mesoderm and endoderm markers are blocked throughout EB differentiation, while early and late ectoderm markers are far less diminished.

In contrast to Tcf3-short KD, Tcf3-long KD resulted in only small changes in early lineage marker expression compared to wild-type ESCs, but blocked expression of later markers. Among the early markers examined, only Gata6 and Vegfr2 expression at day 9 of EB differentiation were down-regulated, while Vegfr2 expression was nearly 2-fold higher in Tcf3-long EBs at day 6 relative to wild-type ESCs. For the ectoderm markers Fgf5 and Nestin, Tcf3-long KD produced a sustained increase in gene expression through day 9, whereas in wild-type EBs, expression was lower on day than day 6. These differences may account for the absence of late lineage marker induction.

Interestingly, although KD of both Tcf3 isoforms produced gene expression changes similar to those induced by Tcf3-short KD, Tcf3-all KD had far less severe consequences on lineage marker expression, even though similar or greater KD of both short and long isoforms was achieved. Tcf3-all KD shows delayed and diminished expression of the endoderm/mesoderm markers Gata4 and Gata6, mesoderm marker Vegfr2, and endoderm marker Afp, but not ectoderm markers Fgf5 or Nestin. Furthermore, both FoxA2 and Bry expression was up-regulated compared to wild-type. Consistent with this observation, both Bry and FoxA2 tissue expression was increased in Tcf3-/- mice, mainly due to axial duplication (Merrill *et al.* 2004). Thus, KD of Tcf3-long was able to restore FoxA2 expression in differentiating EBs, as compared to Tcf3-short KD.

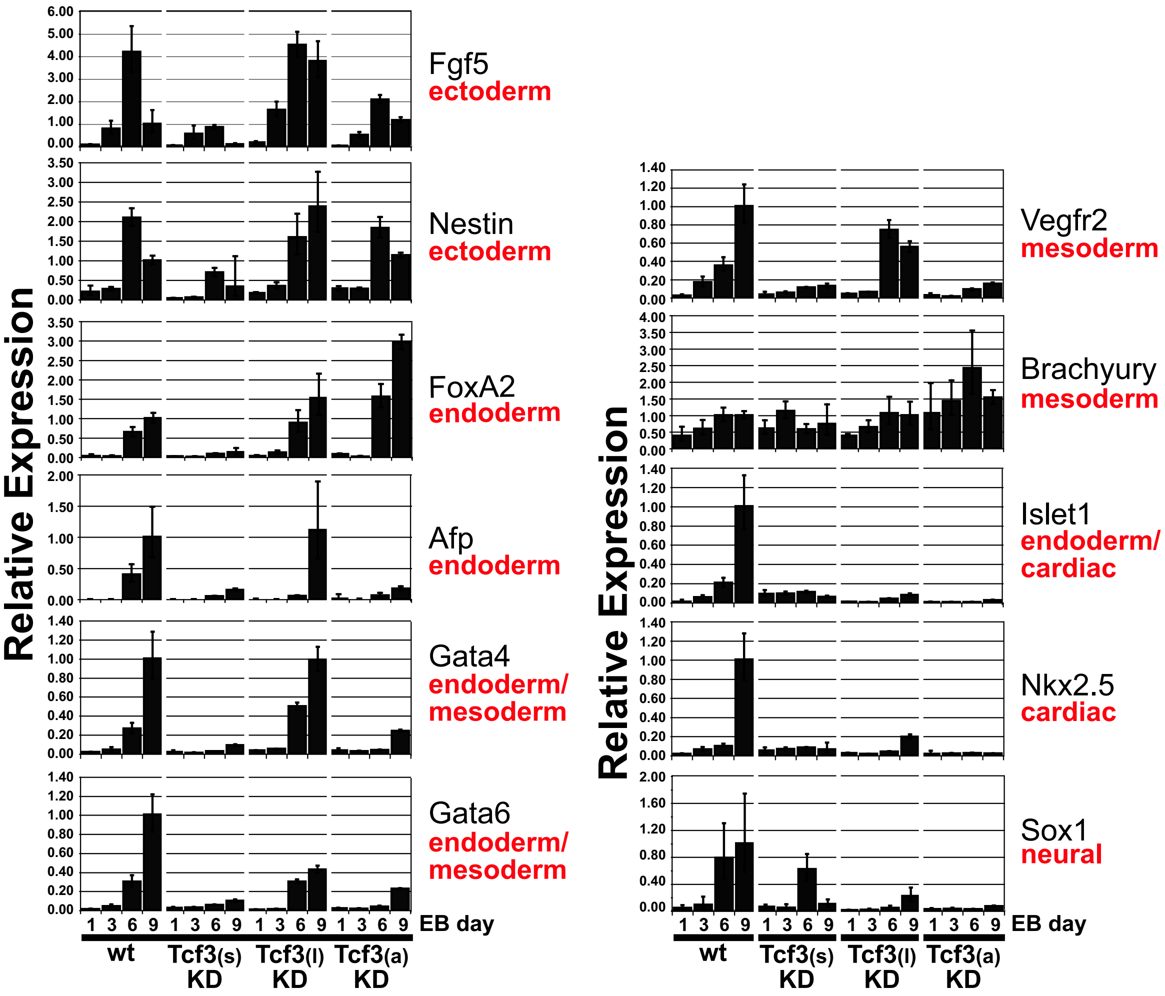


Figure 6.8. Distinct patterns of lineage marker expression for all Tcf3 knockdown lines. qPCR was performed on RNA extracted from multiple days of EB differentiation (days 1-9) for wild-type and Tcf3 isoform KD lines. Expression for each gene is determined relative to two endogenous control genes (Ppia and Rpl7) and relative gene expression changes for all cell lines and time-points are compared to gene expression at day 9 of differentiation, where gene expression was typically most highly induced. The lineages for which each marker is associated with are listed in red under the gene name for each. Values are mean ± SEM of technical triplicates of pooled EB plates (n=96 EBs or greater).

To confirm these results, we derived teratomas by injection of wild-type and shRNA ESCs under the skin of Severe Combined Immunodeficiency (SCID) mice. After 4 weeks of growth, wild-type ESC teratomas developed into cells of all 3 primordial germ layers (e.g., muscle, neural rosettes, cartilage, adipose and epithelial cells), as shown by morphological examination of hematoxylin and eosin stain (H&E) stained sections (Figure 6.9 A). Examination of Tcf3-long KD ESC-derived tumors revealed similar cell structures to wild-type ESCs with a similar predominance in cell types. However, in Tcf3-short and Tcf3-all KD ESC derived-tumors there is an absence in cell structures characteristic of endodermal and mesodermal tissues, but a predominance of cells with neural rosette morphology and organization. These data appear to agree with lineage marker expression with differentiation, supporting the hypothesis that Tcf3-short is essential for mesoderm and endoderm differentiation, while Tcf3-long appears to be dispensable for early lineage commitment. Future immunohistochemical analysis of specific cell types in these teratomas will be important to verify these observations.

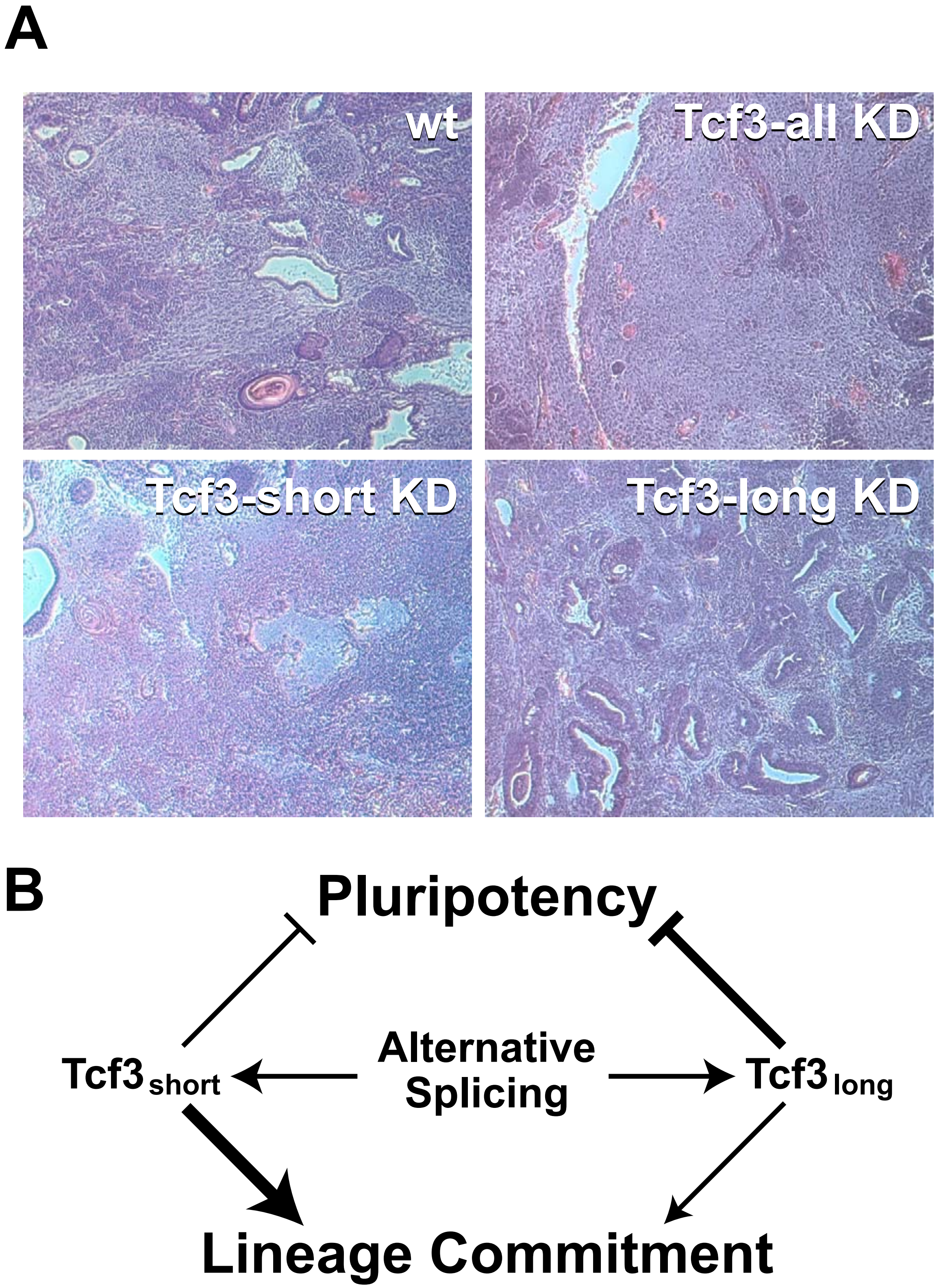


Figure 6.9. Absence of differentiated mesodermal and endodermal

structures with Tcf3-short and complete knockdown. (A) H&E staining of tumors derived by injection of ESCs in to SCID mice for each of the Tcf3 KD and wild-type lines. (B) Proposed model for Tcf3 isoform function in the context of pluripotency and differentiation.

6.4 Discussion

AS and APS have been increasingly recognized as critical mechanisms during development to increase proteomic diversity and resulting signaling complexity. This has been elegantly demonstrated for the AS of genes involved in neonatal to adult cardiac adaptation, synaptogenesis and sex determination (Burgess *et al.* 1999; Hammes *et al.* 2001; Xu *et al.* 2005). With new methods to interrogate alternative exons on a whole-genome scale, we have the opportunity to fully appreciate the contribution and specificity to which these AS and APS contribute to biological complexity.

In this analysis, examination of exon-exon junction profiles using AltAnalyze, led to the identification of 91 putative AS and 15 APS events during the transition of mouse ESCs to differentiated EBs. While the number of AEEs detected for ESC differentiation is far less than distinct cell type comparisons examined (e.g., brain and heart), this effect is likely due to the heterogeneity of EBs, which provides greater specificity for identifying isoforms specifically enriched in undifferentiated cells. This is evidenced by comparison of AEEs for the different AS paradigms, revealing that isoforms regulated during ESC differentiation are relatively restricted to this dataset (Figure 6.1 C).

Isoforms with the most extreme changes detected by RT-PCR corresponded to either pathways of Wnt signaling (Tcf3, Ctnnd1, Map3k7) or cell cycle progression (Smarcb1, Epb4.1, Mark3 and Mxi1). Many of these AEEs altered the composition of the annotated protein domains and functionally important sequences. These data suggest that AS and APS genes may encode unique functional protein products during differentiation that may contribute to alternate pluripotency or differentiation outcomes. To test this hypothesis, we examined the contribution of splice isoforms for the transcription factor Tcf3, a critical component of the pluripotency core transcriptional network that was highlighted by our alternative exon analysis.

Both Tcf3 isoforms were found to produce full-length proteins in ESCs as well as repress expression of Nanog and Tcf-β-catenin reporters. Using isoform-specific RNA inactivation, we found that both isoforms have distinct functional outcomes during the differentiation of ESCs. The ESC-enriched, long Tcf3 isoform appeared to be dispensable for early lineage specification, but the Tcf3-short form was required for the expression of both early (e.g., Nestin, Vegfr2, and Afp) and late tissue lineage markers (e.g., Sox2, Islet1, and Nkx2.5). Similarly, analysis of teratomas derived from KD ESCs showed that Tcf3-short and Tcf3-all KD appeared to suppress endodermal and mesodermal cell structures. Given that these Tcf3 isoforms likely retain their endogenous tissue specificities, these effects might be due to cellular distribution differences in differentiating cells.

These results indicate that Tcf3 isoforms fulfill distinct roles during the transition from pluripotency to differentiation, with Tcf3-short being indispensable for normal differentiation (Figure 6.9 B). While the precise function of Tcf3-long, what co-factors it requires, and its transcriptional targets has not been clearly elucidated in these studies, evidence that it negatively regulates expression of both Nanog and Oct4 in ESCs suggests that its function is critical in these cells. Interestingly, KD of either isoform was sufficient to inhibit differentiation of ESCs in the absence of LIF and further suggests that Tcf3 isoforms inhibit self-renewal in a dose-dependent manner. The different effects of the Tcf3 isoforms could thus be mediated by cell-type specific expression or alternatively through interactions with distinct binding partners. Although we could not identify an equivalent Tcf3-long isoform in human ESCs (supplemental data), the role of Tcf/Lef proteins has not been clearly defined in human ESCs as of yet. Ultimately, this study provides further evidence that Tcf3 is a critical component of pluripotent cells and demonstrates that its post-transcriptional regulation is a significant determinant of mouse ESC differentiation. These results will enable further studies necessary to delinate how precise activities of alternatively spliced gene products shape the processes of self renewal and lineage commitment

6.5 Materials and Methods

6.5.1 Tissue isolation and sample preparation

Developmental conditions examined were mouse ESCs differentiated in to embryoid bodies (EBs) and myometrium from adult virgin mice, mice at 14.5 days (quiescent), 18.5 days (term) of pregnancy, and mice 6hr post-partum.

Disease conditions consisted of mouse adult cardiac ventricles from a model of dilated cardiomyopathy (Redfern *et al.* 2000) compared to MH-tTA littermate ventricles. Tissue comparisons consisted of data obtained from a previous analysis of AS (Sugnet *et al.* 2006) and from mouse cardiac atria and ventricles harvested from wild-type FVBN mice. Mouse E14 ESCs were grown in a monolayer on gelatin-coated culture plates, maintained in medium supplemented with 10% FBS, pyruvate, non-essential amino-acids, β-mercaptoethanol, LIF, and passaged with trypsin. EBs were derived using the hanging-drop method as described (Ivey *et al.* 2008) in 20% FBS to enrich for the mesoderm lineage. The dilated cardiomyopathy model was created using double transgenic mice harboring the tetracycline transactivator (tTA) driven in the heart by a myosin alpha-heavy chain promoter (MH). When expressed, tTA promotes the expression of a modified version of the kappa-opioid receptor driven by tTA-responsive promoter, in the absence of doxycycline (Redfern *et al.* 2000). Cardiac ventricle was harvested at 8 weeks after doxycycline was withdrawal (inducible expression). Total RNA was isolated from snap-frozen cell cultures/tissues using Trizol extraction and purified with the Qiagen RNA purification kit. For microarray sample preparation, the purified total RNA was converted to cDNA using random hexamer primers and Superscript III RNA polymerase (Invitrogen), fragmented by DNase I digestion (Amersham Pharmacia Biotech), and end-labelled with Terminal Transferase, recombinant with DLR-1a (Roche). This cocktail was hybridized to the Affymetrix prototype AltMouse A array (containing only perfect-match probe sets) (Ule *et al.* 2005), (Sugnet *et al.* 2006) according to the manufacturer's instructions. CEL files produced from the resulting Affymetrix DAT image files with MAS5 were used for all downstream analyses. For nearly all-downstream bioinformatics analyses, we wrote a freely available software package named AltAnalyze (figure 6.1 A). This software performs probe set filtering (downstream of normalization and detection probability calculation), calculates AEE scores and probabilities, and performs functional motif analysis.

6.5.2 Normalization and probe set filtering

CEL files for each set of comparisons (limited to two groups) were used to calculate normalized and background-corrected probe set expression values using Robust Multi-Chip Analsyis (RMA)(Irizarry *et al.* 2003) from Bioconductor (http://bioconductor.org/CRAN/). In parallel, to eliminate probe sets which are expressed at background levels and thus may contribute to false alternative exon predictions, detection p-values for each probe set were calculated similar to Absent-Present calls for microarrays containing both perfect-match and mismatch probes (Sugnet *et al.* 2006) directly from the CEL files using a custom Python script. These probe set expression values and detection probabilities were used as input for the ExpressionBuilder module of the program AltAnalyze. In ExpressionBuilder, probe sets corresponding to alternative exons or exon-junctions were excluded that did not have an average detection p-value less than 75% in at least one experimental condition. Likewise, for inclusion of probe sets corresponding to constitutive transcript regions, both experimental groups were required to have the same mean detection p-value thresholds.

6.5.3 Alternative exon analysis

Two previously described algorithms, analysis of splicing by isoform reciprocity or ASPIRE (Ule *et al.* 2005) and a linear regression based method (Sugnet *et al.* 2006) were adapted and used to generate scores for exon inclusion in the AltAnalyze module (supplemental methods). However, for these studies, only results obtained with the linear regression method are reported. Each putative AEE can consist of two differentially regulated exon-exon junctions or an exon and an exon-exon junction that suggest reciprocal splicing or transcription based on the position of these exons in the transcript sequences (Affymetrix). For each AEE, a permutation-based p-value was calculated by permuting the original probe set data for each gene to recalculate a score for all possible permutations of the samples in the two groups. The p-value is based on the rank of the unpermuted score in the distribution of permutation-based scores. AEE scores with a permutation p-value < 0.05, linear regression fold > 2, and constitutive gene expression difference less than 3-fold were selected for downstream analysis. Additional algorithm details are provided with the source code and executable documentation.

6.5.4 Protein inference and functional motif comparison

Prior to exon-exon junction analysis, protein alignments to all reciprocal microarray probe sets were derived using a custom Python script named

LinkEST. LinkEST aligns each probe set sequence to any mRNAs present in the Unigene or Ensembl databases to find the longest matching and non-matching mRNA sequences (or proteins for Ensembl) for corresponding genes. When choosing the longest mRNA sequence, precedence is given to sequences with full-length annotations (e.g., mRNA versus EST). For selected mRNAs matching to a probe set without a recorded protein translation, *in silico* translations were derived. These relationships are used by AltAnalyze to identify protein sequences associated with each reciprocal exon-exon junction probe set. If protein information is missing for one of the two reciprocal probe sets, the longest non-matching protein association (from the probe set with associated protein information) is used as the reciprocals protein association.

Protein domains and functional region sequences, for one or more residues, were obtained from the UniProt and Ensembl databases. These functional annotations were compared between the two probe set aligned protein sequences. If the complete domain-level sequence was present in one but not the other protein sequence, this functional annotation (e.g., kinase domain) was reported for the AEE. For experimentally confirmed AEEs, literature searches were used to identify additional domain-level annotations.

To identify alternative exons containing putative microRNA binding sites, such binding site sequences were extracted from existing resources (PicTar (Krek *et al.* 2005), miRanda (http://www.microrna.org), miRbase (Griffiths-Jones *et al.* 2008), and TargetScan (http://www.targetscan.org)), and searched for in all possible alternative exon sequences. These alternative exons are associated with AEEs identified by AltAnalyze. For unique genes containing regulated functional regions or predicted microRNA binding sites, an over-representation z-score was calculated and reported by AltAnalyze.

6.5.5 Quantitative and Semi-Quantitative AEE Confirmation

For many of the AEEs with the largest linear regression fold changes, RT-PCR was used to confirm isoform expression changes. Reverse transcription and gene/isoform-specific PCR was carried out using One-Step RT-PCR with Superscript III reverse transcriptase (Invitrogen) amplified for 30-40 cycles and resolved on a 2% or 2.5% agarose gel in Tris-acetate-EDTA. Flanking PCR primers to amplify both isoforms in a single reaction or two isoform specific primer sets were designed using a custom implementation of Primer 3 called AltPrimer (http://conklinwolf.ucsf.edu/tools/picoprimer.html). Quantitative PCR using the SyBR green method was used to measure Tcf3 isoform expression for RNAi knockdown cells. Analysis of differential gene expression in ESCs and EBs was conducted using TaqMan analysis (AppliedBiosystems, Foster City, CA). Rpl7 and Ppia were selected as stable reference genes for TaqMan analysis based on predictions made by GeNorm algorithm after examining six genes (Actnb1, Gapdh, Pgk1, Ubb, Rpl7 and Ppia) (Vandesompele *et al.* 2002).

6.5.6 Isoform-specific expression/RNAi in mouse ESCs

Stable isoform-specific knockdown of Tcf3 alternatively spliced isoforms were obtained in E14 ESCs using sequence specific shRNAs, delivered by lentiviral infection. Three 19 mer shRNAs were designed to target the long, exon inclusion isoform (GGATGGTGCCTCCCACATT), the short exon exclusion isoform (CCAGCACACTTGTCCAACA), and constitutive region of Tcf3 (GCACCTACCTACAGATGAA) using overlapping predictions from the program PSICOLIGOMAKER 1.5 (http://web.mit.edu/jacks-lab/protocols/pSico.html) and the Broad Institute mouse hairpin library (http://www.broad.mit.edu/genome_bio/trc/rnai.html). The pSicoR construct (gifts from Tyler Jacks and Miguel Ramolos-Santos labs) was re-engineered to drive expression of mCherry protein using the Ef1α promoter and a puromycin-resistance gene to allow for stable colony selection. Isoform-specific shRNAs were ligated into the pSicoR-Ef1α-mCh-puro construct and cotransfected with viral the packaging plasmids pMDLgpRRE, pRSV_Rev (D. Trono), and pVSV-G (Clontech) (gifts from the Miguel Ramalho-Santos lab) into HEK293 cells using FuGENE6 (Roche) as previously described (Grskovic *et al.* 2007).

Harvested supernatant from viral producing HEK293 cells was filtered through a 0.45um filter, and 100μl incubated with 200,000 mESCs on rotator for 3 hours. Cells were plated onto gelatinized tissue culture plates, grown under feeder-free conditions in the presence of LIF and selected for puromycin-resistant colonies for at least 5 days. Clonal populations of mCherry-expressing mESCs were screened with isoform-specific qPCR primers to select for clones with optimal isoform-specific KD. cDNAs for both the two Tcf3 isoforms were also expressed in Tcf3-/- ESCs on 129/Sv background (GS1) by electroporation or transfection of a linearized pCDNA3-Tcf3 short (pBM58) form (Pereira *et al.* 2006) or long-form construct. The Tcf3-long form cDNA was obtained by removal of a 670 bp fragment by Kpn1-Pml1 (NEB) digestion of pBM58 and insertion of the corresponding 712 bp ESC RT-PCR fragment (sequence verified) from the long Tcf3 isoform.

6.5.7 Transcription reporter assays

For Tcf3 cDNA transcriptional reporter assays, TCF3 -/- GS1 ESCs (30,000/well) were plated on gelatin-coated 24-well plates and grown in the presence of LIF for 24 hr. The expression constructs pCDNA3-Tcf3-short and pCDNA3-Tcf3-long were transfected into ESCs with Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. To detect β-catenin activated transcription, the TOPFlash assay was used. DNA (1.5μg/well) was prepared in duplicate, using the reporter mix (Topflash:pRLCMV, 0.4 μg/well), Renilla as a transfection efficiency control, β-catenin (0.05 μg/well) and TCF3 cDNA (0.1, 0.3, 0.6, and 1 μg/well). To assess Nanog promoter activity, DNA (1.4μg/well) was prepared in duplicate, consisting of the reporter mix (Nanog 4828 promoter:pRLCMV, 0.4 μg/well), and TCF3 cDNA (0.2, 0.4, 0.6, and 1 μg/well). To achieve the same DNA transfection concentrations per well, empty pcDNA3 was added as appropriate.

6.5.8 Western blot analysis

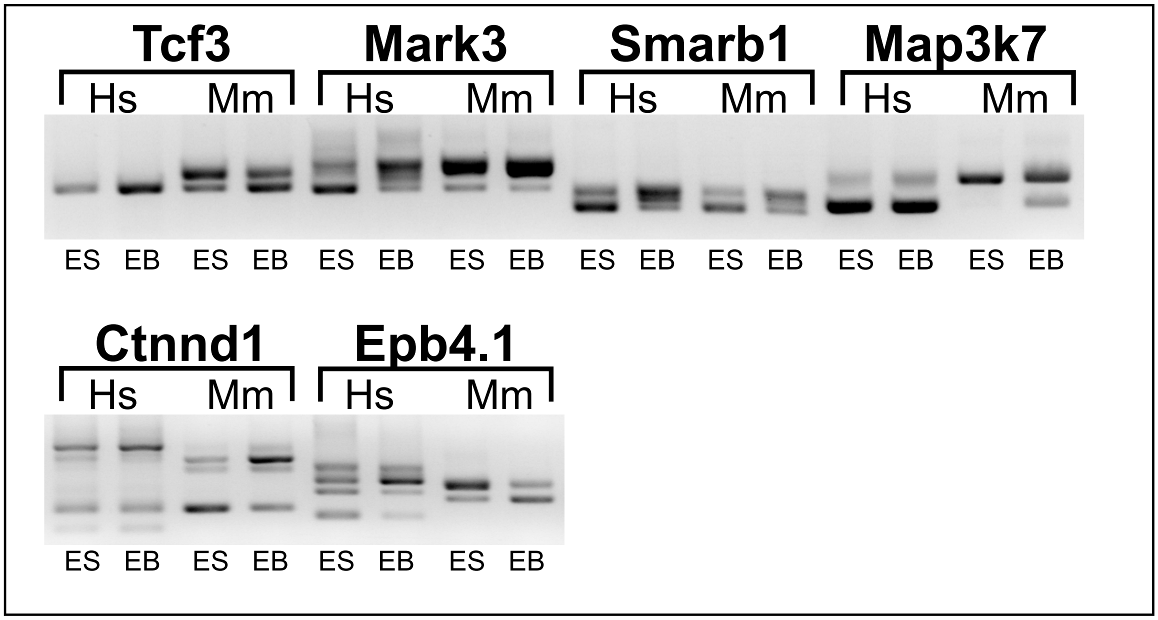
Tcf3 isoform expression was assessed with a Tcf3-specific antibody common to both isoforms. Analyzed proteins consisted of TCF3-/- GS1 ESCs expressing

Tcf3-short or Tcf3-long non-epitope tagged cDNAs, Tcf3 KD cell lines, and wild-type ESCs. Cells were plated and transfected as described above. Cell extracts were lysed with a buffer containing 20 mM Hepes pH 7.4, 1% Tx-100, 150 mM NaCl, 1 mM EDTA, 1 mM EGTA, 10mM sodium pyrophosphate, and protease inhibitors. Since Tcf3 expression is dependent on the efficiently of transfection, lysates were initially run on a small gel to adjust for expression and then run on a large 8% polyacrylamide gel, transferred to a nitrocellulose

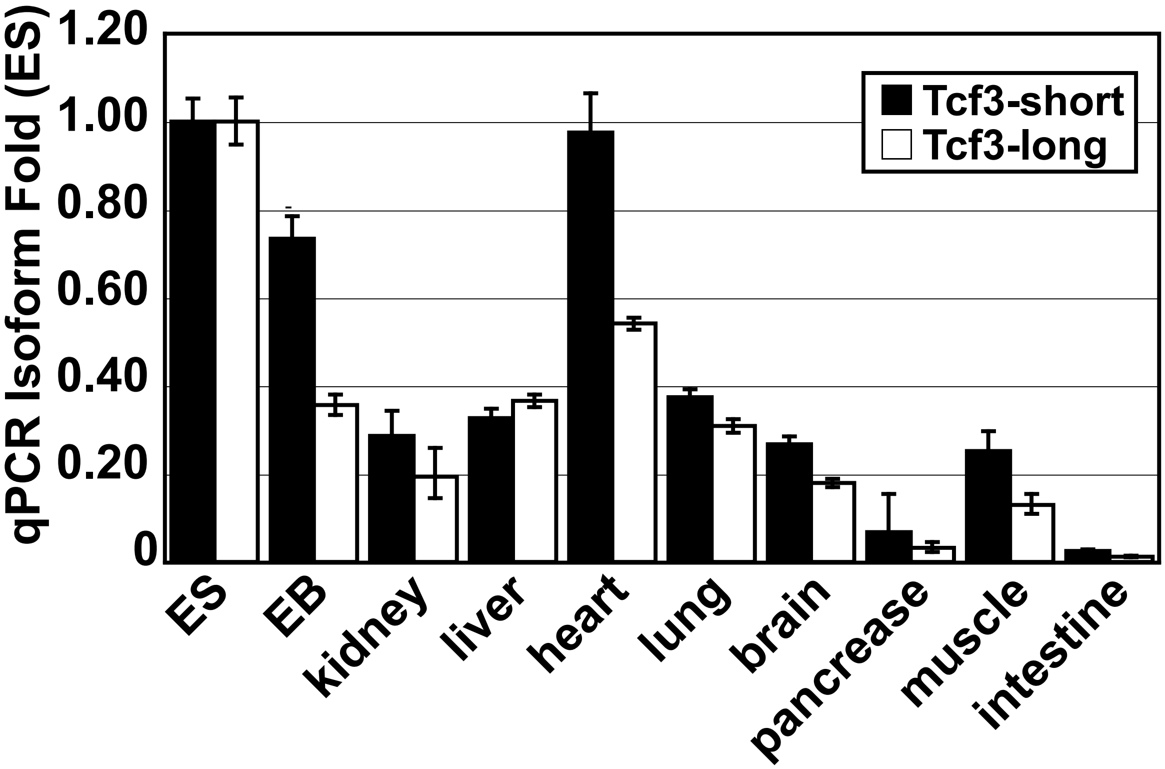
membrane, and probed with a rabbit anti-TCF3 antibody overnight at 4°C.

6.5.9 Teratoma assays

Adult SCID mice were injected with ~1x10^6 E14 wild-type or Tcf3-shRNA isoform ESC line at two adjacent sites on the lower mid-section (n=10 sites per line). Teratomas were observed for all cell lines and were harvested 4 weeks after injection.

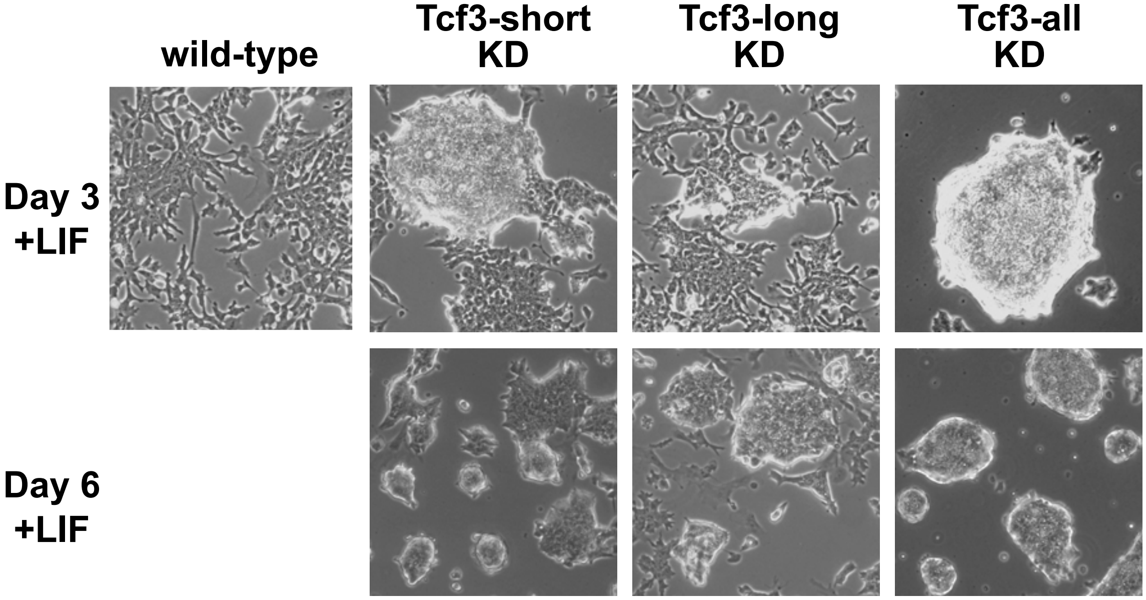


Supplemental Figure 6.1. Conservation of mouse splicing events to human. Several verified splicing events found in mouse (Mm) ESC differentiation by our microarray analysis were examined using orthologous primer sequences in human (Hs) H9 ESCs (ES) and derived EBs. Human RT-PCR products were verified by direct sequencing for Tcf3, Smarcb1, and Mark3.

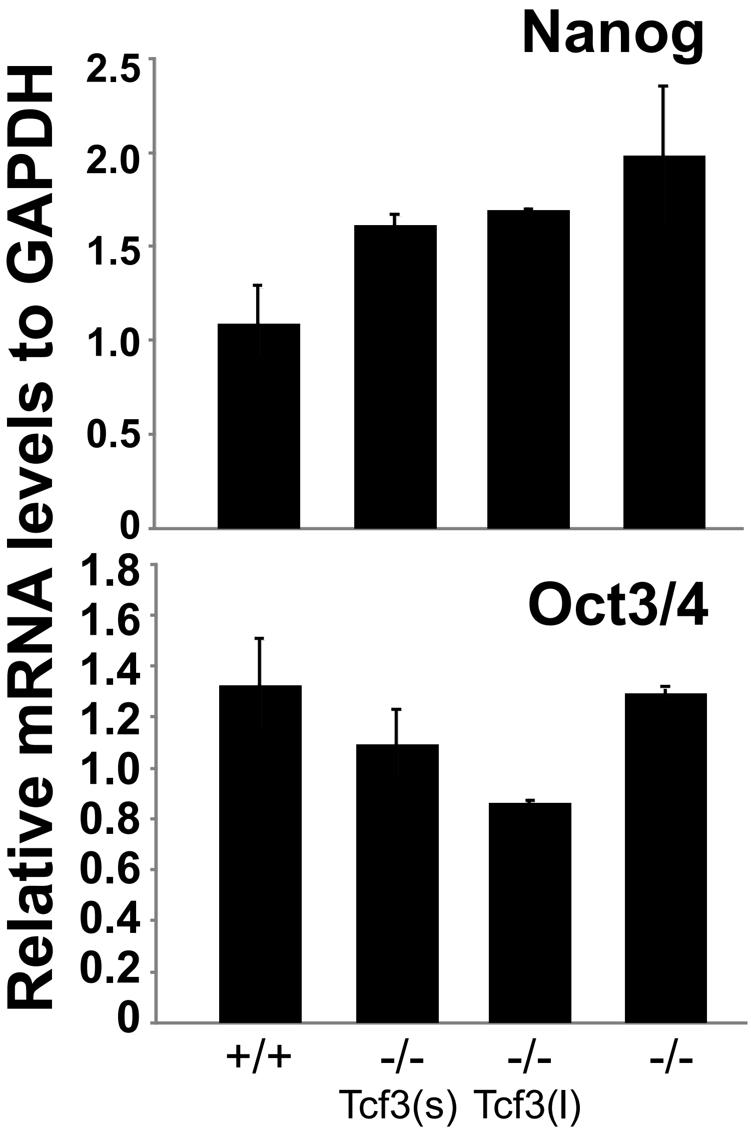


Supplemental Figure 6.2. Cross-tissue expression levels of Tcf3 isoforms. qPCR analysis of both Tcf3 isoforms relative to β-actin as an endogenous control in mouse E14 ESCs and EBs and adult mouse tissues harvested from adult FVBN mice.

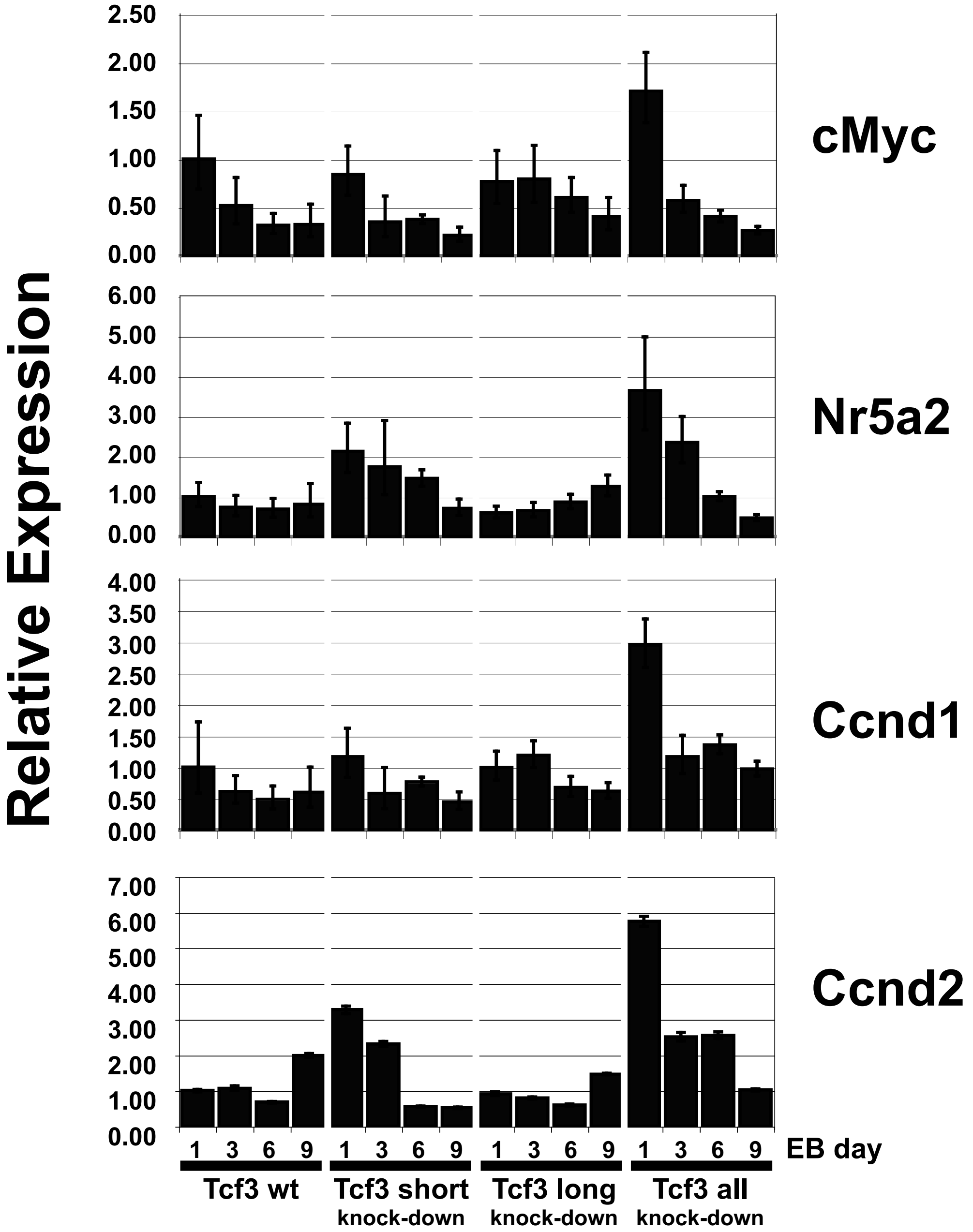
Salomonis_Supplemental_fig.3



Supplemental Figure 6.3. Morphology and growth of Tcf3 knockdown and wild-type cells in LIF containing conditions. See Figure 6.6.



Supplemental Figure 6.4. Oct4 and Nanog expression levels with Tcf3 isoform specific expression in Tcf3 null ESCs. Stable transfected GS1 ESCs with different Tcf3 isoform cDNAs on a Tcf3-/- background were generated and assessed for regulation of Nanog and Oct4 expression.



Supplemental Figure 6.5. Expression of Tcf3 putative transcriptional targets and markers of ESC maintenance with ESC differentiation. Multiple ESC marker genes and putative transcriptional targets of Tcf/Lef proteins were analyzed by qPCR for each for the KD cell lines and wild-type ESCs. See Figure 6.8.

6.6 Supplemental Datasets

Microarray CEL files will be deposited at GEO with acceptance of the corresponding manuscript. Additional analyzed dataset files, indicated as supplemental can be found at:

http://www.genmapp.org/supplemental/Salomonis_2008/mESC_junction/

6.7 References

Bachmann, M., H. Hennemann, P. X. Xing, I. Hoffmann and T. Moroy (2004). "The oncogenic serine/threonine kinase Pim-1 phosphorylates and inhibits the activity of Cdc25C-associated kinase 1 (C-TAK1): a novel role for Pim-1 at the G2/M cell cycle checkpoint." J Biol Chem **279**(46): 48319-28.

Biesiadecki, B. J., B. D. Elder, Z. B. Yu and J. P. Jin (2002). "Cardiac troponin T variants produced by aberrant splicing of multiple exons in animals with high instances of dilated cardiomyopathy." J Biol Chem **277**(52): 50275-85.

Boyer, L. A., T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, D. K. Gifford, D. A. Melton, R. Jaenisch and R. A. Young (2005). "Core transcriptional regulatory circuitry in human embryonic stem cells." Cell **122**(6): 947-56.

Burgess, R. W., Q. T. Nguyen, Y. J. Son, J. W. Lichtman and J. R. Sanes (1999). "Alternatively spliced isoforms of nerve- and muscle-derived agrin: their roles at the neuromuscular junction." Neuron **23**(1): 33-44.

Cavallo, R. A., R. T. Cox, M. M. Moline, J. Roose, G. A. Polevoy, H. Clevers, M. Peifer and A. Bejsovec (1998). "Drosophila Tcf and Groucho interact to repress Wingless signalling activity." Nature **395**(6702): 604-8.

Chu, P. J., J. K. Larsen, C. C. Chen and P. M. Best (2004). "Distribution and relative expression levels of calcium channel beta subunits within the chambers of the rat heart." J Mol Cell Cardiol **36**(3): 423-34.

Cole, M. F., S. E. Johnstone, J. J. Newman, M. H. Kagey and R. A. Young (2008). "Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells." Genes Dev **22**(6): 746-55.

Cooper, T. A. (2005). "Alternative splicing regulation impacts heart development." Cell **120**(1): 1-2.

Curley, M., J. J. Morrison and T. J. Smith (2004). "Analysis of Maxi-K alpha subunit splice variants in human myometrium." Reprod Biol Endocrinol **2**: 67.

Dabertrand, F., N. Fritz, J. Mironneau, N. Macrez and J. L. Morel (2007). "Role of RYR3 splice variants in calcium signaling in mouse nonpregnant and pregnant myometrium." Am J Physiol Cell Physiol **293**(3): C848-54.

Dugast-Darzacq, C., T. Grange and N. B. Schreiber-Agus (2007). "Differential effects of Mxi1-SRalpha and Mxi1-SRbeta in Myc antagonism." Febs J **274**(17): 4643-53.

Duursma, A. M., M. Kedde, M. Schrier, C. le Sage and R. Agami (2008). "miR-148 targets human DNMT3b protein coding region." Rna **14**(5): 872-7.

Griffiths-Jones, S., H. K. Saini, S. van Dongen and A. J. Enright (2008). "miRBase: tools for microRNA genomics." Nucleic Acids Res **36**(Database issue): D154-8.

Grskovic, M., C. Chaivorapol, A. Gaspar-Maia, H. Li and M. Ramalho-Santos (2007). "Systematic identification of cis-regulatory sequences active in mouse and human embryonic stem cells." PLoS Genet **3**(8): e145.

Hammes, A., J. K. Guo, G. Lutsch, J. R. Leheste, D. Landrock, U. Ziegler, M. C. Gubler and A. Schedl (2001). "Two splice variants of the Wilms' tumor 1 gene have distinct functions during sex determination and nephron formation." Cell **106**(3): 319-29.

http://bioconductor.org/CRAN/. "The Comprehensive R Archive Network." from http://bioconductor.org/CRAN/.

Huang, S. C., E. S. Liu, S. H. Chan, I. D. Munagala, H. T. Cho, R. Jagadeeswaran and E. J. Benz, Jr. (2005). "Mitotic regulation of protein 4.1R involves phosphorylation by cdc2 kinase." Mol Biol Cell **16**(1): 117-27.

Imbalzano, A. N. and S. N. Jones (2005). "Snf5 tumor suppressor couples chromatin remodeling, checkpoint control, and chromosomal stability." Cancer Cell **7**(4): 294-5.

Irizarry, R. A., B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs and T. P. Speed (2003). "Summaries of Affymetrix GeneChip probe level data." Nucleic Acids Res **31**(4): e15.

Ishitani, T., J. Ninomiya-Tsuji, S. Nagai, M. Nishita, M. Meneghini, N. Barker, M. Waterman, B. Bowerman, H. Clevers, H. Shibuya and K. Matsumoto (1999). "The TAK1-NLK-MAPK-related pathway antagonizes signalling between beta-catenin and transcription factor TCF." Nature **399**(6738): 798-802.

Ivey, K. N., A. Muth, J. Arnold, F. W. King, R. F. Yeh, J. E. Fish, E. C. Hsiao, R. J. Schwartz, B. R. Conklin, H. S. Bernstein and D. Srivastava (2008). "MicroRNA regulation of cell lineages in mouse and human embryonic stem cells." Cell Stem Cell **2**(3): 219-29.

Katoh, Y. and M. Katoh (2007). "Comparative genomics on PROM1 gene encoding stem cell marker CD133." Int J Mol Med **19**(6): 967-70.

Krek, A., D. Grun, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel and N. Rajewsky (2005). "Combinatorial microRNA target predictions." Nat Genet **37**(5): 495-500.

Kunarso, G., K. Y. Wong, L. W. Stanton and L. Lipovich (2008). "Detailed characterization of the mouse embryonic stem cell transcriptome reveals novel genes and intergenic splicing associated with pluripotency." BMC Genomics **9**: 155.

Merrill, B. J., H. A. Pasolli, L. Polak, M. Rendl, M. J. Garcia-Garcia, K. V. Anderson and E. Fuchs (2004). "Tcf3: a transcriptional regulator of axis induction in the early embryo." Development **131**(2): 263-74.

Miyabayashi, T., J. L. Teo, M. Yamamoto, M. McMillan, C. Nguyen and M. Kahn (2007). "Wnt/beta-catenin/CBP signaling maintains long-term murine embryonic stem cell pluripotency." Proc Natl Acad Sci U S A **104**(13): 5668-73.

Pereira, L., F. Yi and B. J. Merrill (2006). "Repression of Nanog gene transcription by Tcf3 limits embryonic stem cell self-renewal." Mol Cell Biol **26**(20): 7479-91.

Pritsker, M., T. T. Doniger, L. C. Kramer, S. E. Westcot and I. R. Lemischka (2005). "Diversification of stem cell molecular repertoire by alternative splicing." Proc Natl Acad Sci U S A **102**(40): 14290-5.

Redfern, C. H., M. Y. Degtyarev, A. T. Kwa, N. Salomonis, N. Cotte, T. Nanevicz, N. Fidelman, K. Desai, K. Vranizan, E. K. Lee, P. Coward, N. Shah, J. A. Warrington, G. I. Fishman, D. Bernstein, A. J. Baker and B. R. Conklin (2000). "Conditional expression of a Gi-coupled receptor causes ventricular conduction delay and a lethal cardiomyopathy." Proc Natl Acad Sci U S A **97**(9): 4826-31.

Roose, J., M. Molenaar, J. Peterson, J. Hurenkamp, H. Brantjes, P. Moerer, M. van de Wetering, O. Destree and H. Clevers (1998). "The Xenopus Wnt effector XTcf-3 interacts with Groucho-related transcriptional repressors." Nature **395**(6702): 608-12.

Sato, N. and A. H. Brivanlou (2006). "Manipulation of self-renewal in human embryonic stem cells through a novel pharmacological GSK-3 inhibitor." Methods Mol Biol **331**: 115-28.

Sato, N., T. Kawakami, A. Nakayama, H. Suzuki, H. Kasahara and T. Obinata (2003). "A novel variant of cardiac myosin-binding protein-C that is unable to assemble into sarcomeres is expressed in the aged mouse atrium." Mol Biol Cell **14**(8): 3180-91.

Spring, C. M., K. F. Kelly, I. O'Kelly, M. Graham, H. C. Crawford and J. M. Daniel (2005). "The catenin p120ctn inhibits Kaiso-mediated transcriptional repression of the beta-catenin/TCF target gene matrilysin." Exp Cell Res **305**(2): 253-65.

Sugnet, C. W., K. Srinivasan, T. A. Clark, G. O'Brien, M. S. Cline, H. Wang, A. Williams, D. Kulp, J. E. Blume, D. Haussler and M. Ares, Jr. (2006). "Unusual intron conservation near tissue-regulated exons found by splicing microarrays." PLoS Comput Biol **2**(1): e4.

Takao, Y., T. Yokota and H. Koide (2007). "Beta-catenin up-regulates Nanog expression through interaction with Oct-3/4 in embryonic stem cells." Biochem Biophys Res Commun **353**(3): 699-705.

Tam, W. L., C. Y. Lim, J. Han, J. Zhang, Y. S. Ang, H. H. Ng, H. Yang and B. Lim (2008). "Tcf3 Regulates Embryonic Stem Cell Pluripotency and Self-Renewal by the Transcriptional Control of Multiple Lineage Pathways." Stem Cells.

Tyson-Capper, A. J. (2007). "Alternative splicing: an important mechanism for myometrial gene regulation that can be manipulated to target specific genes associated with preterm labour." BMC Pregnancy Childbirth **7 Suppl 1**: S13.

Ule, J., A. Ule, J. Spencer, A. Williams, J. S. Hu, M. Cline, H. Wang, T. Clark, C. Fraser, M. Ruggiu, B. R. Zeeberg, D. Kane, J. N. Weinstein, J. Blume and R. B. Darnell (2005). "Nova regulates brain-specific splicing to shape the synapse." Nat Genet **37**(8): 844-52.

Vandesompele, J., K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, A. De Paepe and F. Speleman (2002). "Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes." Genome Biol **3**(7): RESEARCH0034.

Veeman, M. T., D. C. Slusarski, A. Kaykas, S. H. Louie and R. T. Moon (2003). "Zebrafish prickle, a modulator of noncanonical Wnt/Fz signaling, regulates gastrulation movements." Curr Biol **13**(8): 680-5.

Xu, X., D. Yang, J. H. Ding, W. Wang, P. H. Chu, N. D. Dalton, H. Y. Wang, J. R. Bermingham, Jr., Z. Ye, F. Liu, M. G. Rosenfeld, J. L. Manley, J. Ross, Jr., J. Chen, R. P. Xiao, H. Cheng and X. D. Fu (2005). "ASF/SF2-regulated CaMKIIdelta alternative splicing temporally reprograms excitation-contraction coupling in cardiac muscle." Cell **120**(1): 59-72.

Yamanaka, S. (2008). "Pluripotency and nuclear reprogramming." Philos Trans R Soc Lond B Biol Sci **363**(1500): 2079-87.

Yeo, G. W., X. Xu, T. Y. Liang, A. R. Muotri, C. T. Carson, N. G. Coufal and F. H. Gage (2007). "Alternative splicing events identified in human embryonic stem cells and neural progenitors." PLoS Comput Biol **3**(10): 1951-67.

Yi, F., L. Pereira and B. J. Merrill (2008). "Tcf3 Functions as a Steady State Limiter of Transcriptional Programs of Mouse Embryonic Stem Cell Self Renewal." Stem Cells.

Chapter 7
Consequences of Alternative Splicing in the Myometrium Throughout Gestation

7.1 Introduction

In the previous chapter we examined the alternative regulation of transcripts in multiple systems, including embryonic stem cell (ESC) differentiation, cardiomyopathy, and gestational induced remodeling of the myometrium. From these analyses, we found that the myometrium throughout gestation was predicted to undergo substantial alternative transcript regulation and thus may be an interesting model system to explore the role of alternative splicing (AS) in the regulation of uterine contractile responses of during gestation.

A number of recent studies have shown that distinct protein isoforms are expressed in the myometrium at the interface between contractile quiescence and the induction of labor. These proteins are largely associated with contractile regulation and include the calcium-activated potassium channel Kcnma1, the adenylyl-cyclase activating G-protein Gnas, the cyclic-AMP dependent transcription factor CREM and the transient receptor potential cation channel 4 (Europe-Finner *et al.* 1997; Bailey *et al.* 2000; Benkusky *et al.* 2000; Yang *et al.* 2002). In addition to these proteins, altered expression of the antagonistic splicing factors hnRNP A1 and ASF-SF2 have also been observed in concert with regulation of these myometrial AS events (Pollard *et al.* 2000). Gene expression studies by myself and other members of the Conklin laboratory, suggest that a number of additional mRNA processing regulators may be regulated throughout pregnancy, such as the ASF-SF2 kinase and Clk1, which are down-regulated with uterine quiescence, but not at term pregnancy (Salomonis *et al.* 2005).

In this supplemental analysis, I have performed a relatively detailed study of bioinformatics predictions at the level of AS, downstream functional predictions, and confirmation of AS in the mouse myometrium throughout gestation and postpartum. These results highlight both global and gene level changes by AS predicted to have functional consequences on the transition from uterine quiescence to activation at term.

7.2 Results

7.2.1 Patterns of myometrial AS are similar to those for transcription

To assess the relative extent of alternative exon regulation in the myometrium throughout gestation and postpartum, each of the examined time-points (14.5 and 18.5 days gestation or 6 hr postpartum) were individually compared to non-pregnant mouse myometrium samples in AltAnalyze. This analysis identified 735 unique AEEs corresponding to these comparisons. More than half of these unique exon-exclusion junctions (356) were associated with annotated AS events. Clustering of linear regression folds for these comparisons produced several clusters corresponding to AS events with quiescence, term, and postpartum specific patterns (Figure 7.1). These patterns match those identified in a similar time-course analysis of gene expression in the myometrium throughout gestation, corresponding to “quiescence”, “activation” and “involution” regulated events (Figure 3.1).

7.2.2 Global protein and microRNA functional predictions of “activation” regulated AEEs

A large set of the clustered AS predictions correspond to the activation phase of gestation (regulated selectively at term). Since evidence exists linking AS to transcript and splicing factor regulation with the transition from quiescence to term, we focused on AS predictions from this direct comparison for the remainder of this analysis.

436 unique exon-exclusion junctions were regulated when comparing myometrium at 18.5 days of gestation to 14.5 days of gestation, with the previously established thresholds and algorithms (chapter 6). These results were initially not filtered based on external AS or APS annotations. Examination of functional predictions from the AltAnalyze program (chapter 6) revealed that a large fraction of these AEEs corresponded to predicted changes in protein length (84.4%) and a small fraction that result in altered inclusion of predicted microRNA binding sites (5.5%). Only 40% of these AEEs, however, linked to annotated splicing events (162 AEEs).

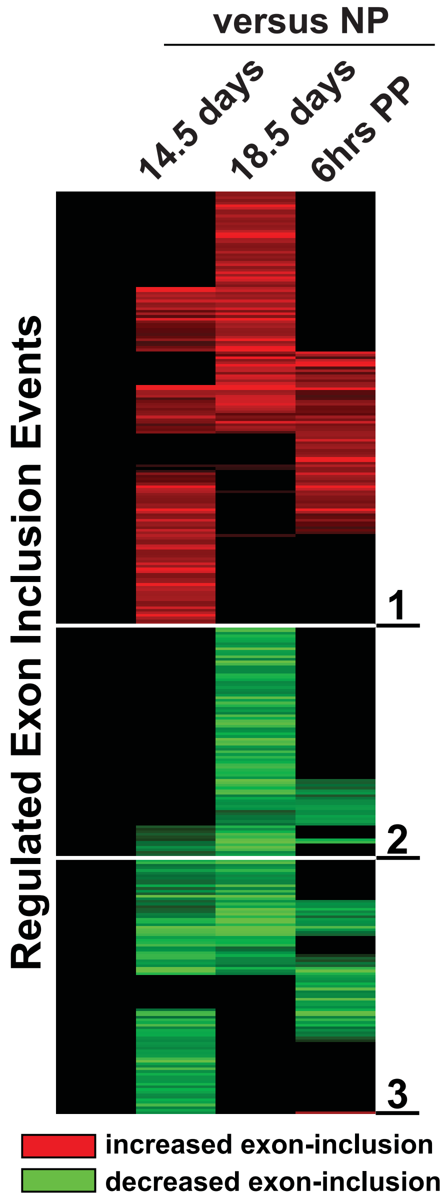


Figure 7.1. Expression clustering of myometrial AS events. Linear regression fold changes for each myometrium time-point relative to non-pregnant animals were clustered using the program HOPACH (chapter 3). All linear regression folds are displayed for 735 unique exclusion-junctions that had evidence of AS and at least a two-fold change in reciprocal isoform expression for any of the comparisons. Three clusters were formed by HOPACH (1-3).

To determine if there are global differences in protein size among alternatively regulated genes, AltAnalyze calculates a mean, median, and standard deviation in protein length for all AEEs, comparing the protein isoforms that are up- and down-regulated in a given comparison. Applying this analysis to all comparisons studied (myometrium and non-myometrium) reveals that isoforms up-regulated in term myometrium have a greater tendency to be longer than those that are down-regulated. This result was true when term myometrium was compared to virgin or 14.5 days of gestation, resulting in a median fold change difference of > 4, mean fold change > 2 and a t-test p= $1.1\text{x}10^{-16}$ in both of these term comparisons, when comparing all up- versus down-regulated isoform protein lengths (Figure 7.2). We next checked to see whether this pattern held up when only analyzing AEEs linked to AS events for term vs. quiescence. Although up-regulated isoforms were still longer on average than down-regulated isoforms, with an equivalent p-value (p=$7.9\text{x}10^{-11}$), the fold difference was diminished, with a mean fold ~ 1.6 fold and median fold ~ 1.7. Thus, the large observed difference in predicted protein lengths may not be due to AS.

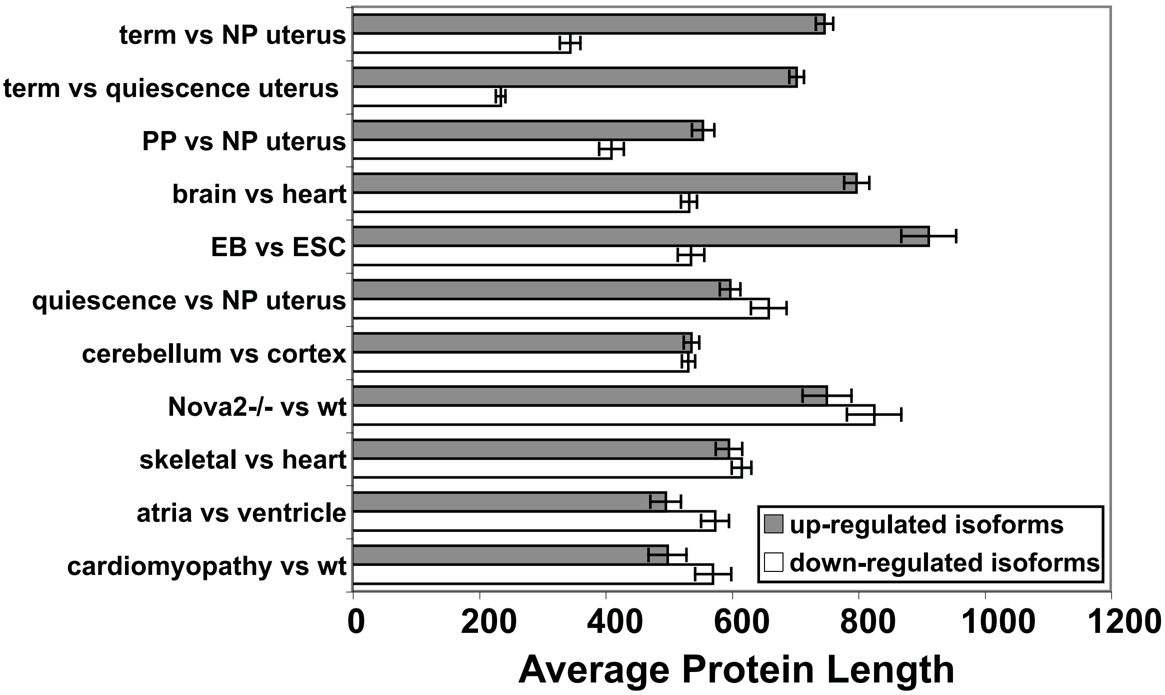


Figure 7.2. Overall differences in protein length between up- and down-regulated isoforms. Inferred proteins from AltAnalyze for the reciprocal up- and down-regulated isoforms were stored in separate lists within this program to determine a mean, standard deviation, and median protein length for all AEEs analyzed. From these data, relative fold changes and t-test p values were derived by comparing the up- and down-regulated lists. Values are mean ± SEM, where the population is the number of unique AEE associated genes.

When protein functional elements (e.g., protein domains) that are over-represented according to AltAnalyze were examined among these AS genes, 4 elements were highlighted: ERM and RNA recognition regions and zinc binding and modified phosphoserine residues, each of which were alternatively regulated in at least 3 unique genes with a z score > 1.96 (Table 7.1). These data suggest that a common set of pathways may be impacted by AS for specific protein/RNA interactions.

7.2.3 Regulation of cytoskeletal/cell-matrix interactions, contraction, and splicing control by AS

To identify biological processes that are impacted by AS in the myometrium with the switch to term gestation, we performed Gene Ontology over-representation analysis with the program GO-Elite (chapter 4). Genes predicted to be regulated by AS for the term-activation comparison corresponded to a number of pathways, including developmental (osteoblast differentiation, synaptogenesis, and embryonic morphogenesis), cell interaction (integrin-mediated signaling, actin-binding, cell-cell junction, and cell matrix adhesion), muscle development and the regulation of splicing (Table 7.2). Interestingly, several of these pathways overlap with those identified from gene expression analyses of these same time-points, suggesting that AS may complement gene expression by regulating distinct components of the same pathways by alternative mechanisms.

Functional Element	***Z Score***	***Unique Gene Count***	***Unique Denominator Gene Count***	***Gene Symbols***
Ez/rad/moesin-IPR000798	5.90	3	18	Epb4.1l2, Epb4.1l3, Rdx
METAL-Zinc	3.89	4	58	Dnpep, Ppp3ca, Ide, Mobk1b
RRM_1-IPR003954	2.99	3	51	Hnrpc, Raly, Tial1
MOD_RES-Phosphoserine	2.20	29	1613	Epb4.1l2, Catnb, Cnot2, Snx3, Eprs, Tmpo, AA536749, Cbx1, Hnrpc, Atp5a1, Oxr1, Ehf, 2610024N24Rik, Ddb1, Epb4.1l3, Tpd52l2, Pex2, Sorbs1, Mef2d, Ptpn1, Ppfibp1, Zfp162, Mlf2, Pde3a, Hnrpa2b1, Rad18, Bcl7b, Arhgap17, Atp13a

Table 7.1. Over-represented functional elements associated with "activation" AS. Protein function elements highlighted by AltAnalyze for myometrium term versus mid-pregnancy AS genes. Only regulated functional elements with 3 or more unique associated genes and an over-representation z score > 1.96 are reported.

GO Name	GO Type	Changed/ Measured	Z Score	Permute P	Gene Symbols
osteoblast differentiation	P	4/13	7.04	0	Cbfb|Ctnnb1|Gnas|Mef2d
synaptogenesis	P	3/9	6.38	0	Agrn|Cadm1|Nrg1
integrin binding	F	3/10	6.01	0.001	Mfge8|Npnt|Spp1
cell-cell adherens junction	C	4/20	5.44	0.0015	Ctnnb1|Dlg1|Dlg5|Sorbs1
microvillus	C	3/12	5.40	0.0035	Ctnnb1|Dcxr|Rdx
chondrocyte differentiation	P	3/13	5.14	0.0025	Ctnnb1|Fgfr1|Mef2d
cell-matrix adhesion	P	4/26	4.60	0.0015	Ctnnb1|Npnt|Sorbs1|Spp1
extrinsic to membrane	C	5/39	4.54	0.0015	Epb4.1l2|Epb4.1l3|Gnas|Mfge8|Rdx
regulation of myeloid cell differentiation	P	3/17	4.35	0.005	Ctnnb1|Gnas|Spp1
apical part of cell	C	4/36	3.66	0.0035	Ctnnb1|Inadl|Rdx|Spp1
spliceosome	C	5/56	3.46	0.0045	Hnrpa2b1|Hnrpc|Raly|Sf1|Srrm1
muscle fiber development	P	3/25	3.35	0.013	Agrn|Myocd|Ppp3ca
regulation of membrane potential	P	3/25	3.35	0.013	Dlg1|Pex2|Ppp3ca
actin binding	F	8/122	3.32	0.004	AA536749|Cald1|Epb4.1l2|Epb4.1l3| Myo1d|Phactr4|Rdx|Scin
basolateral plasma membrane	C	4/42	3.25	0.0095	Cadm1|Ctnnb1|Dlg1|Sorbs1
embryonic appendage morphogenesis	P	3/27	3.17	0.0215	Ctnnb1|Fgfr1|Gnas
branching morphogenesis of a tube	P	3/29	3.00	0.022	Ctnnb1|Fgfr1|Npnt
transcription cofactor activity	F	5/68	2.92	0.0165	Cbfb|Cited2|Ctnnb1|Myocd|Nrg1
translation factor activity, nucleic acid binding	F	4/48	2.91	0.0185	Eif3h|Eif3k|Eif4a2|Tcea1

Table 7.2. Pathway analysis of AS with myometrial "activation". Gene Ontology (GO) terms highlighted by analysis with the program GO-Elite that were predicted to undergo AS in term vs. quiescent myometrium. GO terms are ranked according to z score and genes with predicted AS are listed under "Gene Symbols".

7.2.4 Confirmation of predicted myometrial "activation" AS events

To verify AS events regulated with term-activation, we used the RT-PCR based strategy described for confirmation of ESC differentiation AEEs in chapter 6. Of 12 predicted AS events, five genes were considered confirmed by this method (Vldlr, Cald1, Pde3a, Hnrpa2b1, and 5730555F13Rik (Modulator of estrogen induced transcription)), whereas three produced bands that did not match the predicted sizes, two produced only one of the two predicted bands and two produced the predicted bands but did not display a clear isoform shift (Figure 7.3 A). Thus, more than half (5 out of 9) of the predicted AS events were confirmed within this set.

Among the confirmed AS events, two new predictions represent novel findings with potential implications for uterine activation at term: Cald1 and Vldlr. Cald1 or caldesmon, is an actin-binding protein that encodes for isoforms with either a smooth muscle (h) or non-muscle (l) expression pattern. The l isoform is shorter with a premature splice site in exon 3 that results in a loss of 234 amino acids, corresponding to a spacer sequence of unknown function (Guo *et al.* 2005). Cald1 inhibits the binding of actin to myosin and thus reversibly inhibits contraction. While functional differences between these two isoforms have not been clearly elucidated, we observe down-regulation of the smooth-muscle isoform in term myometrium. This result is intriguing given that up-regulation of h-Cald1 in the gestational myometrium has been associated with suppression of coordinated contractions (Li *et al.* 2003). Although the spliced-out Cald1 region is not annotated as such in the literature, AltAnalyze predicts removal of the

Ensembl annotated tropomyosin domain, which is only present in the h-Cald1 form (Figure 7.3 B).

One of the highest scoring AltAnalyze AEEs from our analysis and most robust verified changes was the AS of Vldlr. Vldlr or very-low-density lipoprotein receptor is a liproprotein receptor that binds to and internalizes triacylglycerol-rich apo-E containing lipoproteins, such as VLDL (Oka *et al.* 1994). AS was predicted by AltAnalyze to down-regulate the longer isoform of Vldlr at term, associated with a protein that gains a cassette-exon and as a result, 28 amino acids, compared to the alternative shorter isoform. Similar to Cald1, the term down-regulated form of Vldlr is the muscle enriched form of the protein (Iijima *et al.* 1998) and was highly enriched during quiescence, but is expressed at similar expression levels to the shorter isoform at term. The spliced-out exon is associated with elimination of a UniProt annotated clustered O-linked oligosaccharide region, according to AltAnalyze, which has been functionally verified from biochemical studies *in vitro* (Iijima *et al.* 1998). Compared to the quiescence enriched longer isoform, the short isoform of Vldlr also undergoes rapid degradation and proteolytic cleavage *in vitro*. Interestingly, the C-terminal regulated cassette-exon contained several microRNA binding site predictions by the algorithm miRNADA (chapter 6), raising the possibility that AS may decrease the likelihood of translational inhibition of this protein at term (Figure 7.3 C).

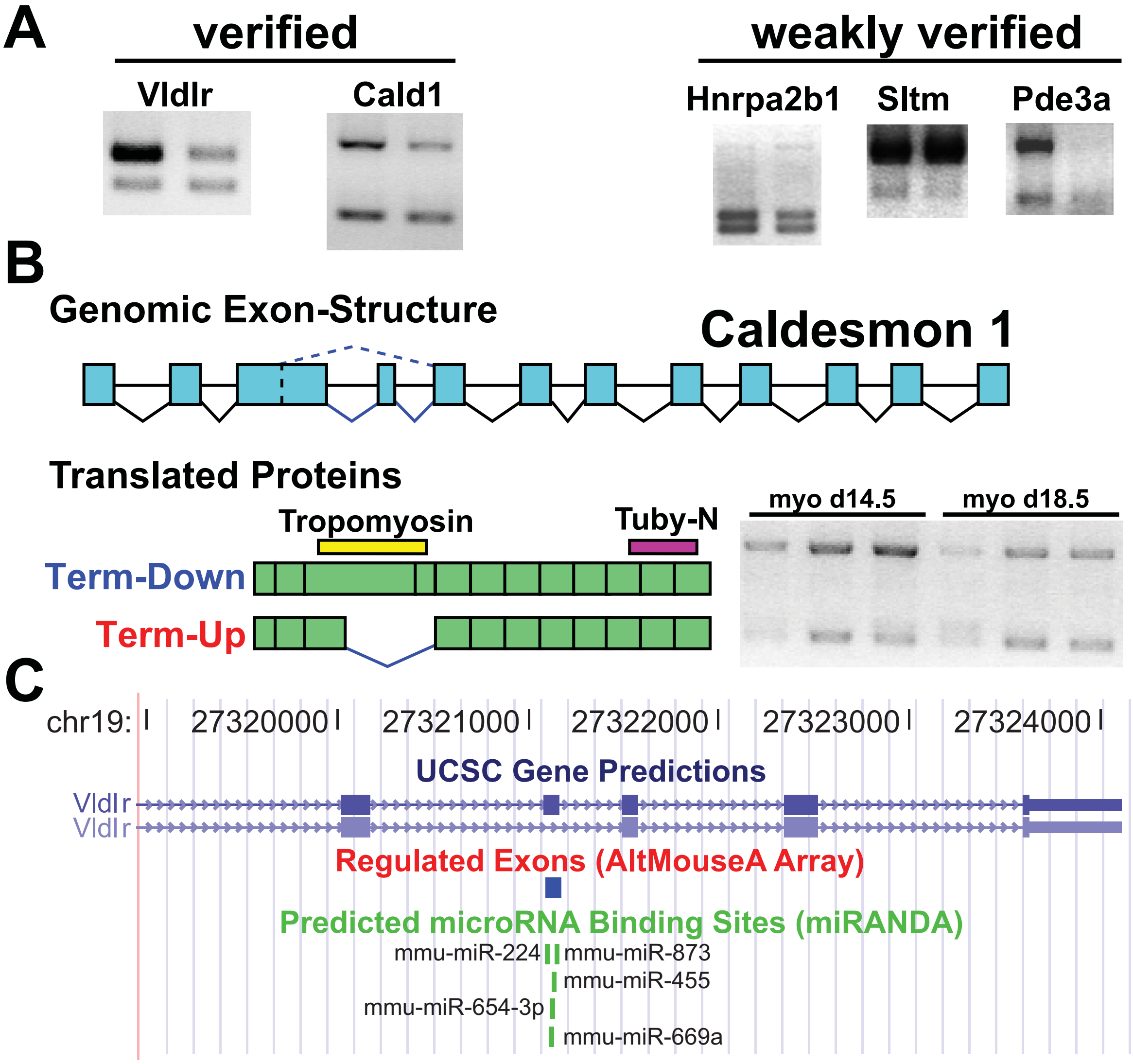


Figure 7.3. Functional analysis of validated "activation" AS. (A) Isoforms with verified AS expression patterns with readily observable and subtle (weak) shifts in isoform expression. (B) AS of the Cald1 gene for exons oriented 5' to 3' (not to scale). Below this exon-structure are depictions of the corresponding proteins derived from AS, where each segment corresponds to an exon in the exon-structure. Above the translated protein diagrams are AltAnalyze predicted protein functional elements as determine by Ensembl and UniProt. To the right of this graph is an RT-PCR image of biological triplicates for quiescence (14.5 days) or

term (18.5 days) myometrial products. (C) UCSC genome browser display of Vldlr isoforms (UCSC gene predictions), down-regulated AltMouse exons, and predicted microRNA binding sites.

7.3 Discussion

The mechanisms regulating the remodeling of uterus prior to the onset of labor still remain largely unknown. Whole-genome microarray studies profiling gene expression changes within the myometrium throughput gestation and postpartum have provided improved insights into possible mechanisms that contribute to both quiescence and term activation. One of the primary results from such studies has been the coordinate regulation of splicing factors throughout gestation (chapter 3). These results complement other studies that demonstrate AS of signaling and transcriptional components with this physiological transition, such as Kcnma1 and Crem.

In the current study, we find that AS is regulated on a large scale in the myometrium throughout gestation and postpartum, impacting over 300 predicted proteins with temporal patterns mimicking those found by conventional microarray studies. Focused analysis of variants regulated with the switch from mid-gestation to term, highlights many of the same pathways identified from these conventional microarray studies, including interactions between extracellular matrix and smooth muscle and cell-cell junction signaling. Interestingly, genes predicted to undergo AS tended to produce longer protein associated variants at term relative to mid-pregnancy. If valid, such changes might impact the expression of such transcripts (if targeted by non-sense mediated decay) or composition in ways that will significantly modify the functional capacity of associated signaling pathways and cellular structural components.

Analysis of a small set of predicted splice variants by RT-PCR indicated that 50% or greater of the predicted variants could be readily confirmed by this method. Two of the most interesting results, splicing of Cald1 and Vldlr, were uncovered from this analysis, each with potential roles in the regulation of uterine contraction. Cald1 acts to suppress coordinate contractions while Vldlr regulates the uptake of Vldl triglycerides and thus impacts lipid metabolism. At term, the secretion of hormones such as oxytocin and prostaglandins stimulates the uterus through activation of down-stream G-protein coupled receptor signaling pathways. Prostaglandins are bioactive lipids produced from arachadonic acid in the cell. Although the role of Vldlr on the regulation of myometrial lipid metabolism has not been studied, our analysis shows that two distinct isoforms of this gene are highly regulated with the transition to term pregnancy. Interestingly, the term down-regulated variant of Vldlr has decreased recycling and secretion kinetics and binds to clustered O-linked oligosaccharides as compared to the down-regulated isoform (Iijima *et al.* 1998). Furthermore, we found a number of predicted microRNA binding sites in the mid-pregnancy enriched form, which may alter the ability of this protein to be regulated by microRNAs.

While both Cald1 and Vldlr genes participate in pathways important for term activation (contractile regulation or lipid metabolism), further study is required to determine whether these splicing changes contribute to the physiology of the uterus or are secondary changes due to splicing factor regulation. Furthermore, more extensive validation of AS events in these myometrial comparisons is required to determine our accuracy of these bioinformatics predictions and thus the functional significance of verified changes. However, such studies provide a useful starting point to understand the global contribution of AS on protein content in the remodeling uterus.

7.4 References

Bailey, J., C. Sparey, R. J. Phillips, K. Gilmore, S. C. Robson, W. Dunlop and G. N. Europe-Finner (2000). "Expression of the cyclic AMP-dependent transcription factors, CREB, CREM and ATF2, in the human myometrium during pregnancy and labour." Mol Hum Reprod **6**(7): 648-60.

Benkusky, N. A., D. J. Fergus, T. M. Zucchero and S. K. England (2000). "Regulation of the Ca2+-sensitive domains of the maxi-K channel in the mouse myometrium during gestation." J Biol Chem **275**(36): 27712-9.

Europe-Finner, G. N., S. Phaneuf, E. Cartwright, H. J. Mardon and A. Lopez Bernal (1997). "Expression of human myometrial G alpha s messenger ribonucleic acid transcript during pregnancy and labour: involvement of alternative splicing pathways." J Mol Endocrinol **18**(1): 15-25.

Guo, H. and C. L. Wang (2005). "Specific disruption of smooth muscle caldesmon expression in mice." Biochem Biophys Res Commun **330**(4): 1132-7.

Iijima, H., M. Miyazawa, J. Sakai, K. Magoori, M. R. Ito, H. Suzuki, M. Nose, Y. Kawarabayasi and T. T. Yamamoto (1998). "Expression and characterization of a very low density lipoprotein receptor variant lacking the O-linked sugar region generated by alternative splicing." J Biochem **124**(4): 747-55.

Li, Y., H. D. Je, S. Malek and K. G. Morgan (2003). "ERK1/2-mediated phosphorylation of myometrial caldesmon during pregnancy and labor." Am J Physiol Regul Integr Comp Physiol **284**(1): R192-9.

Oka, K., K. Ishimura-Oka, M. J. Chu, M. Sullivan, J. Krushkal, W. H. Li and L. Chan (1994). "Mouse very-low-density-lipoprotein receptor (VLDLR) cDNA cloning, tissue-specific expression and evolutionary relationship with the low-density-lipoprotein receptor." Eur J Biochem **224**(3): 975-82.

Pollard, A. J., C. Sparey, S. C. Robson, A. R. Krainer and G. N. Europe-Finner (2000). "Spatio-temporal expression of the trans-acting splicing factors

SF2/ASF and heterogeneous ribonuclear proteins A1/A1B in the myometrium of the pregnant human uterus: a molecular mechanism for regulating regional protein isoform expression in vivo." J. Clin. Endocrinol. Metab. **85**(5): 1928-1936.

Salomonis, N., N. Cotte, A. C. Zambon, K. S. Pollard, K. Vranizan, S. W. Doniger, G. Dolganov and B. R. Conklin (2005). "Identifying genetic networks underlying myometrial transition to labor." Genome Biol **6**(2): R12.

Yang, M., A. Gupta, S. G. Shlykov, R. Corrigan, S. Tsujimoto and B. M. Sanborn (2002). "Multiple Trp isoforms implicated in capacitative calcium entry are expressed in human pregnant myometrium and myometrial cells." Biol Reprod **67**(3): 988-94.

Chapter 8
Summary and Discussion

8.1 Using a systems biology approach to address fundamental biological questions

In this dissertation, I have utilized new genomic technologies as a means to assess two discrete biological questions: (1) identifying the role of global coordinated transcriptional effects with muscle transformation; and (2) building functional correlations to secondary processing events regulating lineage commitment in embryonic stem cells (ESCs).   On the surface, these are very broad questions that may seem to have little relevance to each other. However, when the data obtained from these whole genome experiments is integrated with existing biological knowledge bases (pathways, protein domains, microRNA binding sites, and chromosomal co-localization), we can begin to assess both molecular as well as systems level similarities and differences.

By interrogating multiple time-points, each with physiologically distinct profiles in the myometrium as it transitions through gestation, we can gauge changes in the regulation of specific pathway components over time. This approach is more informative than assessing a single “snap-shot”, which would only include gene expression changes critical to a specific moment (chapter 3). Likewise, analysis of secondary regulatory transcript mechanisms, such as alternative splicing (AS), within the same computational framework, allows us to consider multiple variables that can contribute to gene expression and ultimately protein composition.

Examination of splicing changes that were either unique to differentiating cardiac precursors or shared between distinct developmental programs, led to the identification of a host of factors expressed along pathways critical to these developmental programs. Combining these microarray predictions with additional datasets, including protein domain-level changes and microRNA binding site occurrence, allowed us to gauge novel functional roles for these AS-regulated genes.

To further demonstrate that such predictions could provide new insights into developmental processes, we carried out a detailed study of a single factor, Tcf3, highlighted by the analysis of mouse ESC differentiation AS microarray data. As a result, we find that distinct splice isoforms can have diverse functional roles, with crucial consequences for development.

8.2 Correlating signaling and physiology to genome-wide transcriptional and splicing profiles

The myometrium, as it transitions through gestation to labor and then to postpartum, represents a valuable model system to assess changes in the composition and signaling of a tissue over time. In our analysis of myometrial transcription throughout gestation, we identified coordinated transcriptional events that correspond to each of the major signaling phases during gestation:

quiescence, activation, and involution (Figure 8.1). Gene expression changes localized to these different phases correspond to distinct biological pathways, each providing novel insights into molecular events that are critical to regulate contractile signaling throughout pregnancy as well as transform, maintain, and ultimately digest components in and around uterine myocytes.

During the quiescence phase of gestation, we observe the coordinated up-regulation of serine proteases (granzymes B-G), hypothesized to mediate extracellular matrix (ECM)/cell interactions, focal adhesion/integrin-mediated signaling, and cell growth pathways (Figure 8.2 A). Also observed is the up-regulation of hormone signaling components (GPCRs, GPCR ligands, transcription factors), which participate in contractile relaxation pathways in uterine myocytes (e.g., Gαs or cAMP stimulatory signaling). These changes are accompanied by the up-regulation of regulators of metabolic signaling, such as glucocorticoid metabolism and prostaglandin signaling, highlighting both previously established (Cyp11a, Hsd11b2) as well as relatively novel components (phospholipase A2 inhibitors).

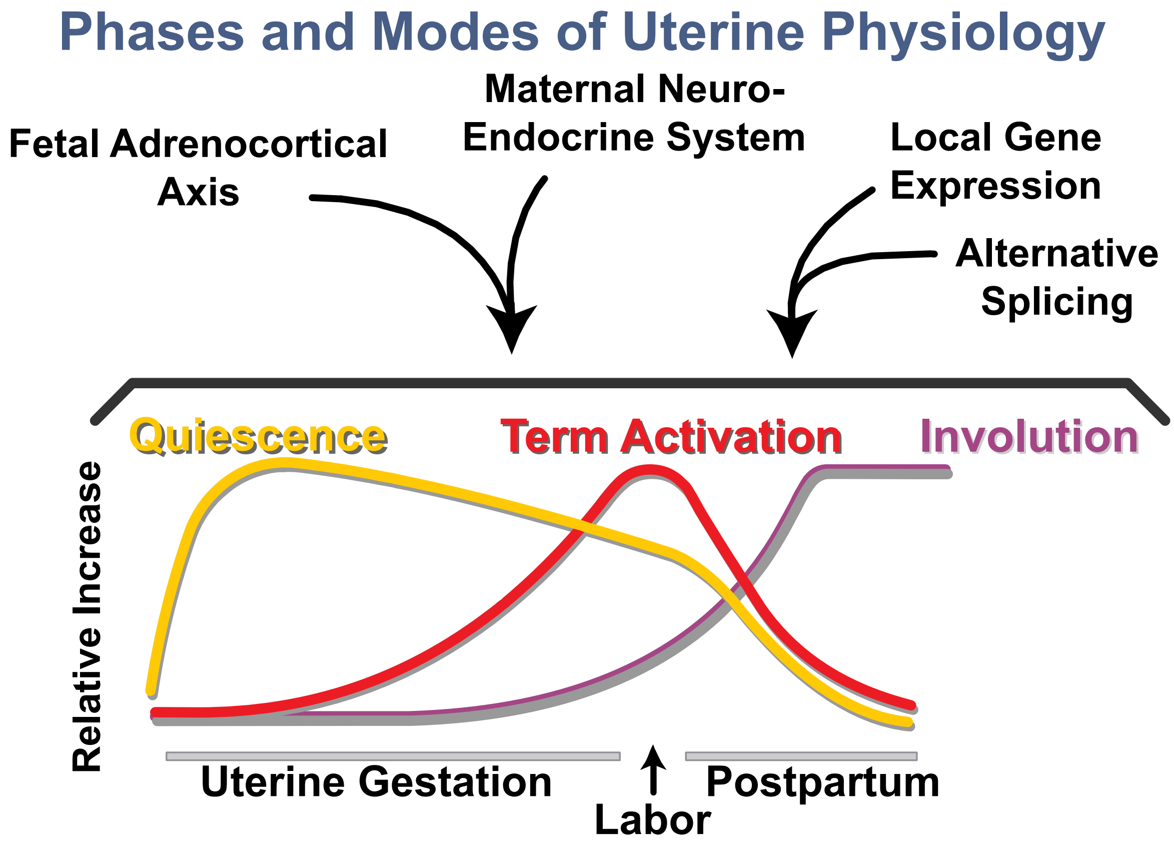
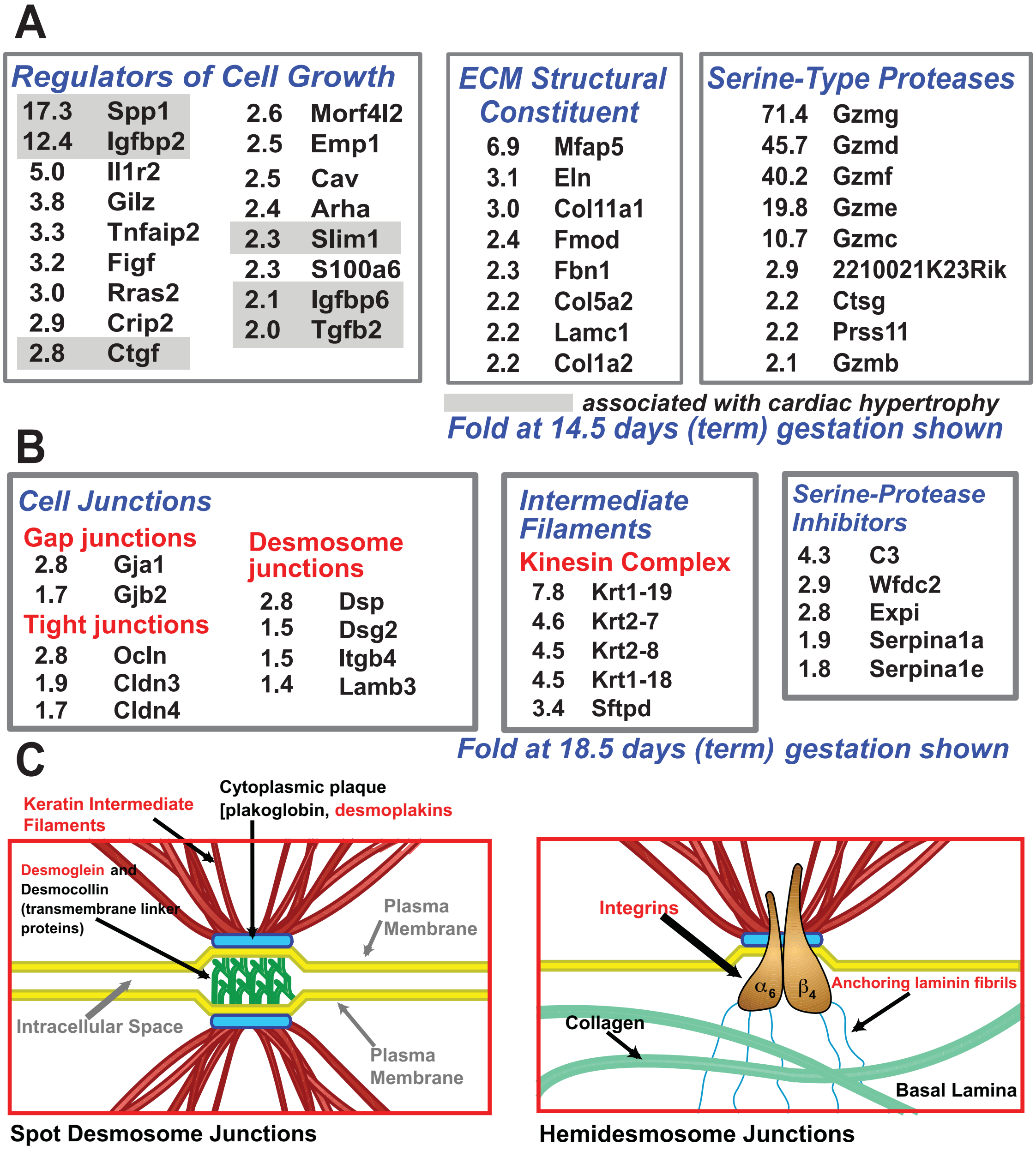


Figure 8.1. Phases of uterine gestation. An illustration of the phases of mouse uterine gestation with intrinsic and extrinsic regulatory factors.



A

Regulators of Cell Growth			
17.3	Spp1 ▨	2.6	Morf4l2
12.4	Igfbp2 ▨	2.5	Emp1
5.0	Il1r2	2.5	Cav
3.8	Gilz	2.4	Arha
3.3	Tnfaip2	2.3	Slim1 ▨
3.2	Figf	2.3	S100a6
3.0	Rras2	2.1	Igfbp6 ▨
2.9	Crip2	2.0	Tgfb2 ▨
2.8	Ctgf ▨		

ECM Structural Constituent	
6.9	Mfap5
3.1	Eln
3.0	Col11a1
2.4	Fmod
2.3	Fbn1
2.2	Col5a2
2.2	Lamc1
2.2	Col1a2

Serine-Type Proteases	
71.4	Gzmg
45.7	Gzmd
40.2	Gzmf
19.8	Gzme
10.7	Gzmc
2.9	2210021K23Rik
2.2	Ctsg
2.2	Prss11
2.1	Gzmb

▨ *associated with cardiac hypertrophy*

Fold at 14.5 days (term) gestation shown

B

Cell Junctions			
Gap junctions		Desmosome junctions	
2.8	Gja1	2.8	Dsp
1.7	Gjb2	1.5	Dsg2
Tight junctions		1.5	Itgb4
2.8	Ocln	1.4	Lamb3
1.9	Cldn3		
1.7	Cldn4		

Intermediate Filaments	
Kinesin Complex	
7.8	Krt1-19
4.6	Krt2-7
4.5	Krt2-8
4.5	Krt1-18
3.4	Sftpd

Serine-Protease Inhibitors	
4.3	C3
2.9	Wfdc2
2.8	Expi
1.9	Serpina1a
1.8	Serpina1e

Fold at 18.5 days (term) gestation shown

C
Keratin Intermediate Filaments
Cytoplasmic plaque [plakoglobin, desmoplakins
Desmoglein and Desmocollin (transmembrane linker proteins)
Plasma Membrane
Intracellular Space
Plasma Membrane
Spot Desmosome Junctions
Integrins
α_6
β_4
Anchoring laminin fibrils
Collagen
Basal Lamina
Hemidesmosome Junctions

Figure 8.2. Regulation of distinct remodeling pathways by gene transcription during uterine gestation. Up-regulated genes are shown for separate remodeling and cell structural components, corresponding to (A) quiescence and (B) term gestation. (C) The interaction of several of these term-regulated cell structural components in the context of cell-cell junctions and cell-ECM contacts.

With initiation of contractile activation at term, we observe distinct changes along several of these same pathways that act to specifically counteract quiescence-induced expression changes. These include the up-regulation of serine-protease inhibitors, contractile activators of the Gαq calcium stimulating pathway, and up-regulation of prostaglandin signaling regulators (Figure 8.2 B). Surprisingly, we also observe the up-regulation of several non-myocyte cytoskeletal structural and signaling regulators including keratins and tight/desmosome junction components, suggesting these interactions may also mediate labor along with an increase in gap junctions between cells (Figure 8.2 C). Similar pathways were also regulated at the level of AS, when term and mid-gestation myometrium were analyzed using splicing sensitive microarrays (chapter 7). These data suggest that both gene expression and AS may act in concert to regulate distinct components of the same pathways.

Pathway analyses of genes with an involution-restricted expression pattern indicate an overwhelming shift towards pathways of protein degradation (proteosome), apoptosis, Wnt signaling, and matrix metalloproteinase activation (http://www.genmapp.org/supplemental/MAPPs/pathways.html). While protein degradation and apoptosis are processes clearly associated with uterine involution, the specific regulators have not been carefully elucidated by either gene-by-gene analyses or through single time-point microarray strategies. Interestingly, by using a custom program named GEMFinder (Gene Expression Module Finder) we identified coordinate regulation of both postpartum matrix metalloproteinases and quiescence serine proteases at the level of both expression and genomic localization (Figure 3.4). For each of these examples, pathway analysis coupled with gene expression clustering and multiple visualization strategies was essential for identifying meaningful biological interactions.

8.3 Linking whole genome alternative splicing profiles to functional predictions

Not unlike the myometrium throughout gestation, totipotent cells of the blastocyst inner cell mass must alter their signaling properties and composition during differentiation. Unlike the myometrium, these cells are non-reversibly committed to any one of hundreds of possible cell fates. While many studies have examined the temporal regulation of gene expression during ESC differentiation (Hailesellasse Sene *et al.* 2007), such data is often limited by a lack of differentiation time-points, homogenous derived adult cell precursors, and ultimately a lack of specific hypotheses to test once gene lists have been generated. While such gene expression datasets can ultimately provide valuable information, mining such data can be arduous given the fact that there are often hundreds to thousands of changes with no clear way to segregate these changes beyond separation by biological category or specific bias of the investigator.

Thus, methods that exploit additional genomic information, distinct from data collected only by the microarray experiment, can often lead to more informative results and testable hypotheses. Examples include searching for the co-occurrence of specific transcription binding sites among regulated genes

(Grskovic *et al.* 2007) and integration with epigenetic data (Boyer *et al.* 2005), (Loh *et al.* 2006) as useful methods to examine and ultimately test specific hypotheses.

AS provides a critical means to increase proteomic diversity, often independent of the gene expression changes. Given that the AS profiles of distinct cell lineages are quite different from one another (Yeo *et al.* 2004), AS may largely define functional changes within distinct cell types during differentiation and thus should be carefully considered when performing genome-wide experiments. To assess the contribution of AS in human and mouse ESCs as they differentiate, we interrogated splicing events using newly developed software in both mixed (mouse) and lineage restricted (human) analyses, with the aim of elucidating functionally relevant isoforms which regulate developmental pathways in differentiating cells as well the maintenance of ESCs.

By using a simple pattern segregation method (ANOVA), we were highly successful in delineating both gene expression and AS events specific to cardiac precursor differentiation or in common to both human neural and cardiac specification. This computational approach identified AS events that could be readily validated (up to 90%). The splicing events identified showed remarkable specificity for either ESCs or cardiac precursors, with often a single isoform showing expression in each cell type. When compared to adult tissue profiles, several of these genes showed restricted patterns to those predicted by our segregation analysis strategy. Functional analysis at the level of protein domains and microRNA binding sites identified several novel functional correlations between developmentally regulated isoforms predicted to impact protein function and/or translational inhibition. Intriguingly, this data suggests that isoforms enriched in ESCs tend to favor pathways that oppose apoptosis and proliferation, while isoforms specifically enriched in cardiac precursors act to modify the composition and expression of proteins to either blunt cardiac inhibitors (HDAC9) or promote contractile signaling (ASPH, SPTBN1).

In addition to these findings, we present new software for the integrated analysis of AS and transcriptional changes. This includes software for: (1) determining gene expression values from whole genome exon, exon-exon junction or conventional array data (ExpressionBuilder); (2) further calculating statistics between biological groups and filtering out probe sets based on detection calls (ExpressionBuilder), (3) calculating splicing scores and aligning results to multiple functional predictions (AltAnalyze) and (4) visualization of the resulting data in the context of known gene structures and splicing events (SubgeneViewer). These applications are being made freely available and open-source to encourage community use and contribution.

8.4 Functional dissection of splice variants for a critical pluripotency and differentiation factor, Tcf3, in mouse ESCs

An independent analysis of mouse ESC differentiation yielded similar pathways regulated by AS, as compared to human ESC differentiation. These included factors involved in the regulation pluripotency (Tcf3, Map3k7), cell cycle control (Smarcb1, Mark3 and Epb4.1), and cardiac physiology (Atp2a2). In the case of

Tcf3, a factor critical for regulating ESC maintenance and differentiation, expression of a novel form of this transcript was highly enriched ESCs, with insertion of 14AA into a critical co-factor binding domain of the protein (Groucho). Given the nature of this splicing event and the fact that most studies of Tcf3 in mouse ESCs utilize the non-ESC enriched form, we chose to explore the relationships of these two isoforms during pluripotency and with differentiation to multiple cell lineages.

To achieve this goal, we utilized isoform-specific RNA targeting and isoform-specific expression in ESCs devoid of endogenous Tcf3, in order to characterize ESCs with expression of one, two, or no isoforms. This analysis strategy, while largely novel, was highly effective in elucidating the role of individual Tcf3 isoforms, suggesting that the ESC-enriched Tcf3 isoform is a potent transcriptional repressor of Nanog and Oct4 in pluripotent ESCs and is not required for early cell fate decisions. Alternatively, the embryoid body (EB) enriched Tcf3 isoform (short form) was required for both early and late lineage commitment steps as determined by quantitative mRNA analysis in differentiating EBs and derived teratomas. This data shows that AS provides a critical switch, which is likely required for early developmental stem cell differentiation.

8.5 Conservation of AS during ESC differentiation

Although distinct microarray technologies were used to assess AS in human and mouse ESCs, several conserved AS events were verified in our analyses (Figure 8.3). Both mouse and human ESCs share a number of properties, including expression of the same core transcription factors (Oct4, Sox2 and Nanog), regulation by Wnt signaling factors and defined factors necessary for reprogramming of adult somatic cells (chapter 6). However, there are also clear differences in the pluripotency pathways in these cells, given that neither the Stat3 nor BMP signaling pathway contribute to pluripotency in hESCs, but do so in mouse (Vallier *et al.* 2005). In our analysis, we verified human and mouse AS events that are predicted to be either common to differentiation (MADD, MARK3, KIF13), specific for cardiac spheroid differentiation (relative to neural precursor differentiation) (ATP2A2, DNML1) or that have un-examined precursor differentiation patterns (Smarcb1). Thus, more detailed exploration of these AS events with ESC differentiation or cardiac specification may yield conserved mechanisms, important for ESC differentiation.

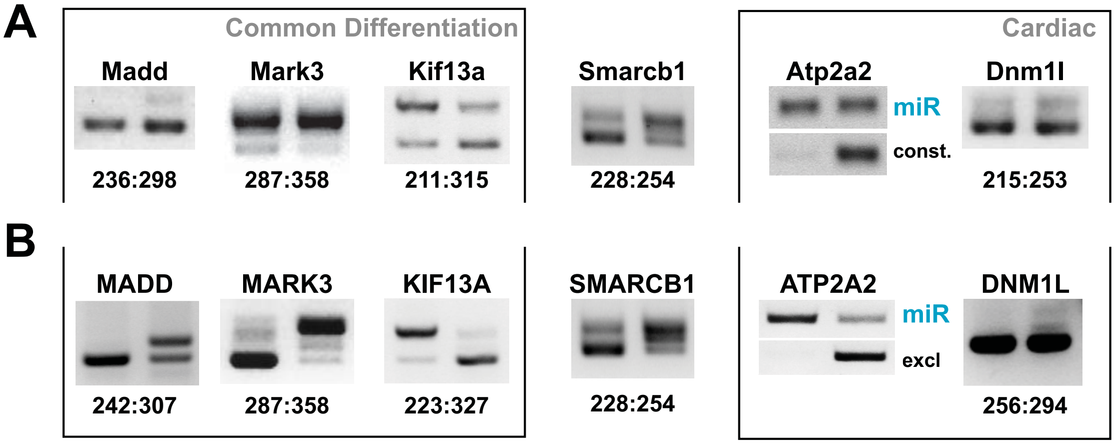


Figure 8.3. Conservation of AS in human and mouse ESCs. (A) Mouse genes examined for AS based common microarray predictions (Mark3, Atp2a2, Dnm1l) or human confirmation (Madd, Kif3a). Genes in the left hand panel are predicted to have a common cardiac/neural differentiation pattern in hESCs, whereas genes in the right panel had a cardiac enriched pattern. (B) Corresponding AS events found from hESC exon-array analyses (MADD, MARK3, KIF13A, ATP2A2, DNM1L) or identified based on mouse exon-exon junction array studies (SMARCB1 – not probed on exon-array). Human validation was in Rex+ hESCs and derived CSs (chapter 5) or H9 ESCs and derived EBs (SMARCB1).

8.6 Common pathways regulating muscle remodeling and lineage commitment

When AS and gene expression profiles are compared for mouse ESC differentiation and myometrial gestational remodeling, we find very little overlap in individual genes regulated (Figure 6.1 and unpublished comparisons). Similar results are also obtained when comparing myometrial expression profiles to human cardiac and neural differentiation datasets (data not shown), with only a small percentage of AS events predicted between human and mouse ESC differentiation datasets (Figure 8.3). While the same genes and splice variants do not appear to be largely regulated in common, ORA analysis indicates that for multiple modes of gene regulation (transcription and splicing), a core set of biological processes is regulated with these developmental and remodeling paradigms. Most significant are tight junction, cytoskeletal and ECM remodeling components along with cell cycle progression and cell growth and contractile signaling pathways. In the myometrium these pathways are regulated at multiple levels (transcription and AS, chapter 6-7), by distinct regulators at different phases of gestation. With AS in hESCs, these same pathways were over-represented among both verified and AS regulated genes. An interesting example is integrin-mediated signaling which was over-represented in both the myometrium at term and among alternative spliced genes with ESC differentiation (mouse and human) (Figure 8.4). Although distinct components were regulated by transcription or splicing, both appear to regulate components that participate in interactions with the ECM and cell-cell contact formation (e.g., tight junctions). Given that both systems are promoting contractile remodeling, these changes may represent distinct modes of achieving this goal. Interestingly, integrin-mediated/focal-adhesion signaling has been implicated as a critical process in both the stretch induced regulation of the myometrium at term to promote contraction and in the specification of ESCs to cardiomyocytes (Hakuno *et al.* 2005; Li *et al.* 2007). Thus, a systems level approach to assess such changes is able to highlight mechanistic similarities that likely require distinct yet complimentary interactions.

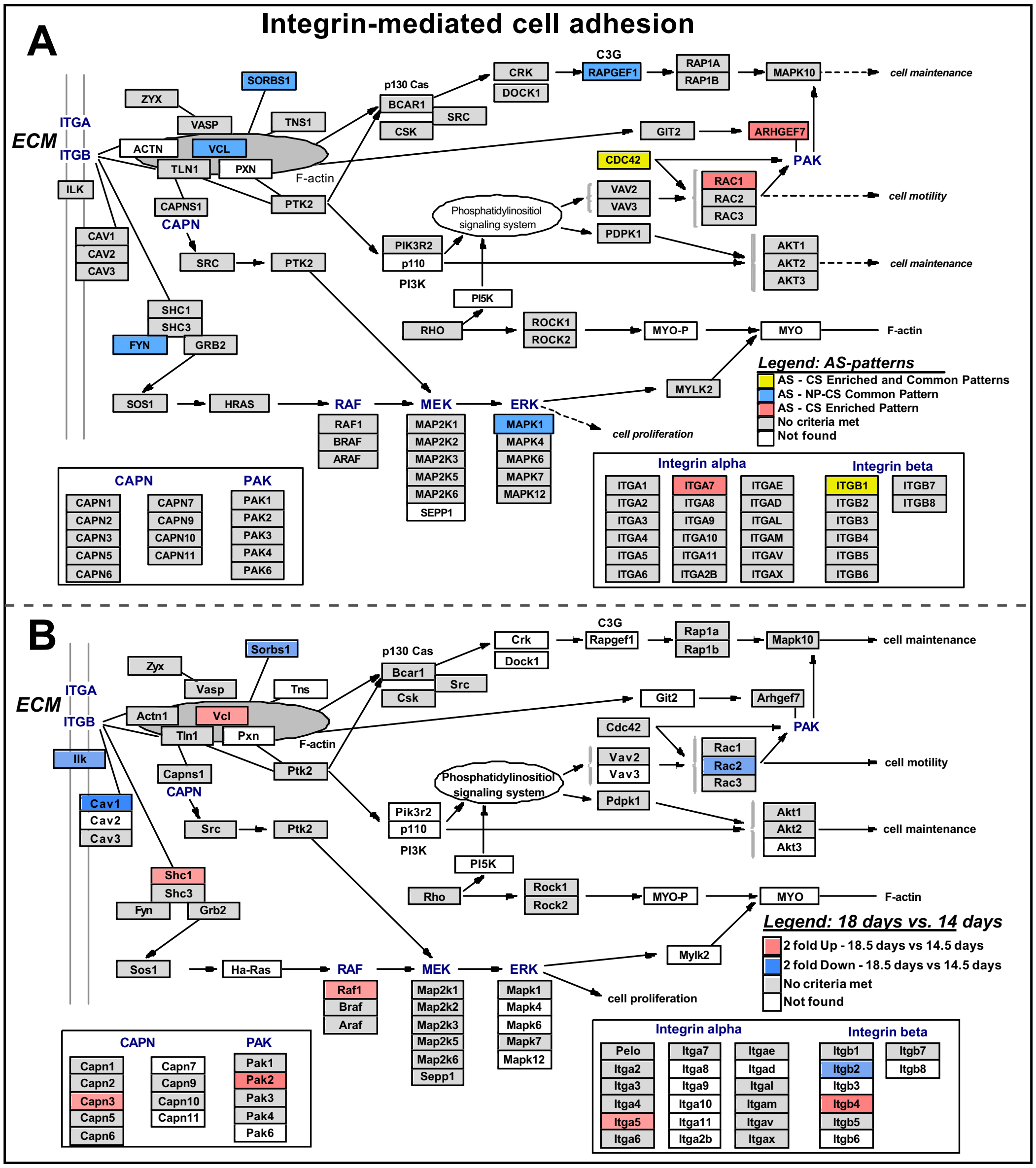


Figure 8.4. Regulation of common signaling pathways in distinct model systems. Integrin-mediated signaling (Wikipathways.org) is shown with data for (A) human genes undergoing AS with differentiation to distinct lineages or (B) mouse genes up or down-regulated at term gestation in the myometrium. In the human pathway, probe sets with evidence of AS having either a common cardiac/neural (blue), cardiac enriched (red) or exons with both patterns (gold) are displayed. In the the myometrium with gestation, genes with a red box indicate up-regulation, whereas those with a blue box indicate down-regulation.

Software Packages Deveoped	Development Team	Distribution	User-base
GenMAPP	GenMAPP	open-source	>10,000
MAPPFinder	GenMAPP	open-source	>5,000
SubGeneViewer	GenMAPP	free	TBA
GO-Elite	NS	free	>50
AltAnalyze	NS	free	TBA
ExpressionBuilder	NS	free	TBA
LinkEST	NS	in-house	NA
GEMFinder	NS	in-house	NA
Onco-Split	NS	free	<10

Table 8.1. Contributed software development projects. Software packages developed as apart of this doctoral thesis. Programs include GenMAPP and MAPPFinder, originally developed by NS with other members of the GenMAPP team, with a current user base of over 10,000. Additional packages include the Gene Ontology and pathway ORA application GO-Elite, AltAnalyze, ExpressionBuilder, and OncoSplit, all of which can be downloaded at (http://conklinwolf.ucsf.edu/informatics/nsalomonis.html). Packages with a user-base indication of TBA were in the process of being posted with the submission of this dissertation. In-house packages were not publicly distributed at the time of this report.

8.7 Why the need for open source software for whole genome analyses

A major component and focus of this thesis has been the development of freely available open-source bioinformatics tools for the analysis of complex, high throughput genomic datasets. These include the tools GenMAPP and MAPPFinder (chapter 2), GO-Elite (chapter 4), AltAnalyze, ExpressionBuilder, LinkEST and SubGeneViewer (chapters 5-7), GEMFinder (chapter 3), and OncoSplit (unpublished collaboration with the Barry Gusterson laboratory in Glasgow Scotland) (Table 8.1).

In each of these cases, myself and other developers have focused on creating novel tools that are free and easy to use by the research community. While it seems fairly obvious to make such tools available to the public, it is not uncommon for researchers to keep such applications entirely in house, requiring interested parties to directly collaborate with that laboratory. As a result, the research community ends up spending more time, money, and energy to re-develop tools that have already been described. This was the case for several AS methods described herein, including the algorithms ASPIRE (Ule *et al.* 2005), LinRegress (Sugnet *et al.* 2006) and the splicing index method (Gardina *et al.* 2006) all re-implemented in AltAnalyze. In addition to being useful for the community, openly providing this software promotes scientific evaluation of the methods, independent of the published report. Open-source software further allows for improvement or update of the original source code and associated databases by the user community directly. Thus, development and maintenance of open-source software projects are crucial to increase the longevity of these resources and encourage community development from within.

8.8 Next steps – Integrative approaches for software development and genomic technologies

A central theme of the work presented here has been the integration of complementary genomic resources, datasets, and new technologies to obtain novel insights into discrete biological transitions. To achieve these goals I have had to develop new applications and computational methods as well as integrate data from multiple resources. Given that the complexity, amount of data collected, and the diversity in biological assays will only increase over time, the necessity for integrative approaches is becoming even more important. By using multiple time-points or conditions in our study design, we have been able to address specific questions that would otherwise be obscured in a sea of data. By combining our data with external bioinformatics resources (mRNA/protein sequence, protein domain, microRNA binding site, and genomic location), we have been successful in identifying new biological mechanisms that regulate gene activity and resulting protein function. At the level of experimental hypothesis testing, we have shown that informatics predictions derived from this strategy can represent critical *in vivo* developmental regulatory mechanisms that can be reasonably validated in living cells. In the coming years, such integrative bioinformatics strategies will be critical in defining new biology from complex cellular processes.

8.9 References

Boyer, L. A., T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, D. K. Gifford, D. A. Melton, R. Jaenisch and R. A. Young (2005). "Core transcriptional regulatory circuitry in human embryonic stem cells." Cell **122**(6): 947-56.

Gardina, P. J., T. A. Clark, B. Shimada, M. K. Staples, Q. Yang, J. Veitch, A. Schweitzer, T. Awad, C. Sugnet, S. Dee, C. Davies, A. Williams and Y. Turpaz (2006). "Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array." BMC Genomics **7**: 325.

Grskovic, M., C. Chaivorapol, A. Gaspar-Maia, H. Li and M. Ramalho-Santos (2007). "Systematic identification of cis-regulatory sequences active in mouse and human embryonic stem cells." PLoS Genet **3**(8): e145.

Hailesellasse Sene, K., C. J. Porter, G. Palidwor, C. Perez-Iratxeta, E. M. Muro, P. A. Campbell, M. A. Rudnicki and M. A. Andrade-Navarro (2007). "Gene function in early mouse embryonic stem cell differentiation." BMC Genomics **8**: 85.

Hakuno, D., T. Takahashi, J. Lammerding and R. T. Lee (2005). "Focal adhesion kinase signaling regulates cardiogenesis of embryonic stem cells." J Biol Chem **280**(47): 39534-44.

Li, Y., C. Gallant, S. Malek and K. G. Morgan (2007). "Focal adhesion signaling is required for myometrial ERK activation and contractile phenotype switch before labor." J Cell Biochem **100**(1): 129-40.

Loh, Y. H., Q. Wu, J. L. Chew, V. B. Vega, W. Zhang, X. Chen, G. Bourque, J. George, B. Leong, J. Liu, K. Y. Wong, K. W. Sung, C. W. Lee, X. D. Zhao, K. P. Chiu, L. Lipovich, V. A. Kuznetsov, P. Robson, L. W. Stanton, C. L. Wei, Y. Ruan, B. Lim and H. H. Ng (2006). "The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells." Nat Genet **38**(4): 431-40.

Sugnet, C. W., K. Srinivasan, T. A. Clark, G. O'Brien, M. S. Cline, H. Wang, A. Williams, D. Kulp, J. E. Blume, D. Haussler and M. Ares, Jr. (2006). "Unusual intron conservation near tissue-regulated exons found by splicing microarrays." PLoS Comput Biol **2**(1): e4.

Ule, J., A. Ule, J. Spencer, A. Williams, J. S. Hu, M. Cline, H. Wang, T. Clark, C. Fraser, M. Ruggiu, B. R. Zeeberg, D. Kane, J. N. Weinstein, J. Blume and R. B. Darnell (2005). "Nova regulates brain-specific splicing to shape the synapse." Nat Genet **37**(8): 844-52.

Vallier, L. and R. A. Pedersen (2005). "Human embryonic stem cells: an in vitro model to study mechanisms controlling pluripotency in early mammalian development." Stem Cell Rev **1**(2): 119-30.

Yeo, G., D. Holste, G. Kreiman and C. B. Burge (2004). "Variation in alternative splicing across human tissues." Genome Biol **5**(10): R74.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses and dissertations. Copies of all UCSF theses and dissertations will be routed to the library via the Graduate Division. The library will make all theses and dissertations accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis or dissertation to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

_________________________________ 6-12-08
Author Signature Date

AltAnalyze Information and Instructions

Version 1.0 Beta

Section 1 - Introduction

1.1 Program Description

AltAnalyze is a freely available, cross-platform application that allows you to take relatively raw microarray data and assess alternative splicing or alternative promoter usage and then view how these changes may affect protein sequence, domain composition, and microRNA targeting. This software requires no advanced knowledge of bioinformatics programs or scripting. All you need are your microarray files along with some simple descriptions of the conditions that you're analyzing.

AltAnalyze is composed of a set of programs designed to (A) organize, filter, and summarize transcript tiling data; (B) calculate scores for alternative splicing (AS), alternative promoter selection (APS) or transcript elongation; (C) annotate regulated alternative exon events; and (D) assess downstream predicted functional consequences at the level of protein functional regions and microRNA (miRNA) binding sites. The resulting data will be a series of text files (results and over-representation analyses) that you can directly open in a computer spreadsheet program for analysis and filtering.

This software is currently compatible with the Affymetrix exon 1.0 ST array and the custom Affymetrix AltMouse A array, however, it may be adapted to support other platforms on a per example basis (contact author) or by other developers.

1.2 Implementation

AltAnalyze is composed of a set of distinct modules written in the programming language Python. Python is a cross-platform compatible language, therefore, AltAnalyze can be run on any operating system that has Python installed or on Windows without Python. Python is bundled with most current Mac and Linux operating systems. For the Windows operating system, a stand-alone executable file is available which does not require installation of any additional software, including Python. The AltAnalyze interface is an interactive graphic user interface and command prompt with easy to use options.

1.3 Requirements

The basic installation of AltAnalyze requires a minimum of 1GB of hard-drive space for all required databases and components. This includes support for all species and currently supported arrays. Future versions will include an option for automated download of databases specific to the user specified array analyses. A minimum of 1GB of RAM and Intel Pentium III processor speed are further recommended. At least an additional 1GB of free hard-drive space is recommended for building the required output files.

1.4 Before Using AltAnalyze

This software requires that the user obtain normalized expression values prior to use, as opposed to raw microarray image files (CEL files). Example methods for obtaining such data include:

1) ExpressionConsole and RMA analysis (Windows only)

2) Affymetrix Power Tools (APT) and RMA analysis
3) Bioconductor and RMA analysis

We recommend using a method that can produce both the expression file (containing probe set and expression values for each array in your study) and a detection above background (DABG) p-value file (containing corresponding detection p-values for each probe set). ExpressionConsole, is a free application from Affymetrix that outputs both of these file types when normalizing array files (http://www.affymetrix.com/products/software/specific/expression_console_software.affx). This application has an easy to use graphic user interface (GUI) and excellent documentation. For non-Windows operating systems, the program APT is also available to perform RMA and DABG using a command line interface. For APT download and documentation, see: http://www.affymetrix.com/support/developer/powertools/index.affx.

In additional to expression analysis, users can optionally install the programs APT and R (http://www.cran.org). APT is necessary if the user wished to include MiDAS statistics when performing an exon-array analysis. APT for different operating systems and configurations can be found at: http://www.affymetrix.com/support/developer/powertools/changelog/PLATFORMS.html. R is optional when performing an exon-junction array analysis using the algorithm Linear Regression with the `rlm` method (***not needed for basic Linear Regression***). Further directions are provided to interface MiDAS and R with AltAnalyze under the respective algorithm descriptions.

1.5 Help with AltAnalyze

Additional documentation, help, and user questions are available at the AltAnalyze website or at the AltAnalyze Google Groups user forum: http://groups.google.com/group/alt_predictions

Section 2 – Running AltAnalyze

2.1 Preparing your data for the first time

After downloading and extracting AltAnalyze to your computer, prepare the following four tab-delimited text files:

1) Experiment file *(required)* - Expression dataset with all probe sets and expression values analyzed (prefix = “exp.”)
2) Statistics file - DABG p-value dataset with all probe sets and p-values analyzed (prefix = “stats.”)
3) Groups file *(required)* - Table of array sample names (in files 1 and 2), arbitrary ordered group numbers (1, 2, 3 and so on), and group names (e.g., control = 1, stimulated = 2, cancer = 3) (prefix = “groups.”)
4) Comparisons file *(required)* - Table of group comparisons (2, 1 and 3,1) (prefix = “comps.”)

Expression and Statistics files

The first two files (experiment and statistics) are automatically produced when analyzing exon-array data from Affymetrix’s ExpressionConsole. If using this software, it is recommended that you analyze all probe sets (full) as opposed to a subset of probe sets (e.g., core). These two files will need to be renamed “exp.experiment.txt” and “stats.experiment.txt”, where “experiment” is the name of your dataset (e.g., brain-tumor_comparisons). These files can include any number of different experimental conditions and samples. **Although the statistics file is not required, it is highly recommended in order for the program to help eliminate false positive predictions.**

Group file

The third file (groups) assigns each sample or column in the first two files to a biological group. This file is organized as such:

array_file_name	***Group_number***	***group_name***
wt1.CEL	1	normal
wt2.CEL	1	normal
cancer1.CEL	2	cancer
cancer2.CEL	2	cancer
drug1.CEL	3	drug
drug2.CEL	3	drug

The group number should start with 1 and follow sequentially. The group name should be something meaningful for you (no spaces in the name). This file should be named “groups.experiment.txt”.

Comparison file

Finally, the last is the comparison file that tells AltAnalyze which pair-wise comparisons to compute on. This simple file contains two columns, your numerator group number and your denominator group number. For example, if you only have three groups and two comparisons then the file would have two rows for each comparison:

2	***1***
3	***2***

Here, 2 corresponds to your experimental group (e.g., cancer) ,1 your control group (e.g., normal), 3 is another experimental group (e.g., drug). You can have as many comparisons as you like, as long as the group numbers appropriately correspond to those in the Group file. This file should be named "comps.experiment.txt".

Where to save these files

These files should all be saved to the directory named "ExpressionInput" under the appropriate array type directory (e.g., exon).

2.2 Running Analyses

PC Directions:

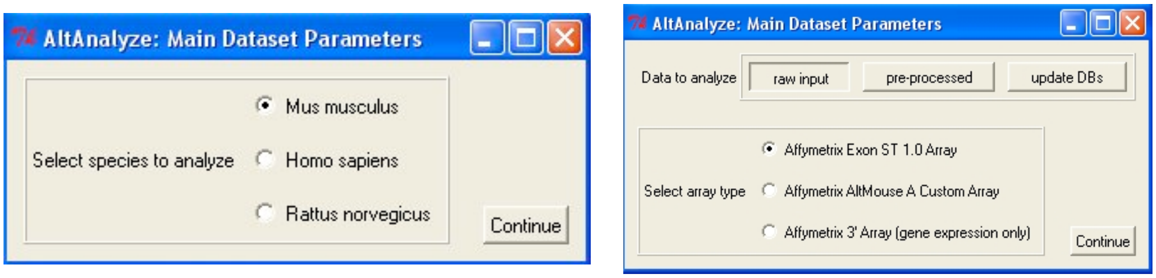
Once the input files have been saved to the appropriate directory, open the executable file named "AltAnalyze.exe" in the AltAnalyze program directory. This will open a set of user interface windows where you will be presented with a series of program options (see next set of direction).

Mac or Linux Directions:

Once the input files have been saved to the appropriate directory, open a new terminal window. On a Mac, the program "Terminal" is accessible from "Applications/Utilities". To run the program, change directories until you are in AltAnalyze folder (e.g., cd Desktop/AltAnalyze) and then run AltAnalyze with the command "python AltAnalyze.py". This will open the AltAnalyze user interface where you will be presented with a series of program options (see next set of direction). If any GUI support files are missing on that computer, command line options will available presented instead through the terminal.

Program Options:

1) Select species and array-type - After starting AltAnalyze, the user will be prompted to select the species analyzing. Compatible species are read from the file “Config/species.txt”. After selecting a species, select the microarray platform used (e.g., Affymetrix 1.0 ST exon array). Only arrays listed as compatible for those species in the file “Config/arrays.txt” are shown. Additional species and array types can be added to these configuration files if the appropriate support files are included.



AltAnalyze: Main Dataset Parameters
Continue
AltAnalyze: Main Dataset Parameters
Data to analyze
raw input
pre-processed
update DBs
Continue

Select species to analyze
● Mus musculus
○ Homo sapiens
○ Rattus norvegicus

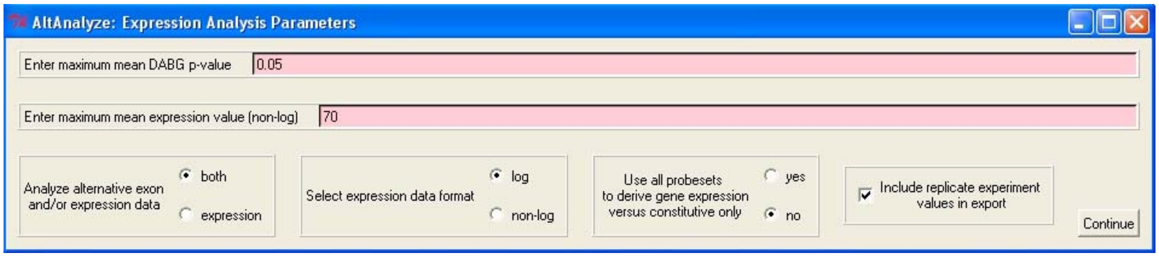
Select array type
● Affymetrix Exon ST 1.0 Array
○ Affymetrix AltMouse A Custom Array
○ Affymetrix 3' Array (gene expression only)

Figure 1.1. AltAnalyze Main Dataset Parameters Menu. A) Options for selecting species and restricting analyses based on compatible arrays. B) Options for data analysis and possible arrays for the selected species.

2) Select input file type – If the user is analyzing their data for the first time in AltAnalyze select the option “raw input”. This option will create a gene-expression summary file and filtered probe set expression files. If AltAnalyze has previously built these files, select the “pre-processed data”, tab. Selecting this option will substantially reduce run-time. Select “Update DBs” to build or update a new set of databases by automatically downloading files from the AltAnalyze website.

3) Select expression analysis parameters - Next, the user will be prompted to use a set of default parameters or customize these options. The first window will provide options for gene expression summarization. In addition to exon and exon-junction expression inputs, conventional, 3’

arrays can be analyzed with these options. Selecting “expression” for the option “Analyze alternative exon/or expression data” will only perform these expression analyses and will skip any alternative exon analyses (if compatible with the array type). Defaults are stored as files in the “Config” folder with the prefix “default-”. All displayed options, including those found in the default files, are read from the file “Config/options.txt”. The user can modify these defaults by editing these files.


AltAnalyze: Expression Analysis Parameters

Enter maximum mean DABG p-value: 0.05

Enter maximum mean expression value (non-log): 70

Analyze alternative exon and/or expression data: ● both ○ expression

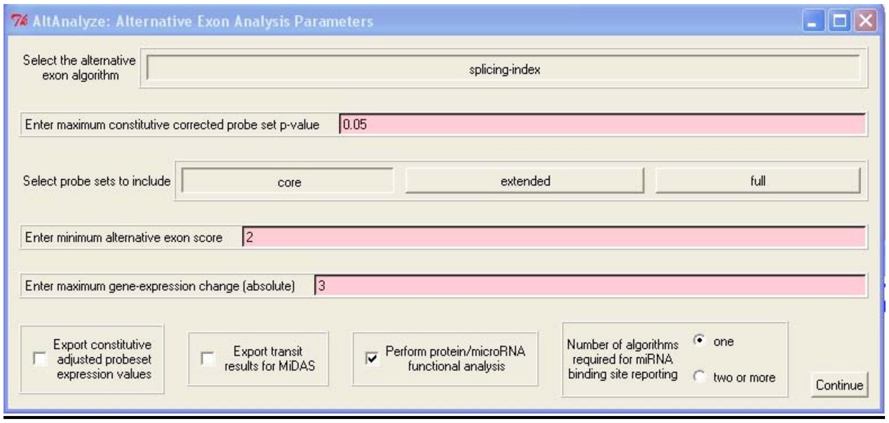
Select expression data format: ● log ○ non-log

Use all probesets to derive gene expression versus constitutive only: ○ yes ● no

☑ Include replicate experiment values in export

Figure 1.2. Expression Analysis Parameters. Statistical thresholds and analysis options for performing gene expression summarization and probe set filtering prior to alternative exon analysis.

4) <u>Select alternative exon analysis parameters</u> – If using an exon-level microarray (e.g., Human Affymetrix 1.0 ST exon array), the user will be presented with specific options for that microarray (see two possible interfaces below). These options include alternative exon analysis methods, statistical thresholds, and options for additional analyses (e.g, MiDAS). If the option “Export transit results for MiDAS” is selected, AltAnalyze will export input files for the program APT (“AltResults/MiDAS”) along with commands to build the MiDAS output files. When APT analysis is complete (outside of AltAnalyze), the user can select “continue” in AltAnalyze to incorporate these statistics into the analysis.



AltAnalyze: Alternative Exon Analysis Parameters

Select the alternative exon algorithm: splicing-index

Enter maximum constitutive corrected probe set p-value: 0.05

Select probe sets to include: core | extended | full

Enter minimum alternative exon score: 2

Enter maximum gene-expression change (absolute): 3

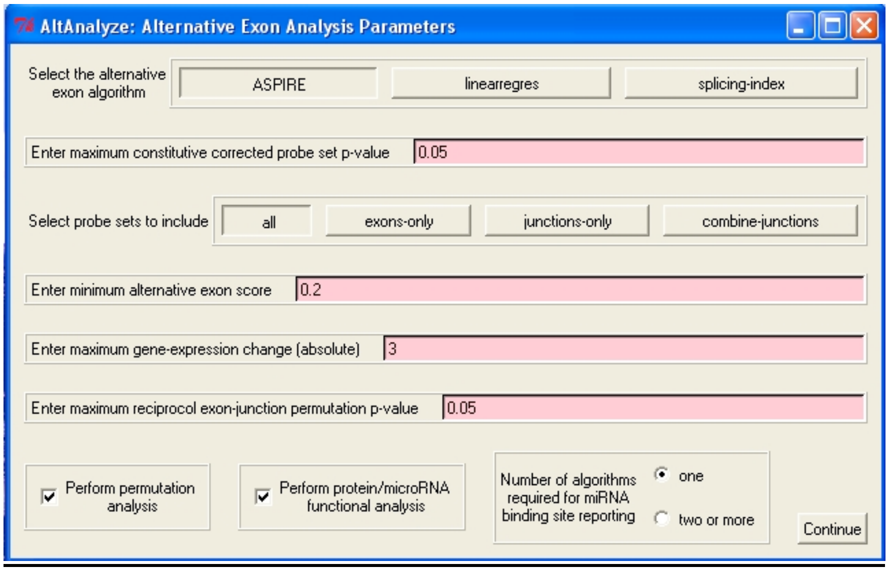
□ Export constitutive adjusted probeset expression values

□ Export transit results for MiDAS

☑ Perform protein/microRNA functional analysis

Number of algorithms required for miRNA binding site reporting: ● one ○ two or more

Continue



AltAnalyze: Alternative Exon Analysis Parameters

Select the alternative exon algorithm: ASPIRE | linearregres | splicing-index

Enter maximum constitutive corrected probe set p-value: 0.05

Select probe sets to include: all | exons-only | junctions-only | combine-junctions

Enter minimum alternative exon score: 0.2

Enter maximum gene-expression change (absolute): 3

Enter maximum reciprocol exon-junction permutation p-value: 0.05

☑ Perform permutation analysis

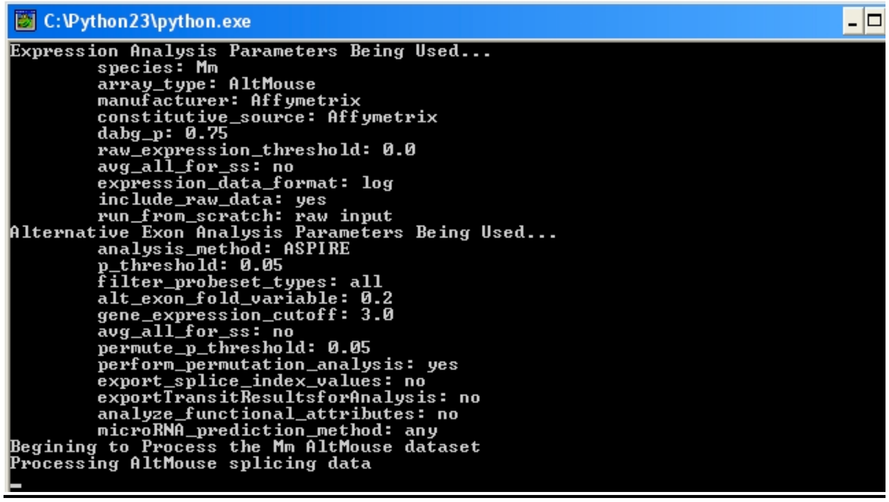
☑ Perform protein/microRNA functional analysis

Number of algorithms required for miRNA binding site reporting: ● one ○ two or more

Continue

Figure 1.3. Alternative Exon Analysis Parameters. Different options for the selected microarray, (A) for 1.0 ST exon array and (B) for AltMouseA, when analyzing alternative exon-level data. Options include alternative analysis methods (e.g., splicing-index, ASPIRE, linearregress) and filtering probe sets based on annotation type or other optional statistics (e.g., MiDAS or permutation analysis). Default options are selected.

After selecting “continue” in this last window, the GUI will close and AltAnalyze will begin performing the selected analyses. While the AltAnalyze program is running, several intermediate results files will be created. The terminal window (see below) will indicate the progress of each analysis as it is running. When finished, AltAnalyze will prompt the user that the analysis is finished. A report is exported with each run containing a summary of overall statistics and analysis progress to "AltResults/AlternativeOutput/summary_report.txt".



```
Expression Analysis Parameters Being Used...
        species: Mm
        array_type: AltMouse
        manufacturer: Affymetrix
        constitutive_source: Affymetrix
        dabg_p: 0.75
        raw_expression_threshold: 0.0
        avg_all_for_ss: no
        expression_data_format: log
        include_raw_data: yes
        run_from_scratch: raw input
Alternative Exon Analysis Parameters Being Used...
        analysis_method: ASPIRE
        p_threshold: 0.05
        filter_probeset_types: all
        alt_exon_fold_variable: 0.2
        gene_expression_cutoff: 3.0
        avg_all_for_ss: no
        permute_p_threshold: 0.05
        perform_permutation_analysis: yes
        export_splice_index_values: no
        exportTransitResultsforAnalysis: no
        analyze_functional_attributes: no
        microRNA_prediction_method: any
Begining to Process the Mm AltMouse dataset
Processing AltMouse splicing data
_
```

Figure 1.4. AltAnalyze Terminal Status. Initial parameters set by the user after choosing to continue the analysis. This status window (shown for Windows machines) displays the progress of the analyses for the duration of AltAnalyze processes.

2.3 Overview of Analysis Results

AltAnalyze will output two classes of files:

1) Gene expression (GE) summary
2) Exon level summary

Gene expression summary data

The GE summary is a single file that contains all computed constitutive gene expression values from your dataset. The values are derived from probe sets that align to regions of a gene that are common to all transcripts (constitutive) and thus are informative for transcription (unless all probe sets are selected – see “Select expression analysis parameters”, above). Along with the raw gene expression values, statistics for each indicated comparison (mean expression, folds, t-test p-values) will be included along with gene annotations for that array. This file is analogous to the results file you would have with a typical, non-exon microarray experiment and is saved to the folder “ExpressionOutput”.

Exon-level summary data

These results are produced from all probe sets that may suggest alternative splicing, alterative promoter regulation, or any other variation relative to the constitutive gene expression for that gene (derived from comparisons file). Each set of results correspond to a single pair-wise comparison (e.g., cancer vs. normal) and will be named with the group names you assigned (groups file). Four sets of results files are produced in the end:

1) Probe-level - Probe set/exon-level statistics, AS/APS annotations, and functional feature predictions (protein, miRNA binding site).
2) Gene-level – Summary of probe-level data file.
3) Domain-level – Over-representation analysis of gene-level domain/residue modifications due alternative regulation.

4) miRNA binding sites - Over-representation analysis of gene-level. predicted miRNA binding sites present in alternatively regulation exons.
5) Summary statistics file – Number of genes alternatively regulated compared to differentially expressed and summary protein association information (e.g., mean regulated protein length).

Each file is a tab delimited text file that can be opened, sorted and filtered in a spreadsheet program. These files are saved to the folder “AltResults/AlternativeOutput”, all with the same prefix (pair-wise group comparisons). AltAnalyze will analyze all pair-wise comparisons in succession and combine the probe-level and gene-level results into two additional separate files (named based on the splicing algorithm chosen).

Probe- and Gene-Level Result Files

The probe-level file contains alternative exon data for either one probe set (exon-array) or reciprocal probe sets (junction array). This includes:

- Gene and probe set annotations (e.g., description, symbol, probe set ID, probe set exon ID, transcript clusters, associated Ensembl/UCSC exons, ordered exon-region IDs).
- Raw expression data for the regulated probe set.
- Constitutive gene expression changes and baseline expression.
- Statistical results (e.g., splicing-index score and p-value, MiDAS p-values, probe set p-value).
- Alternative exon annotations (e.g., splicing-events, alternative promoters, alternative annotation confidence score).
- Protein- and miRNA-level associations (e.g., associated IDs, sequence, pattern of regulation, regulated domains/miRNA binding sites).

The gene-level file contains a summary of the data at the gene level, with each row representing a unique gene. This file also includes:

- Gene-ontology and pathway information for each gene extracted from any Affymetrix CSV annotation files for that species present in the directory “AltDatabase/Affymetrix/*species*”.

Protein Feature and miRNA Binding Site Result Files

Over-representation analyses, (files 3 and 4) have the same structure:

- Column A is the name of the protein feature or domain.
- Column B is the over-representation z-score (see Section 3 - Algorithms) for all unique genes aligning to the feature that are alternative regulated by the analysis.
- Column C is the z-score for just those unique genes aligning to the feature, where that feature is considered ***up-regulated*** or ***included*** in the numerator of the biological comparison (e.g., cancer).
- Column D is the z-score for just those unique genes aligning to the feature, where that feature is considered ***down-regulated*** or ***excluded*** in the denominator of the biological comparison (e.g., normal).
- Columns E-G have the number of unique genes regulated corresponding to columns B-D.
- Columns H-J have the gene symbols corresponding to the unique genes listed in columns E-G.
- Column K has the total number of unique genes measured on the array aligning to the feature.

Section 3 – Algorithms

Multiple algorithms are available in AltAnalyze to identify individual probe sets (for exon arrays (EA)) or reciprocal probe sets (exon-exon junction array (JA)) that are differentially regulated relative to constitutive gene expression changes. These include the splicing index method (EA and JA), MiDAS (EA and JA), ASPIRE (JA) and Linear Regression (JA).

3.1 Default Methods

Because some optional algorithms require installation of outside tools (APT and R), these algorithms are not selected as AltAnalyze default options. As mentioned in early sections, the default options are listed in the folder "Config" as "defaults-expr.txt", "defaults-alt_exon.txt", and "defaults-funct.txt".

defaults-expr.txt	Default expression analysis options (Figure 1.2)
defaults-alt_exon.txt	Default alternative exon analysis options (Figure 1.3)
defaults-funct.txt	Default functional analysis options (Figure 1.3)

These options correspond to those found in the configuration file "options.txt". The user is welcome to modify the defaults and theoretically even the options in the "options.txt" file, however, care is required to ensure that these options are supported the by the program. Since AltAnalyze is an open-source program, it is feasible for the user to add new species and array support or to do so with the AltAnalyze support team. A basic modification is the addition of new species to the "species.txt" file and conventional 3' Affymetrix microarrays for expression analyses. These only require the addition of an Affymetrix CSV annotation file to the appropriate species “Affymetrix” directory, in the folder “AltDatabase”.

The default algorithms for the currently supported arrays are as follows:

Exon	splicing-index (score > 2 and t-test p<0.05), no MiDAS
AltMouse	ASPIRE (score > 0.2 and permute p<0.05)
3’ array	NA

3.2 Algorithm Descriptions

Splicing Index Method

This algorithm is described in detail in the following publications:

(Srinivasan *et al.* 2005) (Gardina *et al.* 2006). In brief, the expression value of each probe set for each array is converted to log space (if necessary). For each probe sets examined, its expression (log2) is subtracted from mean expression of all constitutive aligning probe sets for that array and to calculate a constitutive corrected log expression ratio (subtract instead of divide when these values are in log space). This ratio is calculated for each microarray sample, using only data from that sample. To derive the splicing-index value, the group mean ratio of the control is subtracted from the experimental. This value is the change in exon-inclusion (delta I or δI). A t-test p-value is calculated (two tailed, assuming unequal variance) by comparing these ratios for all samples between the two experimental groups. A negative δI score of -1 indicates a two-fold increase in the adjusted expression of a probe set in the experimental versus control group.

MiDAS

The MiDAS statistic is described in detail in the white paper: www.affymetrix.com/support/technical/whitepapers/exon_alt_transcript_analysis_whitepaper.pdf. This analysis method is available from the computer program APT, mentioned previously. APT uses a series of text files to examine the expression values of each probe set compared to the calculated constitutive expression for that gene, based on multiple probe sets. Since AltAnalyze derives constitutive probe sets different than other methods (which often just look at all probe sets for that gene), AltAnalyze creates it's own unique gene identifiers (different than the Affymetrix transcript clusters) that correspond to each Ensembl gene ID. These relationships are stored in the following files along with the probe set expression values:

meta-Hs_Exon_cancer_vs_normal.txt	Relates probe set to gene
gene-Hs_Exon_cancer_vs_normal.txt	Gene expression values (non-log)
exon-Hs_Exon_cancer_vs_normal.txt	Probe set expression values (non-log)
Celfiles-Hs_Exon_cancer_vs_normal.txt	Relates sample to group
commands-Hs_Exon_cancer_vs_normal.txt	Contains user commands for APT
probeset-conversion -Hs_Exon_cancer_vs_normal.txt	Relates arbitrary gene IDs back to Ensembl

When the user selects the option “Export transit results for MiDAS”, AltAnalyze first exports data for all probe sets (not just indicated by “Select probe sets to include” – Figure 1.3) to these files for all pair-wise comparisons, in succession. Once exported, AltAnalyze will try to open APT (command prompt) and prompts the user to follow the necessary steps to export MiDAS p-values. These include, opening the file “commands-dataset.txt” in the “AltResults/MIDAS” for each pair-wise comparison and pasting the two lines of generated code into APT. These contain instructions to change the directory to the one containing these files, “AltResults/MIDAS”, and analyzing these files using the MiDAS algorithm. MiDAS will create a folder with the pair-wise comparison dataset name and a file with MiDAS p-values that will be automatically read by AltAnalyze and used for statistical filtering. To continue AltAnalyze once building these files (takes approximately 30 seconds per dataset), simply hit return in AltAnalyze to run the full analysis on you’re pair-wise comparisons using the user specified parameters. The MiDAS statistics will be clearly labeled in the results file for each probe set. Note: this statistic will be used to filter splicing-index results based on the user defined “minimum constitutive corrected probe set p-value” (Figure 1.3).

ASPIRE

For exon-exon junction microarray data (e.g., AltMouseA), the algorithm “analysis of splicing by isoform reciprocity” or ASPIRE was adapted from the original report (Ule *et al.* 2005) for inclusion into AltAnalyze. This algorithm uses the expression of probe sets aligning to two competitive exon-exon junctions, or one exon-exon junction and an exon along with constitutive expression values calculated as described with the splicing-index method. These probe set relationships were derived using the Affymetrix exon or exon-exon junction names (e.g., E1-E3 and E2-E3 or E1-E3 and E2), obtained by the Affymetrix AltMerge transcript assembly program (Wheeler 2002). For exon-exon junctions and exons aligning to the same gene, recipricol probe set pairs were extracted using the `ExonAnnotate_module.py` program in AltAnalyze using the identifyPutativeSpliceEvents() function. Such splicing events are further classified as mutually-exclusive (mx-mx) or exon-inclusion/exon-exclusion (ei-ex). Mutually-exclusive splicing events represent an exchange of one exon for another (e.g., E2-E4 and E1-E3). Similar to the splicing-index method, for each reciprocal probe set, a ratio was calculated for expression of the probeset (non-log) divided by the mean of all constitutive aligning probe sets (non-log), for the baseline and experimental groups. The ASPIRE δI was then calculated for the inclusion (ratio1) and exclusion (ratio2) probe sets, as such:

R_{in} = baseline_ratio1/experimental_ratio1
R_{ex} = baseline_ratio2/experimental_ratio2
I_1=baseline_ratio1/(baseline_ratio1+baseline_ratio2)
I_2=experimental_ratio1/(experimental_ratio1+experimental_ratio2)

$in_1= ((R_{ex}-1.0)*R_{in})/(Rex-Rin)$
$in_2= (R_{ex}-1.0)/(R_{ex}-R_{in})$
$\delta I = ((in_2-in_1)+(I_2-I_1))/2.0$

If ($R_{in}>1$ and $R_{ex}<1$) or ($R_{in}<1$ and $R_{ex}>1$) and the absolute δI score is greater than the user supplied threshold (default is 0.2), then the δI is retained for the next step in the analysis. If designated by the user, this next step will be a permutation analysis of the raw input data to determine the likelihood of each ASPIRE score occurring by chance alone. This permutation p-value is calculated by first storing all possible combinations of the two group comparisons. For example, if there are 4 samples (A-D) corresponding to the control group and 5 (E-H) samples in the experimental group, then all possible combinations of 4 and

5 samples would be stored (e.g, [B, C, G, H] and [A, D, E, F]). For each permutation set, ASPIRE scores were re-calculated and stored for all of these combinations. The permutation p-value is the number of times that the absolute value of a permutation ASPIRE score is greater than or equal to the absolute value of the original ASPIRE score (count) divided by the number of possible permutations that produced a valid ASPIRE score ((R_{in}>1 and R_{ex}<1) or (R_{in}<1 and R_{ex}>1)). If this p-value is less than user defined threshold, or count<2 (since some datasets have a small number of samples and thus little power for this analysis), the reciprocal probe sets are reported in the results file.

Linear Regression

When working with the same type of reciprocal probe set data as ASPIRE, a linear regression based approach can be used to produce similar results. This method is based on a previously described approach (Sugnet *et al.* 2006). This algorithm uses the same input as ASPIRE (junction comparisons, constitutive adjusted expression ratios). To derive the slope for each of the two biological conditions (control and experimental) the constitutive corrected expression of all samples for both reciprocal junctions is plotted against each other to calculate a slope for all samples belonging to the same biological group (e.g., control) using the least squared method. In each case, the slope was forced through the origin of the graph (model = y ~ x – 1 as opposed to y ~ x). The final linear regression score is the $\log_2$ ratio of the slope of the experimental group divided by the slope of the baseline group. The same permutation analysis used for ASPIRE is also available for this algorithm.

Note: For previous published analyses (Sugnet et al. 2006 and Salomonis et al. in preparation), linear regression was implemented using the algorithm `rlm`, which is apart of the R `mass` package from bioconductor. The Python R interpreter `rpy`, was used to run these analyses (which requires installation of R). To use this option, select "linearregres-rlm" under "select the alternative exon algorithm" (Figure 1.3).

Over-Representation Statistics

A z-score is calculated to assess over-representation of specific protein features and miRNA binding sites found to overlap with probe sets that are alternatively regulated according to the AltAnalyze user analysis. This z-score is calculated by subtracting the expected number of genes with a specific protein feature or miRNA binding site meeting the criterion (e.g., alternatively regulated with the user supplied thresholds) from the observed number of genes, and dividing by the standard deviation of the observed number of genes (Doniger *et al.* 2003). This z-score is a normal approximation to the hypergeometric distribution. This equation is expressed as:

$$z = \frac{(observed - expected)}{std.deviation(observed)} \qquad z = \frac{\left(r - n\frac{R}{N}\right)}{\sqrt{n\left(\frac{R}{N}\right)\left(1-\left(\frac{R}{N}\right)\right)\left(1-\frac{n-1}{N-1}\right)}}$$

n = All genes associated with a given element
r = Alternatively regulated genes associated with a given element
N = All genes examined
R = All alternatively regulated genes

Section 4 – Using External Programs with AltAnalyze

While AltAnalyze is a largely a stand-alone program, some statistical analyses can be included that depend on external applications. These require prior installation of these tools using operating system specific binaries or installers and properly interfacing them with AltAnalyze.

4.1 Using APT to Perform MiDAS

Affymetrix Power Tools can be used to normalize user expression data and perform statistical analyses for alternative exon analysis. Here we discuss how to perform the MiDAS statistical analysis and incorporate these results into AltAnalyze.

Installing APT

Go to the APT download site to find the proper installation and instructions for your operating system at: http://www.affymetrix.com/support/developer/powertools/changelog/PLATFORMS.html. An example default directory to install APT on Windows is "C:/Program Files/Affymetrix Power Tools/". Once installed you can test to see that the APT installation worked by opening the APT Command Prompt, available on Windows from "Start>All Programs>Affymetrix Power Tools>APT Command Prompt". This will open a command prompt window that you can enter APT designated commands (see APT Help). Once verified, open the “Affymetrix Power Tools” folder on your hard-drive and find the versioned folder (e.g., “apt-1.4.0”), then the folder “bin” and find the file named “apt-vars.bat”. Now, open the file “Config/ default-files.csv” and see if the location listed next to the entry “APT”, under the column header “Location” is the same as the location of the “apt-vars.bat” file on your hard-drive. If not, change it to the new location. This will allow AltAnalyze to automatically open the APT command prompt when finished generating the MiDAS output files.

Running APT

In the initial GUI window “Alternative Exon Analysis Parameters”, if you select the option “Export transit results for MiDAS”, AltAnalyze will adjust its initial parameters such that it exports a series of input files for MiDAS analysis (see Section 3 – Algorithms). One of these files, has the prefix “commands-“ and has the command line for running MiDAS on these exported files to. These p-values will be stored in a new folder in the “AltResults/MIDAS” directory, with the name of the selected dataset. Once created and AltAnalzye is prompted to continue the analysis, these p-values will be included in that analysis for filtering.

If the default “apt-vars.bat” file location has been properly saved in the “default-files.csv” configuration file, the APT command prompt will automatically open when all of the MiDAS input files have been written. The first step is to change the directory that APT is looking in to the “AltResults/MIDAS” directory. This can be most easily accomplished by opening the “commands-“ file for any of the exported comparisons, copying the first line, which has the command for changing the APT directory to the “AltResults” folder, and pasting this into the APT prompt window. In Windows, this is accomplished by right-clicking on the APT window and selecting paste and then hitting the return key. If APT does not raise any issues with this command, you can proceed to create the MiDAS p-value file by pasting the second line in the file that begins with “apt-midas -c celfiles-“. This line instructs APT to use the files created by AltAnalyze to calculate MiDAS p-values. This can be repeated (no need to change the directory again) using other pair-comparison files. More information on this calculation can be found on the APT website.

Configuring R

Although installation of R is not required for any of the standard AltAnalyze analyses, for users who wish to use more advanced statistics, it will be necessary. Currently, the only statistic that requires installation of R is the linear regression method `rlm`. `rlm` is a regression statistical method apart of the R package `MASS`. This method is preferred by some users over the alternative linear regression method provided by default in AltAnalyze. Both methods produce very similar statistics, with only a few probe sets differing between threshold parameters out of hundreds of results. However, since the `rlm` method was used in the published linear regression analyses and some users may wish to replicate these results, this algorithm is available.

To run the option “linearegress-rlm” from the “Alternative Exon Analysis Parameters” window, you will need to install a compatible version R (only version

2.1 has been extensively tested). Along with R, the user may need to install the R statistical package `MASS` from within the R environment. R is interpreted by Python using the Python program `Rpy`, which should be packaged with the compiled versions of AltAnalyze but not the source code (http://rpy.sourceforge.net/download.html). Whether dealing with a compiled or source version of AltAnalzye, if Python reports that it cannot find the current version of R, the user may need to update the computers Environment Variables setting Path (Windows only). This is accessed by opening "Control Pannels>System Properties>Advanced>Environment Variables" and selecting the Variable "Path" and entering the path location of R (for example, ";C:\Program Files\R\rw2010;" – **No spaces before or after ;**) at the end of this list. Contact AltAnalyze support if problems persist.

Section 5 - Software Infrastructure

5.1 Overview

The core of AltAnalyze consists of two programs, ExpressionBuilder and AltAnalyze, which can be used in tandem or separately. The ExpressionBuilder component builds constitutive gene expression summary files as well as filters the probe set expression data prior to alternative-exon analysis. The AltAnalyze module performs all of the alternative-exon analyses, minus MiDAS p-value calculations.

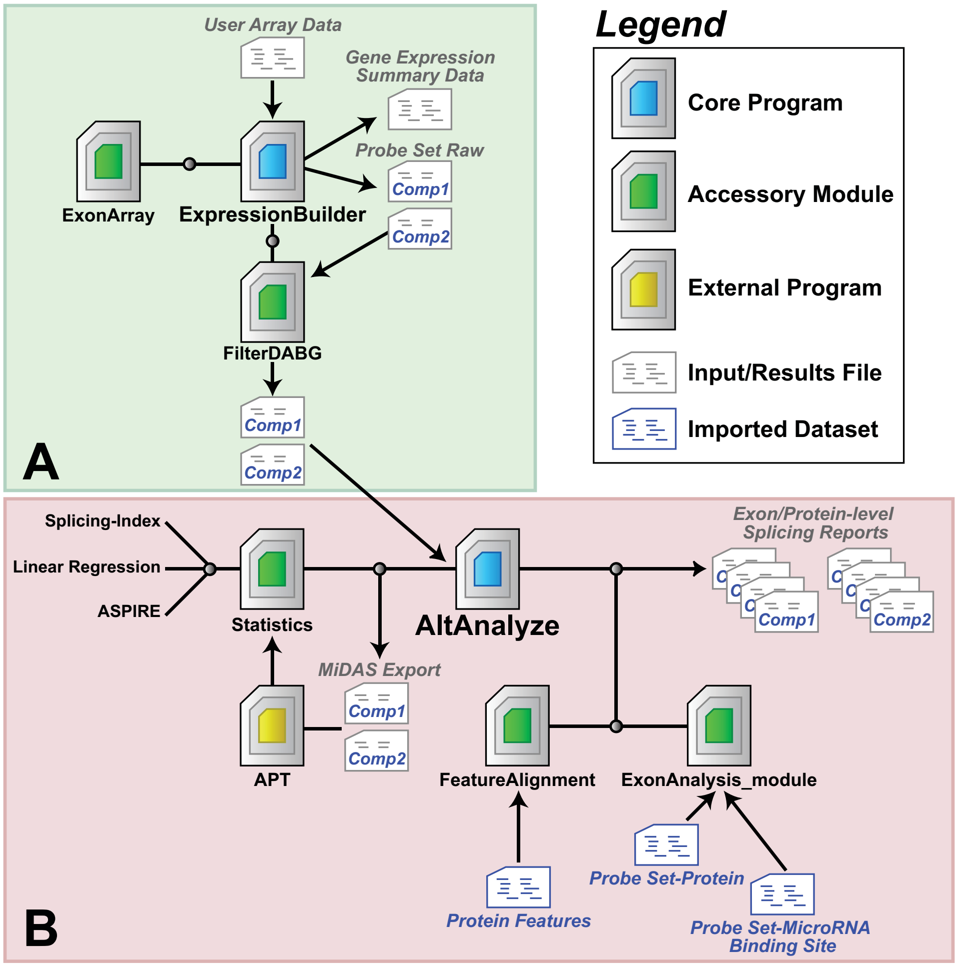


Figure 5.1. AltAnalyze Analysis Pipeline. Pictorial overview of the processing flow for AltAnalyze is depicted. The transparent green box highlights functions performed by the ExpressionBuilder function of AltAnalyze whereas the transparent red box highlights the AltAnalyze function. (A) User microarray data (probe set expression values and detection p-values) are imported into AltAnalyze via the ExpressionBuilder module, which separates data for different biological array groups into user designated pair-wise comparisons (e.g., cancer vs. normal). For each pair-wise comparison, probe set expression values and detection p-values are exported to separate files, and then analyzed by the module FilterDABG to exclude probe sets with poor detection parameters. The resulting files are stable inputs for alternative exon analysis. In parallel, a gene expression summary file is produced with Ensembl gene level expression (based on constitutive probe set expression) for each gene and array along with summary statistics (average, fold, and t-test p-value for all pair-wise comparisons) and annotations. (B) Using the ExpressionBuilder pair-wise comparison files, AltAnalyze re-calculates constitutive expression values for all probe sets linked to transcripts, evaluates changes in probe set expression relative to constitutive (Statistics module), and links probe sets with “significant” changes to aligning alternative protein sequence and predicted changes in protein and microRNA binding site architecture (ExonAnalyze and FeatureAlignment modules). The result is a series of probe set and gene summary files along with over-representation statistics for the regulation of protein and microRNA binding site features. Optionally, probe set and constitutive expression values can be exported to the external application Affymetrix Power Tools to calculate additional alternative exon statistics to be included in the AltAnalzye analysis.

5.2 ExpressionBuilder

The ExpressionBuilder program is principally designed to perform the following tasks:

1) Import user expression data from tab-delimited files.
2) Compare the imported probe set level expression data to provided expression detection probabilities (only when applicable).
3) Organize your data according to biological groups and comparisons (specified by the user from custom text files).
4) Calculate gene transcription levels for all Ensembl genes from exon or exon-exon junction array data.
5) Export raw transcription values along with folds, t-test p-values, and gene annotations for all genes and all user indicated comparisons.
6) Export the exon- or junction-level data for all pair-wise comparisons (exon array analyses are restricted to two conditions).
7) Filter the resulting exon or junction data using expression probabilities specific for the two pair-wise comparisons and user-defined thresholds (Figure 1.2).

Tasks 1-7 are all performed in order when beginning an AltAnalyze analysis. This set of processes is performed by the ExpressionBuilder program. You will notice that detection probabilities are assessed in two distinct steps (2 and 7). In step 2, import of detection p-values are for the purpose of calculating a transcription intensity value only for those constitutive probe sets (present in all or most transcripts) that show detection above background (DABG), since some probe sets will not work as well as others. If no probe sets have a DABG p-value less than the default or user supplied threshold (for at least one sample in your dataset), all selected probe sets will be used to calculate expression. If the user species to use all probe sets to determine gene expression, then all probe sets meating the filtering thresholds will be averaged to obtain this value.

In step 7, the probe set DABG p-values are examined to determine if probe sets should be included or excluded alternative splicing analysis. This step is important in minimizing false positive splicing calls. False positive splicing calls occur when an exon or junction probe set is not differentially expressed when transcription is not detected and thus can result in a transcription-corrected exon value that appears to be alternatively regulated. When ExpressionBuilder outputs the initial file containing the pair-wise comparisons of expression values, it does the same thing for DABG p-values. ExpressionBuilder uses these files to determine to determine if for a constitutive aligning probe set, both groups have a mean DABG $p<0.05$ or for a non-constitutive probe set, if one group has a DABG $p<0.05$ (default options), using the FilterDABG module. The probe sets passing the user-defined filters are exported to a new file that is ready to use for splicing analyses (“AltExpression/*array_type*/”).

Runtime of ExpressionBuilder is dependent on the number of conditions and array type being analyzed (>10 minutes for Affymetrix exon 1.0 ST arrays). If multiple comparisons are present in a single expression file, input files for AltAnalyze will all be generated at once and thus runtimes will take longer. You can skip this option if re-running an alternative-exon analysis on previously filtered ExpressionBuilder results (Figure 1.1 B), as long as the expression filtering parameters are the same.

***Note 1**: While this pipeline is mainly for use with exon or junction arrays, it is also compatible with a standard 3' Affymetrix microarray dataset to calculate folds, t-test p-value, and assign annotations to this data. This is useful when you have many comparisons in your dataset and you don't wish to manually calculate these values.*

***Note 2**: You do not need to run ExpressionBuilder if you have an alternative way of building AltAnalyze input files. To do so, your file headers for each array must have the name "group:sample_name", where your group names are different for each group and the denominator group is listed first and the numerator is listed second. Below the header line should only be probe set IDs and log2 expression values.*

5.3 AltAnalyze

The AltAnalyze program is the central module in the AltAnalyze pipeline. This software imports the filtered expression data and performs all statistical and functional analyses. This program will analyze any number of input comparison files that are in the "AltExpression" directory for that array type. The main analysis steps in this program are:

1) Import exon or junction annotations, to determine which probe sets to analyze and which correspond to known AS or APS events.
2) Import probe set-protein and probe set-miRNA associations.
3) Import protein functional annotations and corresponding sequence from Ensembl and UniProt domain-level annotation files (built outside of AltAnalyze – see the section LinkEST).
4) ***(junction array only)*** Identify which reciprocal junction-junction or exon-junction pairs to analyze.
5) Import the user expression data for the pair-wise comparison.
6) Store data for all probe sets corresponding to either a constitutive exon or selected annotations (e.g., associated with a splicing event), along with the group membership of each value (e.g., cancer vs. normal).

7) Calculate a constitutive expression value for each gene and each sample (used for splicing score later on). <u>*OPTIONAL*</u>: If the user selected a cut-off for constitutive fold change, then genes with an absolute fold change greater than this threshold will be removed from the analysis.
8) Calculate a splicing score and t-test p-value from the probe set and constitutive expression values. This calculation requires that splicing ratios are calculate for each sample (exon/constitutive expression) and then compared between groups. For exon arrays, the splicing index (SI) method is calculated for each probe set. For junction arrays, ASPIRE, Linear Regression can be used with the pre-determined reciprocal junctions or alternatively are calculated for individual probe sets using the SI method.
9) ***(junction array only)*** <u>OPTIONAL</u>: Performs a permutation analysis of the sample ASPIRE input values or Linear Regression values to calculate a likelihood p-value for all possible sample combinations.
10) <u>*OPTIONAL*</u>: Exports input for the Affymetrix Power Tools (APT) program to calculate a MiDAS p-value for each probe set. If using this option, AltAnalyze will pause while you follow the simple directions to get APT to generate this file (see details under MiDAS analysis).
11) <u>*OPTIONAL*</u>: Exports constitutive adjusted probe set expression values for external applications (e.g., clustering).
12) Retain only probe sets that meet the scoring thresholds for these statistics (splicing score, splicing t-test p, permutation p, and MiDAS p).
13) For remaining probe set link these identifiers to matching protein sequences. For exon arrays, probe sets are matched to the best matching protein (derived from an mRNA containing the probe set sequence) and the best non-matching (derived from an mRNA NOT containing the probe set sequence but corresponding to the same gene). For junction arrays, this same method is used if only one of the reciprocal probe sets aligns to an mRNA otherwise, the best matching proteins for both reciprocal probe

sets are used. If a probe set or reciprocal junctions align to two reciprocal proteins, the following steps are performed by the module ExonAnalyze:

a) Identify all protein functional region annotations corresponding to all protein IDs for that gene (e.g. kinase domain or serine phosphoserine).
b) For each functional annotation and the sequence that corresponds to it, search for the sequence within the two reciprocal protein sequences. If a functional sequence is found in one but not the other protein sequence, store this functional annotation along with which protein ID it is missing from. Note: functional annotations consist of one or more amino acids that comprise a functional sequence. If this sequence is less than 6AA, it is expanded to 6AA using flanking sequence.
c) Next, the two reciprocal protein sequences are compared to regionally where they are different (N-terminal, C-terminal, or middle). If the change is restricted to the middle of the protein (not within 12AA of either terminus) then this difference is referred to as Alt-coding. If the N-terminal sequences of both proteins are the same but the overall protein length one is half or less of the other, this change is annotated as truncation.
d) Store the functional annotations, protein sequences, and IDs differencing between reciprocal proteins.
e) Next, determine if any probe sets regulated are among those containing putative miRNA binding sites. Probe set to miRNA binding site annotations are pre-determined using the programs `MatchTargetPredictions.py` and `ExonSeqSearch.py` and stored in a local file for AltAnalyze to access.
f) Store the miRNA binding site names, sequences, and source of the prediction. These predictions are also stored with the direction of the fold change/alignment of the probe set.

14) Determine which unique genes contain regulated protein functional annotations or miRNA binding sites. Perform over-representation analysis for unique genes with a common regulated protein functional annotation or miRNA binding site, compared to all annotations examined.
15) Export over-representation statistics (miRNA binding site and protein feature) to the “AlternativeOutput” folder of “AltResults”.
16) ***(junction array only)*** Import splicing and exon annotations for regulated exons corresponding to each set of reciprocal probe sets (e.g., for E1-E3 compared to E1-E2, E2 is the regulated exon).
17) For ExonAnalyze annotations, reformat the direction/inclusion status of the annotation. For example, if a kinase domain is only found in a protein that aligns to a probe set, but was down-regulated, then the annotation is listed as (-)Kinase-domain, but if up-regulated is listed as (+)Kinase-domain.
18) Export the results from this analysis.
19) Summarize the probe set or reciprocal junction data at the level of genes and export these results (along with Gene Ontology/Pathway annotations).
20) Export overall statistics from this run (e.g., number of genes regulated, splicing events).
21) ***(junction array only)*** Combine and export the saved probe set and gene files for each comparison analyzed, to compare and contrast differences.

When finished AltAnalyze will have generated four primary files.

1) name-scoringmethod-exon-inclusion-results.txt
2) name-scoringmethod-exon-inclusion-GENE-results.txt
3) name-scoringmethod-ft-domain-zscores.txt
4) name-scoringmethod-miRNA-zscores

Here, “name” indicates the comparison file name from ExpressionBuilder, composed of the species + array_type + comparison_name (e.g.

Hs_Exon_cancer_vs_normal), scoringmethod indicates the alternative exon analysis method used (e.g., SI) and the suffix indicates the type of file.

The annotation files used by AltAnalyze are pre-built using other modules with this application or through external software not included (e.g., Ensembl API perl scripts, and SQL). Although the user should not need to re-build these files on their own, advanced users may wish to modify these tables manually or with programs provided (see Section 6 - Building AltAnalyze Annotation Files for more details).

For protein-level functional annotations, this software assumes that if an exon is up-regulated in a certain conditions that the functional region (e.g., protein domain) is also up-regulated and indicates it as such. For example, for exon array data, if a probe set is up-regulated (relative to gene constitutive expression) in an experimental group and this domain is found in the protein aligning to this probe set, in the results file this will be annotated as (+) domain. If the probe set were down-regulated (and aligns as indicated), this would be annotated as (-) domain. The opposite is true if a protein feature aligns to the non-matching protein.

Section 6 –Building AltAnalyze Annotation Files

6.1 Splicing Annotations and Protein Associations

A number of annotation files are built prior to running AltAnalyze that are necessary for:

1) Organizing exons and introns from discrete transcripts into consistently ordered sequence blocks (`UCSCImport.py` and `EnsemblImport.py`).
2) Identifying which exons and introns align to alternative annotations (`alignToKnownAlt.py` and `EnsemblImport.py`).
3) Identifying probe sets with likely constitutive annotations (`ExonArrayAffyRules.py`).
4) Identifying which probe sets align to which exons and introns (`ExonArrayEnsemblRules.py`).
5) Extracting out protein sequences with functional annotations (`ExtractUniProtFunctAnnot.py`, EnsemblAPI_script.perl).
6) Identifying miRNA binding sites (`MatchTargetPredictions.py`)
7) Matching miRNA binding site sequence to probe set sequence (`ExonSeqSearch.py`).
8) Matching probe set sequence to cDNA and EST sequences (`LinkESTSeq.py`).
9) Identify the longest matching and non-matching mRNA for each probe set and associated/predicted protein sequences (`LinkESTSeq.py`).

These annotation files are necessary for all exon and junction array analyses. Junction array analyses further require:

10) Matching reciprocal junction probe sets to annotated exons or introns (`JunctionArray.py`, `EnsemblImport.py` and

`JunctionArrayEnsemblRules.py`), creating a file analogous to (4) above.

11) Matching reciprocal junction probe sets to miRNA binding sites (`JunctionSeqSearch.py`), creating a file analogous to (7) above.

All of the associated Python programs were written specifically for AltAnalyze. With the creation/update of these files, the user is ready perform alternative exon analyses for the selected species and array type. Since many of these analyses utilize genomic coordinate alignment as opposed to direct sequence comparison, it is import to ensure that all files were derived from the same genomic assembly.

Note: Although all necessary files are available with the AltAnalyze program at installation and such files can be updated automatically from the AltAnalyze server, users can use these programs to adjust the content of these files, use the output for alternative analyses, or create custom databases for currently unsupported species.

6.2 Building Ensembl-Probe Set Associations

Exon Arrays

Affymetrix exon 1.0 ST arrays are provided with probe set sequence, transcript cluster, genomic location, and mRNA count annotations. Each of these annotations is used by AltAnalyze to provide detailed sub-gene associations. Although transcript clusters represent putative genes, the AltAnalyze pipeline derives new gene associations to Ensembl genes, so that each probe set aligns to a single gene from a single gene database. This annotation schema further allows AltAnalyze to determine which probe sets align defined exons regions (with external exon annotations), introns, and untranslated regions (UTR).

To begin this process, Ensembl exons (each with a unique ID), their genomic location, and transcript associations are downloaded for the most recent genomic assembly using the BioMart server (http://www.ensembl.org/index.html). This file is saved to the directory "AltDatabase/ensembl/*species*/" with the filename "*species*_Ensembl_transcript-annotations.txt". Since Ensembl transcript associations are typically conservative, transcript associations are further augmented with exon-transcript structure data from the UCSC genome database (http://www.genome.ucsc.edu), from the file "all_mrna.txt" (Downloads>*species*>Annotation database>all_mrna.txt.gz). This file encodes genomic coordinates for exons in each transcript similar to Ensembl. Transcript genomic coordinates and genomic strand data from UCSC is matched to Ensembl gene coordinates to identify genes that specifically associate with Ensembl genes with the Python program `UCSCImport.py`. Unique transcripts, with distinct exon structures from Ensembl, are exported to the folder "AltDatabase/ucsc/*species*" to the file "*species*_UCSC_transcript_structure_filtered_mrna.txt", with the same structure as the Ensembl_transcript-annotations file.

Once both transcript-structure files have been saved to the appropriate directory, `ExonArrayAffyRules.py` calls the program `EnsemblImport.py` to perform the following steps:

1) Imports these two files, stores exon-transcript associations, identifies exon regions to exclude from further annotations. These excluded exons signify

intron retention (overlapping with two adjacent spliced exons) and thus are no longer stored exons but as retained introns regions.

2) Assembles exons from all transcripts for a gene into discrete exon clusters. If an exon cluster contains multiple exons with distinct boundaries, the exon cluster is divided into regions, which represent putative alternative splice sites (e.g., region 1, 2, 3). These splice sites are explicitly annotated downstream. Each exon cluster is ordered and number from the first to the last exon cluster (e.g, E1, E2, E3, E4, E5), composed of one or more regions. These exon cluster and region coordinates and annotations are stored in memory for downstream probe set alignment in the module `ExonArrayAffyRules.py` (e.g, E1-1, E1-2, E2-1, E3-1).
3) Identifies alternative splicing events (cassette-exon inclusion, alternative 3' or 5' splice sites, alternative N-terminal and C-terminal exons, and combinations there of), for all Ensembl and UCSC transcripts by comparing exon cluster and region numbers for all pairs of exons in each transcript. Alternative exons/exon-regions and corresponding exon-junctions are stored in memory for later probe set annotation and exported to summary files for creation of databases for the Cytoscape exon structure viewer, SubgeneViewer.

Upon completion, `ExonArrayAffyRules.py`:

1) Imports Affymetrix exon 1.0 ST probe sets genomic locations, transcript cluster, and mRNA counts from the Affymetrix probeset.csv annotation file (e.g., HuEx-1_0-st-v2.na23.hg18.probeset.csv) for all probe sets. Although transcript clusters will not be used as primary gene IDs, these are used initially to group probe sets.
2) Transcript cluster genomic locations are matched to Ensembl genes genomic locations (gene start and stop) to identify single transcript clusters that align to only one Ensembl gene for the respective genomic strand. For transcript clusters aligning to more than one Ensembl gene,

coordinates for each individual probe set are matched to aligning Ensembl genes, to identify unique matches. If multiple transcript clusters align to a single Ensembl gene, only probe sets with an Affymetrix annotated annotation corresponding to that Ensembl gene, from the transcript.csv file (e.g., HuEx-1_0-st-v2.na23.hg18.transcript.csv) are stored as proper relationships. This ensures that if other genes, not annotated by Ensembl exist in the same genomic interval, that they will not be inaccurately combined with a nearby Ensembl gene. If multiple associations or other inconsistencies are found, match probe set coordinates directly to the exon cluster locations derived in `EnsemblImport.py`.

3) Once unique probe set to Ensembl genes associations have been defined, constitutive probe sets are identified using the Affymetrix mRNA counts provided in the program `ExonArrayEnsemblRules.py`. The mRNA counts are distributed based on the types of mRNAs they align to (full-length, Ensembl, and EST), where the probe sets with the largest number of high quality mRNA associations are chosen as constitutive. Probe sets for a given gene are ranked based on the number of: A) Ensembl; B) full-length; and C) EST transcripts associated associated with the probe set, in that order, where multiple associations are required for each annotation type. If all probe sets have the same number of Ensembl and full-length transcript associations, then the number of EST aligning are compared. If no difference in these mRNA assignments exists, no constitutive probe sets for that gene are annotated (and thus not analyzed at the level of alternative exons).

4) Each probe sets is then aligned to exon clusters, regions, retained introns, and splicing/exon annotations for that gene. In addition to splicing annotations from `EnsemblImport.py`, splicing annotations from the UCSC genome annotation file “knownAlt.txt” (found in the same server directory at UCSC as “all_mRNA.txt”) are obtained using the program `alignToKnownAlt.py`. If a probe set does not align to an `Ensemblmport.py` defined exon or intron and is upstream of the first

exon and is downstream of the last exon, the probe set is assigned a UTR annotation. All aligning probe sets are annotated based on the exon cluster number and the relative position of that probe set in the exon, based on relative 5' genomic start (e.g., E2-1). This can mean that probe set E2-1 actually aligns to the second exon cluster in that gene in any of the exon regions, if it is the most 5' aligning.

5) These probe set annotations are exported to the directory "AltDatabase/*species*/exon" with the filename "*species* _Ensembl_probesets.txt" (typically less than half of all probe sets from the array).

Junction Arrays

For the exon-junction array AltMouse, the same process is applied to the highlighted exon(s) from all pre-determined reciprocal probe sets, exported by the program `ExonAnnotate_module.py`. A highlighted exon is an exon that is considered to be regulated as the result of two alternative junctions. For example, if examining the exon-junctions E1-E2 and E1-E3, E2 would be the highlighted exon. Alternatively, for the mutually-exclusive splicing event E2-E4 and E1-E3, E2 and E3 would be considered to be the highlighted exons (actual exons spliced in or out). To obtain the genomic locations of these exons, sequences for each are obtained from a static build of the mouse AltMerge program (March 2002) (`ExonAnalyze_module.py)` and searched for in fasta formatted sequence obtained from BioMart for all Ensembl genes with an additional 2 kb upstream and downstream sequence (`JunctionArray.py` and `EnsemblImport.py`). This allows for the export of an exon-coordinate file analogous to the exon probeset.csv file. The main difference in this file is that

AltMouse gene to Ensembl ID associations are obtained by comparing gene symbol names and external GenBank accession numbers in common, as opposed to coordinate comparisons, and constitutive exon annotations are directly lifted from the “Mm_Ensembl_transcript-annotations.txt” file, obtained from BioMart. Unlike the exon array, these constitutive exon annotations are not used to determine which probe sets are most likely constitutive, since specific probe sets have been designed for this array to probe predicted constitutive features, each aligning to multiple exons. The resulting highlighted exon file is named “Mm_Ensembl_AltMouse_probesets.txt” and is saved to “AltDatabase/Mm/AltMouse”, with the same structure as its exon array analogue.

6.3 Extracting UniProt Protein Domain Annotations

The UniProt protein database is a highly curated protein database that provides annotations for whole proteins as well as protein segments (protein features). These protein feature annotations correspond to specific amino acid (AA) sequences that are annotated using a common vocabulary, including a class (feature key) and detailed description field. An example is the TCF7L1 protein (http://www.uniprot.org/uniprot/Q9HCS4), which has five annotated feature regions, ranging in size from 7 to 210 AA. One of these regions has the feature key annotation “DNA binding” and the description “HMG box”. To utilize these annotations in AltAnalyze, these functional tags are extracted along with full protein sequence, and external annotations for each protein (e.g., Ensembl gene) from the “uniprot_sprot_*taxonomy*.dat” file using the

`ExtractUniProtFunctAnnot.py` program. This program produces two files ("uniprot_feature_file.txt" and "uniprot_trembl_sequence.txt") that are saved to the appropriate "AltDatabase/uniprot" species directory. FTP file locations for the UniProt database file can be found in the file "Config/Default-file.csv" for each supported species. To improve Ensembl-UniProt annotations, these relationships are also downloaded from BioMart and stored in the folder "AltDatabase/uniprot/*species*" as "*species*_Ensembl-UniProt.txt", which are gathered by `ExtractUniProtFunctAnnot.py` at runtime to include in the UniProt sequence annotation file. These files are saved to "AltDatabase/uniprot/*species*" as "uniprot_feature_file.txt" and "uniprot_sequence.txt". Runtime is approximately 5 minutes (not including downloads).

6.4 Extracting Ensembl Protein Domain Annotations

In addition to protein features extracted from UniProt, protein features associated with specific Ensembl transcripts are extracted from the Ensembl database. One advantage of these annotations over UniProt, is that alternative exon changes that alter the sequence of a feature, but not its inclusion will be reported as a gain and loss of the same feature, as opposed to just one with UniProt. This is because protein feature annotations in UniProt only typically exist for one isoform of a gene and thus, alternation of this feature in any way will result in this feature being called regulated. Although an Ensembl annotated feature with a reported gain and loss can be considered not changed at all, functional differences can exist do to a minor feature sequence change that would not be predicted if the gain and loss of the feature were not reported.

Three separate annotation files are built to provide feature sequences and descriptions, “Ensembl_Protein”, “Protein”, and “ProteinFeatures” files (“AltDatabase/ensembl/*species*”). The “ProteinFeatures” contains relative AA positions for protein features for all Ensembl protein IDs along with feature annotations and source. The “Protein” file contains AA sequences for each Ensembl protein as well as transcript start and stop (base pairs – not used by AltAnalyze). The “Ensembl_Protein " provides Ensembl gene, transcript, and protein ID associations. Data for these files is extracted using a custom Perl script that interfaces with the Ensembl Perl application programmer interface (API). The main feature annotation sources in these files are Prosite and Pfam, which provides a description similar to UniProt. As an example, see: http://ensembl.genomics.org.cn/Homo_sapiens/protview?db=core;peptide=ENSP00000282111, which has similar feature descriptions to UniProt for the same gene, TCF7L1. In AltAnalyze, protein features, descriptions, and amino acid locations are used to store the amino sequences associated with the particular feature. Features with a size less than 6AA are expanded to 6AA using flanking sequence and are stored in memory so they can be queried against full-length protein sequences corresponding to the different isoforms predicted to be regulated based on the array data. Future versions of AltAnalyze may contain methods for directly extracting this information via Python modules, however, in the interim, updates of these files can be obtained using the update feature in

AltAnalyze, which downloads current builds of these files for new genomic assemblies. Runtime is approximately 2 days (entries are grabbed one by one over the internet), but will be extracted by alternative methods in the future (e.g., local Ensembl SQL database or static release data dump files).

6.5 Extracting miRNA Binding Annotations

To examine the potential gain or loss of miRNA bindings sites as the direct result of exon-inclusion or exclusion, AltAnalyze uses putative miRNA sequences from multiple prediction algorithms. These binding site annotations are extracted from the following flat files:

- TargetScan conserved predicted targets (http://www.targetscan.org/cgi-bin/targetscan/data_download.cgi?db=vert_42). Gene symbol and putative miRNA associations are extracted (no sequence). The primary gene ID, gene-symbol, is linked to Ensembl based on BioMart downloaded gene-symbol to Ensembl gene annotations. A sequence file is available at this site, but only designates putative seed sequence location.
- Miranda human centric predictions with multi-species alignment information is obtained from target predictions organized by Ensembl gene ID (http://cbio.mskcc.org/research/sander/data/miRNA2003/mammalian/index.html). A larger set of associations is also pulled from species-specific files (http://www.miRNA.org/miRNA/getDownloads.do), where gene symbol is related to Ensembl gene ID. Both files provide target miRNA sequence.
- Sanger center (miRBase) sequence is provided as a custom (requested) dump of their version 5 target predictions (http://miRNA.sanger.ac.uk/targets/v5/), containing Ensembl gene IDs,

miRNA names, and putative target sequences specific for either mouse or human.

- PicTar conserved predicted targets were provided as supplementary data (Supplementary Table 3) at http://www.nature.com/ng/journal/v37/n5/suppinfo/ng1536_S1.html, with conservation in human, chimp, mouse, rat, and dog for a set of 168 miRNAs. For mouse, human gene symbols were searched for in the BioMart derived “Mm_ Ensembl_annotation.txt” table after converting these IDs to a mouse compatible format (e.g., TCF7L1 to Tcf7l1).

Ensembl gene to miRNA name and sequence are stored for all prediction algorithm flat files and directly compared to find genes with one or more lines of miRNA binding site evidence using the program `MatchTargetPredictions.py`. The flat file produced from this program (“combined_gene-target-sequences.txt”) is used by the program `ExonSeqSearch.py` to search for these putative miRNA binding site sequences among all probe sets from the “*species*_Ensembl_probeset.txt” file built by `ExonArrayEnsemblRules.py` and probe set sequence from the Affymetrix 1.0 ST probe set fasta sequence file (Affymetrix) or the reciprocal junction highlighted exon sequence file (see section 6.2). Two resulting files, one with any binding site predictions and another required to have evidence from at least two algorithms, are saved to “AltDatabase/*species*/*array_type*/” as “*species*_probeset_miRNAs_any.txt” and “*species*_probeset_miRNAs_multiple.txt”, respectively.

6.6 Inferring Protein-Probe Set Associations

To obtain associations between specific probe sets and proteins, the program `LinkESTSeq.py` was written. This program takes the consensus probe set sequence for all probe sets examined by AltAnalyze ("*species*_Ensembl_probeset.txt" file), similar to `ExonSeqSearch.py`, and searches for a match among Ensembl mRNA transcript fasta formatted sequences (BioMart) and Unigene mRNAs and ESTs (ftp://ftp.ncbi.nih.gov/repository/UniGene/Homo_sapiens/Hs.seq.all). To link Unigene IDs to Ensembl, Ensembl-Unigene relationships are from downloaded from BioMart. For junction array probe sets, only a 100% sequence match is allowed, (matches may not occur do to polymorphisms between sequence sources and genomic assemblies). For exon arrays, two types of matches are acceptable; A) complete probe set match or B) last 25 or first 25 base-pair match of the probe set to the mRNA. All matches and non-matches are stored along with all mRNA transcript sequences for each accession number (built files are typically up-to 4GB in size!), to the "AltDabase/*species*" directories.

Once all sequence searches are complete, this program will search through all probe set-mRNA matches to find the mRNA with the longest sequence match and non-match for each probe-set, for different types of sequences (Ensembl, cDNA, or EST). The non-match is any mRNA for the gene associated with a probe set that doesn't contain a match to that sequence.

When complete, the longest matching and non-matching mRNAs are stored for all three sequence types. Ensembl associations are considered the highest quality sequence match, followed by cDNAs, and lastly ESTs. Although this process should ideally find the longest matching protein sequence, that search is too time-intensive, requiring determination of protein sequence length for each mRNA, for millions of mRNA sequences. However, since Ensembl is a smaller database (<100,000 mRNAs/species) and protein associations are pre-computed, protein length is also considered for choosing the longest probe set matches for probe sets linked to Ensembl proteins. At this point, most probe sets should have sequence matching data. However, to augment these associations, for probe sets without matches, mRNA associations predicted from `ExonArrayEnsemblRules.py`, are searched for and verified among UCSC mRNA transcript sequences from the file "mrna.fa" (found in the same server directory at UCSC as "all_mRNA.txt"), since these transcripts can be missing from the other sequence files. Finally, the "best", longest mRNA-probe set associations are written to the interim file "*species*_probeset-mRNA_relationships.txt", by determining if these associations first exist in Ensembl, next in any cDNAs and finally any ESTs. A lower-quality mRNA will be selected, only if it is 20% longer than a higher-quality mRNA. A goal for future versions of this program is take into account protein sequence of the longest mRNAs and try to identify matching and non-matching isoforms with the smallest number of amino acid changes between them (most specific, conservative prediction).

At this point, only two mRNAs are matched to each probe set (matching and non-matching). To align these mRNAs to proteins, the following is performed:

- Link mRNA to existing protein IDs based on relationships from Ensembl (*species*_EnsProt-Annotations.txt), RefSeq (*species*.protein.gpff), UniProt (uniprot_sequence.txt), or EntrezGene (“gene2accession.txt”).
- Link protein IDs to protein sequence using Ensembl (*species*_EnsProt_sequence.txt), UniProt (uniprot_sequence.txt), RefSeq (*species*.protein.gpff), and NCBI (rel*version*.fsa_aa.txt) sequence files.
- Predict protein sequence for mRNAs without protein annotations or protein sequence, using mRNA sequence from the Ensembl, Unigene sequence, and UCSC (mrna.fa) sequence match analysis.
- Save all valid protein-probe set associations where both a match and non-match exist to the files “probeset-protein-dbase.txt” and “SEQUENCE-protein-dbase.txt” in the directory “AltDatabase/*species*/*array_type*/”.

Putative protein sequences are derived using the function “BuildInSilicoTranslations()”, which uses the BioPython module to translate an mRNA based on all possible start and stop sites. This data is used to identify the longest putative translation that also shares either the first or last 5 AA of its sequence with the N-terminus or C-terminus (respectfully) of a UniProt protein. The N-terminal and C-terminal comparisons are only performed if there multiple protein predictions for a single mRNA with similar predicted protein lengths (within a 30% difference) that have evidence of a frame-shift. The choosen putative protein ID is named “*mRNA accession*-PEP”. The resulting files needs to be built for every new array analyzed and with every new genomic assembly (do to annotation and sequence changes at the respective databases).

6.7 Required Files for Manual Update

Below is a list of all external files referenced in the above build strategies, that are required when either building annotations for a new array or manually updating the existing annotations. To have the automated download pull down all specified files, choose the option “Update DBs” when beginning AltAnalyze. From there on, you will be presented with several options in the standard command prompt/terminal. To find or change the download location of any automated downloads, see the file “Config/Default-file.csv”.

BioMart (Manual) Downloads

Exon and Junction Array

- *species*_Ensembl-Unigene
- *species*_Ensembl-annotations
- *species*_Ensembl_transcript-annotations.txt
- *species*_Ensembl-UniProt.txt
- *species*_EnsProt_sequence.txt
- *species*_EnsProt-Annotations.txt
- *species*_ensembl_cDNA.fasta.txt

Junction Array

- Mm_gene-seq-2000_flank

NCBI (Automated) Downloads

Exon and Junction Array

- *species*.seq.all
- *species*.protein.gpff
- gene2accession.txt
- rel*version*.fsa_aa.txt

UCSC (Automated) Downloads

Exon and Junction Array

- *species*.seq.all
- Ensembl-annotations
- all_mrna.txt
- mrna.fa
- EnsProt_sequence
- EnsProt-Annotations
- ensembl_cDNA.fasta

UniProt (Automated) Downloads

Exon and Junction Array

- uniprot_sprot_*species/class*.dat.gz
- uniprot_trembl_*species/class*.dat.gz

To see example file structures for any BioMart files, you can download the existing files manually at: http://conklinwolf.ucsf.edu/informatics/AltAnalyze/AltDatabase/.

Development Team

Main Developer: Nathan Salomonis

Investigator: Bruce R. Conklin

SQL and API Scripting: Alexander R. Pico

References

Doniger, S. W., N. Salomonis, K. D. Dahlquist, K. Vranizan, S. C. Lawlor and B. R. Conklin (2003). "MAPPFinder: Using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data." Genome Biol. **4**: R7–R7.12.

Gardina, P. J., T. A. Clark, B. Shimada, M. K. Staples, Q. Yang, J. Veitch, A. Schweitzer, T. Awad, C. Sugnet, S. Dee, C. Davies, A. Williams and Y. Turpaz (2006). "Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array." BMC Genomics **7**: 325.

Srinivasan, K., L. Shiue, J. D. Hayes, R. Centers, S. Fitzwater, R. Loewen, L. R. Edmondson, J. Bryant, M. Smith, C. Rommelfanger, V. Welch, T. A. Clark, C. W. Sugnet, K. J. Howe, Y. Mandel-Gutfreund and M. Ares, Jr. (2005). "Detection and measurement of alternative splicing using splicing-sensitive microarrays." Methods **37**(4): 345-59.

Sugnet, C. W., K. Srinivasan, T. A. Clark, G. O'Brien, M. S. Cline, H. Wang, A. Williams, D. Kulp, J. E. Blume, D. Haussler and M. Ares, Jr. (2006). "Unusual intron conservation near tissue-regulated exons found by splicing microarrays." PLoS Comput Biol **2**(1): e4.

Ule, J., A. Ule, J. Spencer, A. Williams, J. S. Hu, M. Cline, H. Wang, T. Clark, C. Fraser, M. Ruggiu, B. R. Zeeberg, D. Kane, J. N. Weinstein, J. Blume and R. B. Darnell (2005). "Nova regulates brain-specific splicing to shape the synapse." Nat Genet **37**(8): 844-52.

Wheeler, R. (2002). "A method of consolidating and combining EST and mRNA alignments to a genome to enumerate supported splice variants " Algorithms in Bioinformatics: Second International Workshop, Springer Berlin / Heidelberg **Volume 2452/2002**: 201–209.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses and dissertations. Copies of all UCSF theses and dissertations will be routed to the library via the Graduate Division. The library will make all theses and dissertations accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis or dissertation to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature

6-12-08

Date