uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Paul M. Krzyzanowski
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Cellular and Molecular Medicine)
GRADE / DEGREE

Department of Cellular and Molecular Medicine
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Predicting Molecular Markers of Development and Regeneration Through Integrative Analysis of High-Throughput Genomics Data

TITRE DE LA THÈSE / TITLE OF THESIS

Michael Rudnicki
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

Miguel Andrade
CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

Ilya Ioschikhes

Madds Kaern

Steven Jones
B.C. Cancer Research Center

Theodore Perkins

Gary W. Slater
Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

Predicting molecular markers of development and regeneration through integrative analysis of high-throughput genomics data

Paul M. Krzyzanowski

This Thesis is submitted to the
Faculty of Graduate and Postdoctoral Studies
in Partial Fulfillment of the Requirements
for the Degree
Doctor of Philosophy
in Cellular and Molecular Medicine

Department of Cellular and Molecular Medicine
Faculty of Medicine
University of Ottawa

Your file Votre référence
ISBN: 978-0-494-73883-2
Our file Notre référence
ISBN: 978-0-494-73883-2

Preliminary Pages

Permissions

Regarding the use of material from "Paul M. Krzyzanowski and Miguel A. Andrade-Navarro. Identification of novel stem cell markers using gap analysis of gene expression data. Genome Biology (2007)" in this thesis:

"Under the terms of BioMed Central's Open Access Charter, all research articles are made available and publicly accessible via the Internet without any restrictions or payment by the user. PDF versions of all research articles in BioMed Central are available for download. This is a convenient way for users to print high quality copies of articles. As part of our Copyright and License Agreement, research articles may be reproduced without formal permission or payment of permission fees. ... Reproduction of figures or tables is permitted free of charge and without formal written permission from the publisher or the copyright holder, provided that the figure/table is original, BioMed Central is duly identified as the original publisher, and that proper attribution of authorship and the correct citation details are given as acknowledgment. "

BioMed Central (BMC) Reprints and Permissions

http://www.biomedcentral.com/info/about/reprintsandperm

Abstract

In biology and medicine, the expression of particular gene transcripts and proteins is often used to identify particular cell types, developmental states, or pathogenic states to which they are unique. In this context, such transcripts and proteins are termed molecular markers. Molecular markers are indispensable in many facets of scientific research and can ultimately be used to reveal the biology behind their functions as such. However, in many cases, knowledge of which markers are ideal for a given task is still an elusive goal. In particular, developmental and stem cell research strongly relies on both protein and nucleotide molecular markers for numerous aspects as cell state and identity is very important in these fields.

To identify protein-coding transcripts which may indicate novel stem cell marker proteins, a database of heterogeneous murine stem cell microarray data was analyzed using a novel clustering method. Analysis of transcripts enriched in markers demarcating undifferentiated and differentiated cells revealed that aggregate functions could be defined for markers of differentiated cells, but not for those in undifferentiated cells. In addition, the results indicated that genes expressed in mammalian stem cells and their immediate derivatives have common ancestors in the sea urchin, suggesting that mechanisms governing mammalian stem cells were established very early in evolution.

To identify novel non-coding RNA markers in a muscle system, a ncRNA prediction approach enabling the detection of longer ncRNAs was used to identify miRNA containing sites in the mouse genome. The use of EST data was determined to improve ncRNA prediction strategies during the development of this method by evaluating the performance of miRNA prediction by combining information regarding

structured RNA loci with EST data according to a miRNA biogenesis model. The methodology described herein was employed to identify and validate novel ncRNAs exhibiting potential control over Myf5 during myoblast differentiation.

Acknowledgements

Like most others, this thesis would not have come to be without the help and guidance of many people. Though practically every person I know had a hand in this through their actions or omissions, there are many I would like to specifically thank.

Firstly, I have had the luck of having two mentors throughout my graduate studies. First and foremost, I would like to thank Miguel Andrade-Navarro whom I have had the honour of working with throughout the time of my graduate studies. I thank him for his guidance and mentorship throughout the past several years, scientifically, but more importantly personally. He allowed me to explore many areas of knowledge that have shaped my view of the world. Part way through my studies, Miguel saddened our group by announcing that he would be leaving Ottawa, taking advantage of an opportunity in Berlin, Germany. This brought my second supervisor into frame, Michael Rudnicki, who complemented Miguel's mentorship by showing me that boldness, courage, and following one's gut to achieve their goals have as much a place in scientific research as in any other modern industry. Having a good hypothesis and following it with the requisite scientific rigor is only one piece of the puzzle.

In addition to these two individuals, I would not have achieved this milestone without the help of the many colleagues whom I've met in the last few years. Gareth Palidwor and Chris Porter tolerated the many times I asked for their second opinions and advice on many aspects of this work. Advice from Carol Perez-Iratxeta was indispensable during the completion of my PhD. Finally, thanks go to Enrique Muro and Matt Huska, who left Ottawa to join Miguel in Berlin.

I have also been blessed with numerous friends who helped me arrive here through their many ways of showing their support. Their efforts truly preserved my sanity. To avoid heated arguments they are, in alphabetical order: Rashed Abuodeh, Christopher Deans, Daniel Halloran, Andrew Jones, Paul Krzepisz, Iain McKinnell, Sean O'Connor, and Feodor Price.

I would like to thank my family for their belief in my being able to accomplish this, especially to Ewa & Graeme Johnson who kept me very aware that family was nearby amidst dark and snowy winter nights in Ottawa. My parents-in-law, Phyllis and James Battersby had a huge hand in making me feel welcome every time I returned to Toronto to visit. Many other extended family members also helped in ways too numerous to mention here, especially Jason and Megan Battersby.

Most significantly, I would like to thank my mother Elizabeth for always believing in me and reminding me to accomplish greater things in life. Without her hard work and sacrifice throughout the years I would not have gone very far; almost thirty years of gentle but consistent pushing can achieve miracles. I hope I have made you very proud.

Finally, I would thank my wife Melissa. Without you, I would have given up on everything long, long ago. You have shown me a glimpse of what true happiness really is and where it lies – far beyond what my tunnel vision of the past years could see. Now that this work is finished, I hope we can build that life together for the two of us – and not for anyone else. I look forward to spending many post-PhD years hand in hand with you.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

60mer	60nt oligonucleotide microarray probe
A	adenosine
Alu	Alu repetitive element
APP	Amyloid Precursor Protein
AUC	Area Under Curve
B-cell	B lymphocyte
BED file	UCSC genome annotation track
BLAST	Basic Local Alignment Search Tool
bp	base pair
canFam2	Canis Familiaris (Dog) genome, version 2
CD (gene prefix)	Cluster of Differentiation
cDNA	complementary DNA
CEL file	Affymetrix GeneChip® feature intensity file
CLL	Chronic Lymphocytic Leukemia
COE/knot	collier/knot
COUP-TF2	alias for Nr2f2
CpG	cytosine-guanine dinucleotide
CPU	Central Processing Unit
Ctdsp2	carboxy-terminal domain, small phosphatase 2
C-terminal	carboxy-terminal
Cx43	Connexin 43
CYP	Cytochrome P450 protein
-D (suffix)	differentiated
Dab2	Disabled homolog 2
dbEST	Database of Expressed Sequence Tags
dC	deoxycytidine
dG	deoxyguanosine
DGCR8	Di George Critical Region 8

Dmd	Dystrophin
DMEM	Dulbecco's modified Eagle's medium
DNA	deoxyribonucleic acid
dsDNA	double stranded DNA
dsRNA	double stranded RNA
Ebf	Early B-cell factor
EGFL7	Epidermal growth factor like domain 7
ENCODE	Encycopedia of DNA elements (consortium)
ES	embryonic stem
ESC	embryonic stem cell
EST	Expressed Sequence Tag
Ezh2	Enhance of Zeste 2
Fgf5	fibroblast growth factor 5
FISH	Fluorescence In-Situ Hybridization
FPR	false positive rate
GCRMA	GC-corrected Robust Multichip Average
gDNA	genomic DNA
GEO	Gene Expression Omnibus
Gja1	Gap junction protein, alpha 1
GO	Gene Ontology
GTP	Guanine-5'-triphosphate
GTPase	GTP phosphatase
H1N1	Influenza 1 virus, H1N1 subtype
hEB	human embryoid bodies
HEK293	Human Embryonic Kidney 293 cells
hES	human embryonic stem (cell)
hESC	human embryonic stem cell
hg18	human genome, version 18
HSC	hematopoietic stem cell

Hsp	heat shock protein
ICM	inner cell mass
ID	identifier
IFNβ	Interferon-β
IL-7	Interleukin 7
iPS cell	induced pluripotent stem cell
kb	kilobase
kcal	kilocalorie
Klf4	Kruppel-like factor 4
KLS cells	c-Kit+Lin-Sca-1+ cells
LNA	Locked nucleic acid
MaSC	mammary stem cells
Mast	mast cells
MCK	Muscle Creatine Kinase
MEF	Myocyte Enhancer Factor
Mef2	Myocyte enhancer factor 2
mESC	murine embryonic stem cells
MFE	Minimum Free Energy
μg	microgram
MHC	Myosin Heavy Chain
miRNA	microRNA
mm8	Mus musculus (mouse) genome, version 8
mm9	Mus musculus (mouse) genome, version 9
MOE430	Affymetrix GeneChip® Mouse Expression Set 430 microarray
mol	Mole
MRF	Myogenic Regulatory Factor
Mrf4	Myogenic regulatory factor 4
mRNA	messenger RNA
Msx1	msh-like homeobox 1

Myf5	Myogenic factor 5
MyoD	Myogenic Differentiation gene
MySQL	MySQL relational database or query language
Nanog	Nanog homeobox
NCBI	National Center for Biotechnology Information
ncRNA	non-coding RNA
NFE	Normalized Minimum Free Energy
Nr2f1	nuclear receptor subfamily 2, group F, member 1
Nr2f2	nuclear receptor subfamily 2, group F, member 2
nt	nucleotide
Oct4	Pou5f1
-OH	hydroxyl group
oligo	oligonucleotide
Osteo	Osteoblast
PAM	Prediction Analysis for Microarrays
Pax	paired box
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
pg	picogram
PHP	PHP: Hypertext Preprocessor
piRNA	piwi RNA
POU	Pit-l, Oct-l/2 and unc-86 conserved domain
Pou5f1	POU domain containing, class 5, transcription factor 1
pre-miRNA	precursor miRNA hairpin
pri-miRNA	primary miRNA transcript
R	R Langauge
Rab	Ras-related in brain (protein family)
RACE	Rapid amplification of cDNA ends
RefSeq	Reference Sequence (NCBI)

RISC	RNA-induced silencing complex
RMA	Robust Multichip Average
rn4	Rattus norwegius (rat) genome, version 4
RNA	ribonucleic acid
RNApol	RNA polymerase II
rRNA	ribosomal RNA
RT-PCR	reverse transcription polymerase chain reaction
RUNX	runt related transcription factor
S/N ratio	Signal to Noise ratio
SAGE	Serial Analysis of Gene Expression
SAM	Significance Analsis of Microarray
SC	stem cell
Sca-1	Stem cell antigen 1
Serpin	Serine Protease Inhibitor
SINE	Short Interspersed Repetetive Element
siRNA	small interfering RNA
SMART	Simple modular architecture research tool
Sox2	SRY-box containing gene 2
ssDNA	Single stranded DNA
SVM	Support Vector Machine
Svp	Seven up
T	thymidine
T-cell	T lymphocyte
Thy-1	Thymus cell antigen 1
tRNA	transfer RNA
U	uridine
-U (suffix)	Undifferentiated
UCSC	University of California Santa Cruz
UHN	University Health Network

UTR	Untranslated Region
Vangl2	Van Gogh-like 2

Chapter 1 General Introduction

1.1 General Overview

The general theme of this thesis covers the development of methods used to predict candidate molecular markers in the forms of either translated proteins or transcribed nucleotide sequences. It is organized into three major sections:

In the first section, we hypothesized that stem cells exhibit unique patterns of gene expression. Therefore, to identify potential protein markers with which to distinguish different types of stem cells, publicly available murine microarray gene expression data from StemBase (Perez-Iratxeta et al., 2005a) was analyzed. To this end, a clustering method was developed and used to identify markers, and the contexts in which some might function was investigated. The work and results surrounding this theme was published in 2007 (Krzyzanowski and Andrade-Navarro, 2007).

Secondly, the development of ncRNA markers was explored by using miRNA prediction as a framework. In this wholly computational section of the thesis (Chapter 3), the value of using published EST data to refine ncRNA prediction methods was investigated. It was determined that ESTs can increase the quality of miRNA predictions using a variety of methods currently employed in the literature.

Finally, to further demonstrate how EST data can be leveraged in ncRNA research, miRNAs with potential expression in developing muscle or myoblast tissues were predicted and subsequently validated with a custom tiling microarray and follow-up experiments. In addition to providing evidence for putative novel myoblast expressed miRNAs, this work identified several candidate ncRNA markers which exert control over Myf5 transcript levels, a gene governing normal myogenic differentiation. These results

justify further investigation into the roles of ncRNAs during muscle development.

1.2 Molecular Markers

1.2.1 Introduction

In the context of biomedical sciences, a molecule is defined as a marker if it enables a distinction between different traits of interest to be made. Oftentimes, the marker itself is not the subject in a line of inquiry, but it is associated with an underlying characteristic that one is interested in. In addition, multiple markers can generally be associated with any particular property. Thus, a molecular marker is an object that is correlative, but not necessarily causative, to some underlying property.

Molecular markers can be any molecular species that help to differentiate between different types of cells, subcellular structures, or cell states (e.g. developmental, stress related, pathological). Therefore, the distinction may be spatial, as in cases of different tissues and their underlying cell types, but it also may be temporal, as in populations during different stages of a biological process, such as the cell cycle or growth in an organism. When probed with a cognate RNA molecule, DNA sequences can act as markers and be used to track specific chromosomes. Expressed mRNA molecules and proteins can be used for a multitude of marker-like purposes, a major one being the prediction/detection of pathological conditions. This is well exemplified by the volume of work devoted to finding markers of cancerous cells and their varying degrees of aggressiveness (Burton et al., 2002; Coate et al., 2009; Golub et al., 1999; Perou et al., 2000; Skotheim et al., 2003; Sorlie et al., 2001).

1.2.1.1 Proteins as markers

In its most basic form, a protein marker is a marker composed of or derived primarily from peptide sequences. Many well known biological markers are simply proteins with uniqueness to a particular subcellular structure or cell type. Protein markers are in turn useful for a variety of purposes, such as indicating the onset of cell differentiation or lineage commitment (e.g. loss of Oct4/Sox2 to signal loss of embryonic stem cell pluripotency (Boyer et al., 2005); Gain of Brachyury during mesoderm formation (Wolpert, 2002); IL-7 receptor expression in the lymphoid-restricted lineage during hematopoiesis (Kondo et al., 2003)), cell sorting (Isolation of mammary stem cells using $CD49f^{+}$ (Stingl et al., 2006);Isolation of human HSCs based on $CD34^{+}$/Thy-1^{+} (Baum et al., 1992)), or microscopic visualization (Vangl2 in cells undergoing polarized division (Montcouquiol et al., 2003); Alzheimer's Disease pathology determined by observing APP and phosphorylated Tau (Bossy-Wetzel et al., 2004)).

Protein markers have several advantages, a major one being that the presence of a protein marker is closely related to its functionality as a protein. In other words, the presence or absence of a protein marker in a certain location can infer whether or not a correlated cellular process is occurring. The detection of protein species is typically achieved by antibody hybridization. This allows for detection of specific species, at times being able to resolve between proteins with different post-translational modifications (such as phosphorylated and non-phosphorylated variants of the same peptide), which for functional purposes may be completely different. On the other hand, an antibody-based protein detection strategy may be at a disadvantage when antibody generation is either impossible or yields a product with insufficient specificity.

Antibody-free detection of specific peptides can be achieved by mass spectrometry, but usually this method forgoes obtaining information about cellular localization, and also has the general disadvantage of requiring destruction of cells within a sample.

1.2.1.2 Nucleotide sequences as markers

Like proteins, nucleotide sequences can also act as markers in many contexts. In comparison to protein markers, one large advantage of nucleotide based markers is that their recognition is usually based on some method of recognizing the marker (target) sequence using a complementary probe based on sequence and not on any particular structural features. Classic examples of their application are detection of specific gDNA sequences by PCR, mRNAs by RT-PCR, and chromosomal gene mapping by FISH (Lodish, 2000; Ried et al., 1992). The sequence based recognition between probe and target allows for detection of nucleotide species with high sensitivity and specificity. On a practical level, sequence based probes can be chemically synthesized very rapidly and inexpensively.

These qualities of nucleotide markers permit numerous uses. Clinical uses benefit from the rapidity at which nucleotide probes can be developed, illustrated by the development of PCR based methods to detect H1N1 influenza virus, which had great potential to trigger a pandemic crisis in 2009 (Bolotin et al., 2009; Chidlow et al., 2010; Pabbaraju et al., 2009; Panning et al., 2009). In most cases, sensitivity and/or specificity levels for the H1N1 variant were high, with one of the most recent reports citing levels of 98.8% and 100%, respectively (Chidlow et al., 2010). Methods of detecting target marker sequences with such high levels of sensitivity and specificity are also required for non-diagnostic uses such as biological research involving high-throughput microarrays.

Well designed nucleotide microarray probes can, and indeed must, detect the presence of specific RNA species amongst all other competing molecules, given that state of the art microarrays can contain nucleotide probes numbering in the millions (Gardina et al., 2006; Steemers and Gunderson, 2005). The *in situ* detection of nucleotide markers is also useful for many purposes. For example, fluorescence in-situ hybridization (FISH) can visualize alterations in genomic DNA to identify female carriers of mutations causing Duchenne Muscular Dystrophy (Calvano et al., 1997; Voskova-Goldman et al., 1997). In a research setting, *in situ* hybridization has been extensively used to visualize RNA transcript localizations in whole organisms, a technique which has proved useful in revealing systems behind embryonic patterning in several species.

1.3 Stem cells

In this thesis, potential markers associated with different types of stem cells will be examined, which can ultimately be useful to characterize and purify these types of cells. The following section provides a background on stem cells, on some of their currently defined markers, and on StemBase, a database of stem cell gene expression data that was used to detect novel markers.

One of the most active biomedical research fields concerns stem cells: identifying them within organisms and tissues, characterizing their life cycles, and elucidating the genetic networks that control them. A stem cell can be defined as a type of cell with the capacity to divide and produce one or more daughter cells with capacity to undergo identical cell division events. The daughter stem cells are, by definition, identical in capacity to that of the parent cell. A pool of stem cells therefore possesses self renewing capacity. Through symmetric division and production of daughter cells of equivalent

potency, a population of stem cells can expand exponentially. Stem cells can also undergo asymmetric divisions to produce cells committed to differentiate into more restricted sets of cell types. Collectively, the property of cells endowed with the capacities for self-renewal and ability to differentiate into multiple downstream derivatives is generally referred to as "stemness". In part by virtue of these properties, continued development of stem cell related knowledge is believed to yield significant advances in human health, from regenerating tissues and organs to eradicating cancers.

The most primitive type of stem cell is an Embryonic Stem (ES) cell, which is derived from the inner cell mass (ICM) of a developing embryo (Figure 1-1). These cells are hypothesized to have the capacity to form any cell in the organism and are termed pluripotent. One of the earliest changes in identity of ES cells is marked by their differentiation into the three major cell lineages: mesodermal, ectodermal, and endodermal. Once this transition has taken place the cells are termed multipotent and can generate downstream progenitors of each lineage but are generally committed to a single path. The appearance of each of the three lineages is signaled by specific markers. For example, Brachyury is a key marker of the mesodermal lineage (Mitsui et al., 2003); Fgf5 is a marker of the primitive ectoderm (Cappuccio et al., 2006); while Gata4 and Dab2 are markers for primitive endoderm (Masui et al., 2008). Thus, the use of markers can be used to distinguish between different cell types and cell lineages.

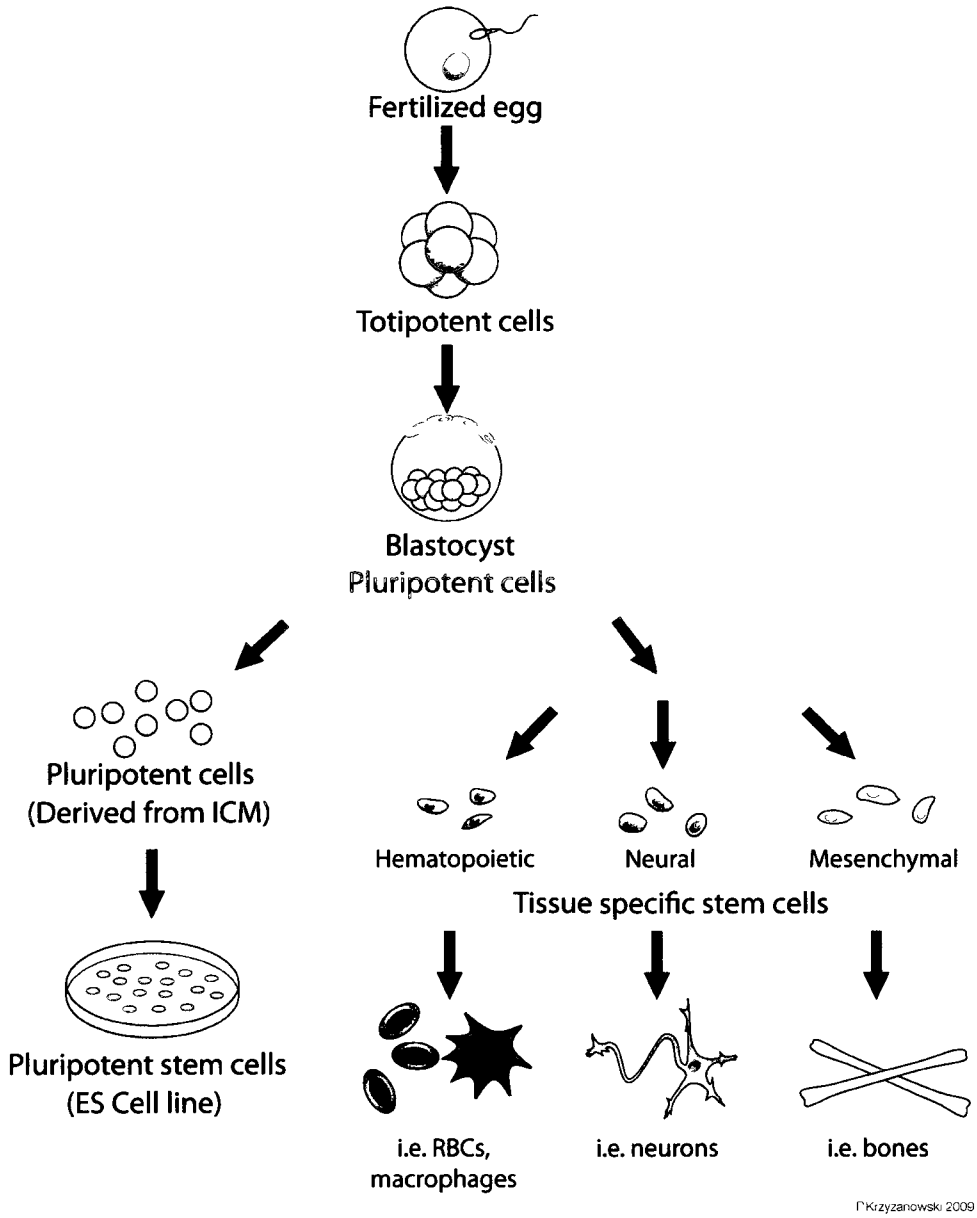

Figure 1-1: Examples of different stem cell types in mammalian development.
Following the blastocyst stage of development, tissue-specific stem cells arise and have functions in different areas of mammalian development (right). The Inner Cell Mass (ICM) of the blastocyst can be used to produce lines of cultured pluripotent stem cells (left).

Adult stem cells persist throughout the lifetime of the organism and function in normal and regenerative processes. For instance, hematopoietic stem cells residing in the bone marrow provide a life-long reservoir to produce amounts of all mature blood cell types that are limited only by the lifespan of the organism (Kondo et al., 2003). Muscle stem cells (To be further described in Section 1.5) also provide critical capacity for regeneration of muscle tissues following injury in both acute and pathological cases (Le Grand and Rudnicki, 2007; McKinnell et al., 2005). In order to better understand these processes, there is a critical need for sensitive markers that can discern stem cells and their downstream derivatives both from surrounding tissue cells and from each other. The dynamic nature of cells during growth and regenerative processes makes this need persist during different stages of stem cell existence, especially during their progression through differentiation. Efforts to identify stem-cell specific markers by high-throughput methods such as analyses of transcript or protein expression may also lay the foundation for research into the mechanisms responsible for maintaining stemness.

1.3.1 Stem cell markers

As with any cells possessing unique properties, stem cells can be identified using molecular markers, and different types of stem cells generally have different markers or sets of markers associated with them. However, although specifically identifying stem cells using molecular markers is a worthy goal, defining mechanisms important for maintenance of stemness is one of several other key scientific endpoints. In other words, identifying stem cell markers can contribute towards advancing biological knowedge of mechanisms responsible for stemness. Certain cellular markers may be shared by two or more types of stem cells: for instance, Pax6 is a marker of both neural progenitor stem

cells and retinal stem cells (Watson et al., 2009; Zaghloul and Moody, 2007). The identification of the same markers in different stem cell types may be used to infer that similar underlying biological mechanisms are active in both cell types through functional conservation.

The identification or introduction of stem cell markers can contribute greatly towards resolving outstanding scientific questions. For instance, the hematopoietic system is one with numerous markers defined for both stem cells and subpopulations of cells, and is one that has had several scientific dilemmas solved over the years, in part due to periodic advances in the knowledge of HSC associated markers and how to introduce exogenous ones if required. The clonal origin of hematopoietic stem cells was acknowledged in 1963 when abnormal karyotypes of irradiated bone marrow were used as markers to help identify cell colonies derived from single cells (Becker et al., 1963). It was not until much later, when the development of retroviral marker technology enabled experiments to follow the progeny of transplanted HSCs, that additional research would reveal the nature of single mammalian stem cell division and fate selection (Jordan and Lemischka, 1990). In 1994, fractionation of HSCs based on positive or negative expression of progenitor and lineage markers (e.g. presence or absence of c-kit, Mac-1, CD4, Sca-1, and other genes expressed in cells downstream from HSCs, generally grouped as Lin[1]) confirmed models of HSC differentiation that existed for 20 years (Cairnie et al., 1976; Morrison and Weissman, 1994). Today, manipulating HSCs for clinical purposes is routinely performed by cell sorting according to characterized HSC and lineage markers.

---

[1] Short form for Lineage.

As ideal stem cell markers are uniquely associated with the subject cell at hand, reducing or eliminating the possibility that potential markers are also associated with other, non-specific, cell types is necessary. To account for this, current modern searches for novel markers (also referred to in the literature as 'biomarkers') exploit high-throughput techniques to screen a multitude of candidates across numerous different conditions or cell types. One of the most common strategies to begin searching for novel cell markers is to analyze large libraries of microarray data, an approach used for tumour classification (Agrawal et al., 2002; Golub et al., 1999; Sorlie et al., 2003) and identifying tissue specific marker expression (Su et al., 2004; Zhang et al., 2004).

1.3.2 StemBase

In order to compare sets of gene expression data with one another, it is imperative to consider and control numerous sources of variability, as each one has an influence on the uniformity of a gene expression data set as a whole. The types of microarrays, sample preparation methods, equipment used to scan array, and final data analysis methods are all common sources of variation in expression array data sets (Irizarry et al., 2005; Quackenbush, 2006) and all must be controlled. In an ideal situation, large sets of microarray data would be identically prepared and processed, in the same laboratory or core facility by the same personnel. Nonetheless, some systematic bias can remain.

StemBase (Perez-Iratxeta et al., 2005a) is a database of high-throughput data sets derived from stem cells and their associated cell samples, such as differentiated derivatives or support cells (Figure 1-2). Growing from its inception in approximately 2003, StemBase gradually involved numerous Principal Investigators across Canada from the Canadian Stem Cell Network. Several of the goals set for the StemBase initiative

were to utilize high-throughput technology to identify genes with the ability to identify individual types of stem cells, to increase our understanding of how genes interact to control stem cells, and to identify novel uncharacterized genes critical for stem cell function (Perez-Iratxeta et al., 2005a). Identification of marker genes within each of these contexts by studying the similarities and differences between stem cells across tissues, organisms, and developmental stages would become contributory elements towards fulfillment of each of these goals.

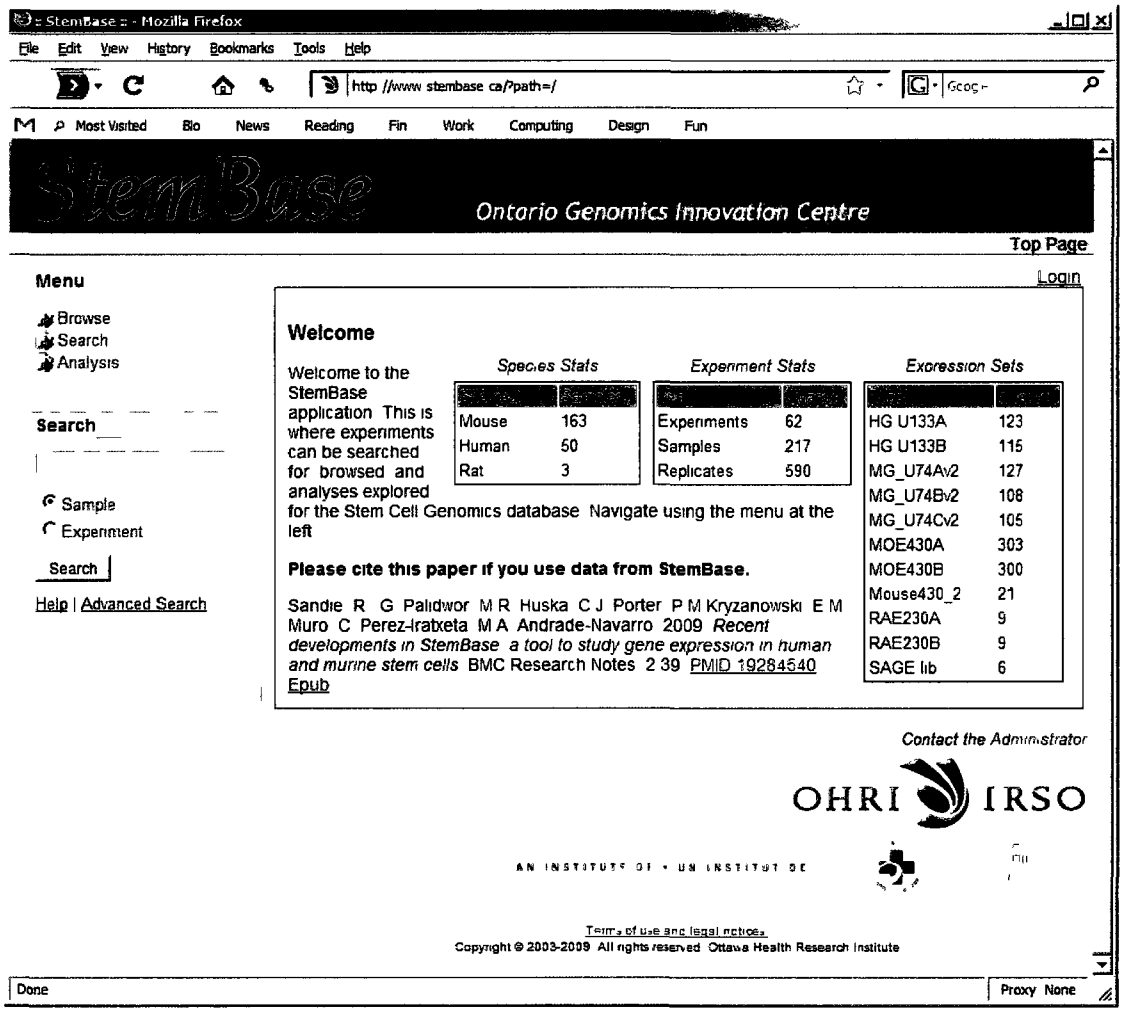

Figure 1-2: StemBase landing page

Part of the uniqueness of the StemBase expression array database is due to

microarray preparation being performed at a single facility with technical staff undergoing strict quality control measures to ensure uniformity of the large volumes of data produced (Perez-Iratxeta et al., 2005a). In fact, the facility at one point became one of the largest contributors of data to the publicly available Gene Expression Omnibus (GEO) database maintained by NCBI[2].

The data itself predominantly consists of Affymetrix microarray experiment data sets and the majority of data is of mouse origin. The quality of StemBase microarray data has been validated through its use by several third parties for many stem cell related projects, and has been used to: Identify self-renewal genes together with high-throughput siRNA knockdown data (Hu et al., 2009); compare profiles of human oocytes, hES cells, and somatic tissues in another large microarray dataset (205 samples) (Assou et al., 2009); analyze profile differences between undifferentiated and differentiated cells (Doherty et al., 2008); and construct a dedicated hESC gene expression atlas with data generated from seven other groups (Assou et al., 2007).

1.4 MicroRNAs

Later chapters of of thesis involve the computational detection of microRNAs (miRNAs) and their experimental verification as possible markers in muscle stem cell differentiation. This section provides some background on miRNAs, their prediction methods, and an overview of miRNA markers in muscle development and regeneration.

1.4.1 Background

Over the past decade, miRNAs have emerged to be important and powerful

[2] The Ontario Genomics Innovation Centre was the 7th most significant contributor on January 18th, 2006.

factors in most biological processes. Although distinct from protein coding mRNAs in their mode of operation, their primary known function is to repress the translation of target mRNAs, and thus may be considered 'translation factors' in analogy to transcription factors. Of relevance to cell growth and differentiation, miRNAs have been shown to function during key transitions between cell states, facilitating transcriptional reprogramming, and are therefore entitled to equivalent attention as possible markers of stem cells and potentially providing insight into mechanisms of stemness.

miRNAs are small (~21nt) species of non-coding RNA molecules that can regulate cognate mRNAs exhibiting sequence complementarity. (For extensive review of their biogenesis, see for example (Bartel, 2004)). In brief, miRNAs are first transcribed as long primary transcripts containing hairpin-like structures (pri-miRNAs) (Figure 1-3). Hairpins in pri-miRNAs are excised by Drosha endonuclease to form pre-miRNAs (Han et al., 2006; Yeom et al., 2006), which are then exported into the cytoplasm. This excision process is dependent on function of DGCR8 (Wang et al., 2007b).

In the cytoplasm, exported pre-miRNAs are processed by the Dicer ribonuclease to form short dsRNA duplexes. These duplexes ultimately supply the RNA Induced Silencing Complex (RISC) complex with mature miRNA strands according to the property of thermodynamic asymmetry between the two termini of the strands (Hutvagner et al., 2001; Lund et al., 2004; O'Toole et al., 2006). Mature miRNAs are thought to coordinate rapid transcriptional transitions by repressing large numbers of mRNA transcripts in parallel (Krek et al., 2005). Based on phylogenetic evidence, it is thought that mechanisms of microRNA based regulation arose before vertebrate speciation (Tanzer and Stadler, 2004).

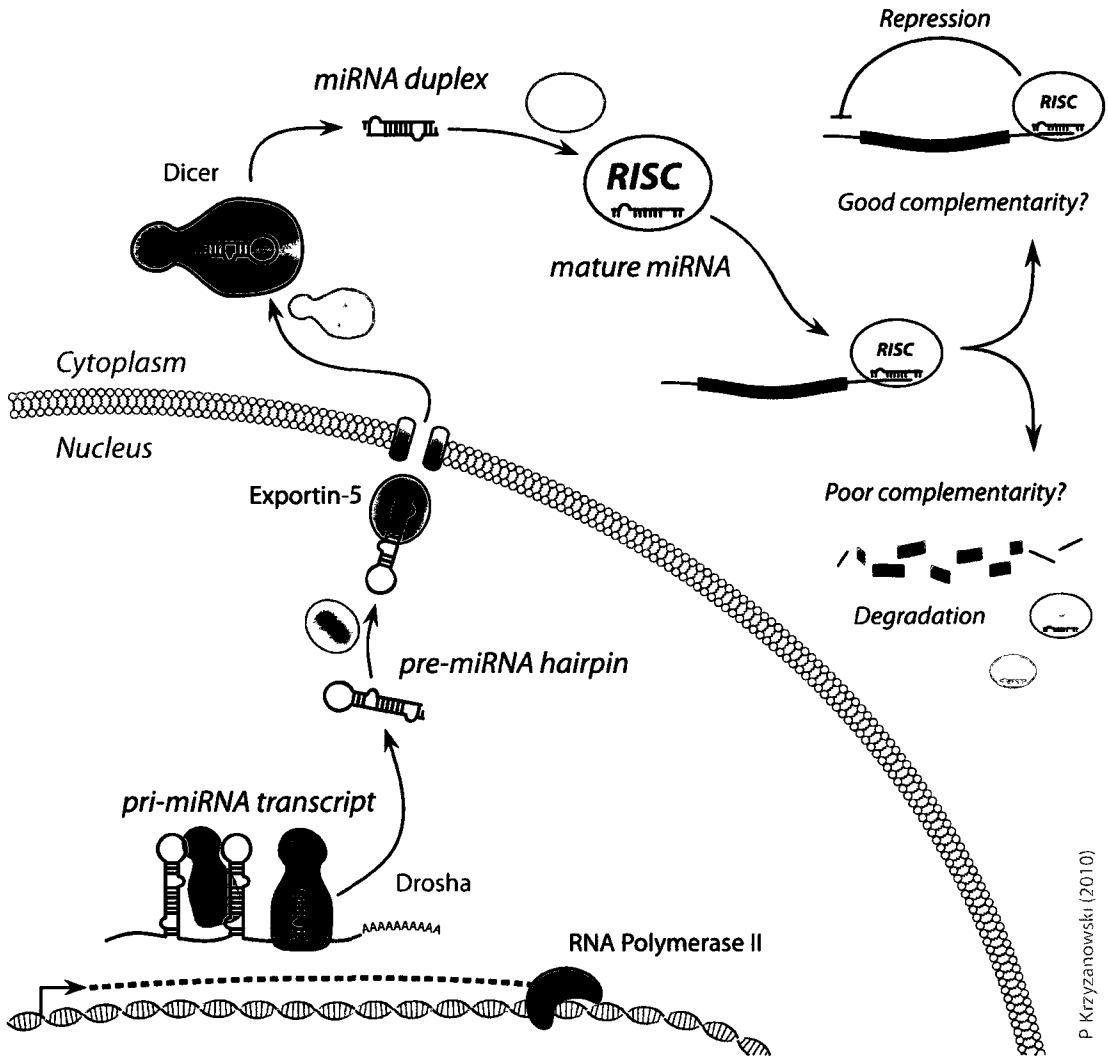

Figure 1-3: Overview of the miRNA processing pathway.
MicroRNAs are transcribed as long polyadenylated transcripts by RNA Polymerase II producing primary miRNA (pri-miRNA) transcripts. Pri-miRNA transcripts contain one or more RNA hairpin structures which are recognized and cleaved by the Drosha RNA nuclease (red) to liberate pre-miRNA hairpins. Pre-miRNAs are recognized and exported out of the nucleus by Exportin-5 (purple), where they are recognized by the Dicer ribonuclease (Blue). Dicer cleaves off the terminal loop structure at approximately 21 nucleotides from the double stranded end of the hairpins. miRNA duplexes supply the RISC containing complex with one mature miRNA strand that is used to target transcripts for repression. It is believed that transcripts with good complementarity with mature miRNAs are generally post-transcriptionally repressed while those with poor complementarity are degraded.

Due to their ability to control numerous target transcripts simultaneously,

miRNAs have emerged as regulators of gene expression in an increasing number of biological processes involving development and tissue regeneration (e.g. (Huang et al., 2008; Joglekar et al., 2007; Nguyen and Frasch, 2006)), such as differentiation control (e.g. mir-124a in neurons (Conaco et al., 2006); mir-1 & mir-133 in myoblasts (Chen et al., 2006)); regulation of cell lineage commitment (e.g. mir-155 in hematopoiesis (Georgantas et al., 2007)); and regulation of apoptosis (miR-1 & miR-133 in cardiomyocytes (Xu et al., 2007)).

The activities in many of these important biological processes have yet unexplained nuances which might be related to a current lack of knowledge regarding involvement of non-coding RNA genes. For example, identifying miRNAs required for these processes would in turn permit identification of their targets, and as the interplay between miRNAs and their targets is a direct observation of one part of cellular signaling pathways, greater understanding of specific biological processes can be obtained. Ultimately, this knowledge of basic biological processes may potentially be leveraged to control pathological processes and generate desirable outcomes such as improving human health.

1.4.2 MicroRNAs as markers

As discussed in Section 1.2, miRNAs are one type of nucleotide marker that can be used for *in situ* staining, disease marker development, and general research purposes.

The *in situ* detection of miRNA, like that of mRNA, is an important visualization tool available to molecular biologists (Figure 1-4). In this purpose, what distinguishes miRNA is that instead of observing locations of transcript transcription (but not

necessarily translation) as is the case with mRNA, one can directly visualize locations where gene repression is occurring. Knowledge of miRNA targets can help directly infer which mRNAs, and hence proteins, are being affected. The directness of the miRNA-target interaction eliminates secondary considerations when detecting mRNAs or proteins, such as whether visualized mRNAs are being transcribed or whether the proteins detected are active or otherwise require co-factors for their inferred activity (which may not be visualized by the same antibody).

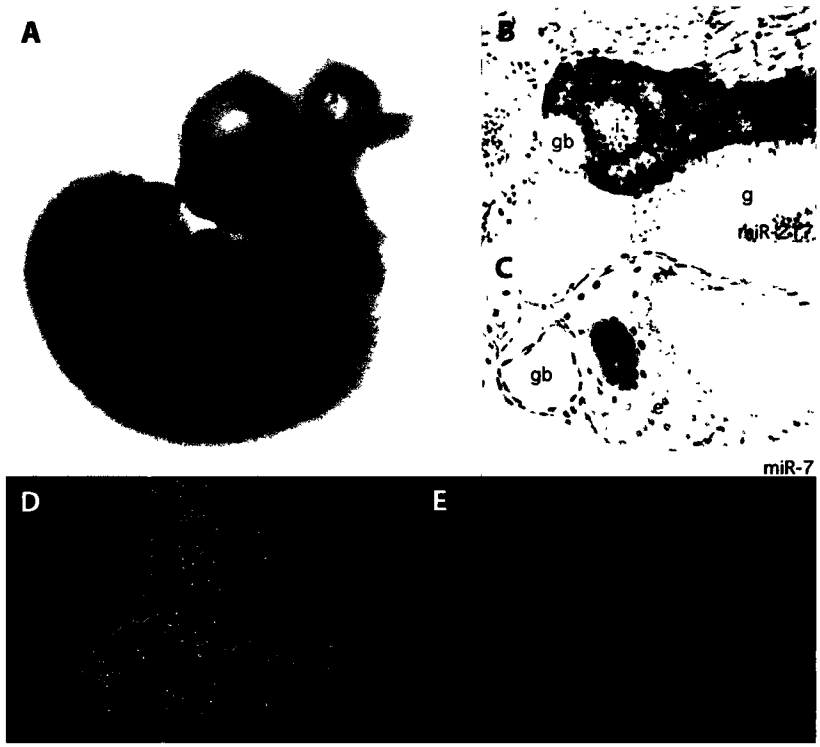

Figure 1-4: miRNA being used for in situ blotting and visualization.
miRNA can be used to mark cells and biological structures at high resolution. Mir-206 staining developing somites in a mouse embryo (A) (Wheeler et al., 2007); mir-217 (B) and mir-7 (C) expression in zebrafish exocrine pancreas and islet cells, respectively (Wienholds et al., 2005); In situ detection of miR-21 accumulation (green) in human bladder tumor (D) and normal (E) frozen tissue sections (Stenvang et al., 2008).

Initially, miRNA *in situ* hybridization was challenging due to the small size of mature miRNAs, which reduced the sensitivity of detection due to low annealing

strength. This problem was compounded by the occasional miRNA with low abundance. The development of locked nucleic acid (LNA) probes overcame this limitation by displaying over 10-fold higher hybridization efficiencies in Northern blotting assays and specificities that were intolerant of even single nucleotide mismatches (Valoczi et al., 2004). As with other nucleotide marker probes, LNA probes can be chemically synthesized according to a desired sequence very rapidly. In the result, detection of most miRNAs is generally possible making them very viable molecular markers.

These characteristics proved to be useful in research, yielding profound insights into transcriptional programs. For example, a 2005 study used LNA probes to examine the localization of 115 known miRNAs during early zebrafish development (Wienholds et al., 2005). It was observed that over two thirds (68%) were expressed in a tissue specific manner, furthermore these observations were not observed before at least partial differentiation in each of these locales had begun. This indicated that many miRNAs function as tissue specific markers at the resolution of the organism, and implied that miRNAs were dispensible during early development, at least at the level of individual tissues. The implication of these findings is that miRNAs may not be required to specify lineage choices in multipotent cells, but are critical at later stages to maintain processes involved in actively differentiating tissues and in the static tissues eventually formed.

In a clinical research context, miRNA markers can also be used to study and diagnose disease states, particularly in diseases of transcriptional dysregulation such as cancer. Stenvang *et al.* illustrated that distinct punctuate loci expressing mir-21 can be seen in bladder carcinoma, a miRNA marker established to play a role in identifying invasive phenotypes in this cancer (Neely et al., 2008; Stenvang et al., 2008) (Figure

1-4D to Figure 1-4E). In other investigations of mir-21 in cancer, broad roles in generating apoptosis resistant and invasive phenotypes have been revealed for this miRNA in glioblastoma and hepatocellular carcinomas (Chan et al., 2005; Meng et al., 2007). These examples of mir-21 illustrated the positive regulation of miRNA associated with diseases. In contrast to positive markers, the loss of miRNA expression can also mark carcinomas. For instance, let-7 miRNA was observed to be lost in a broad panel of cancer cell lines (Hammond, 2006) while in specific examination of cells derived from chronic lymphocytic leukemia (CLL) patients showed that mir-15/16 is commonly lost (Calin et al., 2002). In addition to functioning as cancer markers, loss of miRNAs can be exploited to produce therapeutic strategies. Observations of lost let-7 expression in cancer cells led to the engineering of oncolytic viruses containing let-7 target sites within viral protein coding transcripts (Edge et al., 2008). Normal cells with normal let-7 levels attenuate viral replication, but accelerated replication of the oncolytic viruses was observed in cells with abnormally low let-7 miRNA levels, leading to let-7 miRNA dependent cell killing. These are examples of how knowledge of normal and diseased miRNA marker expression patterns make miRNA based cancer markers candidates for rapid translational development as diagnostic tools or as active therapeutics.

1.4.3 MicroRNA prediction

Most, if not all, functional molecules capable of being generated in the cell have at one point or another experienced demand for methods to predict them *ab initio* from genomic sequences. Different methods have been developed for protein coding genes found in prokaryotes (i.e. GLIMMER (Salzberg et al., 1998)) and eukaryotes (i.e. GENSCAN (Burge and Karlin, 1997)), with distinct approaches being taken for ncRNAs

(i.e. QRNA (Rivas and Eddy, 2001) and Infernal (Nawrocki et al., 2009)). As a special case of ncRNAs, miRNA prediction has taken various forms over the past decade, but the general theme that the majority of methods follow is that 1) miRNA producing sites in a genome contain simple, unbranched RNA hairpin structures and 2) these hairpin structures have some specific properties that make them distinct from hairpins that arise at random. miRNA prediction studies are increasingly also using other sources of information such as evolutionary conservation data and sequence fragments derived from cloning of small RNAs. The use of Expressed Sequence Tag (EST) data for miRNA prediction will be explored in Chapter 3.

Arrival at a major consensus in miRNA gene prediction methods is yet to be achieved, and numerous methods have been published on the subject. Most miRNA prediction methods rely on the abilities of pre-existing algorithms implemented in popular software packages to rapidly predict RNA structure from primary sequence (Hofacker et al., 1994; Hofacker et al., 2004; Zuker, 1989). The results these algorithms generate are generally the most energetically favourable RNA structures within the sequence, or sequence windows, provided to the software as input. Generally speaking, the assumption being made is that folding kinetics of miRNA hairpins is not unlike kinetics of all other RNA molecules: thus, developing methods to filter output of pre-existing RNA folding algorithms is more practical than to define completely novel miRNA-like RNA folding algorithms.

One such method is MiRscan, which was introduced in 2003 (Lim et al., 2003). This tool was devised to score RNA hairpins generated from the well known RNAfold software, which is part of the Vienna RNA software suite of programs (Hofacker et al.,

1994). The innovative addition that MiRscan provided was to distinguish predicted RNA hairpins likely to harbor mature miRNAs from background RNA hairpins (which would include other structured ncRNAs and false-positive RNA structures) using seven characteristics, including: base pairing properties, interspecies conservation (limited to *Caenorhabditis elegans* and *Caenorhabditis briggsae* in the Lim *et al.* (2003) study), and geometrical aspects such as observing symmetrical RNA bulges in opposite arms of predicted RNA structures. Occurrence rates for each characteristic were converted to log-odds ratios and combined into comprehensive scores for each assessed putative mature miRNA in pre-miRNA hairpins, and the maximum score obtained for a hairpin was reported. A threshold for the score was established using a training set of positive miRNAs to calibrate individual parameters, in the end yielding a method so that both known and some predicted miRNAs were scored highly.

Another method was motivated by the observation that some miRNAs are found in clusters (Sewer et al., 2005). In this 2005 study, transcripts containing multiple consecutive predicted RNA stem loops were examined with a support vector machine (SVM) classifier trained on known human miRNAs, tRNAs, rRNAs, and mRNAs based on 40 features which, in summary, still distilled into two factors: geometrical features and sequence composition of predicted RNA hairpins. This approach was therefore fundamentally similar to the MiRscan study previously described. Another 2005 study implemented a genome-wide scan for miRNA hairpins, applying criteria such as RNA hairpin energetic stability, structural features (hairpin stem and loop lengths, free energy per nucleotide, counts of matching base pairs and unpaired bulge sizes) (Bentwich et al., 2005). However, this latter study also also considered conservation criteria using MultiZ

alignment data from the University of California Santa Cruz (UCSC) database (Karolchik et al., 2003; Kent et al., 2002; Kuhn et al., 2009; Schwartz et al., 2003). The genome-scale approach of the Bentwich *et al.* study also set a lower limit for the number of human miRNAs at 800, which was approximately quadruple the number of known and predicted human miRNAs at the time of publication. This would not have been possible without efficient methods to predict simple RNA structures like hairpins.

1.5 Muscle development and markers thereof

Muscle tissues are responsible for generating mechanical motion which is critical for biological processes required for life in higher organisms. The development and regeneration of muscle fibers is dependent on processes regulating gene transcription and controlling differentiation of stem cells. Many known genes, both protein coding and RNA based, are markers of critical junctures during these processes and play important, indispensible roles. In Chapter 4, the prediction of novel miRNAs that may be active in these processes will be discussed.

1.5.1 General development and regeneration

Three major divisions of muscle tissue in adult mammals are: skeletal muscle, generally involved in voluntary movements; cardiac muscle, which provides force to permit blood flow in the circulatory system; and smooth muscle, which is involved in many involuntary actions in the gastrointestinal tract, the iris of the eye, and within arteries and veins. Both graceful and catastrophic failures in any one of these muscle tissue types commonly result in diseases of varying severity. Most muscular diseases are degenerative, emerging with advancing age or due to a congenital defect, and in both cases the capacity for muscle regeneration is often impaired.

In higher vertebrates, most adult skeletal muscles are derived from the cells of somites, early embryonic structures that form from the paraxial mesoderm on either side of the neural tube (Christ et al., 1977). The dorsal parts of somites, called the dermomyotome, will form the dermatome and myotome later in development, giving rise to dermal layers of the back and the musculature of the limbs and trunk, respectively (Christ and Ordahl, 1995). The ventral part of each somite forms the sclerotome which, with time, is destined to yield the cartilage and skeletal structures in the core of the body such as the vertebral column and ribs.

Restricting mesodermal precursor cells to myogenic fates requires the corrdination of signals from surrounding cells. One of the first signals of cells being committed towards myogenic lineages arises in muscle precursor cells arriving at limb buds, which begin to express the myogenic regulatory factor (MRF) Myf-5 prior to any other MRF (Tajbakhsh and Buckingham, 1994). In conjunction with MyoD, the expression of these two factors marks the initial stage of cell commitment towards mature skeletal muscle, the formation of myoblasts. This is also the point in embryogenesis at which the transcriptional programming of myogenic cells becomes similar to that observed in both stationary and regenerating muscles throughout the lifespan of the organism.

To generate the volume of cells required for a normal phenotype, active proliferation of myoblast cells occurs during the embryonic stage, promoted in part by additional transcription factors such as Pax3 and Msx1 (Houzelstein et al., 1999; Odelberg et al., 2000). The latter stages of differentiation are marked by the activation of MyoD, Myogenin, Mrf4, Mef2, and finally muscle creatine kinase (MCK) and myosin

heavy chain (MHC). In the final stages of differention, fusion of myoblasts results in the production of multinucleated myofibers, which when aggregated into bundles become the familiar structure of mature muscle tissues. A reservoir of satellite cells is reserved from undifferentiated myoblasts (Figure 1-5), which ends up residing in the basal lamina between individual myofibers, and will provide the necessary capacity for muscle tissue regeneration throughout life.

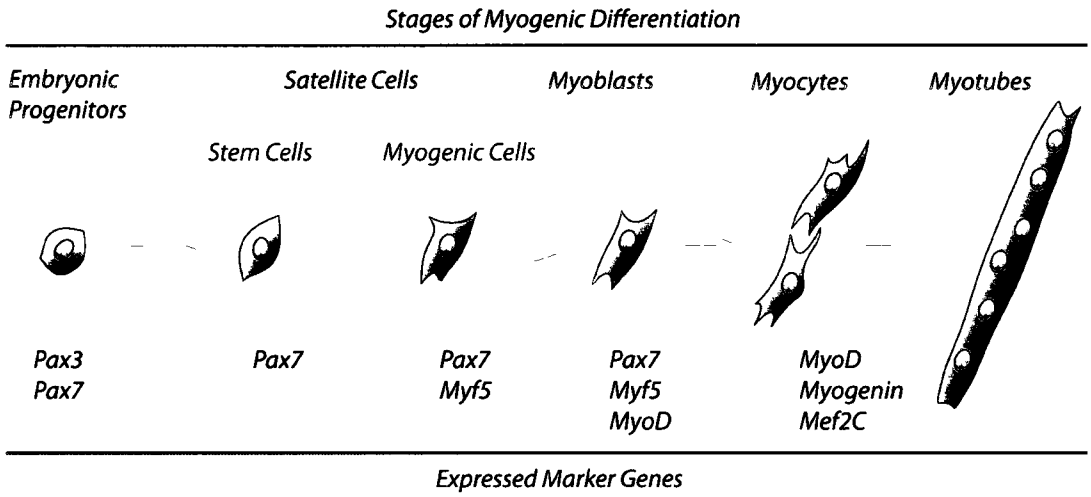

Figure 1-5: Stages of myogenic differentiation.
Various stages of cells in the myogenic differentiation hierarchy are shown, from multipotent embryonic progenitors (left) to terminally differentiated and fused myotubes (right). Key marker genes expressed at each point are shown. Based on data from (Rudnicki et al., 2008).

In the adult, regeneration of adult muscle tissue follows a similar multi-step differentiation process. As the most undifferentiated cells, Pax7 expressing satellite cells form a reservoir of cells competent to regenerate muscle (Seale et al., 2004). Without the expression of functional Pax7, satellite cells are conspicuously absent and muscle regeneration is obviously impaired. Initiating the regenerative process, activation of quiescent $Pax7^{+}$ satellite stem cells is marked by Myf5 activation. $Pax7^{+}/Myf5^{+}$ myogenic cells proliferate and enter the myoblast stage upon MyoD induction.

Continuing differentiation, myoblasts upregulate myogenin and Mef2C while downregulating Myf5 and Pax7 as they transition into myocytes. Similarly to what is seen during embryonic development, expression of myosin heavy chain (MHC) is a characteristic indicator that myocytes are fusing into myotubes and terminally differentiating.

Thus, the successful completion of both muscle development and regeneration depends on several transitions of transcriptional programming. Although key effectors at many stages have long been known to be protein-based transcription factors such as Myf5, MyoD, and Pax7, with the advent of miRNA based regulation, several equivalently indispensable non-coding RNAs have also been implicated in stages of muscle development.

1.5.2 MicroRNA roles in muscle

As is the case in numerous other developmental processes, miRNAs participate in various stages of muscle tissue growth and maintenance. microRNAs were initially reported to have roles in skeletal muscle development utilizing mouse, drosophila, and zebrafish models, and currently three major muscle-specific miRNAs (mir-1, mir-133, and mir-206) have been identified to increase in abundance during muscle cell differentiation (Brennecke et al., 2005a; Nguyen and Frasch, 2006).

The interplay between muscle miRNAs and MRFs is well documented. Mir-206 can be activated by MyoD expression (Rosenberg et al., 2006) and remains active in terminally differentiated muscle fibers (Anderson et al., 2006). The induction of myogenic differentiation in C2C12 cells leads to mir-1 upregulation (Nakajima et al.,

2006), an activation event that likely requires Myf5 as it is required for expression of mir-1 and mir-206 in the developing myotome of chicks (Sweetman et al., 2008). Interestingly, although Myf5 is one of the key regulators of early muscle development, it itself has been shown to be post-transcriptionally regulated by miRNAs in neurons (Daubas et al., 2009), raising the possibility that the same can occur in a myogenic context. FGF4 signalling in developing somites of chicken embryos has been reported to decrease mir-206 abundance (Sweetman et al., 2006). Thus, muscle related miRNA regulation is not likely a completely separate pathway distinct from protein based myogenic genes networks.

The necessity for miRNAs in myogenic processes has also been established. The most stark example is that of mir-1, which when absent during the development of *Drosophila* larvae, results in completely malformed musculature (Sokol and Ambros, 2005). In mammals, mir-206 is nearly absent in proliferating porcine satellite cells but is greatly induced during murine C2C12 differentiation (Kim et al., 2006; McDaneld et al., 2009; Rao et al., 2006). Though a mouse phenotype for mir-206 absence has not been reported, interference with this miRNA maintains cells in a cycling state (Kim et al., 2006), implying that it is a key requirement for differentiation to proceed. Another more complex example of miRNAs partaking in myogenic processes involve embedded miRNAs within several myosin transcripts (van Rooij et al., 2009). In this system, Myh6, Myh7, and Myh7b contain intronic miRNAs mir-208a, mir-208b, and mir-499, respectively, which function to repress repressors of distinct gene programs. Two of the functions imputed from the interplay between these miRNAs are to adapt cardiac gene expression in the adult in response to physiological stimuli and to affect the balance

between slow and fast muscle fiber gene programs in skeletal muscle. Despite these concrete examples, many of the precise roles for miRNAs during stages of differentiation remain to be revealed.

Although many contributors to cell transitions during myogenic differentiation are known, to our knowledge, no miRNAs have been identified that specifically regulate Myf5 or Pax7 expression changes in a myogenic context. As identifying key regulators capable of exerting specific biological functions during muscle regeneration is critical for the development of treatments for debilitating diseases such as muscular dystrophy, the discovery of novel miRNAs capable of controlling myogenic gene expression would help achieve this goal. Identifying such miRNAs might lead to technologies that enable control of the balance between intermediate myogenic cell populations, potentially leading to the development of therapeutic strategies to combat degenerative disease.

1.6 Objectives, Goals, and Hypotheses

The work presented in this thesis addresses the following three specific hypotheses and thus is organized into three aims as follows:

1.6.1 Known coding transcripts represent uncharacterized stem cell markers

Advancement in the stem cell field depends on knowledge of appropriate molecular markers for various categories and subtypes of stem cells. A database of heterogeneous stem cell microarray data (StemBase) was analyzed with a novel way of classifying stem cells in order to identify groups of protein coding transcripts with transcriptional evidence of distinguishing between distinct stem cell types and their differentiated derivatives.

1.6.2 ESTs increase the quality of ncRNA predictions

Currently published ncRNA prediction methods rely on a combination of genome-wide analyses of predicted RNA annotations and high throughput validation. To our knowledge, the existing body of work has not investigated the potential of using high-throughput transcriptional annotations to guide ncRNA prediction pipelines. As an example of high-throughput transcriptional annotations, the value of EST records in ncRNA prediction was investigated using miRNAs as an example of ncRNAs.

1.6.3 Novel miRNAs are involved in myogenic differentiation

miRNAs have documented ability to influence specific myogenic protein markers and the overall behaviour of myoblasts during differentiation. However, the number of currently known myogenic miRNAs is not extensive. To expand the number of myogenic miRNA markers available for muscular research and to contribute to

knowledge pertaining to muscle biology, a miRNA prediction pipeline was employed to identify novel miRNAs affecting one or more key regulators of myogenic transcriptional transitions such as Pax7, Myf5, and MyoD1.

Chapter 2 Identifying protein coding transcripts as candidate stem cell markers

2.1 Preface

Sections of this chapter were reproduced from the following publication:

Paul M. Krzyzanowski and Miguel A. Andrade-Navarro. **Identification of novel stem cell markers using gap analysis of gene expression data.** Genome Biology; 8(R193); 2007.

Author contributions: PMK and MAA designed the project. PMK performed the analyses. PMK and MAA analyzed data and drafted the manuscript.

2.2 Abstract

We describe a method for the detection of marker genes in large heterogeneous collections of gene expression data. Markers are identified and characterized by the existence of demarcations in their expression values across the whole dataset which suggest the presence of groupings of samples. We apply this method to DNA microarray data generated from 83 mouse stem cell related samples and describe 426 selected markers associated with differentiation to establish principles of stem cell evolution.

2.3 Background.

Gene expression microarrays allow thousands of transcripts in a cellular sample to be quantified simultaneously. (For reviews of the technology and applications, see (Heller, 2002; Stoughton, 2005)). Continuing improvements in microarray technology, in terms of transcript density, technical robustness, and cost, have led to widespread usage of arrays in experiments. The size of single studies has grown and can encompass the analysis of up to hundreds of arrays simultaneously (Su et al., 2002; Wang et al., 2005; Zhang et al., 2004). This vast explosion of reusable data being generated has resulted in efforts being directed to produce expression data repositories where the data are curated and presented in an ordered fashion (Barrett et al., 2005; Perez-Iratxeta et al., 2005a; Sherlock et al., 2001). The large number of data points makes such resources an exceptional source of biological information.

Some common uses of gene expression data are the identification of co-regulated genes across many samples (Eisen et al., 1998), identification of differentially expressed genes in samples of interest (Tusher et al., 2001), and more recently, analysis of alternative splicing (Johnson et al., 2003; Pan et al., 2004; Yeakley et al., 2002) and genome-wide surveillance of transcription (Bertone et al., 2004; David et al., 2006; Shoemaker et al., 2001). They can also be used to identify marker genes associated with specific sets of samples. As distinguishing features, such markers can be used as diagnostic tests for disease (Sieuwerts et al., 2006; Wang et al., 2004), or for the identification and purification of particular cell types (Inacio and Fonseca, 2004; Singh et al., 2004). The identification of multiple markers for a particular phenotype may also reveal biological mechanisms by which certain genes act in concert.

A simple method to identify marker gene candidates is to identify genes that are differentially expressed between a set of control samples and samples from a condition of interest. A two-state comparison can be made and genes associated with each type of sample can be identified and used as markers. Current gene expression databases typically contain data from many types of samples, and this heterogeneity provides the potential for more powerful analyses. One can, for example, identify transcripts that are specific to a sample (or samples) of interest, or make novel comparisons of different combinations of transcription profiles. The increased size of the databases also increases the number of possible two-state comparisons exponentially, which poses a computational problem. Overcoming this problem requires the aid of a computational method.

We have developed a methodology that uses large heterogeneous gene expression datasets to identify genes that can function as markers. In summary, we examine the distribution of expression values of each probe set to identify gaps. These gaps can be used to partition the database into groups of low- and high-expressing samples, which suggest the existence of distinct subpopulations of samples. We then score other probe sets based on their ability to reproduce these database partitions. The characteristics of samples in each database partition identify the context in which genes may act as markers, which aids in the subsequent evaluation of genes in terms of their putative marker roles.

In this work, we illustrate our methodology in the analysis of a database of stem-cell related DNA microarray samples which we previously developed (StemBase, (Perez-Iratxeta et al., 2005a)). In particular, we study 83 mouse stem cell related samples

analyzed with the Affymetrix MOE430 genechip set, which includes approximately 45,000 probe sets. Unbiased application of the method produces a set of 4,449 cell and tissue markers including 45 of 71 known stem cell markers (69%). Analysis of the markers that segregate six types of stem cells (hematopoietic, mast, mammospheres, osteoblasts, and two embryonic) from their differentiated counterparts suggests 426 high confidence markers, 206 highly expressed in the stem cell and 222 highly expressed in the differentiated counterpart (two being highly expressed in stem cells in some cases, and in the differentiated counterpart in others). Of those 426 markers, 17 are involved in multiple distinct lineages that include at least one non-embryonic cell type; 9 markers are highly expressed in the stem cells, 6 are highly expressed in the differentiated cells, and 2 show opposite variation in different stem-derivative cell pairs. Analysis of the functions of the 222 genes highly expressed in the differentiated cells indicates enrichment of extracellular gene products and enzyme inhibitors (12 genes, 5 of them serpins).

The set of 426 stem cell markers allows us to focus on gene superfamilies which have undergone repeated gene duplication events for a phylogenetic analysis of the evolution of proteins involved in stem cell function. By sequence similarity analysis we identify four such families (nuclear receptors, cytochrome P450, Rab family GTPases, and early B-cell factors) with multiple members in this set. The study of examples from each reveals multiple events of gene duplication along the vertebrate lineage giving rise to genes with a very high degree of sequence similarity, but very different patterns of expression in stem cells. This leads to a hypothesis that many stem-cell related genes expressed in particular tissues arose by duplication and specialization of stem-cell related genes originally expressed in other tissues. Superfamilies with large rates of duplication

in the vertebrate lineage may have functions related to the development of an increasingly complex organism, including the generation and control of tissue-specific stem cell pools.

2.4 Results.

We applied our method to a set of DNA microarray data from 83 samples from mouse stem cells and derivatives. Samples included embryonic, hematopoietic, mammosphere, retinal, neurosphere, adipose, and muscle cells (See Table 2-1). All data were obtained using the Affymetrix MOE430 platform and subjected to quality controls as previously described in (Perez-Iratxeta et al., 2005a).

We used the nonparametric Mann-Whitney U test to identify series of expression values able to produce two-state classifications in the microarray expression data. At a U-score cutoff of 0.9 our method identified 893 different two-state classifications by which to segregate the 83 sample data set. Figure 2-1 displays three exemplar distributions of hybridization values for three probe sets which clearly segregate the data set and should be desirable if selected as markers. Interpretation of the segregation pattern becomes obvious from the distribution itself as the group of samples with low gene expression values is separated from the group of samples with high expression values by a gap. Simplicity of interrogation allowed us to create a web tool accessible via the online version of the published manuscript to query these classifications so that users can determine if a given gene is a marker, the samples selected by a marker, and the markers separating two sets of samples of choice.

Class	Samples	Replicates	SampleIDs
Adipose derived stem cells	1	3	S199
Dermis derived stem cells	1	3	S200
Embryonal carcinoma	4	12	S129, S130, S131, S132
Embryonic	8	23	S219, S220, S164, S165, S166, S167, S168, S169
Embryonic fibroblasts	2	6	S180, S286
Embryonic stem cell differentiation	35	105	S153, S154, S155, S156, S157, S158, S159, S181, S174, S241, S175, S206, S207, S208, S209, S210, S211, S212, S213, S215, S216, S217, S242, S243, S244, S251, S252, S245, S250, S246, S247, S248, S249, S127, S128
Hematopoietic	12	33	S233, S234, S235, S236, S237, S147, S291, S292, S293, S294, S295, S296
Mammary	5	15	S255, S256, S311, S312, S313
Muscle derived stem cells	3	7	S274, S184, S197
Neural	3	10	S271, S272, S198
Osteoblast differentiation	7	21	S185, S186, S188, S190, S192, S194, S196
Retinal derived stem cells	2	3	S232, S240

Table 2-1: Set of mouse samples selected for our analysis from StemBase.
Various stem cells (SC) are represented in the data set. All original data including sample and experiment descriptions are accessible from the StemBase homepage.

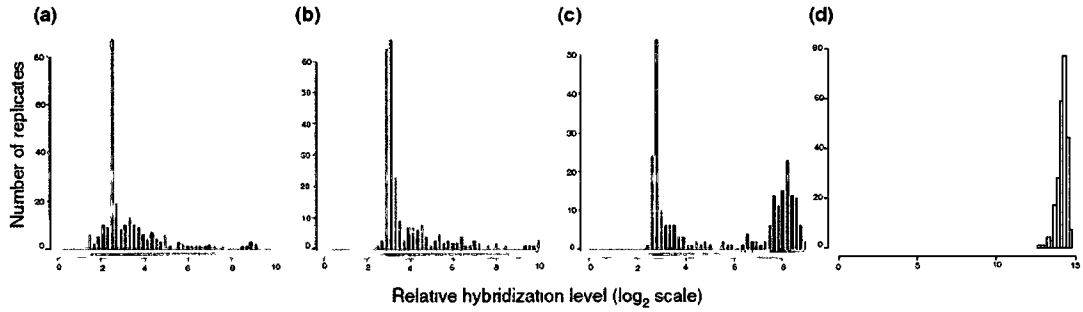

Figure 2-1: Distributions of hybridization values for probe sets.
Each histogram depicts the number of replicates (from a total of 241) with a given hybridization value for a given probe set. For illustrative purposes, we display the distribution of hybridization values of three probe sets selected by the gap method as markers corresponding to (a) a known neural stem cell marker (Nestin, probe set 1449022_at), (b) a novel stem cell marker encoding a protein of known function (phospholipase Pla2g7, probe set 1430700_a_at) which we observed as upregulated in bone marrow mast cell precursors and in undifferentiated mammospheres, and (c) a novel stem cell marker corresponding to an uncharacterized transcript upregulated in undifferentiated V6.5 and J1 mESC (2410146L05Rik, probe set 1460471_at). Details on these three cases can be obtained through our online webserver by viewing group numbers 73, 497, or 265, respectively. (d) For comparison, we show the distribution for a housekeeping gene (Eef1a1, probe set 1424635_at), a ribosomal translation elongation factor protein, which is always expressed and was not classified as a marker in our analysis. For robustness, the segregation of the samples (indicated by red and blue bars in the three marker distributions for the down and up-regulated groups, respectively) is derived by analysis of the global set of patterns (see Methods) and might not correspond perfectly to the distribution observed here. See for example a single replicate in the distribution of Nestin which is left of the gap in the distribution but was (correctly) associated to the upregulated (blue) group.

2.4.1 Properties of the set of potential markers

Classifications contained varying numbers of probe sets. Figure 2-2A shows that patterns with small number of markers are more numerous. The 893 classifications also separated different numbers of samples into groups. Most patterns assigned small numbers of samples into "upregulated" groups (e.g. the gene was highly expressed in three samples vs. 80). Over 80% of the patterns separate 10 or fewer samples from the remainder of the database as a highly expressed group (see Figure 2-2B). The complete set of putative markers associated to the 893 patterns includes 10,401 probe sets, or

approximately 25% of the probes on the microarray, which is intuitively quite a large fraction

We expected that patterns defined by smaller numbers of marker genes would be more likely to include genes important for stem cell related functions To investigate this, we examined the distribution of known stem cell markers in this set and whether the method was able to preferentially select them within small clusters

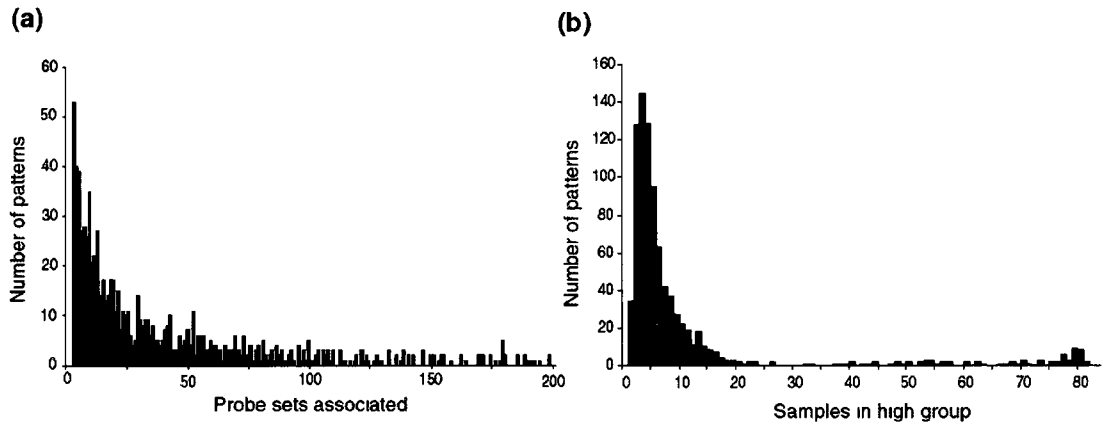

Figure 2-2 Properties of the set of 893 patterns
(a) Number of patterns with a given number of probe sets associated with a score of 90% (b) Number of patterns with a given number of samples segregated in the high expression group

2.4.2 *Known stem cell markers in dataset*

We investigated whether known stem cell markers were identified in our data set in order to understand the properties and usefulness of the selected patterns in the context of stem cell research

By examination of the literature, we selected 88 marker genes that represented the variety of stem cell types in our data set (see Table 2-2) We identified the corresponding entries in the Entrez Gene database for 72 of the 88 marker genes In seven of the

remaining cases we were unable to definitively identify the correct gene through ambiguity of the provided gene identifier (for example, "laminin" could indicate one of several possible genes), and in nine cases the identifier could not be related to any entry in the database (for example, "neuralstemmin"). Of these 72 EntrezGene IDs, 71 had at least one associated probe set on the Affymetrix MOE430A/B chip set. Our set of 10,401 "potential marker" probe sets, which contains patterns with a maximum of 200 associated probe sets contained probe sets for 49 of these 71 marker genes (69%).

Reference	Cellular type	Gene name	Entrez Gene	Probe set polled	Pattern	Score	D
(Tropepe et al , 2000)	differentiated retinal	309L	?	-	-	-	-
		Rho1D4 (rhodopsin)	?	-	-	-	-
		D2P4 (rhodopsin)	?	-	-	-	-
		CHX10	Chx10	1419628_at A	Not Found		
		PKC	*	-	-	-	-
		ROM-1	Rom1	1448996_at A	879	0 942	
	Photoreceptor specific homeobox	Crx	Crx	1418705_at A	Not Found		
	Muller glia	10E4	?	-	-	-	-
(Uchida et al , 2000)	hCNS-SC	CD24 -/lo	CD24a	1416034_at A	573	0 987	2
		CD34 -	CD34	1416072_at A	72	0 990	2
		CD45 -	Ptprc	1422124_a_at A	427	0 987	
	human neuronal lineage	N-CAM	Ncam1	1426864_a_at A	318	1 000	
	Neural	CD133	Prom1	1419700_a_at A	573	0 930	2
	Hematopoietic	CD133	Prom1	1419700_a_at A	573	0 930	2
(Kee et al , 2002)	proliferating neural	Ki-67	Mki67	1426817_at A	8	0 991	
(Mitsui et al , 2003)	Trophoblast	Cdx2	Cdx2	1422074_at A	Not Found		
	Ectoderm	Fgf5	Fgf5	1438883_at B	442	1 000	
	Neuroectoderm	Isl1	Isl1	1422720_at A	83	0 931	
	Pluripotent SC	Nanog	Nanog	1429388_at A	390	0 947	
		Oct3/4	Pou5f1	1417945_at A	151	1 000	
		Rex1	Zfp42	1418362_at A	547	0 964	
	Mesoderm	Brachyury	T	1419304_at A	232	0 938	
(Pevny and Rao, 2003)	Neural SC	Hu	?	-	-	-	-
		Neuralstemmin	?	-	-	-	-
		ABCG2	Abcg2	1422906_at A	Not Found		3
		LeX/SSEA-1	Fut4	1455843_at B	771	0 924	
		Musashi (Msi1)	Msi1	1421409_at A	Not Found		2
		Sox-1	Sox1	1438729_at B	542	0 958	
		Sox-2	Sox2	1416967_at A	71	0 986	
(Charge and Rudnicki, 2004)	Muscle specific	MCK, muscle creatine kinase	Ckm	1417614_at A	104	1 000	
		MHC, myosin heavy chain	Myh4	1427026_at A	Not Found		
	myofiber sarcolemma	Dystrophin	Dmd	1417307_at A	835	0 958	
	basal lamina	Laminin	*	-	-	-	-
	Myogenic lineage	Myf5	Myf5	1420757_at A	Not Found		
		MyoD	Myod1	1418420_at A	32	1 000	
	late myogenic lineage	MRF4	Myf6	1419150_at A	Not Found		
		Myogenin	Myog	1419391_at A	104	1 000	
	Satellite cells	Pax7	Pax7	1452510_at A	Not Found		
(Rao, 2004)	Neural restricted precursors	MAP2	Mtap2	1434194_at B	108	1 000	2
		Beta3 tubulin	Tubb3	1415978_at A	Not Found		
(Toma et al , 2005)	SKP	Fibronectin	Fn1	1426642_at A	681	0 984	
		GAP43	Gap43	1423537_at A	363	0 959	
		MAP2	Mtap2	1434194_At B	108	1 000	2
		Nestin	Nes	1449022_at A	73	0 980	2
		p75NTR	Ngfr	1421241_at A	Not Found		-
		Vimentin	Vim	1450641_at A	770	0 977	
(Yano et al , 2005)	HSC	BCRP1	Abcg2	1422906_at A	Not Found		3
	skin SP	BCRP1	Abcg2	1422906_at A	Not Found		3
	epidermal SC-SP	alpha6-integrin	Itga6	1422444_at A	Not Found		2
		beta1-integrin	Itgb1	1426918_at A	13	0 967	2
		keratin 14	Krt1-14	1460347_at A	386	0 983	
		Sca-1	Ly6a	1417185_at A	637	0 929	2
	epidermal SP	CD34-	CD34	1416072_at A	72	0 990	2
		E-cadherin	Cdh1	1448261_at A	48	0 993	
		Keratin 19	Krt1-19	1417156_at A	268	0 989	2
		CD71-	Tfrc	1452661_at A	845	0 958	

Table continues on following page…

Reference	Cellular type	Gene name	Entrez Gene	Probe set polled	Pattern	Score	D
(Inoue et al , 2006)	Muller glia, retinal	Gln synthetase	Glul	1426235_a_at A	573	0 901	
	retinal	syntaxin	*	-	-	-	-
		Pax6	Pax6	1419271_at A	374	0 960	
		rhodopsin	Rho	1425171_at A	Not Found		-
(Strem et al , 2005)	Neural	(NeuN) Neuron specific protein	?	-	-	-	-
		Neuron-specific enolase	Eno2	1418829_a_at A	Not Found		-
	osteoblasts	Alkaline phosphatase	Akp2	1423611_at A	Not Found		-
		BMP2	Bmp2	1423635_at A	Not Found		-
		BMP4	Bmp4	1422912_at A	Not Found		-
		BMP Receptor 1	Bmpr1a	1425492_at A	581	0 939	
			Bmpr1b	1437312_at B	446	0 917	
		BMP Receptor 2	Bmpr2	1434310_at B	602	0 954	
		PTH receptor	Pthr2	1452129_at A	Not Found		-
		Type 1 collagen	*	-	-	-	-
		bone sialoprotein	Ibsp	1417484_at A	Not Found		-
		PTH receptor	Pthr1	1417092_at A	368	0 964	
		RunX-1	Runx1	1422865_at A	295	0 982	
		osteonectin	Sparc	1448392_at A	581	1 000	
		osteopontin	spp1	1449254_at A	737	0 996	
	general stem cell factor receptor	CD117	Kit	1459588_at B	243	0 912	
	Muscle	merosin	Lama2	1426285_at A	799	0 986	
	cartilage related ECM	aggrecan	Agc1	1449827_at A	Not Found		-
		collagen II	*	-	-	-	-
		collagen IV	*	-	-	-	-
		PRELP	Prelp	1416322_at A	231	0 958	
	Adipose derived SC	CD49d+	Itga4	1421194_at A	Not Found		-
		CD106-	?	-	-	-	-
(Kalirai and Clarke, 2006)	Mammary SC	Bmi-1	Bmi1	Not on array		-	
		p21	Cdkn1a	1424638_at A	400	0 967	
		CD49f	Itga6	1422444_at A	Not Found		2
		Cytokeratin 19	Krt1-19	1417156_at A	268	0 989	2
		Sca-1	Ly6a	1417185_at A	637	0 930	2
		Musashi (Msi1)	Msi1	1421409_at A	Not Found		2
		Cytokeratin 5/6	?	-	-	-	-
(Shackleton et al , 2006)	Neural enrichment	CD24	CD24a	1416034_at A	573	0 990	2
	Skin	CD29	Itgb1	1426918_at A	13	0 970	2
(Kohno et al , 2006)	Neural	GFAP	Gfap	1426508_at A	249	0 924	
		Neurofilament	Nefl	1426255_at A	151	0 935	
		Nestin	Nes	1449022_at A	73	0 980	2
	Photoreceptors	Recoverin	Rcvrn	1450215_at A	Not Found		-
	epithelial lineage	Cytokeratin (904-clone 34betaB4?)	*	-	-	-	-
		CK18	Krt1-18	1448169_at A	83	0 917	
		AE1	Slc4a1	1416464_at A	41	1 000	
(Xia et al , 2006)	epithelial lineage	AE3	Slc4a3	1418485_at A	391	0 931	

Table 2-2· Stem Cell markers

Note that some markers may be included in multiple rows of the table (as many as indicated by the number in column 8) since a number of genes have been identified as markers of more than one cell type e g Abcg2 for hematopoietic and neural cells ((Pevny and Rao, 2003, Yano et al , 2005)), or Nestin for neural and skin-derived precursors (SKPs) ((Kohno et al , 2006, Toma et al , 2005)) Some of the markers could not be assigned to an Entrez gene (column 4) either because the marker name was ambiguous (indicated with '*') or absent from the database of gene names (indicated with '?') Patterns (column 6) can be examined online via the web server All patterns for which a polled probe set (column 5) is a marker can also be examined at the web server Note that the chip containing the probe set is indicated by ' A' or ' B' appended to the identifier since some probe sets are included in both the MOE430A and the MOE430B chips

As previously observed, we obtained numerous patterns which segregate small subsets of samples and that are defined by small numbers of genes. To test our hypothesis that these would tend to contain relevant genes (here, genes useful to characterize stem cells) we examined the recall and precision of our method for the 71 known markers as the maximum marker list length was reduced from 200 to 3 (Figure 2-3). Precision measures the number of true positive markers divided by the number of transcripts identified as markers, which recall measures the number of true positive markers identified divided by the total number of known positives. Limiting the list to patterns defined by 63 or fewer markers reduced the total number of probe sets assigned marker roles from 10,401 to 5,848 (44% reduction), while losing only four known stem cell markers; a recall rate of 45/71 (63%) (Point marked with a circle in Figure 2-3). This supports our theory that marker genes are more often contained in small clusters.

The 705 patterns defined by 63 or fewer markers segregated a mean of 9 samples (median = 5) in the upregulated group and were associated with a mean of 21 markers (median = 15). This set associates 5,848 probe sets (4,449 genes) with at least one pattern; approximately 13% of the probe sets on the MOE430 microarray platform. We propose that many of those can be developed into useful markers. We define this as the "selected marker" set.

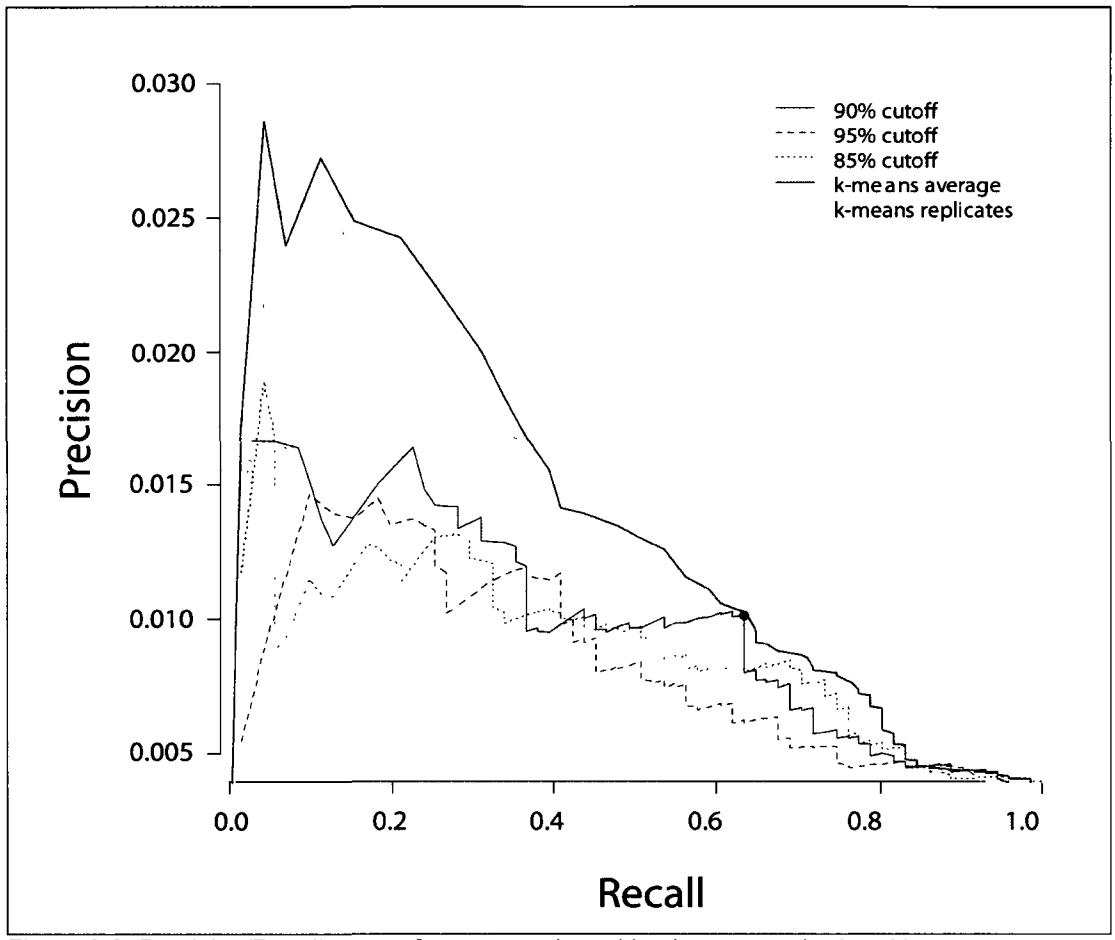

Figure 2-3: Precision/Recall curves for genes selected by the gap method and k-means. Precision/recall curves in red are associated with the gap method, and the point marked with a circle denotes the precision/recall associated with patterns associated with 63 probe sets or less. Grey curves show precision/recall for all replicates of k-means clustering and the average expected precision/recall curve is shown in black. Recall values are based on the 71 stem cell markers defined in Table 2-2, while precision is the fraction of marker genes identified in total number of predicted marker genes.

We compared the performance of our method with a popular method for analysis of gene expression data, k-means, a standard clustering algorithm (See Methods). Both methods performed similarly in producing groups of genes that are expected to be enriched for stem cell markers (Figure 2-3). However, our method differs from a clustering algorithm in that we identify markers that segregate sets of samples whereas clustering algorithms group markers with similar expression patterns. Accordingly, the groups of associated markers produced by the gap method were somewhat different from

the clusters obtained using k-means (mean overlap of 69.8%).

2.4.3 Overview of the selected marker set

To illustrate the variety of patterns identified, Figure 2-4 shows the expression patterns of the 49 probe sets that represent previously known stem cell marker genes identified by the algorithm (yellow), together with other 1,252 genes which were assigned a perfect score by the algorithm (blue). All major divisions in the data set appear clearly defined, with samples related to one of hematopoietic, sphere (a three dimensional aggregate of cells, see(Grenier et al., 2007)), or embryonic sample types forming major groups. For example, hematopoietic samples define many patterns with probe sets identified uniquely within them.

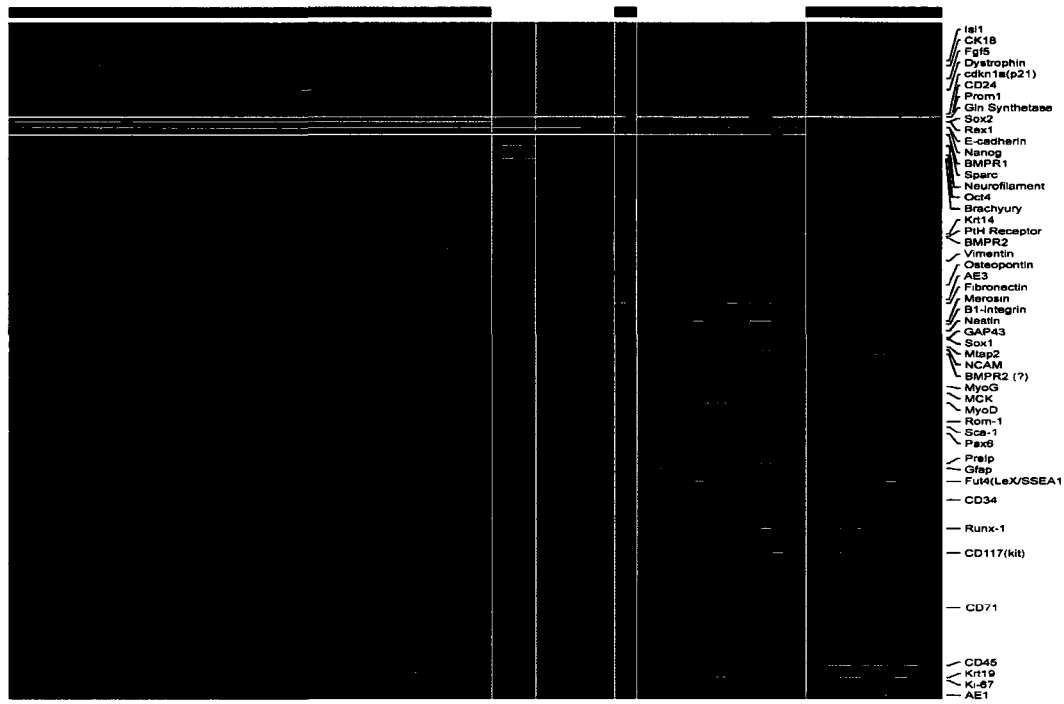

Figure 2-4. Heatmap indicating the distribution of patterns for markers.
Horizontal axis: 241 mouse samples used in this study. Vertical axis: patterns for 1,301 markers either i) scoring 100% if predicted (n=1,252; blue lines) or ii) probe sets belonging to genes ascribed marker roles based on evidence in the literature (n=49; yellow lines). Rows were clustered and diagonalized. Vertical separators were used to distinguish major sample cell types, indicated as coloured groups (top): blue, embryonic cells and derivatives; orange, P19 embryonal carcinoma; grey, osteoblasts and derivatives; purple, neurospheres; yellow, dermal/adipose spheres and derivatives; red, hematopoietic cells and derivates. Gene names are indicated only for the 49 stem cell markers.

2.4.4 GO statistics of the selected marker set

Since the selection of markers was done without reference to the identity of the samples, we would expect to find not just stem cell markers, but also general markers of cell and tissue identity (for example, distinguishing differentiated blood cells from epithelial cells). To investigate in general terms the function of the genes that were selected in the marker set, we collected the 5,848 probe sets defining the 705 selected patterns (with 3 to 63 associated probe sets). These corresponded to 4,449 Entrez Gene IDs, for which we determined the overrepresentation of GO annotations. Using the "potential marker" set of 7,478 genes (which defined groups with up to 200 probe sets) as a reference, we found several significantly enriched functional categories (Table 2-3).

All	Selected set		High in differentiated		GO
N1 (7,478)	N2 (4,449)	P2	N3 (222)	P3	
976	727	2.85×10^{-22}	81	1.60×10^{-16}	Extracellular region
169	131	1.82×10^{-3}	17	9.34×10^{-4}	Extracellular matrix
72	57	n.m.	12	1.06×10^{-3}	Enzyme inhibitor activity
2,061	1,390	8.07×10^{-15}	73	n.m.	Membrane
865	618	2.44×10^{-11}	25	n.m.	Signal transducer activity
587	421	4.00×10^{-07}	15	n.m.	Receptor activity
925	637	8.62×10^{-07}	43	n.m.	Multicellular organismal development
312	235	6.34×10^{-06}	16	n.m.	Defense response
911	614	4.03×10^{-4}	32	n.m.	Cell communication
270	201	4.85×10^{-4}	17	n.m.	Cell adhesion
436	310	5.96×10^{-4}	18	n.m.	Organ development

Table 2-3. Gene Ontology terms of marker sets.
Column labels are as follows: N1 is the number of genes for a Gene Ontology (GO) category in the unselected marker set; N2 and P2 are the number of genes and P value for the smaller marker set; N3 and P3 are the number of genes and P value for the markers highly expressed in differentiated stem cells; GO is the description of the GO term; and GOID is the GO identifier. P values are computed using the 'All' set of markers as background. GO terms displayed if they represent more than ten genes have at least one associated P value below 0.01 and do not overlap more than 80% with another displayed term. (n.m.: Not Meaningful) Table adapted from (Krzyzanowski and Andrade-Navarro, 2007)

Significantly enriched functions (P-value < 0.001) include those that allow cells to interact and respond to environmental cues ('defense response', 'cell communication', 'signal transducer activity', 'receptor activity'), to interact with the immediate neighbourhood of the cell ('extracellular matrix region', 'extracellular matrix', 'membrane', 'cell adhesion'), and functions related to development ('multicellular organismal development' and 'organ development'). Finding these development-related functions is reasonable given that our set of samples is focused on stem cells. The abundance of functions related to the interaction of cells with their environment, on the other hand, may generally reflect that cell identity is largely defined at their surfaces; we would expect to see these functions in the analyses of other collections of gene expression data sampling multiple tissues.

2.4.5 Examination of markers of stem cells differentiation

One distinctive feature of the gap method is that genes are selected based on their ability to define binary groupings of samples. This is meaningful and often desired from the point of view of an experimental researcher. Understanding the significance of the marker becomes simpler as the pattern itself gives a classification of the sample set.

Likewise, the identified sample partitions allow simple, direct, and intuitive ways of manipulating the gene expression data. We illustrate this here by applying a selection procedure to the set of patterns obtained above to focus on markers active in any of several lineages of stem cell differentiation included in our data set.

Generally, we are interested in the properties of probe sets that separate undifferentiated and differentiated sets of stem cells. We chose six pairs of stem cell

samples and their differentiated derivatives from our data set (two embryonic stem cell lines, hematopoietic stem cells, osteoblasts, mammospheres, and mast cell progenitors; see Additional data file 3) and selected probe sets that segregated at least one sample pair with high confidence (99% association score); *i.e.*, the probe set exhibited high expression in the undifferentiated sample and low expression in the differentiated sample, or vice versa. This selection identified 488 probe sets (called the "stem cell related" set) corresponding to 426 genes, of which 206 showed high expression in the undifferentiated sample, and 222 in the differentiated counterpart. Two genes showed high expression in the undifferentiated sample for some cell types, and in the differentiated sample in others: *Ugt1a2* detected by probe set 1426260_a_at, and a gene encoding a hypothetical coiled-coil domain-containing protein detected by probe set 1444761_at. This set of 426 genes included 5 of the 71 known stem cell markers used for benchmarking (*Krt1-14*, *Mtap2*, *Ncam1*, *Spp1*, *Vim*).

Examination of the Gene Ontology terms of the set resulted in a very short list of significant functions (for P-value < 0.01). Separate analyses of genes upregulated in undifferentiated and differentiated states failed to identify GO terms over-represented in the genes highly expressed in stem cells. The only relevant terms for the genes expressed in the differentiated genes (Table 2-3) were related to the extracellular environment ('extracellular region' and 'extracellular matrix'), but with P-values less significant than those for the selected list of 4,449 cell markers. It is known that stem cells often rely on the maintenance of a stable microenvironment (the niche) for physical support and extrinsic cues (for a review, see (Li and Xie, 2005)). The GO term 'enzyme inhibitor activity', not relevant for the larger marker set, appeared as relevant (P-value = 0.001) for

a set of 12 genes, 5 of which belonged to the family of serpins (SERine Protease INhibitors, serpins a1b, a3n, b9, g1, and h1).

Serpins are a large class of proteins that are found in all multicellular eukaryotes and predominantly function as serine protease inhibitors, but can also function as caspase and cysteine protease inhibitors, and in rare cases, as hormone transporters, chaperones, or tumor suppressors. In contrast to eukaryotes, prokaryotic serpins are rare and most serpin-containing prokaryotes have only a single serpin gene (Law et al., 2006). In agreement with our findings, variation in serpin gene expression has been recently observed during differentiation in the myeloid lineage (Kennedy et al., 2007; Missen et al., 2006). Here we give some insight into the functions of the six serpins we identified in our study.

Serpinb9 is an estrogen-inducible caspase inhibitor which is able to inhibit Granzyme B (GZMB)-mediated apoptosis, a key mechanism by which cytolytic lymphocytes are able to destroy target cells. It is expressed at high levels in testis and placenta and may contribute to the ability of immune-privileged cells to evade destruction (Krieg et al., 2001), including embryonic stem cells (Abdullah et al., 2007). Serpins a3n and a3g have been shown to be strongly influenced by LIM-homeobox 2 expression in a hematopoietic system (Siddiqui and Exton, 1992). Serpina3g was previously reported to be highly enriched in HSCs (Terskikh et al., 2001), in concordance with our observations. Similar in function to serpinb9, serpina3n has been implicated in providing protection from Granzyme B mediated cell death in a study of Sertoli cell secreted factors (Sipione et al., 2006). Interestingly, Serpinh1 is one of the exceptions of the serpin family; it is a chaperone molecule which plays a role in the maturation of procollagen (Ishida et al.,

2006). With relevance to stem cell biology, Serpinh1 knock-out ESC produce embryoid bodies with aberrant morphology (Matsuoka et al., 2004). In summary, the abundance of serpins as markers might reflect generalized mechanisms of immune system avoidance activated in cells undergoing differentiation.

The set of cell markers presented here offers a good basis to study principles of stem cell gene function. For example, of the 426 genes selected as markers, a small number, 17 genes, were involved in several of the six lineages selected (at least one of them being non-ESC). Of those, 9 were highly expressed in the stem cells, 6 were highly expressed in the differentiated partners, and only 2 were expressed in both (see Additional data file 3). This indicates not only that a single gene can be involved in multiple stem cell differentiation lineages but that if it does, it will most often follow an expression pattern that is similar across those lineages.

This set of markers also allowed us also to study stem cell evolution. We were interested to know whether there would be a relation between sequence similarity and involvement in stem cell function and involvement in one or many lineages. Large gene families with frequent duplication and reuse are valuable in these investigations as the members will have varying degrees of sequence similarity. Our set of stem cell markers provides a starting point to search for these families. To identify superfamilies within our set of stem cell markers we performed an exhaustive pairwise sequence comparison of the protein sequences of the 426 stem cell markers (See Methods and Additional data file 3). Manual examination of the results to select full length similarity identified four superfamilies containing three or more members: serpins (cluster #18 in Additional data file 3), the nuclear receptor family (cluster #4), the cytochrome P450 family (cluster #17),

and the Rab family GTPases (cluster #10). As serpins were described above, we further investigated members from the additional families in the context of gene evolution in relation to stem cell function and expression pattern (Figure 2-5).

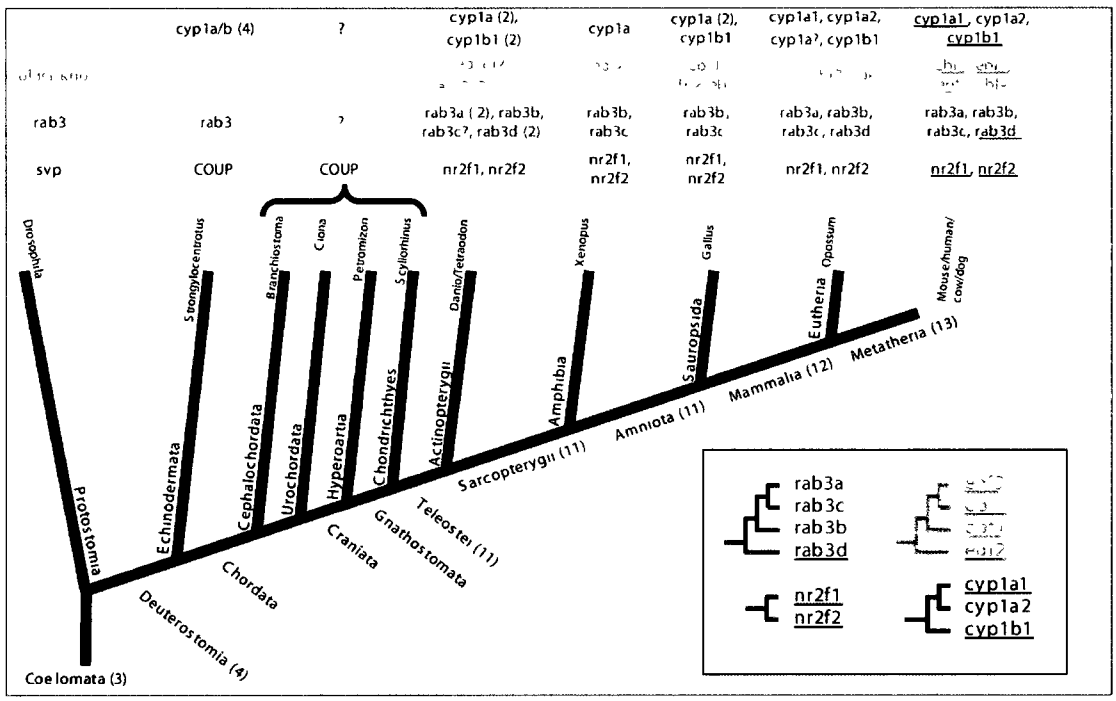

Figure 2-5. Phylogenetic distribution of stem cell markers and their close paralogs in four protein families.

Major taxa along the Coelomata lineage is depicted in bold black text with deduced numbers of genes from these families in parentheses. Phylogenetic relations between species were taken from the NCBI's taxa browser except for the relation between Cephalocordata and Urochordata, which was taken from (Vienne and Pontarotti, 2006). Species names in italics black text. Coloured text indicate paralogs for four families: red for cyp1, orange for ebf, blue for nr2f1, green for rab3. Numbers in parentheses indicate multiple copies of gene (for example, in Actinopterygii genes). Many genes are duplicated in Actinopterygii due to a whole-genome duplication event postulated to have occurred along the ray-finned fish lineage (Christoffels et al., 2004). Most expansion of these four families occurs after divergence of Deuterostomia and before divergence of Teleostei. Underlined genes for mouse indicate genes identified in the selected marker set. The database identifiers of the sequences are given in Additional data file 4. The inset on the lower right corner shows schematic phylogenetic trees for the murine members of the four families. All branches shown had bootstrap values above 0.5. Outlier sequences used for each family (not displayed): D. melanogaster rab3 for rab3, S. purpuratus cytp1 for cytp1, D. melanogaster svp for nrf2f1, D. melanogaster knot for ebf. Image adapted from (Krzyzanowski and Andrade-Navarro).

2.4.5.1 Nuclear receptors - Nr2f2.

The proteins of the family of nuclear steroid and hormone receptors are dimerizing transcription factors characterized by a DNA binding domain and a C-terminal hormone binding domain; they are implicated in cell proliferation, differentiation, and apoptosis (Mangelsdorf et al., 1995). Three members of this family were identified in the set of 426 stem cell markers (*Nr2f2*, *Essrb* and *Rora*). We examined *Nr2f2* in greater detail.

Nr2f2/COUP-TF2 (nuclear receptor subfamily 2, group F, member 2) represses Notch signaling activity in determination of vein identity (You et al., 2005) but it is also expressed in multiple tissues and organs of the embryo and is required for early outgrowth of limb buds (Lee et al., 2004). In our set of samples, probe set 1416159_at, which detects this gene's transcript, segregates the V6.5 differentiated murine ESC (mESC) sample from the rest (Figure 2-6). The 80% identical Nr2f1/COUP-TF1 (nuclear receptor subfamily 2, group F, member 1) (detected by probe set 1418157_at) segregates the samples of retinal spheres, neurospheres, and T1/2 embryonic fibroblasts, but not any of V6.5 or other mESC samples, differentiated or not. By contrast, the probe sets for another close paralog Nr2f6/COUP-TF3 (1460647_a_at and 1460648_at) were not identified as markers.

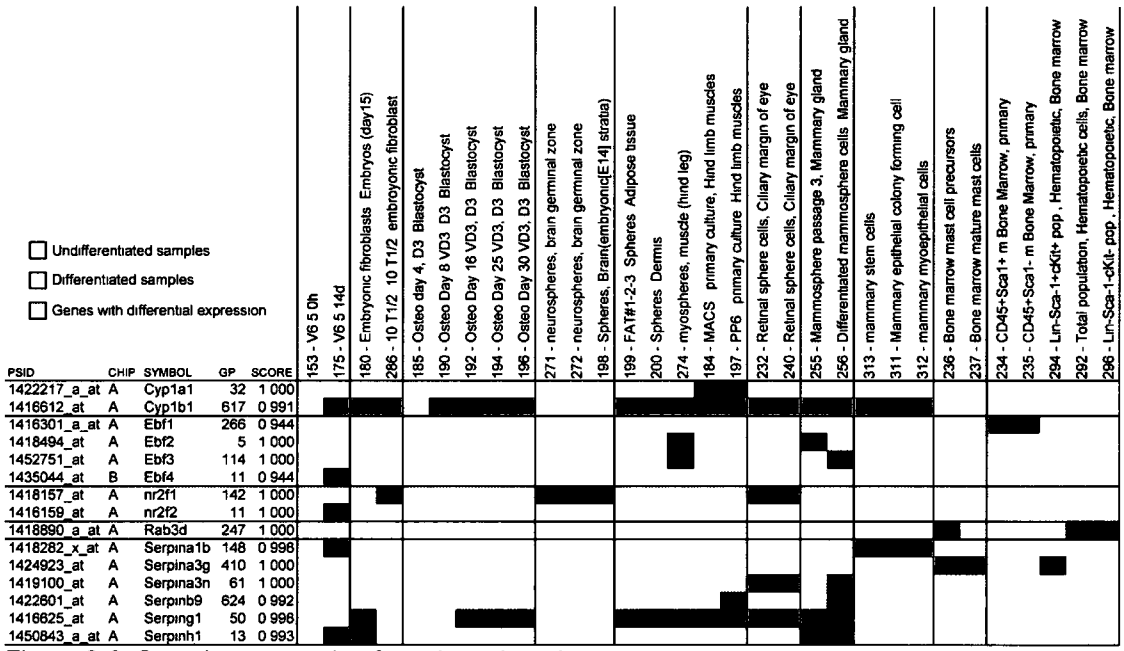

Figure 2-6. Sample segregation for selected markers.
Samples segregated in the high expression group for probe set markers from the nuclear receptor (nr2f1, nr2f2), cytochrome P450 family (cytp1a1, cytp1b1), serpin family (Serpins a1b, a3g, a3n, b9, g1, h1), and Rab GTPase (rab3d) superfamilies. The nr2f1/nr2f2 and cytp1a1/cytp1b1 gene pairs are highly similar in sequence but their expression patterns are notably different. Undifferentiated and differentiated sample pairs are shaded in pink and blue, respectively. Figure adapted from (Krzyzanowski and Andrade-Navarro).

The genes nr2f1 and nr2f2 share a common ancestral gene which is represented in the fly. This Drosophila homolog of *nr2f1* and *nr2f2*, seven up (*svp*), regulates stem cell identity of neuroblasts in order to control the identity of differentiated progeny cells (Kanai et al., 2005). In this case, there is a conservation of the involvement in stem cell function from the divergence between Protostomia and Deuterostomia. Phylogenetic analysis (Figure 2-5) indicates that the gene is conserved as a single copy, possibly until divergence of Gnathostomata. All Teleostei seem to have the duplicated version of the gene. However, the patterns of gene expression in stem cells are very different (Figure 2-6) indicating the specialization of the duplicated copies of the gene.

2.4.5.2 Cytochrome P450 family - Cyp1b1.

Cytochrome P450 proteins (CYPs) are a family of enzymes present in bacteria and Eukarya that participate in the metabolism of exogenous or endogenous chemicals (Bernhardt, 2006; Hannemann et al., 2007). Four members of this family were identified in the set of 426 stem cell markers (*Cyp1b1*, *Cyp24a1*, *Cyp4f18* and *Cyp7b1*). All four were highly expressed in various differentiated cells and expressed at a low level in their undifferentiated counterparts. We performed a detailed analysis of *Cyp1b1*.

Phylogenetic analysis of Cyp1b1 (Figure 2-5) suggests two very close paralogs, Cyp1a1 and Cyp1a2. No equivalent sequence in the *Drosophila* genome (or in any other Protostomia) was identified, but Echinodermata *Strongylocentrotus purpuratus* (purple sea urchin) appears to have an ancestral *Cyp1a/b* gene (with four copies, possibly duplicated after its divergence from Chordates). The ancestral gene appears to have duplicated before divergence of Actinopterygii from Sarcopterygii into the 1a and 1b forms, and the subsequent duplication of the 1a form seems to be absent in Sauropsida (e.g. chicken) but present in all mammals (e.g. opossum). In birds Cyp1a seems to have undergone a separate duplication after divergence from mammals.

In our set, *Cyp1b1* segregates differentiated osteoblast cells and differentiated ES cells from the rest of the data set. Cyp1b1 metabolically activates estradiol (to produce 4-OH estrogens), which are able to induce estrogen receptors, and mutation of Cyp1b1 may stimulate estrogen-mediated carcinogenesis (Sasaki et al., 2003). It has also been suggested that Cyp1b1 is involved in axis control during embryonic development (Stoilov et al., 2004).

By examination of the larger set of markers we can see that Cyp1a1 is a muscle stem cell marker, but its paralog, Cyp1a2, does not behave as a marker in our dataset This is supported by the observation that *Cyp1a2* (-/-) null mutant mice develop normally with just some deficiencies in drug metabolism (Liang et al , 1996) To the contrary, *Cyp1a1* is potentially involved in many cancers and might also have a function in murine embryonic development (Campbell et al , 2005) CYP1A2 is one of the major CYP1 enzymes that catalyze 2-hydroxylation of estrogen (Long et al , 2006), but the substrate of CYP1A1 is not yet known

Cyp1a1 and *1a2* are transcribed from the same bidirectional promoter region (Ueda et al , 2006) Their head-to-head arrangement is conserved in mammalian genomes, which suggests that the genomic organization of these genes is of functional significance The fact that these two genes have different behavior as stem cell markers indicates that there are factors uncoupling their expression

2 4 5 3 Rab family of GTPases - Rab3d

The Rab family of small GTPases are involved in intracellular cell signalling processes, including tethering and docking of vesicles to their target compartment, vesicle budding and interaction of vesicles with cytoskeletal elements (Zerial and McBride, 2001) According to SMART[3] (4 April 2007) there are 66 mouse Rab proteins (defined as containing a Rab domain and no other annotated domain)(Letunic et al , 2009) We identified three members of this family in the set of 426 stem cell markers (*Rab3d*, *Rab31*, *a*nd *Rab38*) *RhoJ* was detected by sequence similarity but discarded after manual examination since it belongs to a different family We performed a detailed

[3] http //smart embl-heidelberg de/

analysis of *Rab3d*. In our set of markers *Rab3d* expression segregates mast cell precursors. None of its paralogs, Rab3a, 3b, and 3c, was identified as marker by our methodology.

The ancestral *Drosophila* gene, Rab3, is expressed in the nervous system (DiAntonio et al., 1993). Echinodermata *S. purpuratus* has only this ancestral gene (Figure 6), but all Teleostei have four copies of the gene, suggesting duplication after the divergence of Chordata and Echinodermata.

In agreement with Rab3 expression patterns in the fly, the four Rab3 paralogs are expressed in mouse brain where they regulate vesicular release; genetic deletion of individual paralogs does not affect viability or fertility in mice, but knock-out of all four genes results in early perinatal mortality (Schluter et al., 2004). However, these genes are expressed elsewhere. For example, Rab3a is detected in acrosomal membranes of mouse sperm (Ward et al., 1999). Rab3d is expressed in the exocrine pancreas and the parotid gland where it is involved in secretory granule maturation (Riedel et al., 2002). Finally, Rab3b and Rab3d are expressed in mast cells (Oberhauser et al., 1994) explaining our observation that Rab3d segregates mast cell precursors.

2.4.5.4 Early B-cell factors – Ebf2 & Ebf3.

In each of the four superfamilies analyzed above (Serpins, nuclear receptors, cytochrome P450, and Rab GTPases), we note that most members displayed the same gene expression behavior along differentiation (being highly expressed in either stem cells or in their differentiated counterparts). However, this is not the general case. If we consider all 49 clusters of protein sequences (Additional data file 3), roughly half (26 of

49, 53%) have some family members highly expressed in stem cells and others highly expressed in the differentiated counterparts; the remaining 23 are highly expressed in stem cells (10), or highly expressed in differentiated cells (13). It is true that some of those clusters might be considering not true paralogs, but this is not necessarily the case. An example is given by the pair Ebf2/Ebf3, which we study here in detail.

The four mouse members of the Early B-cell factors (Ebf) family of helix-loop-helix transcription factors have non-redundant adipogenic potential in multiple cellular models (Jimenez et al., 2007). Interestingly, both Ebf2 and Ebf3 are detected in our set, but with opposite effects. Ebf2 is upregulated in mammospheres and Ebf3 in differentiated mammospheres. The other two mouse members of this family, Ebf1 and Ebf4, are detected in the selected marker set. Ebf1 (probe set 1416301_a_at) is identified as segregating several bone marrow samples. Ebf4 (probe set 1435044_at) is identified as segregating differentiated V6.5 mESC (Wang et al., 2002). This family constitutes an example of genes arising by duplication of an ancestral gene (represented in *Drosophila* by the COE/knot transcription factor, which is involved in Hedgehog patterning (Vervoort et al., 1999) and control of hematopoiesis (Krzemien et al., 2007)), with multiple and varied stem cell related functions arising as the gene duplicates.

2.5 Discussion

We have developed an unsupervised approach for identifying coordinately acting biomarkers using heterogeneous microarray data, which can be generalized to any set of gene expression data regardless of the platform on which it was generated. This method is a ground-up approach which first determines the extent of the information available in a set of array data through a discretization step to classify samples and initialize patterns.

Genes are then associated with patterns based on the presence of a clear demarcation threshold in their expression values which can reproduce the classification of samples. These genes can therefore act as markers for the samples segregated in the subset.

Microarray technology has followed a trend towards increased feature density and increased coverage, and is customarily used to ask large exploratory questions. Our development of this method was motivated by the desire to present experimental groups with information that clearly shows which genes differentiate samples of interest in the larger context of many samples in the database (which genes are up-regulated in my samples of interest?), and what other samples are similar to these samples of interest (in which other samples are these genes up-regulated?).

By identifying probe sets which exhibit bimodal expression level patterns, we directly cater to researchers wishing to assess the significance of particular genes in conditions of interest through methods such as PCR or northern blotting. The interactive graphical presentation of expression values as distributions is exceptionally effective for assessing the ease with which a proposed marker gene might be validated, and the samples which it can be used to identify.

In the application we have presented, our approach focuses on probe sets. As genome annotations have changed, probe set annotations may be altered, changing the probe-set to gene mapping, or the annotated gene functions. This is understood to be an intrinsic problem in array design (Perez-Iratxeta and Andrade, 2005), and many probe sets on Affymetrix arrays are of unknown identity or sparsely annotated. However, these obstacles do not render a probe set useless for the purposes of marker identification, as

each probe set has a target sequence that is fixed at the time of array design. This sequence can be used to detect the molecular RNA species that is acting as a marker. Further, the obvious inference when using cDNA microarrays is that the associated genes can function as markers at the protein level. Identification of a marker probe set implies the existence of a molecule (mRNA or protein) that can potentially be used as a molecular marker to discriminate between cell types. Even if probe set annotations are missing or inaccurate, follow-up experiments can be performed to validate or identify the molecules involved and determine whether development of an mRNA or protein molecular marker is feasible.

Our gap method stands apart from existing methods of microarray clustering. Standard clustering algorithms such as hierarchical or k-means clustering have two impediments commonly raised. The first is that both methods group genes based on global similarity of expression patterns. That is a restrictive test, which is not necessarily useful in the search for a marker. Secondly, and more importantly, both methods assign each gene and condition to a single cluster, which may not always be desirable. Proteins can have multiple functions in different contexts. In general, these clustering methods are best used when trying to identify co-regulation or co-expression of genes in a series of samples that are relatively homogeneous. In contrast, we are striving not to identify strict co-expression of many genes on a global level, but rather sets of genes with expression level thresholds that demarcate similar sets of samples in a heterogeneous microarray data set. This selection procedure is more appropriate to the selection of markers.

Previously established methods also study gene expression values across sets of samples to identify biomarkers. Our method has some differences from those methods

that we believe make it more useful for some applications. The methods from Pepe et al. (2003) (Pepe et al., 2003), SAM (Tusher et al., 2001), and PAM (Tibshirani et al., 2002) require the investigator to separate the samples into two classes; this might be appropriate for simple situations but it would be cumbersome for a dataset with multiple cell types and conditions, especially if a researcher is exploring novel ways to arrange the data. Our unsupervised classification method is more appropriate for such a case.

Like our method, the method presented by Beattie and Robinson (Beattie and Robinson, 2006) can produce patterns in an unsupervised manner. However, the binary patterns are obtained through an analog to digital transformation via a one-dimensional clustering step. We believe that our use of a threshold value to discover binary patterns is more intuitive for experimental biologists, and will simplify the development of following validation studies. As explained above, a clear demarcation of expression level between the two groups influences the decision for follow up more than the overall structure of the data in each state. Ideally, we would like to present candidates that also exhibit large fold-changes with high statistical significance. The logic of using a threshold value that is directly related to the relative mRNA concentrations represents a margin of safety for the detection of differential expression, and is commonly illustrated by the use of a "two-fold rule" in the literature. Put simply, a biologist is more likely to follow up on a candidate marker that has clear separation of expression level, over other candidates that may have better statistical support, but would pose a challenge for validation due to a tight range of expression.

Furthermore, the approach from Beattie and Robinson (Beattie and Robinson, 2006) assembles clusters by combining genes which generate identical binary patterns.

Our observations indicate that the level of noise in gene expression data must be accounted for and some degree of fuzziness must be allowed in any analysis. Essentially, one stray sample should not ruin the association of two genes across a large number of samples (here, 83). Logically, the possibility that such stray hybridizations are encountered increases with increasing data set size. One such example is illustrated in Figure 2-1A. We believe that cluster formation by a measure of discrimination which is more resilient to small degrees of overlap in the distributions (due to experimental error) will yield more fruitful results, and that adhering to matches between digital patterns for cluster formation may be unnecessarily strict. The original data can be always retrieved and a closer examination can be used to verify whether the classification suggested by each probe set is true. This allows genes to be detected as markers of similar patterns.

To illustrate the usefulness of our methodology in cases where unsupervised patterns with resilience to noise are needed, we applied the method to a database of 83 samples from a variety of mouse stem cells and derivatives. Our results indicate that this method produced a list of candidate markers (the 'selected marker set') which was five-fold enriched for a set of 71 known stem cell markers: we identified 45 of these while reducing the total number of probe sets under consideration from 45,137 to 5,848.

To verify the performance of our method, we clustered the data set using k-means, a standard clustering algorithm (see Methods), and found that the precision and recall values were similar. The overlap of results with the k-means clusters was appreciable (69.8%). However, our method cannot be considered as simply a duplication of the results of k-means, as our method groups genes by their ability to segregate the database and not by the similarity of their patterns of expression. Another important

difference is that our method generates groupings of samples while identifying markers; something that k-means lacks. If one were to choose to develop marker genes from the results generated with k-means, a subsequent analysis of the expression profile in each k-means cluster would be required to identify the samples in which the cluster of genes might be more highly expressed. This can be done with a single set of genes but would require an additional automated method to process all k-means clusters if a meta-analysis were desired (such as our identification of genes related to differentiation in multiple cell types). Thus, the gap method presented here provides important additional data that is useful for a variety of subsequent explorations.

The selected marker set of genes was enriched in genes with particular features: products located on the exterior of the cell, and functions related to cell communication and differentiation (Table 2-3). This could be expected since our collection of data contains samples from diverse cells and tissues, and cell identity is mostly related to the cell surfaces and to the ways cells interact with their environment, including communication with other cells. In addition, since our dataset is enriched for differentiating tissues a number of major functions related to development appeared in the list.

Although the selected set of markers is somewhat biased towards the objectives of analysis of stem cells, it also includes markers that, for example, distinguish one lineage from another (e.g. adult versus embryonic expressed genes). Our methodology facilitates focusing attention towards markers for samples that one is interested in. Here, we focused a more detailed analysis on markers segregating at least one of a short list of six differentiation lineages from the set of 83 samples. The use of these 83 samples as

background increases the likelihood of identifying markers that are specific for samples one is interested in.

A total of 426 genes were identified as stem cell related markers for the six differentiating lineages. Functional analysis found fewer statistically over-represented functions than in the selected set of markers. Analysis of the 222 genes segregating the differentiated samples (Table 2-3) found that the only over-represented Gene Ontology (GO) annotations also significant for the selected set of markers were the association of genes to the extracellular region and the extracellular matrix. It is known that extracellular interactions are important between stem cells, their progeny, and their immediate environment in the maintenance of the stem-cell state, control of stem cell populations and associated progeny (Li and Xie, 2005). The one GO annotation over-represented only in this set was enzyme inhibition (for 12 genes). This annotation is represented in the list by five members of the serpin superfamily. A literature search suggests that these genes may have a role in suppression of immune system effects against differentiating cells. The absence of other functional enrichments suggests that there may be no specific gene function that endows cells with the property of 'stemness', in the same way that there appears to be no single stemness gene for all stem cells (Vogel, 2003).

This set of 426 markers of stem cell differentiation allowed us to make some general observations regarding stem cell function and evolution.

Firstly, we observe that if a gene is a marker for differentiation in multiple lineages, then it will act the same way in most cases. Of the 17 genes identified as

markers in multiple lineages, only 2 were found to be highly expressed in one differentiated population and an in another undifferentiated population.

To examine whether patterns of expression in stem cells were retained for gene homologs, we studied gene families represented in the set of markers. Identification of 49 clusters of related protein sequences indicated that gene expression behavior was not conserved within those families since more than half of the clusters (26) included genes with conflicting gene expression segregation properties. However, we also wanted to study how expression patterns in stem cells were conserved with protein sequence identity. For this we looked for the largest families within the set which reflect a large number of gene duplication events and therefore offer multiple levels of sequence identity between their members. We selected four families with two to four true paralogs in the marker set: nuclear receptors, cytochrome P450s, Rab GTPases, and Early B-cell factors. These superfamilies share general functional activities, but have obtained additional functional or tissue specificities through mutation and selection following gene duplication. Phylogenetic analysis of one example from each family (Figure 2-5) led to the study of the evolution of a total of 13 genes. An ancestral gene existed for each of the four families prior to divergence of Deuterostomia, with 3 being present in the Protostome *D. melanogaster* (*knot*, *rab3*, and *svp*). *Svp* and *knot* are involved in stem cell differentiation. The 2 genes arising from duplication of *svp*, and the 4 genes arising from duplications of *knot*, are all detected as markers in our selected marker set. Similar observations are made for the gene, *Cyp1a/b*; this gene underwent two duplications to produce a family of three genes, with two involved in differentiation and identified in our set of markers. We propose that the ancestral versions of this gene (e.g. in *S. purpuratus*)

are involved in differentiation. The Rab3 family illustrates a case where duplication of an ancestral gene (which has synaptic functions in *D. melanogaster*) produces *Rab3d*, which has a role in mast cell development, and *Rab3a/b/c*, which we do not identify as differentiation-related markers. The four murine genes conserve some neural function as the ancestral *Rab3*, but they seem to have obtained other functions and tissues in which they are expressed. Based on the results above we can hypothesize that the duplication of a gene involved in a development related function is likely to result in genes also involved in development.

Our second observation is that, in contrast, the range of expression is not necessarily conserved between close paralogs. *Ebf2* and *Ebf3*, for example, are highly expressed in undifferentiated and differentiated mammospheres, respectively. Other cases can be observed in Figure 2-6. We did not identify redundant markers, that is, paralogous genes with the same segregation properties.

Our third observation is that clusters of genes with stem cell related functions appeared during a window of evolutionary time. Expansion of gene families associated with developmental functions is demonstrated by the number of paralogs of each family present in different organisms (Figure 2-5). Of 13 genes from the four families, three are present in *Drosophila* (*svp*, *Rab3*, and *knot*), four in *S. purpuratus* (*COUP*, an ancestral version of *Cyp1a/b*, *Rab3*, and an ancestral *Ebf*), and probably 11 in Teleostei. By the time of the Metatherian divergence, all 13 genes are present and there are no subsequent duplications. A substantial portion of gene family expansions was complete by the divergence of Teleostei from Chondrichthyes, which agrees with a previous phylogenetic analysis of genes involved in mouse embryonic stem cell differentiation (Sene et al.,

2007). The implication of this is that the use of model organisms such as *D. rerio* (zebrafish) and *Xenopus* in stem cell research may yield insights that can be translated into mammalian systems, provided that the appropriate paralogous genes are chosen for study. Our analysis also suggests that the completion of the genomes of members of the Urochordata, Cephalochordata, Hyperoartia, and Chondrichthyes taxa will provide great insight into the evolution of genes involved in the regulation of cellular and tissue complexity, in particular of those genes related to stem cell differentiation.

With this analysis we identified many stem cell specific markers in parallel, allowing us to establish some concepts regarding the evolution of stem cells. This demonstrates the value of the new tools described in this work. The study of genes from our marker lists will allow identification of mechanisms and cell populations that together contribute to stem cell function in a variety of different tissues.

2.6 Conclusions

We have demonstrated a method for detection of markers from heterogeneous collections of samples of DNA microarray data of gene expression. We have applied this method to a highly heterogeneous set of stem cell gene expression data with the objective of detecting markers relevant to stem cells; a specific contextual question. The gap method detected markers through the unbiased generation of secondary data which facilitated directed analysis of the results.

We believe that our method is more appropriate for the identification of targets for biomarker development than standard analytical techniques such as hierarchical clustering, when applied to DNA microarray data. The gap method is generally

applicable to other large heterogeneous datasets where one desires to find markers acting in a small proportion of the samples.

2.7 Materials and Methods

2.7.1 Environment

Manipulation of data and statistical calculations were performed in the R language (Version 2.3.1). Packages implemented for biological applications are available from the Bioconductor project (Gentleman et al., 2004), which runs in the R computing environment.

2.7.2 Source of experimental data

Raw data in the form of Affymetrix CEL files for the appropriate samples were obtained from StemBase(Perez-Iratxeta et al., 2005a). A subset of mouse samples based on the MOE430A/B DNA chip set was selected, encompassing samples from embryonic and adult stem cells and their derivatives. Expression values were generated from CEL files with the GCRMA package (Wu and Irizarry, 2005) implemented in Bioconductor. Initial analyses allowed us to identify some samples whose values were in general outliers (due to non-standard RNA preparation procedures). These were discarded and the expression values were recomputed with GCRMA on the reduced data set. The data set used here contained 241 replicates (unique chips) derived from 83 different samples (of unique biological origin) (Table 2-1). The expression signals were organized in a matrix with columns for samples and rows for probe sets, with the entries in the matrix representing the hybridization level.

2.7.3 Generation of database partitions

To identify possible permutations of partitions for the expression matrix, we used a strategy to identify individual probe sets whose distribution of expression values appeared to have a break in continuity (a gap). This was motivated by the observation

that known marker genes are expressed at two or more obviously distinct levels (see the example for Nestin in Table 2-1A).

Briefly, for each probe set in the expression matrix, we ordered the expression values and calculated the differences between consecutive values. Large difference values suggest a demarcation in the expression values for a particular probe set, which may be the result of two underlying subpopulations of samples. If the difference exceeds a cutoff value, samples on either side of the gap are assigned to two groups (low/high expression), and encoded as binary vectors. The cutoff value used in this analysis was 50%, that is, a 1.5 fold difference, which on a log2 scale translates approximately to 0.6 units. This value was chosen to generate a number of partitions that did not produce an excessive rate of false positive clusters. Note that we consider the possibility of finding more than one gap.

We used majority voting rules to correct each binary vector so that all replicates of any given sample were either 0 or 1. Ties were assigned a 0. For example, if two out of three replicates were assigned a 1, the remaining replicate also was assigned a 1. This ensures that database queries corresponded to the underlying vectors used in the analysis. These corrected vectors were used for scoring probe sets in each cluster, and were also used for generating visualizations on the web server associated to this work .

2.7.4 Identification of markers for each partition

We identified groups of probe sets which can act as markers for each pattern of up/down regulation defined by the binary patterns as follows. For each binary vector (which represents one way of partitioning the database into two), we calculated Mann-

Whitney U test U-scores for each probe set on the microarray. The U-statistics were calculated from the expression values in each group designated as highly expressed versus the expression values exceeding the 90^{th} percentile of the group with low expression. This modified U-statistic is linearly correlated to the partial Area Under the Curve (AUC) described in (Pepe et al., 2003) at a False Positive Rate (FPR) of 10% (data not shown). The advantage of using the U-statistic here lies in its decreased computational cost versus the calculation and integration of the AUC. The U-scores were converted into a percentage of their maximum value for each group. Probe sets were ranked by these scores and those exceeding 90% of the maximum U-score were saved as candidate markers for each binary vector. The choice of this threshold was determined to produce satisfactory results according to the precision/recall curves computed below.

2.7.4.1 Filtering marker lists by size and number of classified samples.

Since our analysis examines thousands of probe sets, the analysis above is likely to identify many non significant markers by chance. A simple way to reduce those is to accept only sample partitions that are identified by multiple probe sets. To establish a lower threshold for the length of marker list to be accepted, we generated marker lists using a randomized version of the expression matrix. The expression values for each probe set were randomly reassigned to samples in order to destroy their biological ordering while individually maintaining their range and distribution. We then generated marker lists for each binary vector as described above. Approximately 98% of patterns formed from the randomized matrix were associated with 2 or fewer probe sets, suggesting that such patterns are likely to arise by chance. Thus, in our analysis of the stem cell data set we decided to report only patterns associated with three or more probe

sets.

We also observed that the patterns with the largest numbers of probe sets identified major groups of tissue specific genes and therefore were not useful since they reproduced obvious sample partitions. For example, a cluster of 242 probe sets was identified that distinguished all blood related samples (hematopoietic and general bone marrow related) from others in the database. We therefore did not report clusters with more than 200 probe sets. This selection resulted in a list of 893 patterns involving a total of 10,401 probe sets for 7,478 genes. We define this as the "potential marker set" (Additional data file 1).

2.7.5 Calculation of precision/recall curves

In order to assess the methods performance in detecting 71 known stem cell marker genes (Table 2-2), we gradually decreased the upper limit of allowable cluster size in the data set and calculated precision/recall values on the sets of clusters that passed the criteria at each step (Figure 2-3). This analysis was used to choose a 90% cutoff for generating clusters with the gap method, and to define a smaller set of markers (from the 705 patterns with 63 or fewer markers), the "selected marker" set (Additional data file 2).

2.7.6 Benchmarking of method with k-means clustering

In order to compare the gap method with an established method of grouping genes, we identified clusters of probe sets with k-means clustering according to their patterns of gene expression. We specified 1,000 clusters, which is similar to the number of clusters generated by the protocol described above, and selected the genes classified in

those groups. The default algorithm for the 'kmeans' function in R was used. We computed the precision/recall values for k-means clustering by gradually decreasing the maximum allowable cluster size, as above. As the k-means algorithm is not deterministic, to estimate the expected precision/recall, we generated a mean precision/recall curve from 100 repetitions of k-means clustering. This curve is compared to curves generated with our clustering method in Figure 2-3.

We compared the overlap between the sets of genes selected by the gap method and k-means. For each of the k-means control runs, we selected the genes from the set of clusters that produced a recall of at least 45 known stem cell markers (this is the approximate recall level in Figure 2-3). The overlap between that k-means iteration and the gap method was computed as the fraction of genes selected by both methods divided by the total number of genes selected by either method. The mean overlap between the gap method and all k-means iterations was 69.8%.

2.7.7 Enrichment of Gene Ontology (GO) annotations

We examined the functions of genes associated with each pattern by characterizing the enrichment of their GO annotations (Ashburner et al., 2000). As many genes are represented on the array by multiple probe sets, we mapped probes to their corresponding Entrez GeneIDs, and calculated functional enrichment on a per gene basis. Affymetrix probe sets were mapped to Entrez GeneIDs using the April 11, 2006 release of NetAffx annotations (Liu et al., 2003). Where probe sets had multiple GeneID mappings the first one was selected since we observed that in the majority of such cases the first identifier tends to be the only one with a published symbol as opposed to one that was automatically generated. Links from GeneIDs to GO annotations were obtained from

on May 9, 2006, and traced up through the GO ontology to identify 'ancestral' terms. We calculated the cumulative hypergeometric P-values for each GO annotation as per (Tavazoie et al., 1999). Raw P-values were then subjected to a Bonferroni correction to take into account the number of GO categories considered. These adjusted P-values are reported in this analysis. The analysis of GO enrichment in the selected and high-in-differentiated sets of markers (Table 2-3) was done using the potential marker set as background.

2.7.8 Representation of markers associated to samples

To generate an overview of the markers associated to different samples by the gap method represented in Figure 2-4, binary patterns were clustered with complete linkage hierarchical clustering based on their Euclidean distances. Major clusters were manually identified from the hierarchical tree and assembled into groups. Each row group was then associated with the column that contained the maximum value of the averaged binary vectors in the group. Row groups were then arranged by increasing column numbers.

2.7.9 Analysis of protein families involved in stem cell differentiation

We chose six pairs of samples from our database representing undifferentiated and differentiated states of stem cells (Additional data file 3). We selected all probe sets segregating at least one of those pairs with a score exceeding 99%. These probe sets were mapped to RefSeq protein sequence IDs based on NetAffx annotations dated July 12, 2006. The 488 selected probes mapped to 420 RefSeq protein IDs, and these 420 protein sequences were obtained from NCBI on Feb 23, 2007. Pairwise protein BLAST (blastp) (Altschul et al., 1997) was performed on this set of sequences and the expect ('e') values were arranged into a pairwise matrix. Cells with no observed protein hit were replaced

with e = 1, and the diagonal was filled with e = 0. This matrix was converted to a binary matrix by assigning 1 to cells containing an e value larger than 10^{-6} and 0 to the remaining cells, which was then hierarchically clustered in R using binary distances for generation of the distance matrix. Finally, we manually chose three illustrative groups of genes after inspection of sequences and results.

2.7.10 Online Search Engine

Marker databases were created using the MySQL database management system, with a web interface written in PHP.

2.7.11 Additional data files

The following additional data files are available with the online version of the paper.

Additional data file 1: 10,401 probe sets in "potential marker" set. Columns indicate (1 Affy id) Affymetrix probe set identifier, (2 chip) chip set, (3 gene) gene symbol, and (4 patterns). A marker might be associated to multiple patterns and those are encoded in column 4 using the marker server pattern identifier followed by the score for the pattern between brackets, with information for different patterns separated by a colon. Patterns can be retrieved at the marker server using the marker server identifier or they can be queried with the Affymetrix probe set identifier.

Additional data file 2: 5,848 probe sets in "selected marker" set. The meaning of the columns is identical to Additional data file 1.

Additional data file 3: 488 probe sets in "stem cell related" set. Columns indicate

(1 Affy id) Affymetrix probe set identifier, (2 chip) chip set, (3 gene) gene symbol, (4 pattern) marker server pattern identifier chosen. Columns 5-6 are related to the clustering analysis (see Methods): (5 protein) NCBI identifier for the protein sequence used, (6 cluster) label for the protein cluster. Columns 7-18 indicate segregation (value of 1) in pairs of stem cell samples and their differentiated derivatives, respectively: (7 J1-U) and (8 J1-D) J1 mESC, (9 V6.5-U) and (10 V6.5-D) V6.5 mESC, (11 Mast-U) and (12 Mast-D) mast cell precursors, (13 MaSC-U) and (14 MaSC-D) mammospheres, (15 Osteo-U) and (16 Osteo-D) osteoblasts, (17 HSC-U) and (18 HSC-D) hematopoietic cells. Column (19 multi) indicates (value 0/1) 17 genes differentially expressed along multiple stem cell lineages (one at least being non mESC). Samples used were S255 versus S256 for mammary stem cells (undifferentiated and differentiated, respectively), S294 versus five samples (S291, S292, S293, S295, S296) for HSC, S128 versus S127 for J1 ESC, S153 versus S175 for V6.5 ESC, S185 versus S196 for osteoblasts, and S236 versus S237 for mast cells.

Additional data file 4: Identifiers of the proteins used in the phylogenetic analysis. Protein identifiers (GenBank) of the sequences used for the phylogenetic analysis depicted in Figure 2-5. Occasionally, the label used (e.g. Ebf) differs from the gene name in the database. Labels used are derived from the phylogenetic analysis.

Chapter 3 Using EST data facilitates noncoding RNA prediction

3.1 Preface

Sections of this chapter were reproduced from the following publication:

Paul M. Krzyzanowski, Feodor D. Price, Enrique M. Muro, Michael A. Rudnicki, Miguel A. Andrade-Navarro. **Novel muscle miRNA discovery through integration of EST data and predicted RNA secondary structures.** Submitted.

Author Contributions for this chapter: PMK designed the project, performed computational analyses, and drafted the manuscript. EMM assisted in mapping ESTs. PMK and MAA analyzed data. MAR and MAA provided mentorship and direction.

3.2 Abstract

Predicting novel loci yielding non-coding RNAs (ncRNAs) is a topic which has enjoyed a considerable upsurge in interest, with many computational methods being published to predict microRNAs, long ncRNAs, riboswitches and the like. Although all methods strive to improve the quality of predictions generated through various strategies, none, to our knowledge, have explicitly considered the impact of integrating existing cDNA-based Expressed Sequence Tag (EST) data to achieve this goal. To determine whether EST data can assist in miRNA prediction, we evaluated the performance of various methods of miRNA prediction both with and without information combined regarding EST data according to the model of miRNA biogenesis. Our results indicate that ESTs can improve the quality of miRNA predictions and represent a source of data that has been underappreciated in the field of ncRNA discovery.

3.3 Introduction

3.3.1 Non-coding RNA prediction

Interest in non-coding RNAs has increased dramatically since the first reports of their significance. MiRNAs are one class of many ncRNAs which are of scientific importance. Current approaches to discover novel miRNAs vary and range from projects that are predominantly computational (Krek et al., 2005; Washietl et al., 2005a; Washietl et al., 2007) to high-throughput sequencing projects which attempt to identify them empirically (Babiarz et al., 2008; Morin et al., 2008). In general, experiments are designed to utilize complementary information from both angles to generate sets of candidates. In designing a miRNA prediction experiment, both sets of techniques should be used together; the order in which they are applied likely depends on the strengths of the group spearheading the work.

For example, computational prediction methods require experimental validation. An example of this approach is given in Washietl *et al.* (2007), where ENCODE Consortium genomic alignments are analyzed computationally to predict structural non-coding RNAs (Washietl et al., 2007). Visual inspection of results generated computationally was followed by RT-PCR validation of selected candidates, yielding an overall verification rate of 25%. In the Washietl *et al.* study, the validation of several different types of ncRNAs was pursued simultaneously and for reasons such as complexity of RNA structures, this verification rate was expected to differ for various RNA types. A separate group ran a study in *Saccharomyces cerevisiae*, identifying structural ncRNAs, performing follow up validation using a combination of Northern blottings, rapid amplification of cDNA ends (RACE), and the integrations of previously

published tiling array data (Kavanaugh and Dietrich, 2009).

On the other hand, results from experimental techniques such as RNA cloning and sequencing require *a posteriori* computational analysis to provide additional evidence supporting the existence of purported pre-miRNA hairpins. This approach is likely to be favoured by groups where experimental expertise dominates. For example, Ro *et al.* (2007) employed a small RNA cloning approach to identify RNAs in developing mouse testes, separating sample into sub-populations of miRNAs and piwiRNA-like species and focusing on the latter (Ro et al., 2007). The ultimate characteristics of the cloned piRNA-like RNAs was finally confirmed bioinformatically using a combination of size, predicted RNA structure, and other characteristics of the local genomic neighbourhood (e.g clustering of multiple piRNA-like candidates). In 2005, an earlier study of cloned zebra fish small RNAs identified miRNAs by categorizing sequenced small RNAs with an implementation of BLAST sequence alignment versus an ad hoc database of noncoding RNAs (Chen et al., 2005). Here, small RNA sequences were searched against profiles for known RNA families such as rRNA, tRNA, snRNA, snoRNA, scRNA and mRNAs. Any small RNAs not fitting known ncRNA profiles formed a pool from which subsequent validation was pursued on an individual basis.

A hybrid approach to ncRNA prediction lies in attempting to identify novel miRNAs residing within known regions of transcription. One of the most abundant reservoirs of data to facilitate this work is Expressed Sequence Tag (EST) data previously generated to characterize the transcriptome.

3.3.2 Expressed Sequence Tags

Expressed Sequence Tags (ESTs) are nucleotide sequences which are generated by sequencing the ends of cloned sequences from cDNA libraries. cDNA libraries differ from genomic DNA libraries in that DNA inserted into plasmid vectors is derived from genomic regions actually expressed in the samples of interest. There are two main advantages of creating and using cDNA libraries. Firstly, library vectors are enriched in genomic sequences expressed in a sample of interest, permitting exclusion of intergenic genomic regions which may not be of interest to a particular project. Secondly, cDNA library fragments represent coding mRNA sequences that are more or less fully processed, and capable of generating peptide sequences corresponding to functional proteins. This factor is of exceptional utility in cases where mRNA splicing occurs and expression of a protein in a prokaryotic system is desired.

A standard approach in generating cDNA libraries is to isolate and purify mRNA from a sample of interest and purify this RNA fraction using oligo-dT purification columns. This approach excludes non-polyadenylated RNAs such as tRNAs and ribosomal RNAs from the RNA fraction of interest (Lodish, 2000). Once the purified mRNA fraction is isolated, oligo-dT primers are used to synthesize a complimentary DNA strand with a retroviral reverse transcriptase polymerase (Figure 3-1). The original RNA template is eliminated by hydrolysis and a poly-dG sequence is added to the 3'-OH end of the newly synthesized ssDNA strand, permitting the use of oligo-dC primers and DNA polymerase to initiate synthesis of a complimentary DNA strand equivalent to the original RNA molecule.

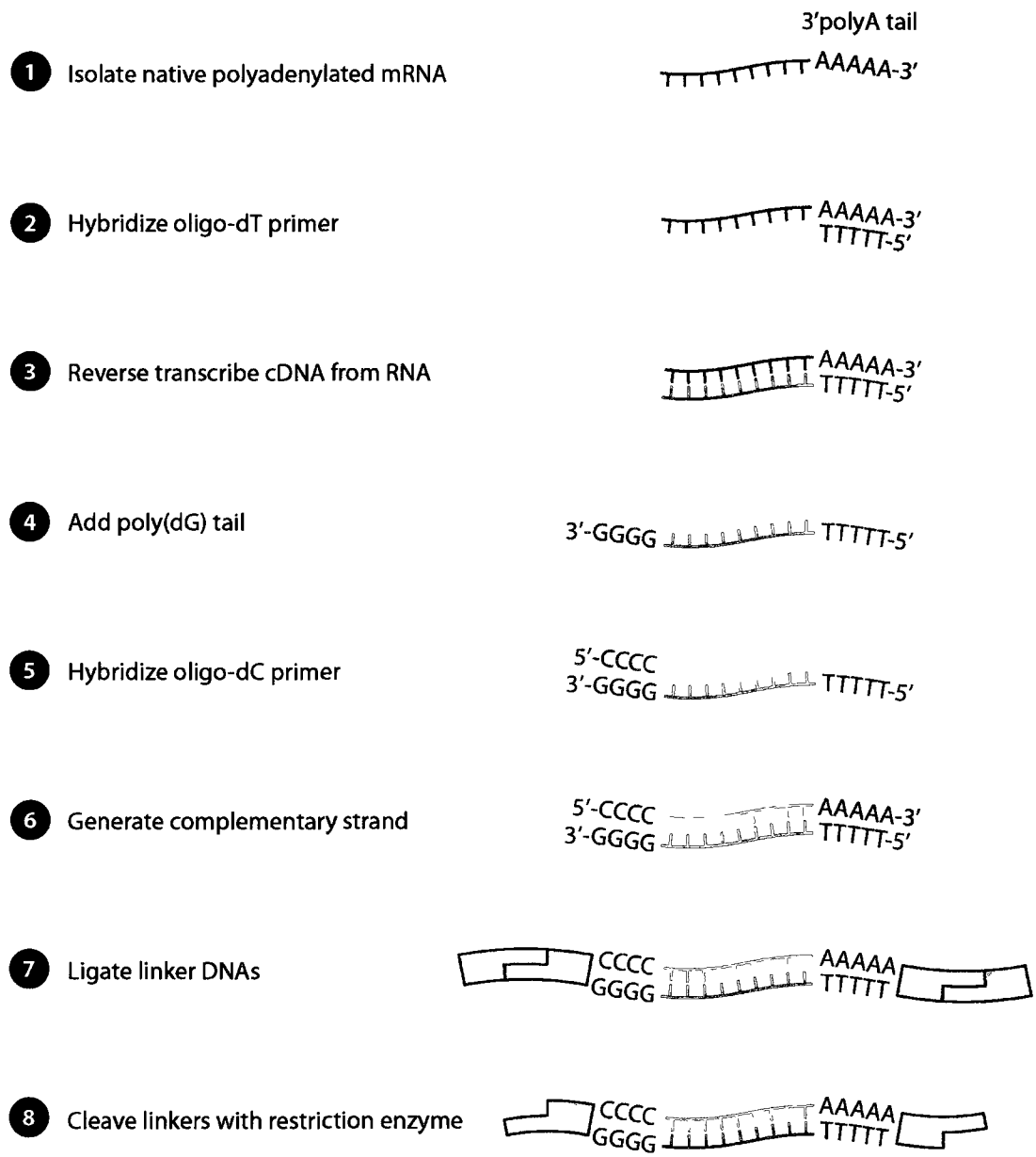

Figure 3-1: Construction of a cDNA library.

To enable insertion of the double stranded cDNA molecules into plasmid vectors, additional dsDNA linker fragments containing restriction enzyme sites are ligated to the blunt cDNA ends. Restriction sites in the linker regions are cleaved to produce "sticky ended" termini (staggered cleavage points) which are cognate to sites within the vector backbones chosen to ultimately store the library. Recombinant vectors are then incorporated into a host organism such as *E.coli*, creating the library where individual *E.coli* colonies typically carry a clone originating from a single mRNA fragment.

Using these and similar techniques, over 60 million EST sequences have been deposited and made publicly available in the NCBI dbEST (Boguski et al., 1993). In addition, many proprietary EST databases may have sequences derived from specialized cDNA libraries (Echenique et al., 2002; Lutfiyya et al., 2007; Michael et al., 2006).

Although EST library creation is a time tested technology with historically large human sequencing initiatives now over a decade old (Adams et al., 1995; Hillier et al., 1996), their usefulness continues to the present day in a variety of applications. EST sequencing is being used to identify cochlear genes potentially involved in the development of hearing disorders (Skvorak et al., 1999), to advance knowledge of diatom genomics for the purpose of improving silicon based nanotechnology designs (Montsant et al., 2005), and to analyze economically important crops and livestock (Patel et al., 2009; Vizoso et al., 2009). Thus, this mature strategy continues to show usefulness in well defined problems.

However, past EST sequencing projects yielded many transcripts which, at the time of creation, were annotated with no known functions. Many ESTs map to regions of

transcriptional activity on genome assemblies which are not predicted to generate protein sequences according to gene prediction algorithms crafted to identify protein coding sequences (Burge and Karlin, 1997; Majoros et al., 2004; Sayers et al., 2009). With the increased understanding of non-protein coding sequences and their important biological roles, transcripts previously failing protein coding gene tests can plausibly be used to identify novel alternative functions according to RNA-based models.

3.3.3 Use of EST data in ncRNA prediction

In order to consider whether EST data can be leveraged for ncRNA prediction, specifically to identify miRNAs, it is important to consider the techniques used to generate EST libraries and their compatibility with miRNA biochemistry. The most direct question to address is whether miRNA transcripts are polyadenylated and thus whether traces of them can be captured in cDNA libraries constructed with poly-T primers, as explained above.

There are, in fact, numerous observations supporting the connection between EST data and miRNA prediction based on polyadenylation. For instance, it is known that pri-miRNAs, the uncleaved precursors of miRNAs bearing mRNA, are polyadenylated in human cells (pri-mir-155) (Eis et al., 2005; Pawlicki and Steitz, 2009), *Arabidopsis thaliana* (pri-miR171a) (Song et al., 2007), and *Drosophila melanogaster* (pri-mir-281) (Xiong et al., 2009). The entire mir-290-295 cluster of miRNAs, thought to play a major role in early embryonic development, is transcribed as a long (~3.1kb), polyadenylated pri-miRNA (Houbaviy et al., 2005).

The parallels between classical mRNA transcription loci and dedicated intergenic

miRNA bearing sites can be strengthened further by considering full transcriptional units. One study concluded that miRNAs spanning up to 10kb can be transcribed as a single primary transcript, with predicted CpG islands near the promoter region and well defined 5' and 3' boundaries (Saini et al., 2007). Prior to the development of this knowledge of miRNA transcription sites, an earlier study successfully amplified several pri-miRNAs using oligo-dT primers (Cai et al., 2004b), providing the most direct information supporting the hypothesis that miRNA derived sequences can be observed in EST libraries.

Several groups have examined EST data in ncRNA prediction, but in each case the strategies designed have been to examine EST sequences directly. At least two groups attempted to identify novel miRNAs by examining ESTs for hairpin-like structures (Li et al., 2006; Zhang et al., 2005). Others have used ESTs to infer miRNA expression by flanking EST RT-PCR (Gu et al., 2006). This latter approach is corroborated by the previously observed correlations in expression between intronic miRNAs and the protein coding transcripts they lie within (Baskerville and Bartel, 2005). This joint expression leads to a tempting hypothesis that has become a common assumption in miRNA research – that the two functional molecules are produced in a 1:1 stoichiometric ratio. However, since exogenous factors can affect splicing/pre-miRNA cleavage efficiencies and respective RNA degradation rates, the ultimate effect on the relative abundances of the two species is difficult to determine and this line of thought is difficult to justify as a hard and fast rule. In fact, one group reported that ESTs illustrate alternative transcription products created via excision of intronic-miRNAs quite nicely, as is the case with mir-126 in the 7th intron of EGFL7 (Kim and Kim, 2007).

Furthermore, most recent high throughput sequencing data shows that mature miRNAs of the same pri-miRNA transcriptional unit, and even of the same pre-miRNA hairpin, can ultimately be present in different ratios (Babiarz et al., 2008). The implication of all these observations is that there is a large amount of interplay between miRNAs, their sequences, the miRNA pathway processing machinery, and most likely many unknown accessory factors that influence various steps an RNA takes along the way from its site encoded within genomic DNA to a fully functioning mature miRNA molecule.

In the absence of an all-encompassing model of miRNA biogenesis, we began our analysis with the hypothesis that currently understood mechanisms of miRNA biogenesis reduce the likelihood of observing full pri-miRNA transcripts in cDNA evidence. We hypothesized that, according to the standard model of miRNA biogenesis, the likelihood of observing full pri-miRNA transcripts in EST data is low, specifically due to the absence of cDNA-like splicing products as expected from long protein coding mRNAs (See Figure 3-2). We therefore investigated whether integrating EST data in a way that considered the cleavage mechanism of miRNA maturation improves predictions generated via several methods. More specifically, usage of poly-T primers in EST library construction increases the likelihood of observing only the 3' end of pri-miRNA transcripts, ending at the 5' terminus of a miRNA hairpin. In addition, since poly-T primers are predominantly used in the construction of EST libraries, the 3' end of processed pri-miRNA transcripts is the one most likely amplified. As many examples of this effect are seen, albeit anecdotally, such as loci of murine mir-145, mir-126 and mir-206 (Figure 3-2 B-D), we endeavored to investigate the utility of adding EST data into a miRNA prediction pipeline.

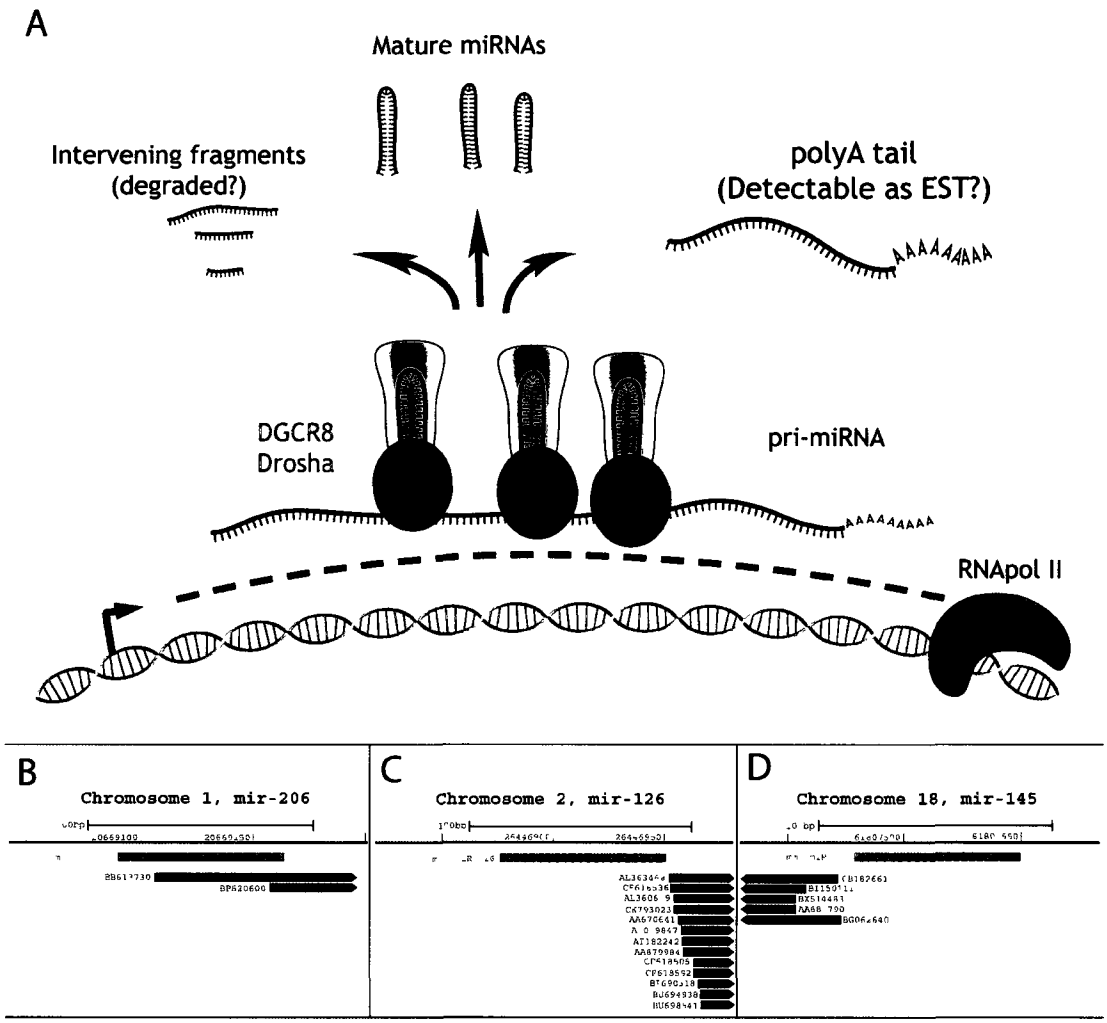

Figure 3-2: Rationale for miRNA-EST model of prediction.
Model of miRNA biogenesis and generation of EST evidence of miRNA expression. RNA polymerase II transcribes the primary miRNA transcript (pri-miRNA), which is cleaved by the Drosha-DGCR8 complex. Cleaved pre-miRNAs are further processed, while intervening fragments are degraded. The remaining polyadenylated 3' fragments of the pri-miRNA may be amplified by polyT primers commonly used in the generation of EST libraries.

To this end, we developed a methodology that combines techniques from miRNA prediction with Expressed Sequence Tag sequences. Our novel reinterpretation of preexisting public EST data identifies traces of miRNAs and provides probable tissues of expression from specific expression data annotated to submitted EST libraries. To date, the usage of ESTs in miRNA prediction has been limited and distinct from our approach.

Furthermore, we propose that the the significance of this effect is general to any ncRNA undergoing cleavage, as illustrated with the discovery of a self cleaving RNA motif in the 3'UTR of Clec2d (Martick et al., 2008).

3.4 Results

3.4.1 EST start points are associated with 3' termini of miRNAs

We first examined the distributions of EST termini around known miRNAs in the human and mouse genomes to evaluate the trend of association between the two types of annotations, and thus to determine whether their use in novel miRNA prediction was plausible. We observed a distinct clustering of EST termini in sequence regions aligned with the 3' ends of known miRNA in the mouse and human genomes, respectively (Figure 3-3). By comparison, this distinct clustering was not observed when examining the 5' termini of known miRNAs.

One important issue in genome wide screens is ascertaining the degree of background signal present. To explore this, the ncRNA-EST analysis was repeated numerous times using randomly shuffled miRNA annotations. Randomization of the annotations is intended to remove any correlations present between known miRNAs and EST products, while estimating a background level of association between the two. The baseline results in Figure 3-3 show that miRNAs are associated with EST above expected background levels. Similar results are maintained when considering different versions of the mouse genome (Figure 3-6).

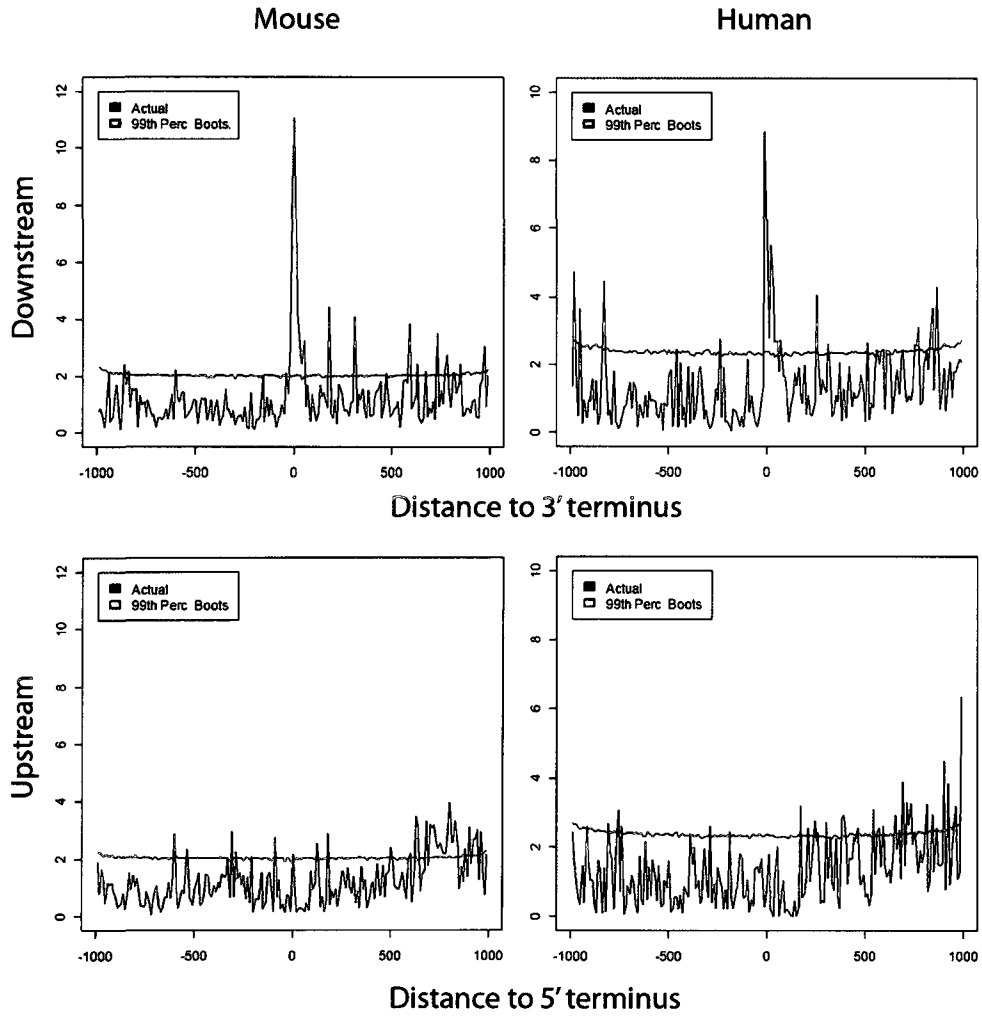

Figure 3-3: EST termini associate with 3' end of annotated miRNA in mouse and human genomes.
Accumulation of EST termini adjacent to 3' ends of known miRNAs in the mouse (mm8) and human (hg18) genomes. We observed a slight increase in signal downstream in the 5' figures which is likely due to EST termini associated with the 3' miRNA terminus and not aligned by virtue of the variable window sizes of known miRNAs. The expected signal with randomly distributed ESTs and miRNAs is shown (red lines; 99th percentile of randomized trials). Additional peaks observed downstream of miRNA 3' termini (most prominent in mm8 in figure) generally arose from miRNAs fortuitously aligned with single downstream ESTs (data not shown).

Together, these observations show that ESTs are specifically associated with the 3' terminus of known miRNAs, supporting our model of concurrent miRNA and EST generation presented in Figure 3-2. With this information, we inferred that miRNA

predictions might be generated by identifying genomic regions with the potential to generate structured RNAs with corresponding EST evidence in proximity of their 3' termini.

3.4.2 ESTs enrich the quality of microRNA predictions when applied to several methods of deriving them

Three different methods were used to investigate whether improved miRNA predictions could be attained by considering potentially structured RNA regions with supporting ESTs.

3.4.2.1 Predictions by RNALfold

One method of generating structured RNA predictions is RNALfold (Hofacker et al., 2004). The RNALfold algorithms identify the most energetically stable RNA structures below a specified size in very long genomic sequences, such as entire chromosomes. After calibrating settings to discard predicted RNA structures unlikely to be miRNAs (See methods for details), we obtained approximately 3.3 million structures in the mouse genome (mm9). In this set of RNA hairpins, it was found that true miRNAs had lower than average normalized Minimum Free Energy (MFE) (Hereafter, normalized MFE is referred to as NFE), in agreement with independently reported results (Bonnet et al., 2004). Known miRNAs exhibited significantly lower NFE than the overall set of genomic hairpins when computed three different ways, the most significant ($P<9x10^{-80}$) being between normalized energies calculated based on the number of nucleotides contained within the duplexed region excluding those in the single stranded tails and terminal RNA loop (Distributions are shown in Figure 3-7). We therefore refined this set of predictions by retaining those with lower than average NFE (under -0.44kcal/mol) and

ESTs within 14nt of the 3' terminus (see Methods for determination of these values). Using this distance between hairpin annotations and ESTs, we determined that there was a significant enrichment in annotation associations as compared with randomly shuffled instances of the same data (Table 3-1). In all cases, positive Signal to Noise (S/N) ratios were generated.

Using these values to filter the hairpin dataset and observing the rate of known miRNA re-prediction, this set of RNA hairpins predictions yielded a high recall rate of almost 85% but a precision rate of only 0.01% when not using EST data. Upon incorporation of EST data, 81 out of 91 (89%) annotated miRNAs with ESTs within 14nt were recalled by using this criteria while the precision rate increased to 0.20%. Therefore, by considering known miRNA annotations recoverable with EST data, we effectively increased the recall rate by approximately 5% while simultaneously increasing the precision by 20 fold (Table 3-2).

Genome	Object type	n objects	n ESTs	Intersecting @14nt	Bootstrap mean	S/N Ratio	Bootstrap P-value
mm8	Raw hairpins	3299070	4561569	41347	33182	1.24	<10-4
	RNAz	79094	4561569	17245	4176	4.13	<10-4
	Cloned	3442	4561569	112	25	4.48	<10-4
mm9	Raw hairpins	3340947	4364783	40872	32496	1.26	<10-4
	RNAz	62057	4364783	12617	3040	4.15	<10-4
	Cloned	3446	4364783	88	18	4.89	<10-4

Table 3-1: ncRNA-EST associations are significantly higher than expected by chance
Associations between structured RNA regions predicted by several methods and existing ESTs are shown. Simple hairpins generated by RNALfold (Raw hairpins), conserved RNA structures (RNAz), and high-throughput Illumina sequenced fragments residing in hairpins ("Cloned"). For details of bootstrapping, please see Methods section.

		mm9					
		Cloned		Hairpins		RNAz	
		Without ESTs	With ESTs	Without ESTs	With ESTs	Without ESTs	With ESTs
A	Number of objects (gross)	3446	3446	3340947	3340947	62057	62057
B	Number of objects (net)	3446	540	3340947	40782	62057	12617
C	Unique miRNAs detected	348	108	393	81	144	35
D	Total miRNA predictable	463	132	463	91	463	91
	Precision (C/B)	10.10%	20.00%	0.01%	0.20%	0.23%	0.28%
	Recall (C/D)	75.16%	81.82%	84.88%	89.01%	31.10%	38.46%

Table 3-2: Considering EST annotations improves rates of miRNA prediction
Rates of known miRNA prediction for three different methods of predicting ncRNAs are shown with and without including known ESTs in the protocol. Methods considered were: Simple hairpins generated by RNALfold (Hairpins), conserved RNA structures (RNAz), and high-throughput Illumina sequence fragments residing in hairpins (Cloned).

3.4.2.2 Predictions by RNAz

In a similar fashion, we generated predictions of structured RNAs exhibiting interspecies conservation between mouse, human, dog and rat using RNAz, a method that considers genomic alignments when generating candidate ncRNAs (Washietl et al., 2005b). RNAz based ncRNA predictions were also associated with EST annoations at a significance level below the threshold calculable by our protocol (Table 3-1). However, the RNAz based results generated a higher Signal-to-Noise ratio than raw hairpins generated by RNALfold, presumably because interspecies conservation eliminated many potential false positives in the set of putative ncRNA hairpins. Applying EST data to predictions using this pipeline increased precision by 22% (0.23% to 0.28%) and recall by 24% (31% to 38%) (Table 3-2). The lower recall rate observed with the RNAz implementation may be due to some known miRNAs being located in genomic regions with lower conservation scores or to some miRNAs being species specific.

3.4.2.3 Predictions derived from high-troughput sequencing data

To further investigate the utility of ESTs data for analysis of small RNA sequence libraries, we examined EST influence on miRNA predictions using a published data derived from a embryonic stem cell library generated with Illumina sequencing technology (Babiarz et al., 2008).

We aligned the sequence fragments deposited in GEO (Accession: GSM314552) to the mouse genome and identified 4239 small (~21nt) sequence windows. After restricting the data to windows predicted to be within putative RNA hairpins, we were left with 3446 potential mature miRNA sequences with experimental support from cloned library sequences (see Methods for details). When these candidate miRNAs were

intersected with EST annotations to retain only loci with both forms of experimental support (e.g. small RNA library and EST), we observed the rates of recall and precision rise from 75.2% to 81.8% and 10.1% to 20.0%, respectively.

Together, the results from the above three approaches strongly argue that improvements in prediction rates of non-coding RNAs, and specifically ncRNAs undergoing processing or cleavage, such as miRNAs, can be attained if future investigations incorporate EST data into their ncRNA prediction protocols.

3.5 Discussion

The data above shows that existing EST records are significantly associated with known miRNA loci. By using miRNAs as a representative class of ncRNAs, these results also suggest that inclusion of EST data into future ncRNA prediction pipelines should be considered.

3.5.1 On the utility of EST data to support ncRNA predictions

We have shown that ncRNA predictions can be refined by including EST data using miRNA predictions as an example. Future development of non-coding gene prediction strategies should not only consider immediate features of genomic sequence such as sequence composition, motifs and RNA secondary structures, but also integrate these factors with knowledge of how transcripts are behaving in specific locales. By taking into account the predicted behaviour of one type of transcript, one can make inferences as to what would appear when examining data produced for an entirely disparate purpose. In the case of miRNAs, this function is served quite nicely with the combination of miRNA biogenesis where we were able to take advantage of the lack of splicing products, and an artifact of the molecular steps occurring during EST generation; namely reverse transcription primed by poly-T primers. Thus, EST proximity was combined with information regarding predicted RNA structures to predict miRNAs.

The use of high-throughput sequencing data (i.e. Generated with Illumina Genome Analyzer sequencers and similar technologies) has become more popular recently with continually decreasing costs. Computational validation of raw sequencing data is still nonetheless required, as appreciable fractions of reads generated from small RNA libraries are likely non-miRNA ncRNAs. For miRNA discovery, a standard

approach is to verify that sequence reads lie within predicted hairpin-like structures. Furthermore, despite the apparently unbiased results produced from a deep sequencing assay, the design of the contributing protocols to isolate small RNAs may themselves be a source of bias, for instance, when small RNA is purified by gel electrophoresis and excision any larger RNA species are discarded and lost. We have shown that including EST data in a precise way aids in focusing attention towards plausible miRNA candidates. It is expected that subsequent validations of strategic candidates with EST support would enjoy higher success rates.

We furthermore expect that including EST data in ncRNA prediction strategies is a generalizable approach for other types of ncRNAs that undergo cleavage during their production or function. This latter point is well illustrated by the Clec2d 3'UTR, which has been reported to contain a self-cleaving hammerhead ribozyme (Martick et al., 2008) with the ribozyme cleavage site in close proximity to a cluster of ESTs (See Figure 7-1 for this example).

One might question why not simply sequence all miRNAs, now that high-throughput sequencing technologies have sufficiently matured so that small RNA library sequencing is practical, provided that enough tissue is available to extract RNA from, and ignoring biases in RNA preparation. Foremost, if the appropriate EST libraries already exist, they offer a means to identify miRNAs without undertaking an additional sequencing project. In some cases RNA also remains limited (e.g. specific sub-sets of stem cells) and a pragmatic course of action still remains: *ab initio* computational methods followed by strategic validation by more sensitive methods.

3.5.2 On the utility of ESTs to support ncRNA validation

The most significant value that ESTs have for miRNA prediction is that they supply a probable tissue where miRNAs can be validated in. ESTs, by their very nature, indicate the context in which the genomic region(s) they map to are expressed. By extension, if sequences in a genomic region are expressed, it is reasonable to assume that genomic features in the vicinity are expressed and possibly functional in the tissues from which the culpable ESTs are derived from. Thus, a predicted miRNA with EST support may be reasonably expected to be expressed and functional in the tissue from which the EST is derived. Using ESTs in this way is investigated later in Chapter 4.

The consideration of ESTs also has a practical aspect in terms of prioritizing ncRNA predictions for manual validation. The differences in miRNA detection precision rates before and after addition of EST data shown in Table 3-2 illustrate that this method facilitates identifying sets of ncRNA predictions containing higher proportions of 'true miRNAs'. Relative increases in precision of 2 fold, 20 fold, and 1.2 fold were observed when examining miRNA predictions generated by small RNA library sequencing and predicting miRNA-like locations using RNALfold and RNAz, respectively. However, in practical terms the most significant result is that observed in the doubling of precision produced when considering EST-supported miRNA predictions derived from sequenced small RNA libraries, which is also the largest absolute increase out of the three. As manual validation of miRNA candidates is normally pursued after they have been identified, these results suggest that prioritizing validation of candidates in proximity to ESTs translates into an approximate doubling of the achievable overall validation rate.

3.5.3 On miRNA prediction

Each miRNA prediction method displayed differing performance, as illustrated by the consistently enriched relationships between ncRNA annotations and EST records which yielded differing Signal to Noise (S/N) ratios (Table 3-1). Analyses of ENCODE Consortium regions using similar methods of ncRNA prediction observed S/N ratios in their experimental setups which ranged between approximately 1.4 to 6.9 (Washietl et al., 2007). The S/N ratios calculated for analyses presented in this thesis fall within the range of independently reported in studies conducting similar analyses.

One striking observation were the impaired recall rates observed when examining data from the RNAz ncRNA prediction method, where rates produced with RNAz ranged between 31-38% versus 75-89% for both other methods. As RNAz requires conserved sequence alignments, the low rate of re-identifying known miRNA annotations could be attributable to known miRNAs displaying either 1) conservation levels below thresholds than those employed in the RNAz pipeline or 2) bona fide uniqueness to the mouse genome.

In addition to the variability in recall rates, differing increases in precision were observed with each ncRNA prediction strategy when adding ESTs to the respective protocols. As previously discussed, the largest and most dramatic relative increases in precision were seen after applying ESTs to ncRNA predictions generated by computational RNA structure predictions, while the largest absolute increase was observed in a small RNA library sequencing approach. Though completely computational ncRNA prediction pipelines are possible, the lower rates of overall absolute precision associated with computational methods of RNA identification versus

those associated with experimental RNA cloning and sequencing approaches are somewhat problematic. Considering the increasing ease of constructing a small RNA library and performing high-throughput sequencing, the drastically higher absolute precision rates suggest that RNA cloning should be performed, when practical, to generate candidate ncRNAs that are truly expressed. The resulting sequence data can be further analyzed and refined computationally to more precisely define the identities of cloned sequences.

Lastly and on a note specific to miRNA prediction, during calibration of RNAfold scripts the binding energies of true miRNA hairpins were observed to be generally below those of average hairpins, in agreement with previous reports (Bonnet et al., 2004). Considering that the DGCR8 co-factor required for Drosha cleavage of pre-miRNA hairpins contains two tandem dsRNA binding domains (Sohn et al., 2007), it is tempting to speculate that lower free energy of formation (ie. stronger RNA hairpin formation) is required to maintain dsRNA regions intact during pre-miRNA generation. This conjecture is supported by observations that Drosha cleavage locations can vary slightly but not stochastically, as the majority of mature miRNA sequences exhibit a dominant terminus when examining results from high throughput sequencing data (See examples in (Grimson et al., 2008) and (Morin et al., 2008)). This makes it tempting to hypothesize that the Drosha/DGCR8 containing complex recognizes a sequence motif that depends on precise and stable alignment of a dsRNA region, the discovery of which would greatly simplify the task of predicting miRNA hairpins. One recent report proposes that a GANNNNGA sequence has such a function in regions flanking miRNAs (Heikkinen et al., 2008) but such claims are currently restricted to miRNAs originating the

Caenorhabitis family. Further investigation into miRNA processing steps is needed.

3.5.4 Conclusion

We have shown that research into novel ncRNAs can be supported by considering EST data. EST sequencing experienced an early surge of interest before genome assemblies and gene prediction software were widely available and they represent an existing pool of data which can be leveraged for novel uses. The development of future non-coding transcript prediction strategies may benefit from reexamining apparently discordant EST data by taking into account transcript behaviors that are distinct from the classical model of ESTs being sequenced from cDNAs which are in turn derived from spliced mRNAs. The conclusions herein are particularly relevant today, when the throughput of sequencing technologies is increasing dramatically. Ongoing investigations should not discount producing EST data which may not directly capture full RNA species of interest, as applying up-to-date knowledge of transcriptome dynamics can help extract novel findings from minimal sets of experiments.

3.6 Materials & Methods

3.6.1 Source data

We used mm8, mm9 and hg18 releases of the mouse genome. 4-way conservation data used in RNAz included human (hg18), rat (rn4), dog (canFam2) and mouse (mm8 or mm9, where appropriate). Where required, miRNA, EST annotations, and multiZ annotations were obtained from the appropriate releases of the UCSC genome annotation datasets (Kuhn et al., 2009).

3.6.2 Generating miRNA predictions

miRNA predictions were generated by three different methods: RNALfold, RNAz, and using published Illumina sequencing data generated from a small RNA library (Babiarz et al., 2008).

3.6.2.1 Generating predictions with RNALfold

To generate miRNA predictions based on RNA structures without conservation, we used RNALfold from the Vienna RNA Package (v1.7), a tool that efficiently scans sequence identifying regions predicted to harbor locally stable RNA structures (Hofacker et al., 2004). We processed both strands of each genome analyzed, screening structures in real-time with filters optimized to reduce false positives.

To calibrate the real-time RNALfold filter, we obtained miRNAs annotated to mm8 and attempted to generate simple RNA hairpins using RNAfold, obtaining 363 miRNA records and adding ±5kb of flanking sequence to each window. We ran RNALfold against these sequences and applied heuristics discarding structures that were branched, did not exhibit a minimum length of bound bases (*s* values tested were 0, 21,

25, 30, 35, and 40 bp), or exhibiting a predicted Minimum Free Energy (MFE) below a given cutoff (ΔG values tested were 0,-10,-20,-30,-40,-50,-60,-70 kcal/mol). In addition, various values of the RNALfold window parameter were tested (L values tested were 150, 200, 250, 300, 350, 400, 600, 800, and 1000 nt). Precision/recall was calculated for all combinations of filter parameters and the settings retained were L = 600, s = 30bp, and dG = -30 kcal/more (See Figure 3-4). Furthermore, we discarded hairpins nested in larger hairpins with identical terminal loop coordinates. For illustration, this stage generated approximately 3.3 million hairpins with mm8.

3.6.2.2 Generating predictions with RNAz

To generate structures with RNAz (v1.0) (Washietl et al., 2005b), records from multiz30way and multiz17way alignments were parsed for mm9 and mm8 versions of the mouse genome, respectively, retaining mouse sequences along with hg18, canFam2, and rn4. The RNAz rnazWindow script was then run on alignments containing a minimum of four sequences to prepare jobs for RNAz. RNAz was run analyzing both alignment strands and retaining candidates with scores above 0.9. The rnazIndex script was used to convert results to BED files encoding structured windows.

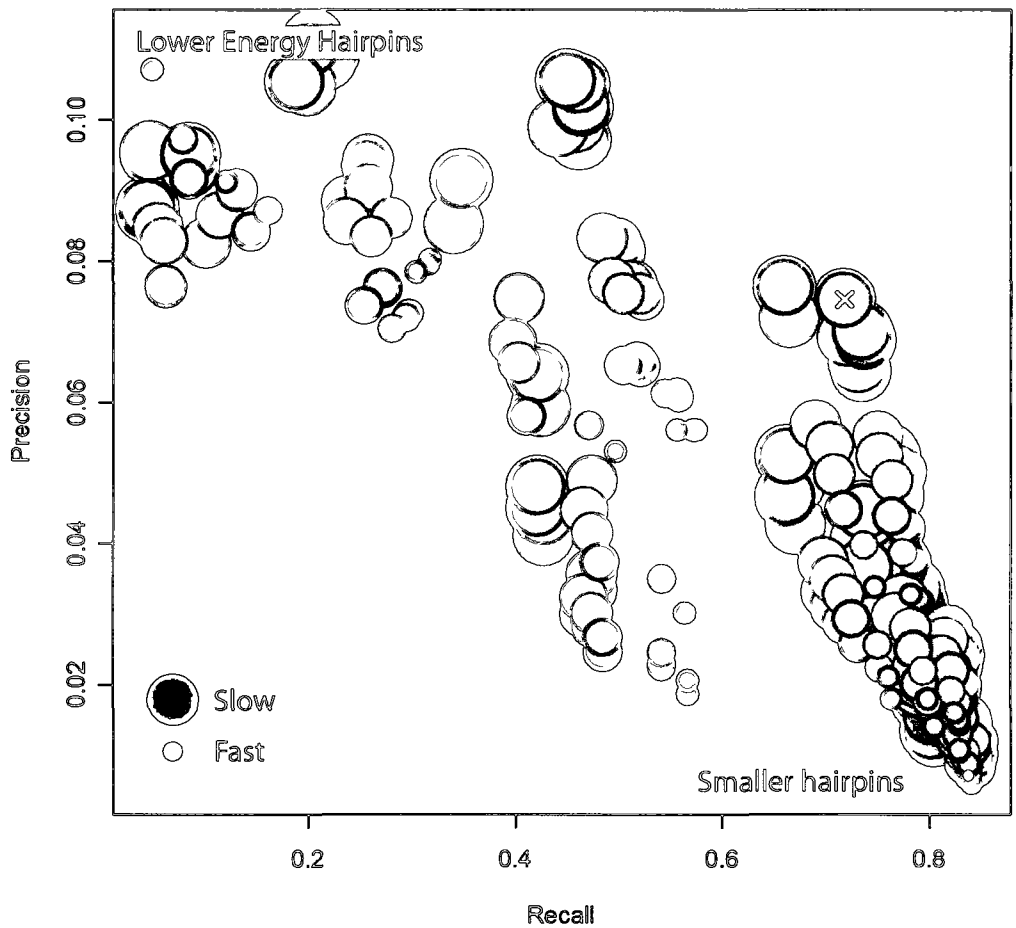

Figure 3-4: Calibration of real-time RNALfold filter parameters.
Precision and recall values were calculated by applying RNALfold on genomic DNA sequences centered on known miRNAs and filtering resultant RNA structures. Filtering parameters used were: hairpin size; hairpin MFE; and RNALfold window size. The point generated with parameters used in this study and corresponding precision and recall are shown with a white X. For computing precision/recall, miRNA-hairpin annotations had a minimum overlap of 80%.

3.6.2.3 Generating predictions with Illumina sequencing

We processed the raw Illumina data provided in GEO record GSM314552 (Babiarz et al., 2008) by aligning all sequence fragments to mouse genomes with novoalign (Novocraft, 2009) and generating wiggle (UCSC .wig) tracks showing total

normalized expression for each nucleotide. Each Illumina fragment had a normalized expression value which was 1 divided by the number of genomic locations it mapped to. We defined miRNA candidate regions by identifying contiguous 21nt sequence windows with a minimum expression value which was also the maximum value within ±15nt. Since this sequencing library was shown to capture non-miRNA species as well (Babiarz et al., 2008), we retained windows which were predicted to reside within a hairpin, predicted with RNAfold (Hofacker et al., 1994).

3.6.3 Observing relationships between ESTs and known miRNAs

To generate the plots illustrating the relationship between miRNAs and ESTs (Figure 3-3 and Figure 3-6), 3' coordinates of microRNA annotations on human and mouse genomes (hg18 and mm8, respectively) were identified and 2kb sequence windows centered on this location were retrieved. Coordinates of EST termini mapping within each sequence window were identified and a score for each relative location was incremented by $1/n$, where n equals the number of ESTs mapped to the miRNA associated sequence window. Pseudocode for this process is in Table 3-3. Scores for all miRNAs were aggregated, binned in 10nt windows, and plotted.

Significance scores for peaks observed in each genome assembly were determined by bootstrapping as follows. Briefly, for each miRNA window, coordinates were shuffled to random locations on the same chromosome and EST termini were identified in ±1kb genomic windows. EST accumulation was scored as described above, and the scores for each genome assembly and test (upstream/downstream of miRNA) were aggregated. The number of times a replicate generated a central peak in excess of the observations was counted and used to determine a P-value.

3.6.4 Generating EST supported miRNA predictions

3.6.4.1 EST intersections with RNALfold

To further screen candidate miRNA hairpins, we filtered hairpin structures based on their normalized ΔG value and the distance to a downstream EST terminus, as true miRNAs were observed to have lower than average normalized MFE than the entire set of genomic hairpins. We thus examined the effect of filtering predicted structures based on precision/recall values computed using NFEs between 0 and -1 kcal/mol and EST proximities between 0 and 200 nt (See Figure 3-5). We retained predicted hairpin structures exhibiting a maximum NFE of -0.44 kcal/mol and with one or more EST termini within the structure, permitting the generation of mature miRNA and up to 14nt downstream of the 3' end.

3.6.4.2 EST intersections with RNAz

RNAz object windows were associated with ESTs if an EST terminus was located within the sequence window plus 14nt as determined above. As loci generated by RNAz did not preserve strandedness, objects were considered on either strand.

3.6.4.3 EST intersections with Illumina data

As genomic coordinates of mapped Illumina sequencing fragments approximate mature miRNA sequences, not pre-miRNA hairpins, we determined the maximal distance between sequencing fragments independently. We examined distances between termini of known mature miRNAs and the extents of their corresponding pre-miRNAs. After calculating upstream and downstream values for each miRNA/pre-miRNA pair, the mean of the larger values was determined to be 51nt. This distance was the cutoff with which sequence fragment/EST intersections were considered.

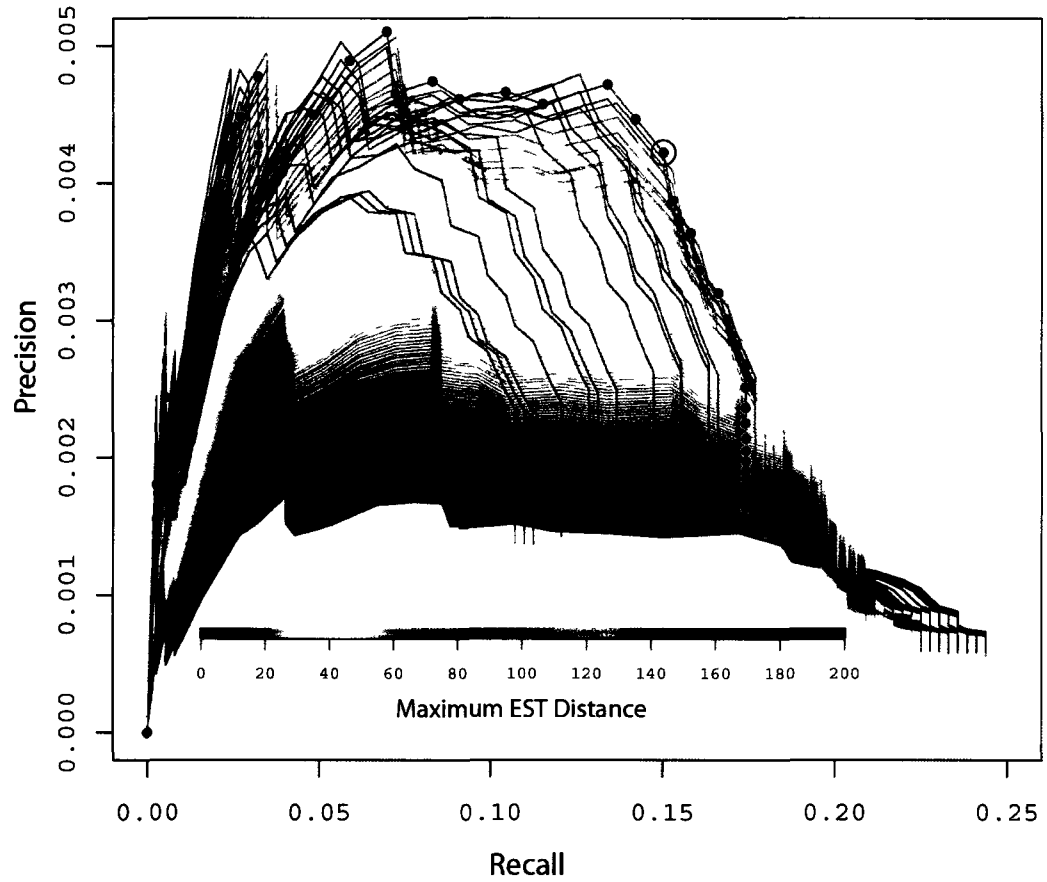

Figure 3-5: Calibration of filter based on Normalized Free Energy and miRNA-EST distances. Cutoffs for Normalized Free Energy (NFE) and miRNA-EST distance were calibrated to filter the set of hairpins generated from the genomewide scan of RNALfold. Precision/Recall curves shown for each miRNA-EST distance (*d*) cutoff, with individual points corresponding to NFE cutoffs at each value of *d*(Black points, moving counterclockwise). NFE values range from 0 to -1 kcal/mol and miRNA to EST distances range from 0nt to 200nt.

3.6.5 Statistical tests

Bootstrapping to estimate significance of the peak graphs was performed by randomizing EST start points in the ±1kb genomic windows and counting the number of times replicates generated a peak at the central ±10nt.

3.6.6 Estimating significance of miRNA-EST associations

Enrichment statistics for EST supported miRNA predictions were determined by bootstrapping. For each bootstrap trial, candidate miRNA annotations were shuffled to random locations preserving chromosomal annotation and strandedness. Candidate miRNAs were computed for each trial by intersecting ESTs according to the protocol appropriate for each miRNA prediction approach, as described above.

3.6.7 Precision/Recall calculations

To calculate the differences in precision and recall between miRNA predictions generated with and without the inclusion of EST data annotations, the rates at which known miRNA annotations could be re-identified with the various miRNA prediction approaches were computed. As known and annotated miRNAs without nearby EST evidence cannot be identified with a prediction method employing EST data, it was necessary to determine which miRNA annotations could be theoretically captured by the various miRNA prediction approaches used.

To adjust for this, the number of known and annotated miRNAs was reduced to include only those with EST annotations in the vicinity. For RNA structural predictions using RNALfold and RNAz, the number of known and annotated miRNAs annotations with ESTs annotations within 14nt of either terminus of structures was determined. For predictions generated with high-throughput sequencing, the number of miRNA annotations with ESTs within 51nt of an annotation terminus was determined (See Figure 3-8). This value was used as high-throughput sequence data aligned to the genome identifies mature miRNA windows, while the available genomic (mm9) miRNA annotations map the pre-miRNA hairpins without specifying the mature miRNA

sequence; accordingly, a larger distance was permitted to account for mature miRNA sequences deriving from the 5' arm of pre-miRNA hairpins. The distributions used to determine this value are shown in Figure 3-8.

For the high-throughput sequencing data from (Babiarz et al., 2008) ("Illumina data"), the unique miRNAs reported as detected were the number of known and annotated miRNA hairpins which also contained at least one Illumina sequencing fragment mapped within the accepted structure as annotated on mm9. For miRNA predictions generated by RNALfold and RNAz, the numbers reported as detected were those exhibiting a greater than 75% overlap between the predicted structures and the known miRNA structures, as annotated on mm9. In all cases of computing predictions either with or without EST evidence, the number of miRNAs predicted were those included in the sets that were predictable by the respective methods, as determined above.

3.7 Supplementary Information

3.7.1 Supplementary Figures

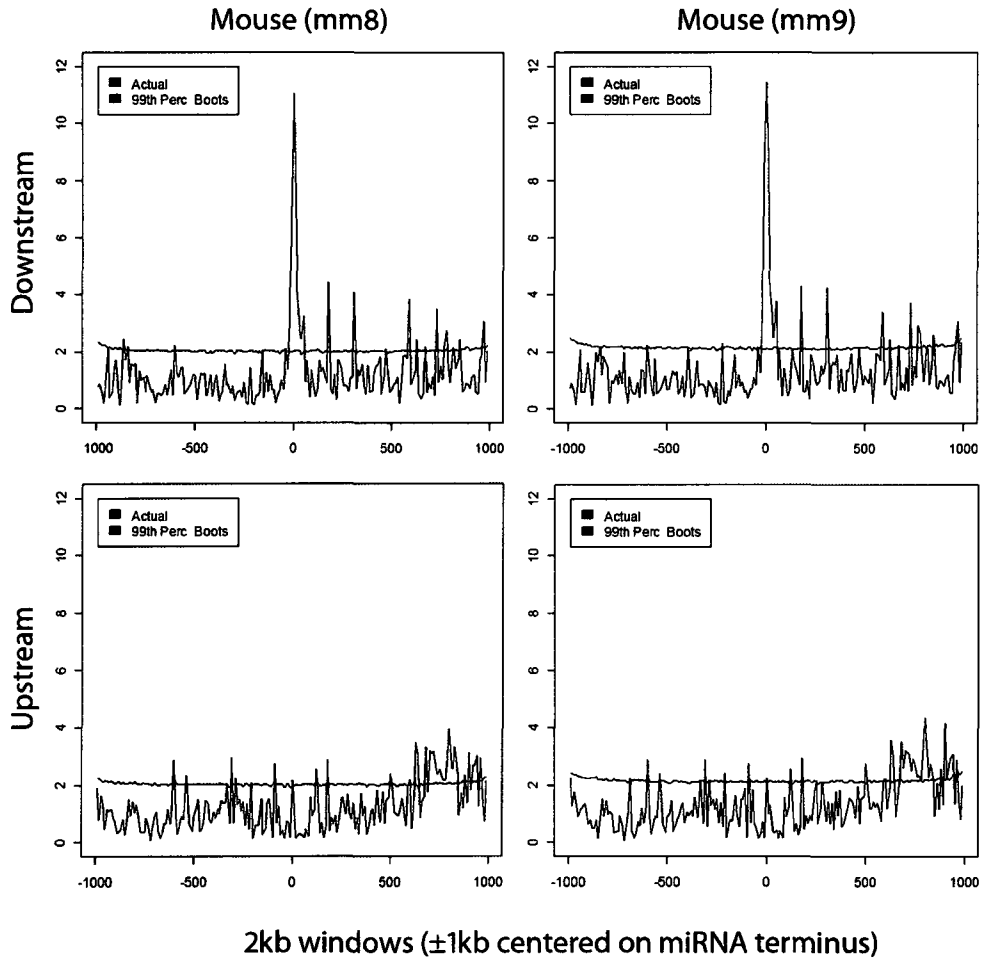

Figure 3-6: Distributions of EST peaks near miRNAs are similar across genome versions. Distributions of miRNA-ESTs are highly similar across different versions of the mouse genome (mm8 and mm9). EST occupancy surrounding miRNAs was calculated as described in Methods.

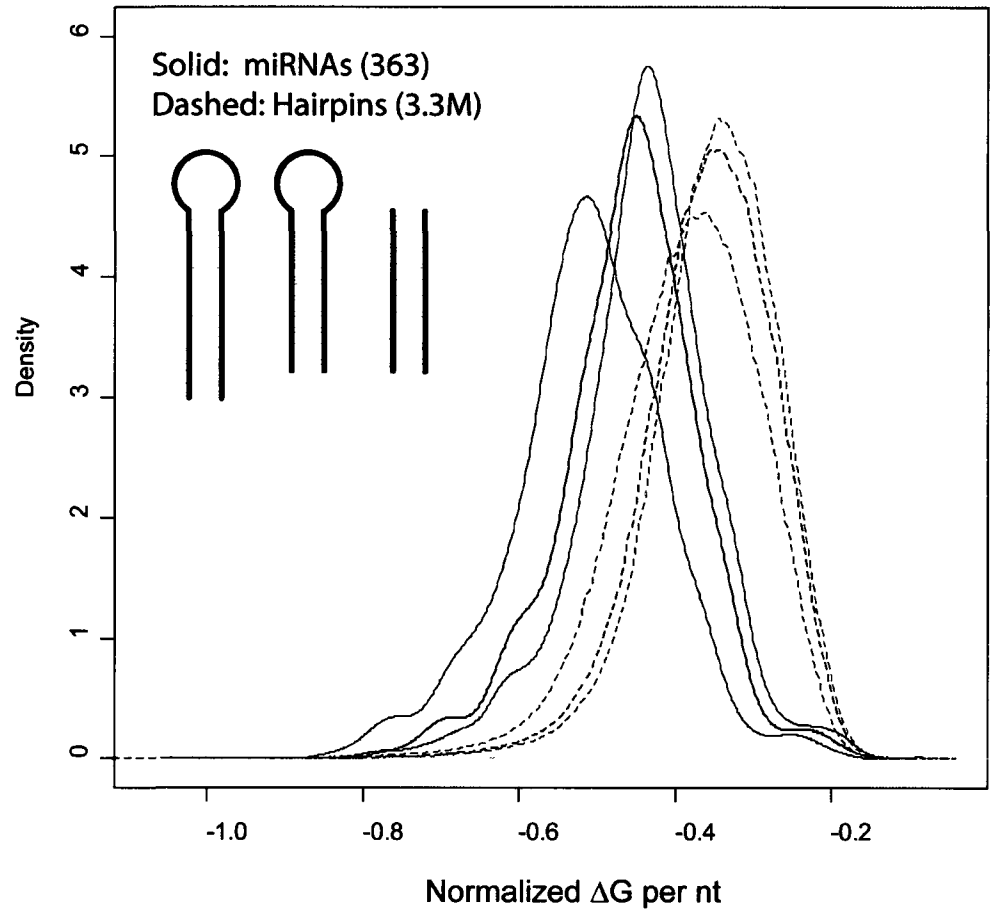

Figure 3-7: Distributions of Normalized Free Energies in sets of miRNAs and set of genome-wide predictions
Three distributions are shown: The red, black, and blue distributions correspond to values derived by normalizing the total MFE by the number of nucleotides in the hairpin sections as highlighted in the inset figure.

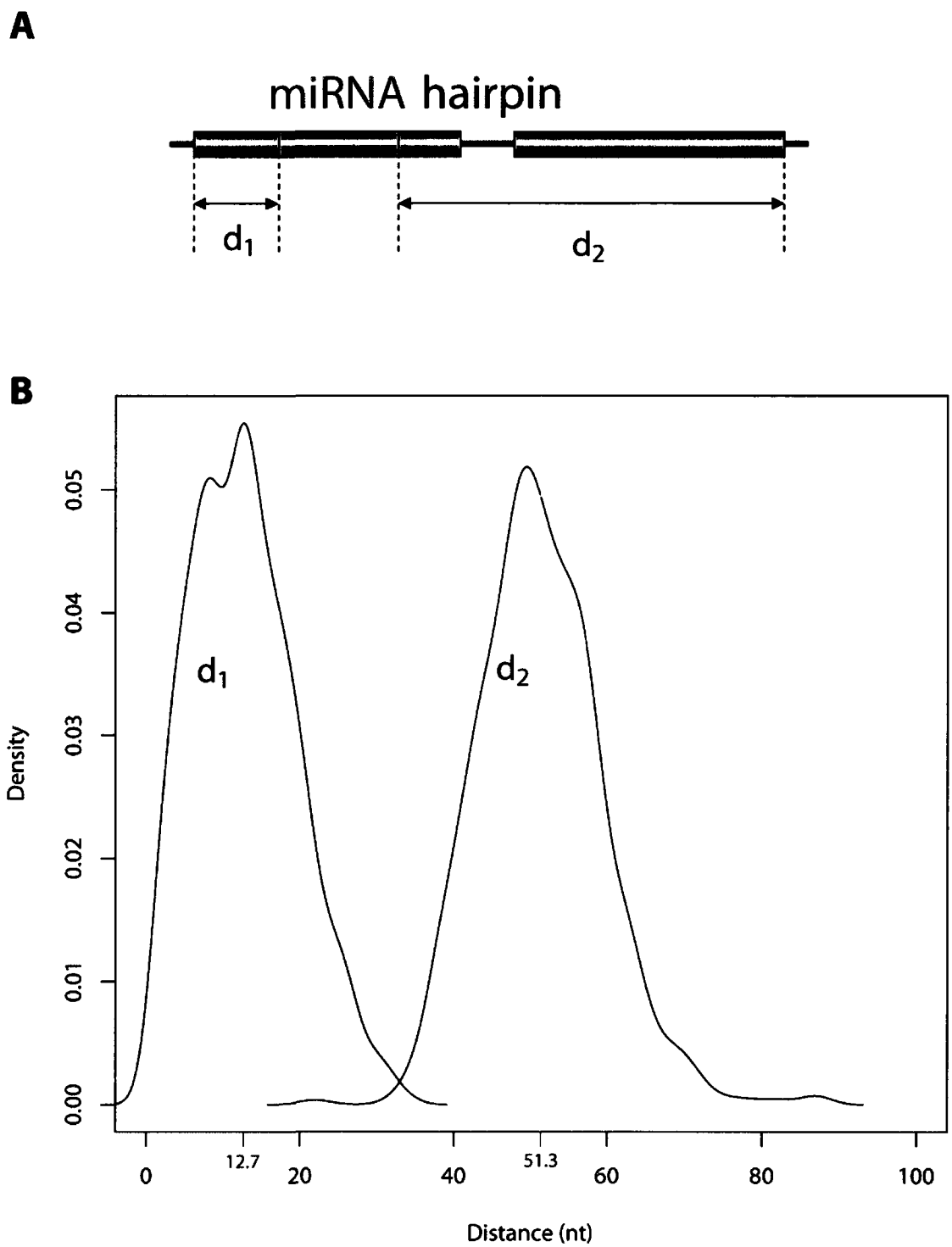

Figure 3-8: Determination of mature miRNA placement within pre-miRNA windows
553 mature miRNA sequences were mapped within their parent pre-miRNA hairpin annotations and distances between termini were measured (A). d_1 and d_2 correspond to the shorter and longer distances, respectively. Distributions of the values are shown in B. The mean of the d_1 and d_2 are shown (12.7nt and 51.3nt, respectively).

3.7.2 Supplementary Tables

```
For each miRNA {
      Shuffle miRNA to new location on chromosome
      Identify coordinates of ESTs ±1kb of terminus (5' or 3')
      Foreach EST terminus {
            Increment relative EST terminus by 1/(n ESTs)
      }
}
```

Table 3-3: Pseudocode for EST terminus finding algorithm.

Chapter 4 Identifying novel microRNA markers in muscle

4.1 Preface

Sections of this chapter were reproduced from the following publication:

Paul M. Krzyzanowski, Feodor D. Price, Enrique M. Muro, Michael A. Rudnicki, Miguel A. Andrade-Navarro. **Novel muscle miRNA discovery through integration of EST data and predicted RNA secondary structures.** Submitted.

Author Contributions for this chapter: PMK designed the project, performed computational analyses, and drafted the manuscript. FDP performed experimental work and provided feedback on experimental results. MAR and MAA provided mentorship and direction.

4.2 Abstract

We used the EST supported miRNA prediction methodology in the previous chapter to identify and validate novel miRNA hairpins exhibiting potential control over either Pax7 or Myf5, which are key regulators governing cell state transitions during myogenic differentiation. To these ends, miRNA predictions with predicted expression in differentiating myoblast cells were generated according to the model previously presented in Chapter 3 and verified by microarray and Northern blotting. Several known and predicted myogenic miRNA hairpins were detected and confirmed. Further investigation demonstrated that several ncRNA hairpins expressed in differentiating myoblast cell culture contribute towards Myf5 reduction, potentially identifying novel ncRNA regulators of myogenic stem cell differentiation.

4.3 Introduction

Known muscle microRNAs have been shown to play critical roles during growth and regeneration. MicroRNAs have also been shown to participate in key regulatory points during differentiation and development. An overview of muscle development and miRNAs involved therein was provided in Section 1.5, "Muscle development and markers thereof". We endeavored to identify novel ncRNAs active during differentiation of muscle tissues using a C2C12 myoblast model.

4.3.1 Objectives, Hypotheses

4.3.1.1 miRNA hairpins active during myogenic differentiation can be identified using combined EST and predicted RNA structure annotations

The EST based miRNA prediction method presented in the previous chapter will be used to identify miRNA-like hairpins that are expressed in differentiating C2C12 myoblasts. The RNA predictions will be validated using an experimental approach which will be microarray based, with further experimental validation to verify positive candidates after analysis of the microarray data.

4.3.1.2 miRNA hairpins expressed in differentiating myoblasts target myogenic regulatory transcripts

Putative miRNA hairpins predicted and validated in the previous section will be examined for potential to target transcripts involved in myogenic differentiation. Novel miRNA hairpins expressed in muscle are hypothesized to show capacity to repress transcripts involved in myogenic regulation, such as Pax7, Myf5, or Myod1.

4.4 Results

4.4.1 A custom microRNA microarray identifies myoblast expressed miRNA-like hairpins

To examine whether novel myogenic miRNAs could be identified, the expression of genomic locations containing putative miRNA hairpins was examined in RNA derived from differentiating myoblasts or mES cells using a custom tiling microarray that we designed to evaluate whether small RNAs are expressed from each locus. To identify potential sites of miRNAs expressed in myogenic cell types, genomic sites with predicted RNA secondary structures were cross-referenced with EST evidence derived from muscle/myoblast and embryonic tissues (For methods, see Section 3.4.2.1). The advantage of this miRNA prediction protocol is that it provides putative miRNA-like RNA loci with annotation concerning their probable tissue(s) of expression based on the biological sources of supporting cDNA library sequences. An overview of various numbers of predictions produced in muscle, myoblast, or other tissues is shown in Table 4-1.

Probable Source	Predictions	Probable Source	Predictions	Probable Source	Predictions
Adipose	43	Hematopoietic	482	Oviduct	29
Adrenal	11	Hippocampus	67	Pancreas	591
Amnion	97	Hypothalamus	103	Parthenogenote	38
B cells	126	Inner Ear	420	Pituitary	91
Bladder	77	Intestine	17	Placenta	375
Blastocyst	157	Islet	168	Pooled Cells	2569
Blood	14	Isthmus	85	Prostate	52
Bone	45	Joints	285	Rathke's Pouches	966
Bone Marrow	36	Kidney	973	Retinal	1457
Bowel	17	Lens	11	Salivary Gland	153
Brain	5201	Leukemia	107	Sarcoma	10
Cancer	404	Limb	166	Sertoli Cells	51
Carcinoma	32	Liver	656	Skin	321
Cardiac	16	Liver Tumour	242	Small Intestine	147
Cecum	26	Lung	406	Spermatids	72
Cerebellum	419	Lung and Heart	25	Spermatocytes	121
Choroid	113	Lymph Node	104	Spermatogonia	52
Colon	406	Lymphoma	149	Spinal Cord	176
Cornea	13	Macrophage	24	Spinal Ganglion	88
Corpora quadrigemina	159	Mammary Gland	906	Spleen	626
Corpus striatum	36	Mammary Tumour	11	Stomach	156
Cortex	94	Mandible	20	Stromal Cells	133
Dendritic	674	Marrow	448	Subfornical Organ	40
Diaphragm	46	Maxilla	18	Sympathetic Ganglion	74
Diencephalon	81	Melanocyte	155	T Cells	51
Ear	23	Melanoma	270	Taste buds	114
Ectoplacental Cone	14	Mesenchymal	66	Testis	995
Egg	259	Mesonephron	18	Thymus	898
Embryo	785	Mixed Cells	35	Thyroid	97
Embryonic Carcinoma	13	Mullerian Duct	49	Tongue	89
Embryonic Stem Cell	874	Muscle	52	Tooth	35
Epididymis	26	Myotubes	144	Total Fetus	254
Erythroblast	32	Neural Stem Cells	176	Trophoblast	88
Eye	967	Nullipotent	25	Tumour	1493
Fertilized eggs	36	Olfactory Systems	184	Unknown Origin	346
Fetus	412	Oocyte	69	Urogenital System	36
Fibroblast	237	Organ of Corti	75	Uterus	17
Gall bladder	24	Osteoblast	66	Vagina	23
Gastric Epithelium	36	Osteoclast	142	Vein	71
Germ cells	131	Otocysts	130	Ventricle	514
Head	558	Ovary	104	Whole Embryo	1661
Heart	398	Ovary and Uterus	24	Wolffian Duct	33

Table 4-1 Overview of predicted miRNA hairpins in different tissues
13952 unique putative miRNA hairpins were annotated in the mouse genome (mm9), based on predictions generated from RNA structures with supporting EST annotations Counts above do not correspond to this value due to some miRNA predictions having ESTs derived from multiple sources, and sources with under 10 predictions being excluded due to space limitations All biological sources not shown are retained in the source database

To generate the design for the custom tiling array, sequences for each site associated with putative myogenic or embryonic stem cell expression were tiled with 60nt windows, overlapping adjacent windows by 20nt. In addition to the experimental probes, multiple control sequences were included and probes designed to detect their expression status on the microarray design (Summarized in Table 4-5). RNA from differentiating C2C12 myoblasts and J1 embryonic stem cells was isolated and submitted for labeling and hybridization to microarrays (UHN Microarray Centre). Raw data was normalized and processed according to previously published methods (For further details on microarray design and analysis refer to Methods starting on Page 150).

We identified high-confidence putative miRNA hairpins by computing the fold enrichment between probes unanimously expressed in either muscle or ES cells. After data normalization and analysis, 148 probe sets with >2 fold mRNA enrichment in muscle RNA versus ES derived RNA were identified, including muscle control probe sets. Included in this set were probes designed against known muscle miRNAs mir-206 (McCarthy, 2008; Sweetman et al., 2008; Sweetman et al., 2006), mir-133a/b, as well as let-7 family miRNAs. In contrast, ES specific control miRNAs derived from the mir-293 cluster were strongly expressed in the RNA samples derived from ES cells over those from differentiating C2C12 cells. Figure 4-1 shows the robust reciprocal expression of the major muscle and ES cell control miRNAs in the correct RNA samples.

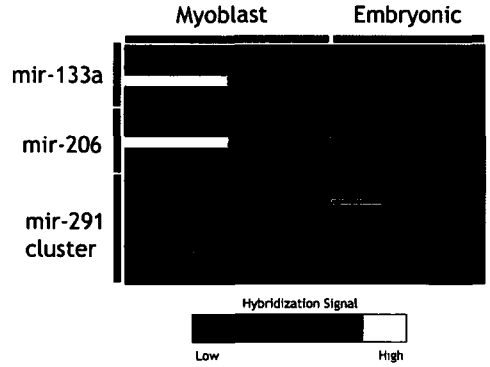

Figure 4-1: Control miRNAs are expressed correctly on microarray.
Heatmap showing correct expression of control miRNA probe sets for muscle (mir-133a and mir-206) and ES cell control (mir-293 containing cluster). Probe sets were composed of multiple probes tiling over each pre-miRNA hairpin sequence, and only probes overlapping mature miRNA sequences yielded expression signals.

4.4.2 Known miRNAs expressed in differentiating myoblast cultures

We re-examined the hairpin sequences tiled on the microarray to identify known miRNAs that were both captured by our prediction method and detected as differentially expressed in myoblasts. Closer inspection showed that several putative miRNA probe sets were tiled over mir-24-2, mir-351, and mir 26a-2, mir-92-2 and mir-7-2 without these miRNAs being selected as control probes. miRNAs potentially upregulated in myoblasts included sequences overlapping mir-24-2, mir-351, and mir-26a-2 while hairpins representing mir-92-2 and mir-7-2 loci were downregulated (Table 4-2).

Hairpin ID	Fold Increase		miRNA	Location
	C2C12	ES		
hp1643988	141.56	-	mir-24-2	chr8:87098910-87099035
hp3202291	108.27	-	mir-351	chrX:49297888-49297981
hp907540	94.12	-	mir-26a-2	chr10:126398456-126398589
hp1980840	-	1.52	mir-92-2	chrX:48986474-48986552
hp85655	-	1.72	mir-7-2	chr7:78761795-78761882

Table 4-2: Hairpins with miRNA overlap exhibiting differential expression on microarray.
Hairpins overlapping miRNAs that were tested on the microarray and exhibited differential expression. Hairpins overlap mm8 annotated miRNAs over at least 33% of their nucleotides.

We manually verified whether the microarray probes yielding these signals were

aligned with mature miRNA sequences. The mature miRNA mir-24-2 sequence was detected at 141 fold higher in C2C12 cells by two independent probe sequences (fold changes of 134.8 and 141.5 for each). The pre-mir-24 hairpin contains another mature miRNA, mir-24-2*, which overlap by one of these two probes. Mature mir-26a sequence, contained in pre-mir-26a-2, was also detected. As well, microarray probes likely detected mature mir-351 sequences due to sequence overlap between probe and target. The only mature miRNA sequence which was expressed at lower levels in myogenic samples and aligned with the correct microarray probe was mir-92a. Probes tiling over the mir-7-2 locus, which yielded hybridization signals slightly higher in ES samples, did not overlap the mature miRNA sequence and the source of this signal is unknown.

To better understand the context by which these microarray signals could have arisen, the genomic contexts of these candidates were examined. Of highest relevance to the observed expression in myoblasts, the miRNA hairpin mapping to mir-26a-2 was found to be intronic to a long splice variant of Ctdsp2 (carboxy-terminal domain, RNA polymerase II, polypeptide A, small phosphatase 2); it appears that a shorter Ctdsp2 variant would be produced if this miRNA were excised from the longer transcript in analogy to mir-126 (Kim and Kim, 2007). This gene is able to induce mesodermal lineage commitment in a *Xenopus* system (Zohn and Brivanlou, 2001). Mir-26a-2 is also known to be expressed during myogenesis, where it represses Enhancer of Zeste (Ezh2) (Wong and Tellam, 2008).

Interestingly, mir-351 is potentially part of a miRNA cluster with mir-503 and mir-322 lying within 2kb upstream. Supporting this possibility are numerous ESTs

mapped in the vicinity, though no contiguous regions of expression have been shown to span the genomic region defined by these three miRNAs. Mir-24-2 is the downstream most member of miRNA cluster containing mir-27a and mir-23a. This cluster is downstream from Zswim4, an uncharacterized Zinc finger containing protein.

4.4.3 Target prediction

To investigate potential roles for the known miRNAs identified in the differentiating myoblast RNA, miRNA target prediction was pursued. The rationale was that mature miRNAs expressed during myogenic differentiation may be repressing transcripts during this process, and that identifying these targets might yield potential functions for the miRNA(s). The TargetScan method of identifying miRNA targets was re-implemented using a custom script shown to produce results that are very highly correlated to the results produced by the original authors (Grimson et al., 2007)(See Methods; 'Replication of published miRNA targeting protocol' on Page 157 for further details).

Potential target transcripts which mature sequences for mir-24-2, mir-26a, and mir-351 might be repressing were predicted, hypothesizing that known myogenic genes would be targets. Rather than consider the targets individually, the scores were combined multiplicatively to identify potential transcripts that might be synergistically repressed by all three (Table 4-3). This method of combining scores penalizes potential targets where one or more miRNAs contribute a low score.

Symbol	Description	miR-24	miR-26a	miR-351
Plod2	Procollagen-lysine,2-oxoglutarate 5-dioxygenase 2 Precursor	-0.3234	-0.7219	-0.2086
Dmd	Dystrophin	-0.352	-0.2867	-0.4483
Tbc1d30	TBC1 domain family member 30	-0.4755	-0.5532	-0.1713
Ret	Proto-oncogene tyrosine-protein kinase receptor ret Precursor (C-ret)	-0.3827	-0.5018	-0.1895
4732454E20Rik	Protein FAM26D	-0.7775	-0.3364	-0.1225
Irf4	Interferon regulatory factor 4	-0.1166	-0.516	-0.5252
Papola	Poly(A) polymerase alpha	-0.5424	-0.229	-0.2491
Zfp697	Zinc finger protein 697	-0.492	-0.3602	-0.1722
Slc25a35	Solute carrier family 25 member 35	-0.3222	-0.2178	-0.4084
St8sia4	CMP-N-acetylneuraminate-poly-alpha-2,8-sialyltransferase	-0.3275	-0.335	-0.2319
Parp14	Poly [ADP-ribose] polymerase 14	-0.401	-0.2517	-0.2463
Mfhas1	Malignant fibrous histiocytoma-amplified sequence 1 homolog	-0.099	-0.5613	-0.4147
Sfxn4	Sideroflexin-4	-0.3599	-0.4244	-0.1393
Hdx	Highly divergent homeobox	-0.3016	-0.4808	-0.1407
Trp53inp1	Tumor protein p53-inducible nuclear protein 1	-0.1942	-0.3591	-0.2874
Kcnh7	Potassium voltage-gated channel subfamily H member 7	-0.1669	-0.4631	-0.2587
2410018C17Rik	dorsal neural-tube nuclear protein	-0.3836	-0.2682	-0.1867
Ceacam1	Carcinoembryonic antigen-related cell adhesion molecule 1 Precursor	-0.2563	-0.1633	-0.441
Mmp16	Matrix metalloproteinase-16 Precursor	-0.3102	-0.4	-0.1485
Tmem161b	Transmembrane protein 161B	-0.3564	-0.213	-0.238
Ttc18	Putative uncharacterized protein	-0.1878	-0.5691	-0.1634
Tsc22d2	TSC22 domain family, member 2	-0.2407	-0.4409	-0.1572
Fbxl19	F-box/LRR-repeat protein 19	-0.2809	-0.6477	-0.0909
Camsap1	Calmodulin-regulated spectrin-associated protein 1	-0.3152	-0.3169	-0.1635
5830411G16Rik	Late secretory pathway protein AVL9 homolog	-0.7576	-0.1321	-0.1574
St8sia6	Alpha-2,8-sialyltransferase 8F	-0.2345	-0.5804	-0.1113
Ceacam2	Carcinoembryonic antigen-related cell adhesion molecule 2 Precursor	-0.2908	-0.1159	-0.442
Bivm	Basic immunoglobulin-like variable motif-containing protein	-0.1627	-0.596	-0.1484
Coq3	Hexaprenyldihydroxybenzoate methyltransferase, mitochondrial Precursor	-0.2912	-0.4987	-0.0974
A530082C11Rik	Solute carrier family 35 member E2	-0.2131	-0.3223	-0.2017

Table 4-3: Top 30 predicted targets for synergistic repression by mir-24, mir-26a, and mir-351. Top 30 targets for synergistic repression by mir-24, mir-26a, and mir-351. Scores for the lowest (best) scoring transcript were retained in cases where the same gene was targeted repeatedly

We investigated the high rank for potential targeting of the Dystrophin (Dmd) transcript. Dystrophin is transcribed from a large genomic region (~3 megabases), contains 79 exons, and undergoes numerous alternative splicing patterns (Sironi et al., 2002). As the second best target, Dystrophin was targeted most robustly by mir-351 (TargetScan score = -0.45). Upon closer inspection, the Dystrophin 3'UTR targeted by all three miRNAs was a longer Dmd transcript (ENSMUST00000114000) which terminates approximately 1kb beyond the main transcription stop site, and which has a single EST as evidence of transcription (CK385805). Despite the low number of ESTs associated with this Ensembl transcript, it is not possible to conclude that the *in vivo* expression of this longer Dmd transcript is proportionally low. When examining targeting of the three mature miRNAs for the main Dystrophin 3'UTR, only mature mir-26a was found to map within it (Data not shown). As different Dystrophin splice variants are expressed in different tissues, with the 3′ region of the transcript undergoing alternative splicing to produce tissue-specific transcripts in brain, cardiac and muscle fibers (Sironi et al., 2002), it is possible that miRNAs may be involved to enforce transcription of some Dmd isoforms over others in certain contexts.

4.4.4 Novel miRNA hairpins are expressed in differentiating myoblasts

To further validate microarray signals associated with putative miRNA hairpins, several candidates with strong expression scores were selected for further analysis to determine whether the expression and molecular size of RNAs responsible for the array expression values were consistent with miRNA presence. To achieve this, Northern blotting was used to visualize the size of RNA species responsible for microarray signals, using the respective microarray probe sequences. Any bands observed in the ~21nt or

60-80nt size ranges would suggest that mature miRNA and pre-miRNA sequences are present, respectively.

Several of the selected probe sets indeed detected expression of small RNAs in the ~20-30nt size range (See Figure 4-2). Positive controls using sequences for mir-133a and mir-206 produced extremely strong signals as these miRNAs are the most highly abundant in muscle tissues. Three predicted miRNA hairpins—hp566131, hp1621227, and hp30491555—were detected at higher levels in myoblasts in the microarray data and Northern blotting confirmed this differential expression in addition to supplying information regarding their size. Mir-26a is a known cardiomyocyte related miRNA which was captured using this prediction protocol and it too showed specific expression in myoblast at a very low level. These results indicate that the microarray signals were due to small RNA species, likely derived from predicted miRNA hairpins, which are expressed in myogenic cells undergoing differentiation at higher levels than embryonic stem cells.

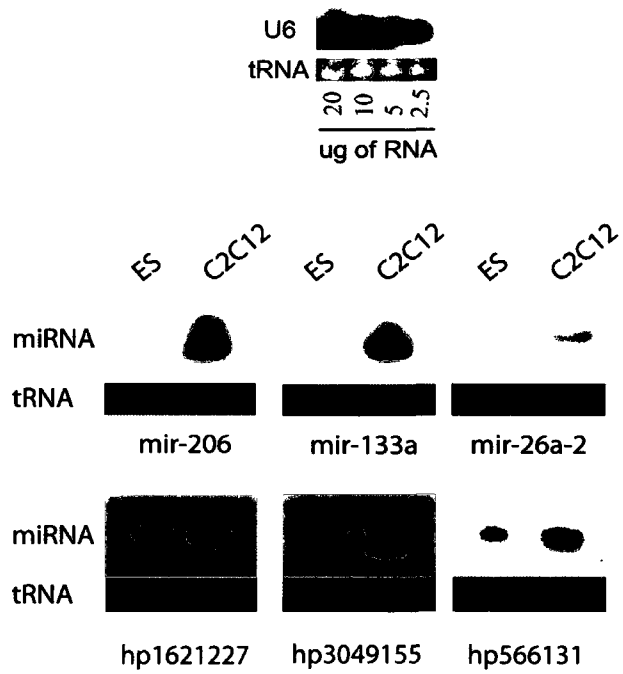

Figure 4-2: Northern blots showing putative miRNAs at same size as known miRNAs.
Top: Control loading of total RNA showing concentration dependent U6 hybridization. Bottom: C2C12 enrichment of various miRNAs and hairpins. Known miRNAs include mir-133a, mir-206 and mir-26a-2 while the others were from those predicted in this study. All Northern blotting courtesy of Feodor Price.

To estimate a validation rate for the EST based miRNA prediction protocol, attempts to validate seven (two known, five novel) out of the 148 miRNA candidates enriched in the C2C12s were made. The resulting data yielded support for four positive candidates (one known, three novel). Tentative validation rates of 50% and 60% can therefore be implied for known miRNAs and novel miRNA hairpins, respectively, although greater numbers of observations would be required in order to increase the robustness of the rates. It is conceivable that the validation rates would fall when considering a larger spectrum of differentially expressed hairpins, but nevertheless these rates compare favourably with previously reported validation success rates of approximately 25% in a larger screen of 175 general ncRNA predictions (Washietl et al.,

2007).

4.4.5 *ncRNA hairpins upregulated in differentiating myoblasts repress Myf5*

To determine whether any predicted RNA hairpin sequences detected by the microarray were active in myogenic signalling networks, we sought to identify candidates that could suppress specific mRNA transcripts. As both known and predicted miRNAs were expressed in differentiating myoblasts, we hypothesized that they might function as repressors of transcripts downregulated during various stages of myoblast differentiation. Specifically focusing on Myf5, Pax7, or MyoD1 as potential targets, we selected putative miRNA hairpins that were both 1) upregulated in the differentiating myoblast samples and 2) exhibited potential to target one or more of the mRNA transcripts.

The miRNA target prediction algorithm employed was the published TargetScan algorithm (Grimson et al., 2007). After replicating the software scripts required to use this algorithm (See Section 4.8.1: Replication of published miRNA targeting protocol, Page 157), we expected to find examples analogous to the mir-206 mediated repression of Connexin43 (Gja1) expression during myoblast differentiation (Anderson et al., 2006), an interaction identified during testing of the miRNA targeting pipeline (See Table 4-8).

Each of the microarray validated RNA hairpins was scored for its ability to target each of the three transcripts independently. We aimed to focus on hairpins demonstrating both high expression in differentiating myoblast culture and strong predicted repression against any one of the transcripts of interest (Figure 4-3).

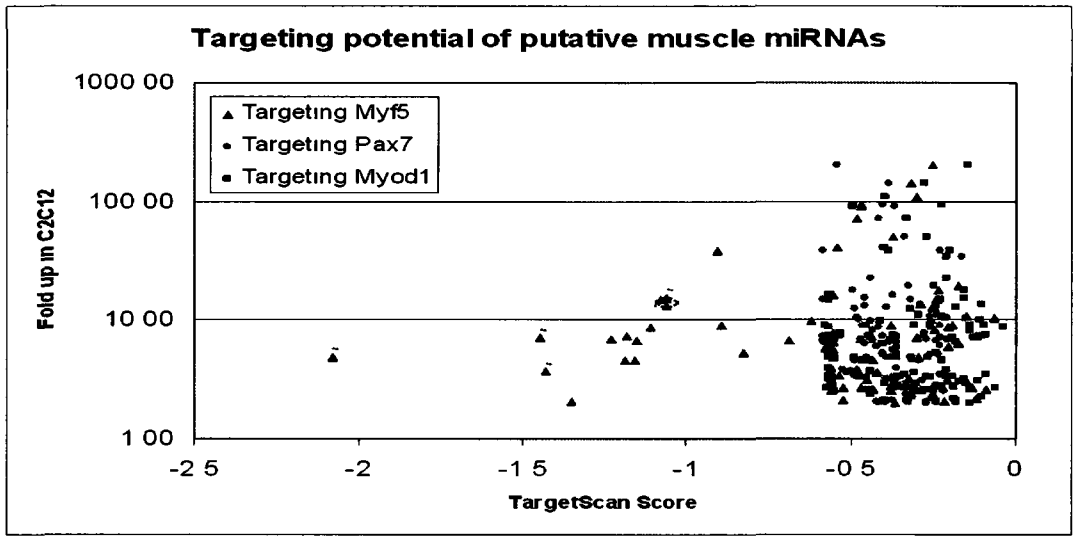

Figure 4-3 Scatterplot of hairpin expressions versus TargetScan scores
Predicted miRNA targeting was predicted for Myf5, Pax7 and Myod1 UTRs for all putative miRNAs exhibiting over 2-fold increases in expression in differentiating C2C12 myoblasts Points enclosed with dashed circle represent hairpins that were followed up Targeting scores were computed according to the TargetScan algorithm

From these data, predicted miRNA hairpins with strong predicted Myf5 repression and strong upregulation in C2C12s were selected for subsequent validation More complete information regarding Myf5 targeting hairpins considered in this phase of the project is in Table 4-14

To test the links between RNA hairpin and Myf5 expression, the full putative miRNA hairpin sequences, including 200nt flanking sequences, were co-expressed with a reporter containing the Myf5 3'UTR in HEK293 cells The flanking sequences were included with the RNA hairpins to avoid discarding sequences in flanking regions thought to be required for proper pre-miRNA excision (Han et al , 2006) As controls,

the 3'UTR of Myf5 was transfected both alone and in conjunction with the mir-206 sequence, which is not known to repress Myf5 transcript levels. In this experiment, the putative miRNA hairpins were expected to reduce levels of Myf5 transcript levels.

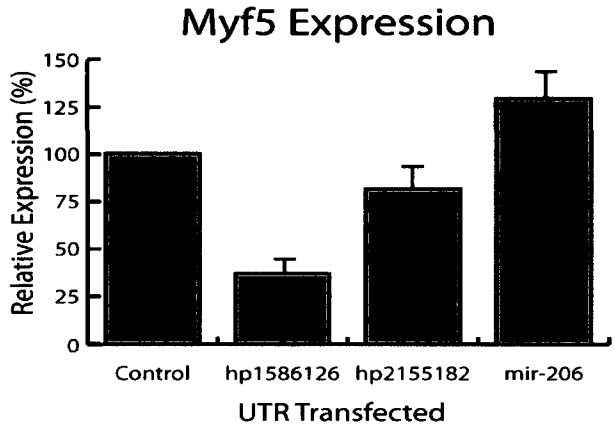

Figure 4-4: Transfection of Myf5 targeting hairpins represses Myf5.
Transfection of Myf5 targeting hairpins represses Myf5. From left to right: Myf5 UTR only; Myf5 UTR + sequence encoding hp1586126; Myf5 UTR + sequence encoding hp2155182; Myf5 UTR + sequence encoding myogenic mir-206. Relative expression normalized to control levels. QPCR data courtesy of Feodor Price (M. Rudnicki Lab, OHRI).

A marked reduction in Myf5 transcript levels was observed in cells ectopically expressing sequences encoding two putative pre-miRNA hairpins hp2155182 and hp1586126 with predicted Myf5 targeting ability. These results suggest that the Myf5 response to the introduced RNA hairpins is reflected through transcript repression. Together with the targeting predictions generated by the miRNA targeting algorithm, these results support the conjecture that the RNA hairpins are putative miRNAs with Myf5 suppression capacity.

To determine the approximate size of RNAs generating signals on the microarray and repressing Myf5, Northern blotting using the differentially expressed microarray

probe sequences was pursued in parallel to the hairpin transfection experiments. In the results (Figure 4-5), small sub 100nt RNAs were observed in both ES and myoblast cells, specifically being upregulated in the latter. In contrast, some larger RNAs with either identity to the probes or cross hybridizing potential exist as background signal in the embryonic samples but are reduced or nearly absent in samples derived from myoblasts. As this background signal is reduced more dramatically than the upregulated small RNA, if cross hybridization was an issue the net observation expected in the microarray data would have been downregulation – which was not observed. Thus, the small RNA species expressed in myoblasts, and not generalized background RNAs, are likely to be responsible for the upregulated signals detected by microarray.

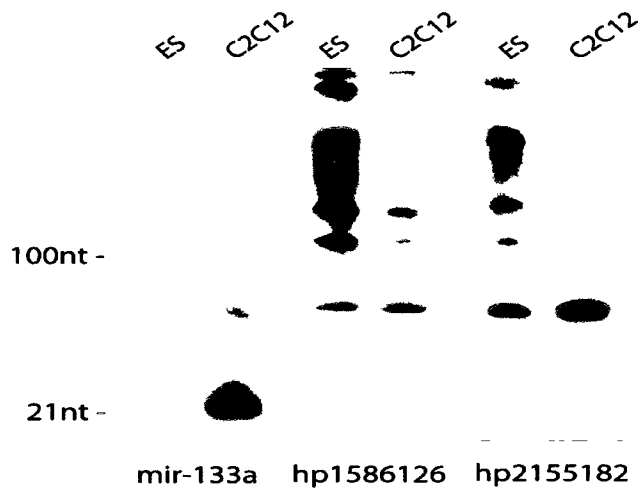

Figure 4-5: Northern blotting of Myf5-targeting hairpins
Detection of mir-133a yields a specific sub 100nt RNA species with little background signal. Two predicted miRNA hairpins (hp1586126 and hp2155182) show upregulation of the smallest RNA species in C2C12 myoblast RNA (marked with triangle) with downregulation of non-specific background signals. RNA on separate gels were aligned using a DNA marker. Data courtesy of Feodor Price (M. Rudnicki Lab, OHRI)

The relative size of the smallest RNA species detected for hp1586126 and hp2155182 compared with the mature mir-133a sequences (Figure 4-5) cannot be definitively established. However, the estimated size for each of the smallest RNA species is below 100nt. Thus, there exists a possibility that the small RNA species detected are larger than fully processed, mature, miRNA fragments (~21nt): possibly pre-miRNA hairpins are being detected in the blotting experiment or another type of small RNA species. The explanation that pre-miRNAs are being detected in the absence of corresponding mature miRNAs requires additional assumptions of a more complicated system regulating miRNA maturation. For example, it is possible that regulation of Dicer-limited production of mature miRNAs from pre-miRNA intermediates occurs in differentiating myoblasts. This type of posttranscriptional control of miRNA genesis has

been observed in an ESC context where Lin28 can inhibit pre-let-7a processing into the mature let-7a, and also in a breast cancer model where pri- and pre-miRNA transcripts from the mir-17..92 containing locus can accumulate prior to being fully processed into mature miRNAs (Castellano et al., 2009; Rybak et al., 2008). A simpler possibility to investigate is that the detected RNAs may be other, larger, non-miRNA ncRNAs with transcript repressing functionality.

The sites from which the two Myf5 repressing hairpins are derived from have some homology to B2 SINE retroelement sequence (Figure 4-6 and Section 4.8.2). B2 SINE elements are generally RNA polymerase III derived transcripts that have been shown to specifically control gene expression by associating with RNA polymerase II complexes and rendering them inactive (Allen et al., 2004b; Espinoza et al., 2007). In spite of this, sequence similarity between these hairpins and the Myf5 UTR could still possibly trigger a miRNA-like downregulation response, as the TargetScan miRNA target prediction algorithm exploits rules based on miRNA-target nucleotide complementarity. Thus, the possibility that cognate SINE sequence elements were contained within the Myf5 3'-UTR was considered. However, upon manual inspection the Myf5 3'-UTR was not seen to contain regions annotated with homology to B2 SINE repeat sequences (Data not shown). This reduces the possibility of a false B2 SINE sequence dependent miRNA-target association between these hairpins and the Myf5 UTR.

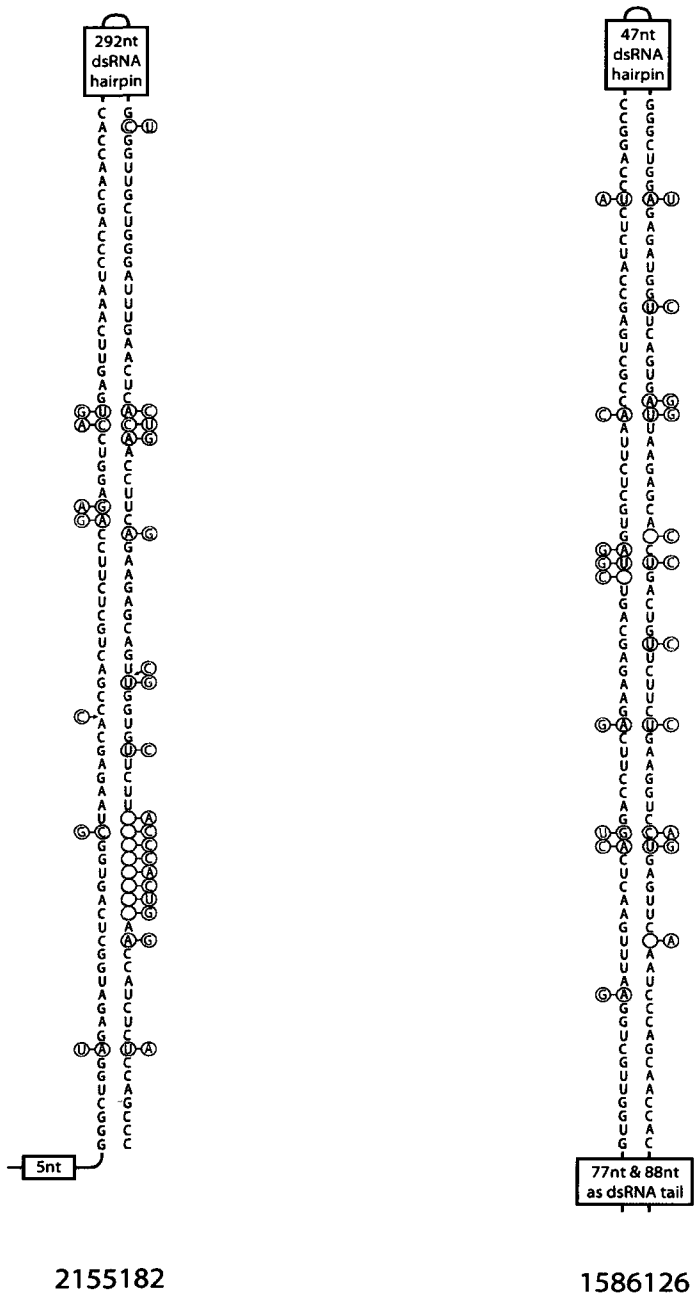

Figure 4-6: Minimum Free Energy structures of B2 sequences contained within two Myf5 repressing hairpins

Sequences derived from hp2155182 (left) and hp1586126 (right) were folded with RNAfold to identify paired bases. Mutations deviating from a reference B2 sequence (Previously described in (Allen et al., 2004b)) are highlighted, with original B2 nucleotides shown in red. Both sequences form unbranched hairpins which extend beyond the paired regions shown.As B2 SINE sequences are present at numerous sites in the genome, the EST sequences supporting each of the hairpin regions were examined for unique association to their respective genomic locations(Table 4-4).

Hairpin	Supporting EST ID	EST sequence Start	EST sequence End	Hit position	% Identity	Sequence Hits Rank	Sequence Hits Total
hp2155182	BE651759	1	631	chr1:63114131-63114761	99.60%	1	1*
	CB587035	1	819	chr1:63113943-63114760	99.60%	1	1*
	BP772987	26	619	chr1:63114168-63114762	99.90%	1	1*
	CJ296315	1	460	chr1:63114302-63114761	100.00%	1	1*
	AA153903	30	472	chr1:63114320-63114762	100.00%	1	1*
	AI615914	30	455	chr1:63114337-63114762	99.80%	1	1*
	CK330888	1	375	chr1:63114387-63114761	99.50%	1	1*
	BM196375	1	374	chr1:63114388-63114761	100.00%	1	1*
	BY662314	1	357	chr1:63114405-63114761	99.50%	1	1*
	AV146851	1	288	chr1:63114470-63114757	98.70%	1	1*
hp1586126	BG080144	9	123	chrX:90438457-90438580	93.90%	3	>200
	BU936244	35	149	chrX:90438457-90438580	93.90%	3	>200
	BG067070	35	149	chrX:90438457-90438580	93.90%	3	>200
	C81606	35	149	chrX:90438457-90438580	93.90%	3	>200

* >200 hits for EST subsequence with B2 SINE homology

Table 4-4: BLAT sequence alignment of of ESTs supporting two putative miRNA hairpins. All EST sequences mapped nearby nearby putative miRNA hairpin hp2155182 align uniquely to the region. However, all hp2155182 associated ESTs possess a sequence with B2 homology which, when isolated, maps to numerous genomic sequences. In comparison, ESTs annotated nearby hairpin hp1586126 aligned to multiple genomic regions. For example, EST BG080144 exhibits 93.90% homology to hp1586126 region, but with the 3rd highest score hit out of over 200 BLAT alignments.

The EST sequences in the region of putative miRNA hairpin hp2155182 map uniquely when considering their entire sequences. Some terminal EST sequences are not included in the alignments as they consisting of mostly polyT regions, presumably being post-transcriptionally added. All ESTs associated with hp2155182 also included a subsequence with homology to the B2 SINE RNA sequence, which when isolated mapped to over 200 genomic regions in each case. On the other hand, EST sequences supporting hp1586126 were found to map to multiple genomic regions due to sequence identity with the B2 SINE element, and although the sequence shows Myf5 repressing function, whether this is the genomic region of origin for this molecule cannot be determined. However, the fact that full-length EST sequences associated with hp2155182 map uniquely to a single genomic region implies that the hp2155182 region is

transcriptionally active, producing either B2 RNA or a miRNA hairpin with some identity to B2 RNA sequence, either of which are able to suppress transcripts containing the Myf5 UTR.

4.5 Discussion

4.5.1 Use of ESTs for ncRNA prediction

The work presented demonstrates the value of including EST evidence for the purposes of identifying ncRNAs, using the miRNA biogenesis pathway as a framework. These results also demonstrate how supporting EST evidence can be used to assist downstream validation steps by helping prioritize genomic locations by probable expression in individual tissues. One of the benefits of using ESTs in this way is that experimental designs are potentially simplified by minimizing sample size requirements by reducing resources required to assess a specific goal. In this study, we were able to focus on RNAs from two biological sources instead of pursuing a systematic screen across many tissues. By comparison, one group interested in generating ncRNA predictions in *Sacharomyces cerevisiae* required an experimental design that tested nine different environmental conditions during their validation stages (Kavanaugh and Dietrich, 2009). The scale of this project is tractable in yeast, but in a higher organism such as mouse, with numerous tissue and cell types, a systematic screen described above would likely be unmanageable. In these cases, including EST based transcriptome data can simplify the project design by either directing ncRNA validation efforts to cell types of greatest interest or by stratifying predictions if pursuing a general systematic screen is still desired.

However, novel technologies enable different approaches to addressing a ncRNA prediction experiment. In contrast to implementing the EST assisted pipeline used here, an alternative would have been to employ one of the many deep sequencing technologies commonly available to sequence a cDNA library generated from small RNAs.

Conceivably, such an approach would identify all possible miRNAs (along with other ncRNAs) within a tissue of interest, however, it is not without its own problems. The most immediate concern in this strategy is that acquiring sufficient quantities of RNA depends on the particular cell type of interest. Constructing a library will be technically challenging if the specific cells of interest are extremely rare. For instance, muscular $Pax7^{+}/Myf5^{-}$satellite stem cells comprise approximately 10% of the adult satellite cell pool (Le Grand et al., 2009), which in turn themselves comprise less than 5% of sublaminar muscle nuclei (Bischoff, 1994). In this case, the cell population of interest composes 0.5% of the gross tissue from which they would be isolated and must yield at least 10 μg of total RNA needed to construct a small RNA library (Morin et al., 2008). Despite the possibility that adequate amounts of RNA are unobtainable, methods do exist to amplify as little as 10 pg of RNA without introducing major distortions in abundance relationships (Iscove et al., 2002). Even with this approach, claims that amplification procedures reduce reproducibility of data do exist and technical biases that are insignificant on a macro scale can be amplified (Day et al., 2007). On a positive note, concerns regarding RNA availability and manipulation are partly mitigated by the extreme sensitivity of deep sequencing technology. To illustrate this point, a single mature mir-206 fragment was observed in the embryonic stem cell derived Illumina sequencing data examined here (Data not shown). The origin of the fragment is speculative; it may have arisen due inherent transcriptional leakiness or from a non-embryonic cell in the sample. Regardless of its source and considering the context of the sample examined, this mir-206 fragment represents experimental noise and relying on fortuitous examples such as this to identify novel ncRNAs is an approach that potentially

creates more problems than opportunities.

The effect of including EST data to enhance RNA fragments identified with high-thoughput sequencing was not experimentally tested using manual validation. Conceivably, this could be accomplished by identifying expressed small RNAs in high-throughput sequencing data, prioritizing those supported by tissue-specific ESTs, and pursuing them by Northern blotting. The results in the preceding chapter suggest that these EST supported ncRNA candidates could be identified at a validation rate double to ncRNA candidates without supporting ESTs.

4.5.2 On other known miRNAs expressed in myoblast samples

Several known miRNAs were detected on the microarray and exhibited differential expression between differentiating myoblasts and ES cells. mir-92-2 and mir-7-2 were transcripts downregulated in myoblast as compared with undifferentiated ES cells and presented some potential to be repressors of transcripts required for myogenic differentiation. However, the microarray signals for these miRNAs were weakly downregulated in myoblasts at 0.66x and 0.58x of the ES cell expression levels, and further a link between the mir-7-2 microarray probe signal and the mature miRNA sequence was not established. mir-19b-2, a third downregulated microRNA which was placed as an ES control and also exhibited reduced expression in differentiating C2C12 cells. Considering these results, a dependency of myogenic differentiation on miRNA downregulation is unlikely to be a strong biological factor.

Interestingly, mir-92-2 and mir-19b-2 are linked in a miRNA cluster situated slightly downstream of the Kis2 ncRNA, a locus which is transcribed as one contiguous

RNA (Landais et al., 2007). The miRNAs in this cluster (inclusive of mir-92-2 and mir-19b-2) demonstrate the ability to repress expression of myosin regulatory light chain interacting protein (Mylip), an ubiquitin-ligase enzyme involved in regulating cell motility (Landais et al., 2007; Olsson et al., 1999). The overexpression of this entire locus appears to be oncogenic, potentially being involved in the development of T-cell leukemia (Landais et al., 2007). The observed reduction of mir-92-2 and mir-19b-2 expression might thus be ascribed to a normal functions in more potent cells such as ES cells in contrast to experiencing regulated reductions in expression during myogenic differentiation.

In contrast, miRNAs upregulated in myoblasts showed robust differential expression. The upregulated miRNAs captured amongst the set of predicted miRNAs provide more information regarding function. Three miRNAs, mir24-2, mir-351, and mir-26a-2, demonstrated comparatively robust increases in expression levels. mir-24-2 was the most strongly overexpressed mature miRNA observed at levels 141 times the level observed in the embryonic stem cell sample. This miRNA is previously known to be upregulated during myogenic differentiation (Sun et al., 2008), and lies in a cluster alongside miRNAs –mir-23a and mir-27a on chromosome 8 upstream of Zswim4 (zinc finger, SWIM domain containing 4), an uncharacterized protein. The mir-24 containing cluster behaves similarly in other cellular contexts with similar effects. For instance, mir-24 is upregulated during differentiation into multiple blood cell lineages during hematopoiesis (Lal et al., 2009) and the entire miRNA cluster is thought to be involved in regulating cell fate choices during hepatic stem cell differentiation (Rogler et al., 2009). While these other observations together suggest that mir-24 is simply another general

differentiation associated miRNA, it has been observed to be specifically expressed in porcine satellite cells in a reciprocal fashion to mir-206, one of the major miRNAs associated with myogenic differentiation (McDaneld et al., 2009). Placing mir-24 in a rare myogenic population may hint at a role in regulating satellite stem cell entry into committed differentiation, possibly by controlling the timing or proportion of cells entering the process.

mir-26a-2 (94x over ES) was also known to be expressed during myogenesis, with a demonstrated function in repressing the polycomb group protein Enhancer of Zeste (Ezh2) (Wong and Tellam, 2008). Ezh2 is known to be involved during early development of the myotome, where it functions to inhibit progress through myogenic differentiation (Caretti et al., 2004). Remarkably, Ezh2 is also active in a system analogous to myogenesis – epidermal differentiation – where it has recently been shown to contribute towards control of differentiation rates throughout multistage epidermal shedding (Ezhkova et al., 2009). In general, the expression of Ezh2 decreases as epidermal layers progress from early to late differentiated counterparts. Ezh2 is also regulated by other miRNAs (mir-101) in other contexts such as bladder carcinoma, (Friedman et al., 2009). These observations might encourage further investigation of Ezh2 as a target of multiple miRNAs.

The known miRNA mir-351 is also upregulated in C2C12 RNA at levels 108-fold over ES. In the mouse, it lies on the X chromosome downstream of mir-503, mir-322, and an uncharacterized transcript with periodic regions of conservation that also appears to undergo splicing (dbEST identifier AK021262). Together with mir-351, these elements may be expressed as unit of a multipart transcript, like Kis2. Interestingly,

further downstream of miRNA-351 are four additional miRNA coding sites (mir-542, mir-450a-2, mir-450a-1, and mir-450b) which do not have any evidence of being transcribed based on known dbEST sequence records. The function of mir-351 and this region is unknown, but mir-351 has been observed in SAGE tags derived from embryonic heart (Ge et al., 2006). mir-351 is also one of several miRNAs upregulated in response to exposure to IFNβ, a signaling molecule involved in eliciting systemic anti-viral response, and can demonstrably attenuate hepatitis C virus replication when introduced on its own (Pedersen et al., 2007). Given the expression in muscular tissues and the potential for mir-351 to be a factor in a well known and general cellular antiviral response raises the possibility that a dual role for this miRNA exists, particularly as it is known that viral infections are detrimental to normal differentiation processes (Basu et al., 2008; Tsutsui, 2009; Waggoner et al., 2007).

Also investigated was whether predicted co-repression of various transcripts by all three mature miRNAs would reveal any putative targets involved in myogenic differentiation. Precedents for miRNAs acting combinatorially exist: mir-84 and let-7 behave synergistically in eliciting *C. elegans* molting arrest (Hayes et al., 2006); multiple cell cycle miRNAs in combination perturb G2/G1 ratios differently from what is observed when individual miRNAs are expressed alone (Ivanovska and Cleary, 2008); and a set of three miRNAs (mir-9, mir-125a, and mir-125b) behave in an additive fashion to repress the neurotrophin receptor trkC in neuroblastoma (Laneve et al., 2007). We reasoned that a possibility existed that the three distinct miRNAs expressed in C2C12s act in concert to exert effects on transcripts comparable to much more highly expressed miRNAs that are solitary actors such as mir-206.

Pursuing this line of thought, possible targets were predicted for cooperative repression by these three miRNAs. One of the top hits was Dystrophin (Dmd), a myotubular structural protein critically linked to the manifestation of Duchenne Muscular Dystrophy (Klamut et al., 1989), which is not expressed during early differentiation of myoblasts until fusion into myotubes has begun (Nudel et al., 1988). Closer inspection into the 3'UTR region of this promising candidate found that the Dmd locus has two known transcriptional termination sites, and that the variant targeted strongly by the three miRNAs was longer than the region covered robustly by known ESTs. When this 3'UTR was excluded, predicted targeting for Dmd was eliminated. If the possibility that this observation is due to spurious targeting is ignored and consider the interaction as probable, it might suggest a case where active miRNAs are specifically responsible in eliminating longer Dmd transcripts created from abnormal termination events. This kind of role for miRNAs has been reported in cases where alternatively spliced or terminated mRNA isoforms are specifically targeted, exerting control over their expression or abundance (Ghosh et al., 2008; Laneve et al., 2007; Legendre et al., 2006).

The only predicted target scoring higher than Dmd, Plod2 (LH2), a collagen cross-linking enzyme (Pornprasertsuk et al., 2004) that has two isoforms with a tissue specific expression distribution in the adult, with strongest protein expression seen in muscle tissues (Salo et al., 2006). Plod2 represents a promising target as the integrity of muscle tissue is strongly dependent on extracellular matrix components and that expression of particular collagen subtypes is associated with myoblasts in different modes of proliferation (Alexakis et al., 2007). What is enigmatic of Plod2 is that transcription of both isoforms is observed in muscle, yet only the longer isoform appears

to undergo translation (Salo et al., 2006). If the shorter isoform is being specifically repressed by miRNAs over the longer one, such an observation would be in contrast to the model of gene dysregulation by miRNA binding site loss via 3'UTR truncation. This model has been used to explain the dysregulation of oncogenic mRNAs (Mayr and Bartel, 2009). The only other tissue reported to express transcripts from both Plod2 isoforms is the brain, where little protein expression is observed. Given the strong expression of Plod2 in muscle and the differential activity of each isoform, combined with potential targeting by three known miRNAs associated with myoblast, argue that exploring miRNA mediated control of Plod2 expression may be promising.

4.5.3 On the identification of Myf5 targeting ncRNA hairpins

We provided evidence to support the existence of novel muscle-specific ncRNAs, including the first small RNAs targeting Myf5 in a myogenic context. It is known that Myf5 is repressed post-transcriptionally in embryonic brain by miRNAs from the mir-17-20 cluster and mir-31 (Daubas et al., 2009) but the question of whether Myf5 is similarly controlled by ncRNAs in muscle tissues is partially addressed by these results.

Myf5 is already known to encourage expression of mir-1 and mir-206 during myogenic differentiation (Sweetman et al., 2008). The identification of novel Myf5 controlling ncRNAs provides insight into systems involved upstream of this process. It is known that levels of MyoD and Myf5 are correlated in an approximately reciprocal fashion throughout the cell cycle and that entrance into committed differentiation occurs at G1 when Myf5 levels are lowest (Kitzmann et al., 1998). A possible native function for the novel Myf5 repressing ncRNAs is to contribute towards control of myoblast entrance into committed differentiation. One possibility is that the ncRNAs contribute to

cycling Myf5 levels. As our data examined mean Myf5 transcript levels in a pool of myoblasts, investigating ncRNA correlations to the cell cycle using synchronized proliferating myoblasts would hint at this possibility and add evidence to a functional link between the two ncRNAs and Myf5 transcript levels. Considering this model of Myf5 repression, it is also possible that loss of one or both of the ncRNAs would lead to an impediment in myoblast entrance into differentiation.

Speculating on the potential disease effects of perturbing the Myf5 repressing ncRNAs is difficult, as the immediate effect of eliminating them would be elevated Myf5 levels. Evidence that extreme alterations in Myf5 levels are well tolerated exists: elevated Myf5 levels in MyoD null mice result in normal phenotypes (Rudnicki et al., 1992) and adult Myf5 null mice are viable, only demonstrating impaired capacity for muscle regeneration (Gayraud-Morel et al., 2007). This effect is due to an elevated rate of differentiation commitment in these mice, which depletes the pool of satellite cell counts. A loss of Myf5-repressing ncRNA function could conceivably de-repress Myf5 levels, diminishing the window of myoblast differentiation and increasing relative numbers of satellite cell counts.

The possibility that the detected ncRNA hairpins are acting as B2 SINE retroelements, as implied by sequence homology, raises several issues. B2 SINE elements are generally RNA polymerase III (RNA Pol III) derived transcripts that have been shown to control gene expression through association with RNA polymerase II (RNA Pol II) complexes, rendering them inactive (Allen et al., 2004b; Espinoza et al., 2007). Repression of RNA Pol II occurs *in trans* as this behaviour can be reproduced *in vitro* using recombinantly expressed and purified B2 SINE sequence. However, B2 SINE

sequences can have proximal RNA Pol II promoter and expressed as RNA Pol II transcripts (Ferrigno et al., 2001).

Polyadenylated RNA Pol III transcribed B2 SINE RNAs have been observed through binding experiments to polyT and polyU oligonucleotides (Bachvarova, 1988; Borodulina and Kramerov, 2008). These results explain how sequences proximal to these elements were captured in EST libraries and therefore considered in this study. Both polyadenylated and unpolyadenylated B2 elements are present in oocytes and differentiated tissues (Bachvarova, 1988) with unpolyadenylated species detected at discrete sizes between 100nt and 180nt and polyadenylated B2 elements being detected in the 200-600nt size range. Furthermore, B2 elements are transcribed and show differential expression between total embryo differentiated brain tissues, with the latter containing 100nt range B2 RNAs at a higher abundance (Bachvarova, 1988). It has also been observed that a 178nt B2 RNA species acts as a general inhibitor of transcription during heat shock (Allen et al., 2004b), suggesting that unpolyadenylated B2 has biological function. It is unknown whether smaller B2 species, like those potentially detected in this study, similarly inhibit RNA Pol II transcription. As B2 SINE elements themselves can have variable numbers of TAAA repeats at their 3' end (Plass et al., 1992), it is conceivable that individual B2 loci might have differing repressive potentials against RNA Pol II. The question of whether the two introduced RNA hairpins are acting to reduce Myf5 transcript levels at the point of transcription as B2 RNAs or post-transcriptionally as miRNAs is unfortunately not addressed by the assays performed above and remains to be determined. It is also possible that the expressed B2 RNA seqences are intermediates in miRNA evolution and that SINE elements represent an

itinerant supply of RNA hairpins. However, given their expression in myoblasts and the ability to modulate Myf5 levels, these RNA hairpins represent possible contributors to transcriptional reprogramming required during differentiation.

4.6 Conclusion

We have demonstrated how a miRNA prediction strategy using Expressed Sequence Tag data can be used to produce tissue specific predictions. This method successfully identified several miRNAs known to be related to myogenic processes and the bioinformatic analyses of targets for these miRNAs yielded potential targets for future experimentation. Furthermore, this approach identified novel Myf5 repressing ncRNAs, either miRNAs or B2 SINE elements, which are potentially agents involved in the control of differentiation in proliferating myoblasts. These ncRNAs represent novel factors that may influence the endogenous capacity of muscle regeneration.

4.7 Materials & Methods

4.7.1 Data extraction & preparation

Source data and annotations were prepared as described in Section 3.6.

4.7.2 MicroRNA predictions

13267 miRNA predictions were generated using the approach described in Chapter 3 using mouse genome sequences (mm8). Source tissue annotations were added to each miRNA prediction based on the source of supporting ESTs associated with them, as derived from dbEST data retrieved on September 9, 2007. A summary of the number of mouse genome (mm9) miRNA predictions in tissues can be seen in Table 4-1. For current purposes, miRNAs predicted to be expressed in muscle (32), myotube (88), embryonic (822), embryonic stem cell (674) were selected. Numbers may differ from those in Table 4-1 as some predictions had supporting ESTs derived from multiple sources.

4.7.3 Design and analysis of miRNA microarray

To detect expression of putative miRNAs, we designed a custom oligonucleotide tiling microarray with 60mer probes (Agilent). Individual 60mer probes were designed to tile over selected known and predicted miRNA hairpins with an overlap of 20 nucleotides. Control probe sequences including positive/negative controls for known miRNAs, other ncRNAs, randomized negative control sequences, and sequences designed against known protein coding genes were designed. A list of control probes is in Table 4-5. After this stage of microarray design, additional space on the microarray was identified for additional putative miRNAs. Accordingly, additional muscle and myotube expressed candidates were added by ranking and prioritizing extra predictions

according to hairpin free energies.

Purpose	Description	Name
miRNA positive controls	muscle miRNA	mmu-mir-1-1 mmu-mir-1-2 mmu-mir-133a-1 mmu-mir-133a-2 mmu-mir-133b mmu-mir-206
	ES miRNA	mmu-mir-134
	Other miRNA	mmu-let-7a-1 mmu-let-7b mmu-let-7c mmu-let-7d mmu-let-7e mmu-let-7f-1 mmu-let-7g mmu-let-7i
miRNA tiling controls	ES miRNA	mmu-mir-290, 291a, 292, 291b mmu-mir-293, 294, 295
	ES miRNA **(bold)**	mmu-mir-**15a**,**16**-1 mmu-mir-15b,**16**-2 mmu-mir-17,18,19a,20a,**19b**1,92 mmu-mir-25,**93**,106b mmu-mir-**96** mmu-mir-**130**a mmu-mir-**130b**, 301b
ncRNA positive controls	ncRNA	snRNA U6 snRNA U1 snRNA U2 snRNA U4 snRNA 7SK
Size fractionation controls	rRNA	rRNA 18S rRNA 28S
Negative control	Unmappable to mm9	Randomized Sequences (x256)

Table 4-5:Table of RNA control sequences included on the miRNA microarray
Individual miRNAs with specific roles included where possible. In cases of clustered miRNAs where one member of a cluster was identified with a specific function (e.g. ES miRNA), additional members of the cluster were tiled as well.

Total and size fractionated RNA from undifferentiated J1 ESC and 3 day differentiated C2C12 cells was extracted with the miRvana miRNA isolation kit (Ambion). Integrity of RNA samples was verified using an Agilent Bioanalyzer. RNA

samples were labeled and hybridized using standard Agilent miRNA array protocols (Wang et al., 2007a) on to custom Agilent 15K arrays (University Health Network, Toronto, Canada) in duplicate.

To normalize the raw array data, raw signal values were processed with the Variance Stabilization Normalization protocol (Huber et al., 2002) previously used with Agilent microarrays (Babak et al., 2005), using the vsn 3.6.0 and biobase 2.0.1 Bioconductor packages under R 2.7.0. 99th percentile signal values of the randomized control sequences were used as a noise threshold for each hybridization. A floor of zero was set for arcsinh values produced by vsn and noise threshold values for each hybridization were subtracted to set the threshold expression levels to zero. Pearson correlations between hybridizations of the same type were highest between replicates, except for one ES J1 sample which experienced signal saturations during hybridization and was discarded (Table 4-6 and data not shown).

	C2C12 Total		C2C12 Small		J1 Total		J1 Small	
C2C12 Total	1.00	0.98	0.80	0.82	0.91	0.68	0.46	0.42
	0.98	1.00	0.80	0.79	0.90	0.69	0.46	0.41
C2C12 Small	0.80	0.80	1.00	0.96	0.59	0.46	0.72	0.63
	0.82	0.79	0.96	1.00	0.61	0.47	0.71	0.65
J1 Total	0.91	0.90	0.59	0.61	1.00	0.74	0.34	0.37
	0.68	0.69	0.46	0.47	0.74	1.00	0.28	0.26
J1 Small	0.46	0.46	0.72	0.71	0.34	0.28	1.00	0.88
	0.42	0.41	0.63	0.65	0.37	0.26	0.88	1.00

Table 4-6: Pearson correlations between replicates of miRNA tiling microarray. All samples were performed in duplicate.

The remaining hybridizations suggested that there are no array wide problems in the data. After further preliminary analysis of the data all samples derived from the same tissue were considered as replicates for further analysis, yielding data for C2C12 RNA and ES RNA samples in quadruplicate and triplicate, respectively, which were subjected

to quantile normalization (Bolstad, 2001). Distributions of signal values during these normalization steps are shown in Figure 4-7.

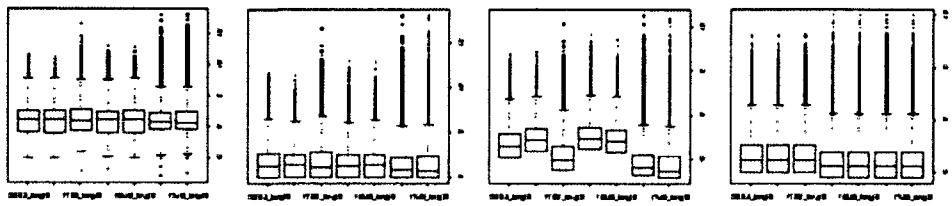

Figure 4-7: Signal value distributions during array normalization.
Left three distributions in each panel are J1 ES samples, and right four are from differentiated C2C12.

Individual probe specific expression values were generated by averaging probe expression values within each tissue, and signal calls (Present/Absent) were generated by labeling probes as present in a tissue if all corresponding signals were positive. To report a single value per probe set, individual probes with unambiguous signal calls were considered and the maximum mean signal value associated an individual probe was reported. We identified high-confidence putative miRNA hairpins by computing the fold enrichment between probes unanimously expressed in either muscle or ES cells.

4.7.4 miRNA target predictions

To predict targets of the miRNA hairpins, we used the TargetScan protocol (Grimson et al., 2007) to identify potential miRNA-3'UTR interactions. The algorithm is described in detail in the original publication but in summary miRNA-target interaction sites are scored based on 1) the type of "seed sequence" (Bases 1-8 of a miRNA) match, 2) the local AU composition, and 3) proximity of the site to the 3'UTR termini. This algorithm was replicated locally with custom scripts which produced results very highly

correlated with published miRNA-target scores ($r > 0.99$). For additional details, see Section 4.8.1 'Replication of published miRNA targeting protocol' on page 157.

To generate scores for miRNA-3'UTR pairs, 3'UTR sequences were retrieved from Ensembl BioMart for the NCBIM37 data set and targeting scores were computed between selected 3'UTRs and putative 21nt mature miRNA sequences using the custom TargetScan scripts described above. In cases where mature miRNA sequences could not be inferred from hairpins, targeting for each 3'UTR was computed for all possible mature miRNA windows (referred to as 'variants') that arose from each pre-miRNA hairpin and reporting the lowest scoring one. To generate scores for combined repression by multiple mature miRNAs, targets for individual mature miRNA sequences were computed separately against each 3'UTR and combined multiplicatively.

4.7.5 Cell cultures

The mouse ES cell line J1 was maintained on a layer of irradiated MEFs (DR4) in DMEM high glucose medium (Invitrogen) supplemented with 15% fetal bovine serum (Hyclone), 1x non essential amino acids, 1x sodium pyruvate, 1x Glutamax, β-Mercaptoethanol, 1x Pen/Strep (Invitrogen), and 1000U ESGRO (chemicon) at 37°C and 5% CO_2 . Prior to RNA isolation ES cells were trypsinized and preplated for 1hr on 0.1% gelatin plates to remove the majority of MEFs. Mouse C2C12 cells and HEK293T cells were maintained in DMEM high glucose medium supplemented with 10% fetal bovine serum (Hyclone) and 1x Pen/Strep at 37°C and 5% CO_2. For differentiation C2C12 cells were grown to confluence and subsequently cultured in DMEM high glucose with 2% horse serum (Hyclone) for 3 days prior to RNA isolation.

4.7.6 Northern Blotting

A subset of biologically interesting miRNA predicting probe sets were identified and validated by Northern blotting. Five micrograms of total RNA were resolved with 12.5% urea-polyacrylamide gels and electroblotted onto Hybond NX (GE). 60mer 5' biotinilated oligonucleotides complementary to short hairpin RNA sequences were used as probes for northern blotting. DNA size marker ladders were used with a minimal size marker of 100nt. The membrane was developed using the light shift chemiluminescent kit with a streptavidin horse radish peroxidase secondary antibody according to the manufacturer's instructions (Pierce). Membranes were exposed to Biomax MR film (Kodak) for visualization.

4.7.7 Myf5 3'-UTR knockdown assays

HEK 293T cells were cultured to 80% confluence in 6 well dishes and transfected with Linear PEI (Polysciences). For knockdown assays a vector containing the 3'UTR of Myf5 (0.5μg) was cotransfected alone or with putative targets hp1586126, hp2155182 or a negative control mir206 (3.5μg) and empty backbone to normalize the total DNA mass to 4ug. After 48 hours cells were washed with PBS and RNA was isolated and subjected to on column DNase digestion using an RNeasy mini Kit as per the manufacturer's instructions (Qiagen). cDNA synthesis was performed using Superscript II reverse transcriptase with random hexamers (Invitrogen). SYBR Green real-time PCR reactions were performed in duplicate using an MX3000p PCR machine (Stratagene, La Jolla, CA) with fold-change normalized against GAPDH. PCR primers for real-time PCR were designed using the online Primer3 software (Rozen and Skaletsky, 2000) with gene specific information obtained from Ensembl. Relative fold change in expression was

calculated using the ΔΔCT method (CT values < 30) and primer specificity was validated by denaturation curve analysis (55–94 °C). Amplification-curve plotting and calculation of cycle threshold (C_t) values were performed using the MX3000p software (v3; Stratagene), with further calculations performed using Microsoft Excel. Each experiment was performed independently on at least three occasions.

4.8 Additional Information

4.8.1 Replication of published miRNA targeting protocol

The TargetScan algorithm for scoring a miRNA-target 3'UTR interaction was implemented according to the protocol in (Grimson et al., 2007). Pseudocode for the scripts is shown in Table 4-7.

```
For each miRNA {
      Identify, classify seed sequence of miRNA
      For each 3'UTR sequence {
            For each site matching seed sequence {
                  Score seed site
                  Score AU content
                  Score binding
                  Determine overall site score
            }
      Retain site scores
      }
}
```

Table 4-7: Pseudocode for TargetScan algorithm.

In order to ensure that the results of our analysis matched those generated by the original authors, published target predictions and source data for Version 4.0 were obtained from the TargetScan website(Lewis et al., 2007), and scripts were designed to reproduce these results as closely as possible (Figure 4-8). The algorithm outputs a score, called a context score, based on several characteristics of a potential site in the 3'UTR of a putative target mRNA.

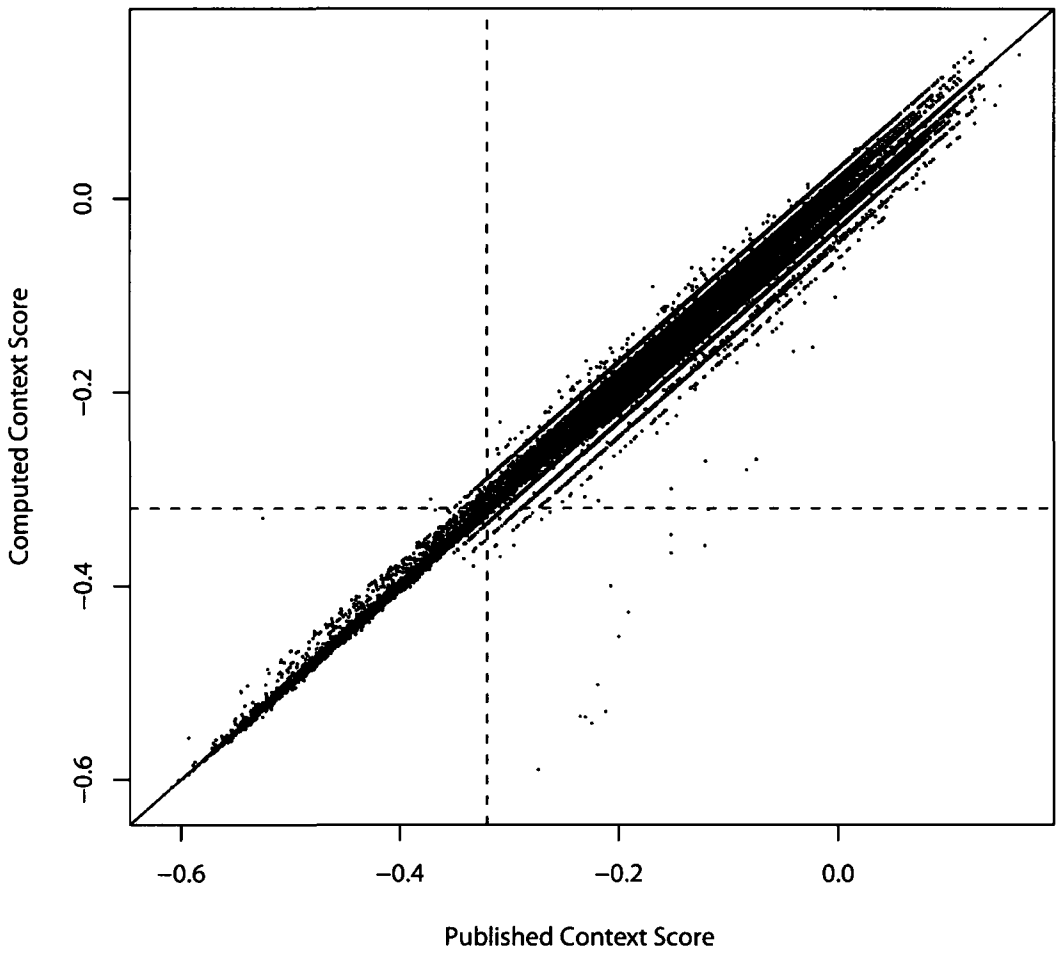

Figure 4-8: Correlation between custom and published implementations of TargetScan. Pairwise correlation between TargetScan context scores of published and re-produced versions. Correlation is higher between points at lower range of (better) context scores below the median score (-0.32, indicated by dashed lines) of demonstrated target interactions in (Grimson et al., 2007). 672131 data points shown (Pearson correlation > 0.998).

As the authors of TargetScan produced compelling results to support the utility of context scores to predicting miRNA-target interactions, the algorithm was used to generate novel predicted miRNA-target sequence interactions that were not provided on the published website. We predicted targets for murine mir-206 using mature miRNA sequences corresponding to the published mature mmu-mir-206 and several mature miRNAs that would occur if the published mature sequence was not precisely correct (Top predictions are shown in Table 4-8). Functionally, this is achieved by shifting the

mature miRNA sequence window within the sequence of the pre-miRNA hairpin. Each mature miRNA sequence arising from a pre-miRNA hairpin was given a variant identifier corresponding to the starting nucleotide position.

Real mature mir-206 (Primary)			
Variant	**Score**	**Probe**	**Symbol**
46	-0.7725	100064_f_at	Gja1
44	-0.5624	100128_at	Cdc2a
48	-0.5418	99039_g_at	Adss
45	-0.3314	94232_at	Ccnd1
45	-0.3124	160273_at	Zfp36l2
47	-0.3006	162964_at	Rgs4
46	-0.2827	103821_at	Cdc6
48	-0.2393	93236_s_at	Tyms
45	-0.2289	160501_at	Kif20a
48	-0.2264	99541_at	Kif11
48	-0.2151	99149_at	Trim59
48	-0.1872	100026_at	Bcat1
46	-0.1616	160162_at	Tagln2
48	-0.1586	100116_at	2810417H13Rik
46	-0.1322	101450_at	Csf1
45	-0.1185	106012_at	Tpx2
45	-0.1086	106256_at	Nuf2
44	-0.1043	101521_at	Birc5
44	-0.0971	96784_at	Anln
45	-0.0896	99564_at	Uhrf1

Table 4-8: Top 20 targets of mature mir-206
The mature miRNA sequence for mir-206 was used to target 3'UTRs derived from the Ensembl NCBIM37 database. In cases where interactions with target transcripts are predicted with multiple times, best scoring hits are shown.

In the original TargetScan publication (Grimson et al., 2007), predicted miRNA-target interactions with context scores ranging from -0.29 to -0.35 (median = -0.32) were demonstrated to yield repression of the target transcript. The top six targets presented in Table 4-8 therefore represent plausible targets according to the criteria used in the original publication. To examine the top hit, Gja1 (Gap junction protein, alpha 1; connexin43; Cx43), we analyzed myoblast differentiation time course microarray data

obtained from StemBase to determine if Gja1 is indeed downregulated as mir-206 is upregulated during myoblast differentiation (Kim et al., 2006; Rao et al., 2006; Yuasa et al., 2008).

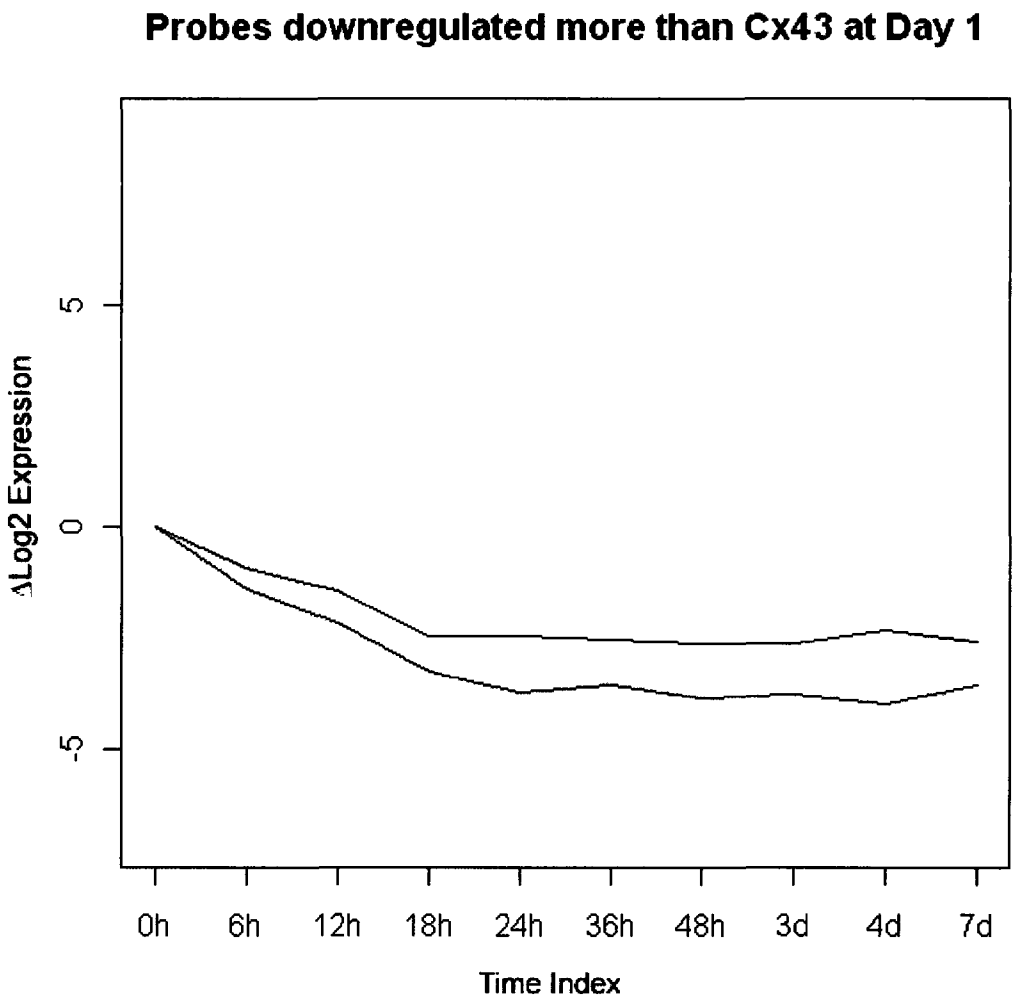

Figure 4-9: Analysis of probe sets downregulated more than Gja1 (Cx43)
Myoblast time course was analyzed to determine concordance of TargetScan results with transcript expression levels. Dark line: Cx43 probe sets; grey lines: probe sets with downregulation greater than Cx43 at 1d. Data from StemBase E233: Time Course in vitro Differentiation of Myogenic Primary Myoblast into Myotubes.

After normalization of raw Affymetrix CEL files according to published protocols

(Krzyzanowski and Andrade-Navarro, 2007) and identifying probe sets detecting Gja1, we visualized both the Gja1 transcriptional decreases and additional decreases of transcripts reduced more robustly. Figure 4-9 shows that both Gja1 probe sets experience reductions until approximately 18-24h post-induction of differentiation.

During this analysis, we found previous literature stating that Gja1 (Connexin43; Cx43) was regulated post-transcriptionally by mir-206 (Anderson et al., 2006). This report showed that Cx43 protein is effectively eliminated at 24 hours into differentiation, but curiously Cx43 mRNA persists until 5 days of differentiation, experiencing a minimal level at 24 hours and then returning to approximately original levels. Their observations are contrary to what we observe in our microarray data which contains a higher resolution of time points and also shows a more simply explained single reduction event which is sustained until 7 days of differentiation.

Together, the miRNA-target predictions generated with our custom script, predicted mir-206—Gja1 repression score and the corroborating data from the microarray time course and the reported Cx43 protein levels in (Anderson et al., 2006) reinforce the use of this implementation of a miRNA target prediction algorithm for our general purposes.

4.8.2 Analysis of B2 RNAs

To determine whether the two predicted miRNA hairpins hp1586126 and hp2155182 could possibly function as B2 RNAs, their sequences were compared to known published B2 RNA sequences to identify any divergence (Table 4-9 and Table 4-10). Deviations between the predicted RNA hairpins and the previously studied B2 sequence were mapped (Figure 4-6) to identify any mutant sites exhibiting correlation. Several regions of sequence covariation exist, implying that a biological requirement for RNA secondary structure has been evolutionarily conserved. The sequence for a mouse B2 RNA gene used was AC020972(GenBank accession; region 121488–121665 of *Mus musculus* chromosome 18, clone RP23-6p18) (Allen et al., 2004b).

```
1586126           1 TAAGTACACTGTAGCTGTTTTTCAGACACCCCAGAAGAGGGCATCAGATC     50

B2_rf             1                                                        0

1586126          51 TCATTACAGATGGTTGTGAGTCACCATGTGGTTGCTGGAATTTGAACTCA    100
                                                  |||||||||||.||||||||||.
B2_rf             1                               gtggttgctgggatttgaactcc     23

1586126         101 GGACCTTCAGAAGAGCAGT-TAGTGCTCTTAACCGCTGAGCCATCTCTCC    149
                    .|||||||.||||||||||| ..||||||||||.||.|||||||||||||.||
B2_rf            24 tgaccttcggaagagcagtcgggtgctcttacccactgagccatctcacc     73

1586126         150 AGGCCCCAGTAGAAGAAAAAAAATTTTTTAAAGATTTATTTAATGTTATT    199
                    ||.||
B2_rf            74 agccc-------------------------------------------      78

1586126         200 TATGGGCTGGAGAGATGGTTCAGTGATTAAGAGCA-CTGACTGTTCTTCT    248
                      .|||||||.|||||||.||||||..|||||||| |.|||||.|||||.
B2_rf            79 --ngggctggtgagatggctcagtgggtaagagcacccgactgctcttcc    126

1586126         249 GAAGGTCCTGAGTTC-AATCCCAGCAACCACATGGTGGCTCACAACCATC    297
                    |||||||..|||||| ||||||||||||||||
B2_rf            127 gaaggtcaggagttcaaatcccagcaaccac                      157

1586126         298 TGTAATGAGATCTGGTGCCCTCTTCTGGTGTGTGTCTGTACTCAAATTAA    347

B2_rf           158                                                      157

1586126         348 AAATAGTATGTTTATTTAT    366

B2_rf           158                       157
```

Table 4-9: Global sequence alignment between hairpin 1586126 and B2 sequences
Full length hp1586126 sequence aligned to sequence corresponding to a B2 inverted repeat composed of an antisense and sense B2 sequence. Matrix: EDNAFULL; Gap penalty: 10; Extension penalty: 0.5.

```
2155182           1 AAGCAGGGCTGGAGAGATGGCTCAGTGGCTAAGAGCA-CCGACTGCTCTT     49
                    |||||||.|||||||||||||||.|||||||| ||||||||||||
B2_fr             1      gggctggtgagatggctcagtgggtaagagcacccgactgctctt     45

2155182          50 CCAGAGGTCCTGAGTTCAAATCCCAGCAACCACATGGTGGCTCACAACCA     99
                    ||..|||||..||||||||||||||||||||||
B2_fr            46 ccgaaggtcaggagttcaaatcccagcaaccac----------------     78

2155182         100 TCTGTAATGGGATCCGATGCCCTCTTCTGGTGTGTCTGAAGACAGCTACG    149

B2_fr            79 --------------------------------------------------     78

2155182         150 GTGTACTGATATACATAAAATAAAATTTATTTAAAAAAAAATGCCCAAAC    199

B2_fr            79 --------------------------------------------------     78

2155182         200 AGTAAAAATTATAGTAAATGTTTAGATATCAAAAATTTAACACAAAGCTG    249

B2_fr            79 --------------------------------------------------     78

2155182         250 AAACTGCTGTGATGTATTTTTTTAAAAAAGATTTATTTATTATTATATCT    299

B2_fr            79 --------------------------------------------------     78

2155182         300 AAGTACACTGTAGCTGTCTTCGGAAGCATCAGAAAAGGGTGTCAGATCCC    349

B2_fr            79 --------------------------------------------------     78

2155182         350 ATTACAGATGGTTGTGAGCCACCATGCGGTTGCTGGGATTTGAACTCACA    399
                                        .|.||||||||||||||||||...
B2_fr            79 ------------------------ngtggttgctgggatttgaactcctg    104

2155182         400 ACCTTCAGAAGAGCAGT-TGGTGTTCTT--------AACCATCTCTCCAG    440
                    ||||||.|||||||||| .||||.||||        |.|||||||.||||
B2_fr           105 accttcggaagagcagtcgggtgctcttacccactgagccatctcaccag    154

2155182         441 CCC    443
                    |||
B2_fr           155 ccc    157
```

Table 4-10: Global sequence alignment between hairpin 2155182 and B2 sequences
Full length hp2155182 sequence aligned to sequence corresponding to a B2 inverted repeat composed of a sense and antisense B2 sequence. Matrix: EDNAFULL; Gap penalty: 10; Extension penalty: 0.5.

4.9 Supplementary Tables

4.9.1 Predicted targets for three known miRNAs

Symbol	Description	Score
Nup210	Nuclear pore membrane glycoprotein 210 Precursor	-1.0609
Wdhd1	WD repeat and HMG-box DNA-binding protein 1	-0.846
OTTMUSG00000002043	hypothetical protein LOC217122	-0.8349
Dmrtb1	Doublesex- and mab-3-related transcription factor B1	-0.7809
4732454E20Rik	Protein FAM26D	-0.7775
Entpd6	ectonucleoside triphosphate diphosphohydrolase 6	-0.7693
Sart1	U4/U6.U5 tri-snRNP-associated protein 1	-0.7617
5830411G16Rik	Late secretory pathway protein AVL9 homolog	-0.7576
Trpm1	Transient receptor potential cation channel subfamily M member 1	-0.755
Bcl2l11	Bcl-2-like protein 11	-0.7433
Ccdc114	Coiled-coil domain-containing protein 114	-0.7323
Pla2g4e	Cytosolic phospholipase A2 epsilon	-0.7283
Arl5c	ADP-ribosylation factor-like protein 5C	-0.7199
Treh	Trehalase Precursor	-0.7169
Nmnat1	Nicotinamide mononucleotide adenylyltransferase 1	-0.7073
2610002M06Rik	Charged multivesicular body protein 1b-2 (Chromatin-modifying protein 1b-2)	-0.6989
N4bp2l2	NEDD4-binding protein 2-like 2	-0.6949
Dhx33	Putative ATP-dependent RNA helicase DHX33	-0.6864
EG619597	galactose-3-O-sulfotransferase 2-like	-0.6753
Il15ra	Interleukin-15 receptor subunit alpha Precursor	-0.6722
Nek4	Serine/threonine-protein kinase Nek4	-0.6674
Klf8	Krueppel-like factor 8	-0.652
Taf4b	TAF4B RNA polymerase II, TATA box binding protein (TBP)-associated factor	-0.648
9030617O03Rik	UPF0317 protein C14orf159 homolog, mitochondrial Precursor	-0.6355
Slc25a44	Solute carrier family 25 member 44	-0.6283

Table 4-11: Predicted targets of mature miRNA miR-24
Top 30 targets for synergistic repression by mir-24. Scores for the lowest (best) scoring transcript were retained in cases where the same gene was targeted repeatedly.

Symbol	Description	Score
Tlr3	Toll-like receptor 3 Precursor (CD283 antigen)	-0.9796
A530053G22Rik	Putative uncharacterized protein	-0.974
Htr1a	5-hydroxytryptamine receptor 1A (Serotonin receptor 1A)	-0.9585
Zdhhc20	Probable palmitoyltransferase ZDHHC20	-0.9424
2810405J04Rik	Protein FAM98A	-0.9085
Syradb	STE20-related kinase adapter protein beta	-0.8749
Rap2c	Ras-related protein Rap-2c Precursor	-0.8686
Dmxl1	DmX-like protein 1 (X-like 1 protein)	-0.8627
Atp11c	Probable phospholipid-transporting ATPase 11C	-0.8471
Tnrc6b	Trinucleotide repeat-containing gene 6B protein	-0.8187
Dock4	Dedicator of cytokinesis protein 4	-0.8075
Pgr15l	G protein-coupled receptor 15-like	-0.7802
0	tet oncogene family member 2	-0.7778
AK220484	cDNA sequence AK220484 [Source:RefSeq peptide;Acc:NP_001077097]	-0.7648
Clgn	Calmegin Precursor	-0.7407
Nab1	NGFI-A-binding protein 1 (EGR-1-binding protein 1)	-0.7308
Ptgs2	Prostaglandin G/H synthase 2 Precursor	-0.7292
Plod2	Procollagen-lysine,2-oxoglutarate 5-dioxygenase 2 Precursor	-0.7219
Tigd4	Tigger transposable element-derived protein 4	-0.7202
Mab21l1	Protein mab-21-like 1	-0.7179
Thnsl1	Threonine synthase-like 1	-0.7096
Lancl1	LanC-like protein 1 (p40)	-0.7073
Rps6ka6	Ribosomal protein S6 kinase alpha-6	-0.7039
Ulk1	Serine/threonine-protein kinase ULK1	-0.7018
Cdk8	Cell division protein kinase 8	-0.6953

Table 4-12: Predicted targets of mature miRNA miR-26a
Top 30 targets for synergistic repression by mir-26a. Scores for the lowest (best) scoring transcript were retained in cases where the same gene was targeted repeatedly.

Symbol	Description	Score
AA409316	Protein FAM83H	-0.8651
C230095G01Rik	Transmembrane protein 72	-0.8481
Arid3a	AT-rich interactive domain-containing protein 3A	-0.7857
Mobp	Myelin-associated oligodendrocyte basic protein	-0.7135
Nrm	Nurim (Nuclear rim protein)	-0.678
Mid1	Midline-1	-0.6639
Sec14l2	SEC14-like protein 2	-0.6484
Sstr3	Somatostatin receptor type 3	-0.6398
2610019A05Rik	STE20-related kinase adapter protein alpha	-0.6294
9830107B12Rik	RIKEN cDNA 9830107B12	-0.6231
Lfng	Beta-1,3-N-acetylglucosaminyltransferase lunatic fringe	-0.6217
Pde4d	cAMP-specific 3',5'-cyclic phosphodiesterase 4D	-0.6187
Trim71	Tripartite motif-containing protein 71 (Lin-41 homolog)	-0.6078
Tmem136	Transmembrane protein 136	-0.6029
Asah3l	Alkaline ceramidase 2	-0.5996
Zfp488	Zinc finger protein 488	-0.5767
Smek1	Serine/threonine-protein phosphatase 4 regulatory subunit 3A	-0.5652
Serpinb5	Serine Protease inhibitor 5 (Maspin)	-0.5643
Crb2	Crumbs homolog 2 Precursor	-0.5631
Hapln3	Hyaluronan and proteoglycan link protein 3 Precursor	-0.5574
Bhmt2	Betaine--homocysteine S-methyltransferase 2	-0.5571
Bmf	Bcl-2-modifying factor	-0.5523
Odf2	Outer dense fiber of sperm tails protein 2	-0.5477
A330008L17Rik	Putative uncharacterized protein	-0.5446
Grk4	G protein-coupled receptor kinase 4	-0.5431

Table 4-13: Predicted targets of mature miRNA miR-351
Top 30 targets for synergistic repression by mir-351. Scores for the lowest (best) scoring transcript were retained in cases where the same gene was targeted repeatedly.

miRNA name	Fold on Array	Best Probe	Best Probe Location	Myf5 Targeting Variant	nSites	sumScores	miRNA Hairpin Location (incl. 200nt padding)
mmu-mir-206	>200	ESMu_95	chr1 20664168-20664227(+)	49	1	-0 2548	Control Probe
3049115	50 28	ESMu_11347	chr12 81641421-81641480(-)	9	1	-0 3743	chr12 81641206-81641696(-)
566131	41 18	ESMu_13874	chr17 29362689-29362748(+)	121	2	-0 5443	chr17 29362521-29363074(+)
2659205	38 13	ESMu_2883	chr2 168003116-168003175(-)	25	3	-0 906	chr2 168002559-168003492(-)
2788691	15 18	ESMu_12309	chr14 44670775-44670834(-)	26	4	-1 0583	chr14 44670530-44671080(-)
3005642	14 64	ESMu_9313	chr10 62728895-62728954(-)	184	4	-1 0764	chr10 62728657-62729353(-)
127082	13 18	ESMu_12442	chr14 59433959-59434018(+)	23	4	-1 0583	chr14 59433713-59434263(+)
1621227	10 86	ESMu_3637	chr3 100439159-100439218(+)	68	1	-0 1509	chr3 100438904-100439391(+)
2366902	8 63	ESMu_8962	chrX 91185563-91185622(-)	148	3	-1 1084	chrX 91185385-91186201(-)
3146230	7 24	ESMu_3055	chr2 181570940-181570999(-)	138	3	-1 1827	chr2 181570484-181571216(-)
611475	7 07	ESMu_11332	chr12 81393387-81393446(+)	224	4	-1 4449	chr12 81392854-81393738(+)
1291182	6 94	ESMu_6065	chr6 117843484-117843543(+)	168	3	-1 2292	chr6 117843200-117844066(+)
1102575	6 76	ESMu_4939	chr5 31732628-31732687(+)	234	5	-1 1534	chr5 31732402-31733232(+)
1586126	4 80	ESMu_8944	chrX 90438458-90438517(+)	193	6	-2 0809	chrX 90438024-90438789(+)
1163346	4 59	ESMu_3350	chr3 66077618-66077677(+)	220	5	-1 1884	chr3 66077374-66078160(+)
892058	4 53	ESMu_10736	chr11 101036342-101036401(+)	42	5	-1 159	chr11 101036034-101036628(+)
2155182	3 63	ESMu_1595	chr1 63114894-63114953(-)	263	6	-1 4301	chr1 63114383-63115225(-)
1434202	2 05	ESMu_4249	chr4 115649900-115649959(+)	96	5	-1 3472	chr4 115649695-115650243(+)

Table 4-14 Data used to guide selection of Myf5 targeting hairpin for validation
Putative miRNAs with more than 2 fold upregulation are shown Predicted Myf5 targeting scores were generated according to the Targetscan algorithm

Chapter 5 General Discussion

5.1 Overview of findings

In the work composing this thesis, several findings of value were identified. The primary aim, identifying transcripts to be prioritized for marker development in stem cell and developmental contexts, was pursued using protein coding transcripts and nucleotide transcripts as frameworks. Results from the protein coding analysis (Chapter 2) yielded insights in to functions of transcripts behaving like markers in differentiated derivatives of stem cell data in the data set. Furthermore, consideration of the entire set of transcripts fitting the model of marker genes identified several homologous families of protein coding transcripts, perhaps hinting at fundamental conservation of processes used in the maintenance of stem cells.

Chapters 3 & 4 explored the potential of transcripts yielding noncoding RNAs as their primary products in their potential as stem cell or developmental markers. The unique and novel contribution in Chapter 3 demonstrated that EST data can be used to improve future ncRNA prediction algorithms. As devising ncRNA prediction algorithms remains an active field, this knowledge can be incorporated into work conducted in the near future. Chapter 4 used this knowledge to investigate ncRNA potentials as markers during myogenic differentiation. The results derived there identified candidates as novel miRNAs and incidentally revealed the possibility of B2 SINE element mediated regulation of transcripts during myogenic regulation – a mechanism previously unreported in this process.

5.2 On markers related to stem cell and development

5.2.1 Difficulty in defining markers of 'stemness'

The analyses of protein and nucleotide markers in Chapter 2 and Chapter 4 identified a common theme with the difficulty in obtaining information with regards to undifferentiated or embryonic stem cells. The difficultly of identifying universal markers for undifferentiated stem cells – thought to be markers of 'stemness' – is a known challenge (Cai et al., 2004a; Evsikov and Solter, 2003; Ramalho-Santos et al., 2002; Vogel, 2003). Current knowledge assigns the capability of stemness induction to several factors – none of which act as solitary factors of stemness – such as Oct3/4, Sox2, Klf4, c-Myc, Nanog, and Lin-28 whose expression can induce de-differentiation of fibroblasts into induced pluripotent stem cells (iPS cells) (Okita et al., 2007; Takahashi et al., 2007; Takahashi and Yamanaka, 2006). The discovery of a single 'stem cell factor' has thus far remained elusive.

In attempts to identify any unusual combination of active gene functions or signaling pathways in the stem cell state, Gene Ontology functions of aggregated markers were analyzed, in particular while generating data for Table 2-3. Despite having several distinct pairs of undifferentiated and differentiated cell types in the data set, biologically functional terms to describe active transcripts in undifferentiated cells could not be identified, implying that no overtly unusual biological activity takes place in stem cells. For example, no bias towards enrichment in transcription factor activity was seen, as one would expect according to a model where cells are being poised for conversion to a differentiated cell state. On the other hand, an increase in significance for genes associated with the extracellular region was observed in transcripts upregulated during

differentiation in pairs of stem-cells and derivates in our data set. Functionally, this may be associated with enhanced requirements for intracellular signaling.

The observations in this study do not exclude identifying specific markers of multipotent states on a tissue-by-tissue basis. For example, the methods in Chapter 2 re-identified the known c-Kit transcript as a marker of c-Kit^{+}Lin^{-}Sca-1^{+}(KLS) hematopoietic stem cells and mast cells, both of which express this antigen at high levels (Challen et al., 2009; Drew et al., 2002). The analysis in Chapter 2 associated also associated some transcripts with specific expression in previously unreported environments. For example, high expression of Serpina3g was also associated with hematopoietic stem cells and mast cells. Following the publication of our analysis (2007), an independent group subsequently associated expression of this protein with the trabecular bone region of bone marrow where the hematopoietic niche is found (Mizukami et al., 2008; Zhang et al., 2003). With time, one would expect that additional individual markers predicted in this thesis can be assigned to specific stem cell related roles. More recently, in 2009 a group studying HC11 mammary stem cell-like cells built upon the results of this study (Williams et al., 2009). By specifically addressing the mammary stem cell system, differential mRNA and protein expression levels were examined and associated with genes from all four protein superfamilies described in Chapter 2 (Cytochrome P450 members, nuclear receptors, Rab GTPases, and serpins). Williams *et al.* also reported a particular example involving the Rab family of GTPases, where expression two Rab members decreased with mammary stem-cell like differentiation while many other homologs were activated. Whether these types of cases represent ancestral gene duplications and divergences for stem cell associated functions remains a problem to be

addressed in the future.

Similarly to the results from the GO analysis of protein coding transcripts, conclusions from examining non-coding transcripts imply that in general, miRNAs responsible for 'stemness' are not extensive. Despite assaying for approximately 300 and 1500 putative miRNAs predicted to be associated with differentiating myoblasts and embryonic stem cells, respectively, predictions associated with the ES cell state were not observed as robustly or often as candidates defined as myogenic in nature. In fact, only 4 hairpins with 10-fold higher expression were observed in ES versus myoblast, a result unlikely due to physical problems with the RNA derived from the ES cell derived RNA, as 3 out of the 4 probes with strongest signals were known ES miRNAs used as positive controls. What makes the lack of signals for novel miRNAs in the ES samples notable is that the number of candidates predicted for ES samples greatly outnumbered those predicted for myoblasts, yet did not yield a proportionally greater number of small RNAs expressed.

However, the lack of expression observed in our set of putative ES miRNAs correlates well with another independent study which was conducted in parallel. In 2008, the results of an unbiased sequencing project were published, exploring the abundances of small RNAs, including miRNAs, in human ES cells and differentiated embryoid bodies (Morin et al., 2008). This large scale RNA sequencing effort identified numerous known and unknown putative miRNAs exhibiting differential expression between human ESCs and associated embyroid bodies, including many that were downregulated while maintaining robust expression in both (Morin et al., 2008). Amongst those detected by Morin *et al.* were human miRNAs known to be positively associated with stemness or

proliferative cell populations such as hsa-mir-302a/b (Card et al., 2008; Lee et al., 2008; Lin et al., 2008), hsa-mir-21 (Chan et al., 2005; Iorio et al., 2005; Yanaihara et al., 2006), and hsa-mir-221/222 (Cardinali et al., 2009). With regards to novel miRNAs, this study was only able to identify 104 potentially novel miRNA sequences, all of which were detected at levels of expression approximately 3 orders of magnitude lower than other known miRNAs associated with pluripotent human cells. These results are despite a sequence read depth of 5-6 million reads from each of the hES and hEB libraries. Together, the demonstrated difficulties in detecting novel ES miRNAs in this study and that of Morin *et al.* (2008), both using independent miRNA prediction methods and organisms, suggest that most, if not all, miRNAs with strong expression and activity in mammalian ES cells have been identified.

Nevertheless, the lack of identifying novel ES related miRNAs may reflect inherent properties of miRNAs. A model that employs miRNAs to help transcriptional changes occur during cellular processes (e.g. differentiation) is more parsimonious than making the assumption that they help maintain static pluripotent cell states by suppressing production of differentiation factors. Mir-145, for example, fits the former model as it was recently identified as a key molecule responsible for inhibiting hESC self-renewal capacity, specifically by reducing Oct4, Klf4 and Sox2 protein levels (Xu et al., 2009). Furthermore, the latter assumption (that miRNAs are specifically used to maintain a cell state) deteriorates when considering that embryonic stem cells deficient in components of the miRNA processing pathway such as DGCR8 and Dicer display aberrations in differentiation and proliferation but nevertheless are still able to persist (Kanellopoulou et al., 2005; Murchison et al., 2005; Wang et al., 2007b). Dicer null ES

cells are even able to continue embryonic development until establishment of a body plan during gastrulation is required (Bernstein et al., 2003). Though evidence suggests that miRNAs are dispensible for self-renewal and potency, it is conceivable that miRNAs are critical when cells need to integrate signals from different cells in their immediate environment. The difficulty in identifying novel ES-related miRNAs and the ability to dispense with miRNA processing in ES culture until the point of gastrulation might imply that miRNAs are part of a basic strategy generally required for higher order cellular organization.

5.2.2 Implications of sea urchin in stem cell research

Exploring the properties of transcripts differentiatially expressed in differentiating cells identified protein coding genes with significant sequence identity (Section 2.4.5). Several protein coding genes were identified as members of larger families of duplicated genes, and individual phylogenetic analyses of the separate families suggested that in all cases divergence occurred at the purple sea urchin, *Strongylocentrotus purpuratus*. This observation has twofold implications. It suggests that the sea urchin might be an appropriate model organism for certain aspects of stem cell research, and it also implies that mechanisms active in various mammalian stem cell types are conserved.

The recent sequencing of the sea urchin genome provided many possible uses for it as a model organism (Sodergren et al., 2006). It contains representative genes of nearly all gene families found in vertebrates, orthologs of many genes involved in animal sensory functions such as vision, hearing and balance, and of medical relevance orthologs of numerous human disease related genes. The overall number of orthologs between sea urchin, mouse, and human is higher than what is found in *Ciona intestinalis*, a chordate

more closely related to mammals and thus an organism that is used to model vertebrate systems (Satoh et al., 2006; Sodergren et al., 2006). The sea urchin genome also has representatives from 97% of human protein kinase subfamilies while containing lower numbers of kinases in absolute numbers and most kinase families having only one member in the urchin (Sodergren et al., 2006). The low numbers of paralogs decrease the chance of genetic redundancies being present. Wnt signal transduction pathway components are also highly represented with the urchin containing homologs for over 95% (Croce et al., 2006) and significant conservation in Wnt target transcription factors (Howard-Ashby et al., 2006). Thus, the sea urchin can reasonably be believed to contain significant overlap with signal transduction pathways in higher vertebrates while simultaneously reducing genetic redundancy in signaling pathways.

Asymmetric division of stem cells to produce both self-renewing populations and commited derivative daughter cells relies on mechanisms employed during embryonic development when cells fated for different lineages are created (Morrison and Kimble, 2006). The sea urchin embryo undergoes asymmetric divisions at an earlier timepoint than higher animals: the 4-cell blastomere stage yields four macromeres and four micromeres. The sea urchin micromeres containing specific expression of Vasa protein (Voronina et al., 2008), which is required for germline cell development in mice and *Drosophila* (Styhler et al., 1998; Tanaka et al., 2000). However, it is thought that Vasa expressing sea urchin micromeres form early stem-cell progenitors that are fated to produce larval and adult somatic cells, not primordial germ cells (Voronina et al., 2008). More puzzling are observations of cells in *Platynereis dumerilii* (A segmented worm) which contain Vasa and other stem-cell specific proteins – these cells are thought to have

dual potential to produce germ and somatic stem cells (Rebscher et al., 2007). These developmental differences suggest that the study of urchin embryogenesis might illuminate how evolutionary processes created mechanisms of pluripotent cell specification between stem cells dedicated to germline stem cells and somatic functions.

At a higher level, the urchin also may be a potential model for work relating to hematopoiesis. In mammals, hematopoiesis involves RUNX family members; all three of which have been involved in the development and formation of various tissues and cancers (Braun and Woollard, 2009). The urchin contains two Runt homologs which exhibit no overlapping expression (Braun and Woollard, 2009; Fernandez-Guerra et al., 2006), and one copy is throught to be associated with proliferating cells (Robertson et al., 2002). The isolated expression of sea urchin Runt copies potentially reduce the possibility of encountering genetic redundancy, especially in light of the fact that all three mammalian Runt copies interact with the same binding partner (Braun and Woollard, 2009). As all the three mammalian Runx members are involved in establishing hematopoietic cell populations from ES cells and also in mature hematopoietic populations for differentiation of platelets, T-cells and B-cells (Braun and Woollard, 2009; Growney et al., 2005; Ichikawa et al., 2004; Okumura et al., 2007), it is possible that *S. purpuratus* may develop into an appropriate model to study Runt activity in the development of hematopoietic systems and provide insights applicable to higher organisms.

Ultimately, the viability of urchin usage as a model system depends several factors including the conservation of mechanisms to controlling cellular potency and the capacity to breed stable genetic backgrounds in the organism. Given the degree of

overlap between the *S. purpuratus* genome and those of higher organisms (Sodergren et al., 2006), the data presented in this thesis suggests that some cases may be identified where sea urchin can model particular mammalian stem cell systems.

5.3 Contributions to ncRNA prediction

5.3.1 Utility of Expressed Sequence Tag data

In Chapter 3 it was shown that the addition of EST data to ncRNA prediction protocols can augment support for predictions thus generated. This is an important contribution on several counts. The first aspect is demonstrating that including even general purpose high-throughput biological data into a computational project can have dramatic influence on outcomes. This can possibly be attributed to the modification of a purely computational analysis into one that is fundamentally more experimental. The fact that the analyis was presented in a linear fashion here, beginning from computational RNA predictions and then adding EST data is a trivial distinction. The analysis could have instead pursued the objectives in reverse order. For example, myogenic ESTs could have been identified and ones with probable RNA hairpins in the vicinity could have been selected for the microarray. Therefore, this method essentially converted the initial computational analysis into one that was firmly rooted in experimental data. The literature has numerous examples of purely computational projects that perform analyses without including some basic experimental data and/or biological assumptions. Including biologically based data could be the cause of the increased quality seen when considering the genome-wide set of known miRNAs.

Secondly, the potentially increased validation rates exhibited in Table 3-2 show that experimental validation of miRNAs, and by analogy other ncRNAs, can be doubled

in the case of examining small RNA high-throughput sequence data. This point is of practical importance for the efficiency of bench work, but more fundamentally provides further support for the argument put forth above. The data from the EST supported small RNA library sequences illustrate quite concisely the benefit of using primary experimental data in a bioinformatic project. In this special case, the data is based on not one but two separate experimentally generated data sets – EST sequences and small RNA sequencing fragments – and the increase in relative precision rates reflects the additional support derived from these two tangible sources. The other two analyses conducted in parallel (Using the RNAfold and RNAz techniques) investigated computational RNA structural predictions, and though the results were conclusive, they were not as dramatic in a practical sense as when using the Illumina data. This doubly supported miRNA prediction protocol clearly highlights the benefit of working within a single coherent model (miRNA biogenesis) and saturating it with as much biological data as practicable. It also hints at a potential synergistic effect of using multiple sources of high-throughput data. With time, this last point may be examined on a broad scale as appropriate data is generated by the scientific community and continues to accumulate, becoming available for computational re-analysis by others.

Lastly, illustrating that existing EST data can be reused for ncRNA prediction obviously increases their value in the future. These results, together with the importance continually demonstrated towards ncRNAs by the scientific community, counters the perception of ESTs as a superceded technology and encourages their continued generation. This work therefore has the potential to spur groups focusing on ncRNA research to construct new EST libraries for that explicit purpose, which by extension will

be made publicly available for re-interpretation by researchers investigating other topics such as the structure of protein coding RNAs and transcriptome mapping. As well, this additional EST data may be used to investigate the synergistic effects observed in miRNA predictions when combining multiple biological sources of data. It is also possible that EST sequence libraries can be generated with alternative primers (e.g. non-polyT) that are selective for subsets of transcripts.

5.3.2 Lack of characterized miRNA motif

Both computational and experimental approaches are complicated by the lack of a characterized motif that can be used to determine whether a RNA hairpin will produce mature, functional miRNA.

There have been numerous attempts to characterize properties of miRNA hairpins in primary transcripts. Detailed examinations of the landscape of binding energy throughout RNA hairpins (Han et al., 2006), number and patterns of bound bases within the hairpins (Saetrom et al., 2006), as well as several complex machine learning approaches – some with hundreds of variables (Helvik et al., 2007) – have been pursued to try and identify some informative details which can separate hairpins what will produce pre-miRNAs (or mature miRNA) from those that will not. The closest repeatable observation is that miRNA hairpins have lower than average binding energy ((Bonnet et al., 2004; Brameier and Wiuf, 2007; Freyhult et al., 2005) and this study), which is likely due to the simple requirement for dsRNA in their structures. The need for improved motifs to identify ncRNAs is underscored by the increasing ability to perform whole-genome scans for RNA structures. A previous study identified approximately 11

million hairpins in the human genome[4] (Bentwich et al., 2005). The magnitude of these predictions is comparable with the results of our raw human genome (hg18) scan, which yielded ~33 million hairpins prior to applying structural screens. Since the implementation of whole genome scans is now attainable by research groups with computational resources of intermediate size, repetition of analysis such as these will soon become more frequent along with examination of structural RNAs.

A definitive miRNA profile could potentially resolve this difficult issue. Over 1300 RNA families have been defined according to sequence alignments and evolutionary models in the Rfam database (Gardner et al., 2009), and as miRNAs have distinct hairpin structures and show interspecies conservation, the development of a universal model which incorporates sequence and structural alignments with sequence covariation seems to be a logical and desireable goal. Indeed, the need for a universal model is notable given that numerous separate profiles for individual miRNAs have been defined in the Rfam database.

A model to describe miRNA bearing hairpins might define recognition sites for for RNA processing at either the level of DGCR8/Drosha or Dicer, or both. However, the Dicer cleavage event seems to depend only on mechanistic recognition of RNA duplexes with an overlapping 'sticky' terminus followed by cleavage at a fixed distance from that end (Macrae et al., 2006). In contrast, observations that miRNA flanking regions are required for miRNA excision suggest that a motif might exist outside of the

[4] The published methods report that 1 kb sequence windows, overlapping by 150 bases, were used to identify these sequences This implies that a method such as RNAfold was used instead of a method like RNALfold The latter was used in this thesis

pri-miRNA hairpin proper (Han et al., 2006; Zeng and Cullen, 2005). These studies have suggested that unpaired sequences in the flanking regions are required for pre-miRNA excising. Our current understanding is that most previous searches for 'miRNA motifs' have been based on pre-miRNA sequences from the Sanger mirBase (Griffiths-Jones et al., 2006), which generally include some pre-miRNA flanking sequence (i.e. flanking a Drosha/DGCR8 site) but upon visual inspection of database records it appears that no systematic method to determine the number of nucleotides to include has been employed. Future attempts to designing profiles to distinguish mature miRNA producing hairpins from those that do not should therefore consider flanking regions in a methodical way, potentially by aligning sequences based on DGCR8/Drosha cleavage points. Once a miRNA hairpin profile is designed, cleavage of pri-miRNA transcripts with the motif present, absent, or in various positions, can be tested by simply observing shifts in the sizes of pre-miRNAs yielded.

5.3.3 Novel muscle ncRNAs and their potential biological roles

The majority of novel ncRNAs identified and validated in Chapter 4 provide evidence for novel miRNA expression during myogenic differentiation. Their demonstrated specific expression in *in vitro* cultures of myoblasts imply that the novel miRNA hairpins have possible activity during this process, and therefore may act as markers in certain contexts within this environment. In addition, this study investigated the functional roles of two of the most highly expressed RNA hairpins, revealing possible non-miRNA ncRNA involvement during myogenesis.

The novel miRNAs contribute towards understanding factors involved during the differentiation of cycling myoblasts, at a time point where expression of Myf5, Pax7, and

Pax3 are reduced. The miRNAs expressed at this developmental stage might interact with these transcripts or with upstream factors controlling their expression. Out of the three genes, Myf5 and Pax3 are known to be targeted by miRNAs. Myf5 is downregulated by several miRNAs, but in a neuronal context (Daubas et al., 2009). During the progress of this work, Pax3 was shown to be targeted by miRNA-27b during embryogenesis and in muscle models (Crist et al., 2009). To our knowledge, miRNA targeting of Pax7 has not been shown, but curiously intramuscular injections of miRNA-1, -133, and -206 can upregulate Pax7 along with other myogenic markers, accelerating regeneration (Nakasa et al., 2009). This continual accumulation of evidence for miRNA based regulation of myogenic genes suggests that deeper miRNA involvement in these processes will be found. Future work examining the expression patterns of miRNAs throughout myogenic differentiation is required to infer target mRNAs. The observations of additional muscle miRNAs in this work support the need to characterize additional myogenic miRNAs beyond those known in the extensively investigated triplet of mir-1, mir-133 and mir-206.

5.3.4 Potential of B2 RNA activity during myogenic differentiation

It was found that two of the miRNA-like hairpins with predicted Myf5 repression in fact were able to repress Myf5 transcript levels, but data regarding their size and sequence homology raise the possibility that they are acting like B2 SINE (Short Interspersed Nuclear Elements) sequences to repress transcription. In addition, pairwise sequence analysis between inverted B2 repeats and both hairpins sequence showed that regions with B2 homology were capable of forming extensive dsRNA regions within the context of larger unbranched dsRNA hairpins (Figure 4-6). Compensatory mutations in

both RNA hairpins maintain base pairing, strongly suggesting that a conserved requirement for RNA hairpin structures exists in these two regions.

The mouse and human genomes are comprised of approximately the same proportion of genomic repeats at approximately 40-45% (Waterston et al., 2002). Interestingly, SINE elements in the human genome are dominated by the Alu repeat while the mouse genome contains several classes of repeats. Despite the different composition, similar types of repeats have accumulated in orthologous regions of both organisms. In comparison with local GC composition, the densities of mouse SINEs are better predictors of human Alu densities in corresponding genomic regions; meanwhile the density of Alu in human is best correlated with occurrence of the homologous mouse B1 element which is derived from a common ancestral sequence (Quentin, 1994; Waterston et al., 2002). On the other hand, murine B2 SINE elements are related to tRNA genes (Daniels and Deininger, 1985) and are generally transcribed by RNA Pol III, but can also reside within classical pre-mRNA transcripts, and thus can also be transcribed by RNA Pol II (Li et al., 1999; Sakamoto and Okada, 1985). B2 elements also exhibit constitutive expression in a wide variety of tissues and are best known by their induction as part of mammalian responses to heat shock (Li et al., 1999), viral infection (White et al., 1990; Williams et al., 2004), and chemotherapeutics (Rudin and Thompson, 2001). They are therefore used to elicit large-scale transcriptional responses to external signals. However, in contrast to simply being a passive defense mechanism, evidence suggesting that B2 SINE elements exert important effects on cell biology during normal conditions is accumulating.

The trans-acting function of B2 and Alu elements via direct interference with

RNA Pol II transcription in likely their best known function (Allen et al., 2004b; Mariner et al., 2008). This trans-acting activity can be used to explain the behaviour observed when examining the two Myf5 repressing RNA hairpins identified in differentiating myoblast. The transcriptional blocking function of B2 operates at the pre-transcriptional level: When present during preinitiation complex assembly in in vitro reconstitution experiments, B2 RNA can interact with RNA Pol II and inhibit the formation of functional initiation complexes (Espinoza et al., 2004). This transcription blocking activity can itself be inhibited by the addition of antisense oligonucleotides that block B2 (Allen et al., 2004b), further linking the effects to B2 RNA being present in the system. In addition, murine B2 RNA can repress RNA Pol II transcription in both human and mouse systems (Allen et al., 2004b; Espinoza et al., 2004). Despite the apparent presence of B2-like RNAs in the microarray experiment, transcription of the mir-206 control miRNA was observed. Myf5 is required for mir-206 expression (Sweetman et al., 2008), arguing against observing a simple generalized inhibition of RNA Polymerase II transcription by the two hairpins with B2 homology, suggesting that they may play a more complex roles than simply RNA Pol II inhibitors.

Adding to a proposed function in C2C12 cells, B2 SINE elements, like other functional RNAs, can be polyadenylated to extend their lifetimes (Borodulina and Kramerov, 2008). The B2 SINE elements contain variable poly(A) tails which are postranscriptionally added and can extend up to 200nt (Borodulina and Kramerov, 2008). These modification can affect their half lives, as polyadenylated B2 RNAs have shown to extend to over 6 hours as compared with approximately 40 minutes for unpolyadenylated counterparts. The variability of free B2 SINE lifetimes suggests that some level of

control over periods in which trans-acting SINE mediated responses function exists.

As previously mentioned, trans-acting B2 RNAs are associated with general heat shock and cellular trauma. The inference from this is that B2 is therefore a systemic blocker of transcription, but many transcripts, specifically those upregulated during heat shock such as *hsp70*, are exempt from the blocking effects of B2 RNAs. While the system governing which batteries of transcripts are selectively downregulated by this process has not been elucidated, the estimates of the number of B2 SINE copies in the mouse genome range between 100 and 350 thousand (Krayev et al., 1982; Waterston et al., 2002). It is therefore possible that some B2 RNAs have evolved differential affinities for RNA Pol II or other polymerase complex subunits and thus have different propensities to control different subsets of transcripts, including during regulated transcriptional changes during growth and differentiation. Curiously enough, there is interplay between Alu repeats, heatshock, and miRNA activity in muscular tissues. Heat shock induced transcription of Alu repeats by RNA Pol III occurs prior to HSP70 protein induction (Liu et al., 1995), however it is unknown whether HSP70 induction is dependent on Alu activity. Once heat-shock transcripts are induced, the muscle tissue miRNA mir-1 represses HSP60 and HSP70 production post-transcriptionally to control their protein levels in rat ventricular cells under both native conditions and under oxidative stress (Xu et al., 2007). This appears to be evidence of a more intricate system relating miRNAs, small ncRNAs, and general control of transcription via heat-shock like responses.

In terms of gross skeletal functions, impaired regenerative capacity in aged skeletal muscles has been associated with deficiencies in heat shock protein (HSP)

production (McArdle et al., 2006). Most critically, it has been shown that the application of heat stress can increase the number of Pax7 positive satellite cells in uninjured rat skeletal muscle, enhancing general protein synthesis and regeneration following cardiotoxin induced muscle injury (Kojima et al., 2007). The molecular mechanism linking elevated HSPs to increases in $Pax7^{+}$ cells during this process is unclear. However, it is known that depletion of cytoplasmic B2 RNAs can prolong the S phase of rat hepatocytes and in turn prolong their growth phases (Crone et al., 1999). Given that the entry of proliferating myoblasts into myogenic differentiation is sensitive to the stage of the cell cycle (Kitzmann et al., 1998), it is conceivable that B2 SINE elements might be involved in the control of this process.

5.3.5 B2 SINE elements as contributors to miRNA evolution

There is a possibility that the B2 SINE elements expressed in myoblasts represent intermediates in the evolution of novel ncRNAs. All models of miRNA generation during evolution necessarily require a source of RNA hairpins, regardless of whether they arise from pre-existing RNA hairpins, such as B2 RNAs, or by chance from completely random sequences. Once the expression of regions with dsRNA forming potential is established, evolutionary forces to shape miRNA-target interactions can refine their function. Repetitive sequences such as B2 SINEs can provide this source of regions with dsRNA potential, which in turn form novel pri-miRNA sequences if expressed. While this repeat associated process has been implicated in the formation of some mammalian miRNAs (Smalheiser and Torvik, 2005), a model of miRNAs arising through the adaptation of random RNA hairpins has also been proposed (Svoboda and Di Cara, 2006). The argument for this generalized mechanism which includes RNA hairpin

adaptation is supported by the ability of the Dicer and Drosha ribonucleases to cleave dsRNAs with various sequences. The inability to find a strict consensus sequence for either of these enzymes (or their parent complexes) may be indicative of such an evolutionary mechanism operating. Furthermore, relative importance of the small 6-8nt 'seed sequence' within mature miRNA on target recognition (Brennecke et al., 2005b) indicates that a very small number of mutations are needed to further refine miRNA targeting once a random RNA hairpin is expressed in a desirable or evolutionarily fortuitious context.

A third alternative method of miRNA genesis assumes a model where the evolutionary source of miRNA hairpin sequences are the target genes themselves. In *Arabidopsis*, the creation of miRNAs targeting specific genes has been linked to the production of inverted duplications of target gene sequences (Allen et al., 2004a). However, generating the possibility of parallel repression of numerous target transcripts by a single miRNA is difficult to explain with this model, unless multiple target genes mRNAs have sequence identity to facilitate establishment of this effect.

Another aspect of evolutionary interplay between transposable repeat sequences, miRNAs, and miRNA targets can be shown using a case of a cluster of miRNAs found in the human genome. On chromosome 19 lies a large cluster of miRNAs exhibiting RNA Pol III dependent transcription using Alu as an upstream promoter element (Borchert et al., 2006). In this case, the entire cluster of miRNAs, spanning approximately 100kb, is interspersed by Alu repeats on the anti-sense strand. Both miRNAs and Alu elements can be transcribed by RNA Pol III. Furthermore, and perhaps most intriguingly, is the observation of RNA Pol III termination sequences (TTTT) in the 5' arms of the miRNAs.

Since the negative strand Alu repeats terminated at these overlapping sites contain sequences of the miRNAs themselves, the authors propose that Alu RNAs might provide a source of novel miRNA target sites if retrotransposed into sequences of unrelated coding genes. Indeed, independently conducted computational analyses do show that Alu elements residing within gene transcripts can supply miRNA targeting sites for many well characterized miRNAs (Smalheiser and Torvik, 2006). This mechanism potentially explains how expansion of miRNA-target interactions might occur.

Returning to the question of miRNA biogenesis, if the possibility that B2 SINEs are indeed intermediates in miRNA evolution is considered, two questions can be addressed using a system that considers muscle associated B2 SINE element and/or miRNAs. If an evolutionary process is occurring, myogenically expressed B2 SINEs can be expected to have 1) greater than expected sequence similarities to each other as compared with B2 SINE sequences in general, or 2) greater than expected similarity to other key myogenic transcripts. Data to address both problems could be collected through isolation and cloning of myogenically expressed B2 fragments followed by massively parallel sequencing of the resulting library.

The first point, that myogenic B2 elements are expected to have similarity to each other, could result due to duplication of genomic regions containing both B2 sequences and sequences that promote myogenic expression. It is known that B2 SINE elements can occasionally provide mobile Pol II promoters, permitting transcription of genomic locations that were previously not translated (Ferrigno et al., 2001). Therefore, it is possible that duplication of myogenically expressed SINE elements can occur as long as selective pressures on the integration event are positive or neutral. In cases where

myogenically expressed B2 SINEs are duplicated by this mechanism and retained, descendant copies would likely be myogenically expressed as well. The consequence of this process is that the duplicated B2 SINE sequences have high identity to each other and thus could be observed as a distinct clade on a phylogenetic tree of murine B2 SINE elements.

The second point, regarding myogenically expressed B2 SINE hairpins having greater identity to myogenic transcripts, would be the most direct evidence to show that evolution of B2 SINE elements into miRNAs occurring. The expectation is that some convergence between sequences in B2 SINE elements and myogenic transcripts requiring repression is occurring and that this effect can be observed as lower than average edit distances between the B2 RNA sequences and muscle related transcripts as compared with B2 sequences in general. These observations would be a compelling example of evolution in action.

5.4 General conclusions

Despite developing knowledge of genomic sequences and annotations describing their landscape, the functions of most biological components remain little understood. Painstakingly converting genomic data to usable knowledge by serially analyzing individual components is impractical; thus, generating high-throughput data sets and methods to strategically analyze them are required. This work fulfills this function in two major areas: protein coding gene transcriptomics and non-coding RNA prediction.

The methods developed herein can be generalized to situations other than those presents. The marker detection algorithm applied against mRNA expression data can be

applied against quantitative data describing the abundance of any molecules, such as protein fragments, ncRNAs or even metabolites. The observed association between EST data and miRNAs can be extended to the analysis of other ncRNAs that are polyadenylated and cleaved. In addition, the demonstrated approach of integrating predicted RNA structures with sequencing data to interpret RNA sequencing data can be reapplied to future RNA-seq experiments which are technologically current. This method also produced several novel miRNAs with expression in differentiating myoblasts, representing potential markers of differentiation during this process.

Finally, the analysis of outlier putative miRNAs expressed in a myogenic system revealed data to suggest the involvement of B2-RNA mediated transcriptional control during myogenic differentiation. This analysis also raised the possibility that expressed B2-RNA hairpins represent intermediates in miRNA evolution, which is a hypothesis that needs to be evaluated for future investigation.

Chapter 6 References

Abdullah, Z., Saric, T., Kashkar, H., Baschuk, N., Yazdanpanah, B., Fleischmann, B.K., Hescheler, J., Kronke, M., and Utermohlen, O. (2007). Serpin-6 expression protects embryonic stem cells from lysis by antigen-specific CTL. J Immunol *178*, 3390-3399.

Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D., White, O., *et al.* (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature *377*, 3-174.

Agrawal, D., Chen, T., Irby, R., Quackenbush, J., Chambers, A.F., Szabo, M., Cantor, A., Coppola, D., and Yeatman, T.J. (2002). Osteopontin identified as lead marker of colon cancer progression, using pooled sample expression profiling. Journal of the National Cancer Institute *94*, 513-521.

Alexakis, C., Partridge, T., and Bou-Gharios, G. (2007). Implication of the satellite cell in dystrophic muscle fibrosis: a self-perpetuating mechanism of collagen overproduction. American journal of physiology *293*, C661-669.

Allen, E., Xie, Z., Gustafson, A.M., Sung, G.H., Spatafora, J.W., and Carrington, J.C. (2004a). Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. Nature genetics *36*, 1282-1290.

Allen, T.A., Von Kaenel, S., Goodrich, J.A., and Kugel, J.F. (2004b). The SINE-encoded mouse B2 RNA represses mRNA transcription in response to heat shock. Nature structural & molecular biology *11*, 816-821.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W.,

and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research *25*, 3389-3402.

Anderson, C., Catoe, H., and Werner, R. (2006). MIR-206 regulates connexin43 expression during skeletal muscle development. Nucleic acids research *34*, 5863-5871.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics *25*, 25-29.

Assou, S., Cerecedo, D., Tondeur, S., Pantesco, V., Hovatta, O., Klein, B., Hamamah, S., and De Vos, J. (2009). A gene expression signature shared by human mature oocytes and embryonic stem cells. BMC genomics *10*, 10.

Assou, S., Le Carrour, T., Tondeur, S., Strom, S., Gabelle, A., Marty, S., Nadal, L., Pantesco, V., Reme, T., Hugnot, J.P., *et al.* (2007). A meta-analysis of human embryonic stem cells transcriptome integrated into a web-based expression atlas. Stem cells (Dayton, Ohio) *25*, 961-973.

Babak, T., Blencowe, B.J., and Hughes, T.R. (2005). A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription. BMC genomics *6*, 104.

Babiarz, J.E., Ruby, J.G., Wang, Y., Bartel, D.P., and Blelloch, R. (2008). Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. Genes & development *22*, 2773-2785.

Bachvarova, R. (1988). Small B2 RNAs in mouse oocytes, embryos, and somatic tissues. Developmental biology *130*, 513-523.

Barrett, T., Suzek, T.O., Troup, D.B., Wilhite, S.E., Ngau, W.C., Ledoux, P., Rudnev, D., Lash, A.E., Fujibuchi, W., and Edgar, R. (2005). NCBI GEO: mining millions of expression profiles--database and tools. Nucleic acids research *33*, D562-566.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281-297.

Baskerville, S., and Bartel, D.P. (2005). Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA (New York, NY *11*, 241-247.

Basu, A., Jain, P., Gangodkar, S.V., Shetty, S., and Ghosh, K. (2008). Dengue 2 virus inhibits in vitro megakaryocytic colony formation and induces apoptosis in thrombopoietin-inducible megakaryocytic differentiation from cord blood CD34+ cells. FEMS immunology and medical microbiology *53*, 46-51.

Baum, C.M., Weissman, I.L., Tsukamoto, A.S., Buckle, A.M., and Peault, B. (1992). Isolation of a candidate human hematopoietic stem-cell population. Proceedings of the National Academy of Sciences of the United States of America *89*, 2804-2808.

Beattie, B.J., and Robinson, P.N. (2006). Binary state pattern clustering: a digital paradigm for class and biomarker discovery in gene microarray studies of cancer. J Comput Biol *13*, 1114-1130.

Becker, A.J., Mc, C.E., and Till, J.E. (1963). Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. Nature *197*, 452-454.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., *et al.* (2005). Identification of hundreds of conserved and nonconserved human microRNAs. Nature genetics *37*, 766-770.

Bernhardt, R. (2006). Cytochromes P450 as versatile biocatalysts. J Biotechnol *124*, 128-145.

Bernstein, E., Kim, S.Y., Carmell, M.A., Murchison, E.P., Alcorn, H., Li, M.Z., Mills, A.A., Elledge, S.J., Anderson, K.V., and Hannon, G.J. (2003). Dicer is essential for mouse development. Nature genetics *35*, 215-217.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., *et al.* (2004). Global identification of human transcribed sequences with genome tiling arrays. Science (New York, NY *306*, 2242-2246.

Bischoff, R. (1994). The satellite cell and muscle regeneration. In Myology (New York, McGraw Hill Companies), pp. 97-118.

BMC. BioMed Central Reprints and Permissions (BioMed Central).

Boguski, M.S., Lowe, T.M., and Tolstoshev, C.M. (1993). dbEST--database for "expressed sequence tags". Nature genetics *4*, 332-333.

Bolotin, S., Robertson, A.V., Eshaghi, A., De Lima, C., Lombos, E., Chong-King, E., Burton, L., Mazzulli, T., and Drews, S.J. (2009). Development of a novel real-time reverse-transcriptase PCR method for the detection of H275Y positive influenza A H1N1 isolates. Journal of virological methods *158*, 190-194.

Bolstad, B.M. (2001). Probe Level Quantile Normalization of High Density Oligonucleotide Array Data.

Bonnet, E., Wuyts, J., Rouze, P., and Van de Peer, Y. (2004). Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics (Oxford, England) *20*, 2911-2917.

Borchert, G.M., Lanier, W., and Davidson, B.L. (2006). RNA polymerase III transcribes human microRNAs. Nature structural & molecular biology *13*, 1097-1101.

Borodulina, O.R., and Kramerov, D.A. (2008). Transcripts synthesized by RNA polymerase III can be polyadenylated in an AAUAAA-dependent manner. RNA (New York, NY *14*, 1865-1873.

Bossy-Wetzel, E., Schwarzenbacher, R., and Lipton, S.A. (2004). Molecular pathways to neurodegeneration. Nature medicine *10 Suppl*, S2-9.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., *et al.* (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. Cell *122*, 947-956.

Brameier, M., and Wiuf, C. (2007). Ab initio identification of human microRNAs

based on structure motifs. BMC bioinformatics *8*, 478.

Braun, T., and Woollard, A. (2009). RUNX factors in development: lessons from invertebrate model systems. Blood cells, molecules & diseases *43*, 43-48.

Brennecke, J., Stark, A., and Cohen, S.M. (2005a). Not miR-ly muscular: microRNAs and muscle development. Genes & development *19*, 2261-2264.

Brennecke, J., Stark, A., Russell, R.B., and Cohen, S.M. (2005b). Principles of microRNA-target recognition. PLoS biology *3*, e85.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. Journal of molecular biology *268*, 78-94.

Burton, E.C., Lamborn, K.R., Feuerstein, B.G., Prados, M., Scott, J., Forsyth, P., Passe, S., Jenkins, R.B., and Aldape, K.D. (2002). Genetic aberrations defined by comparative genomic hybridization distinguish long-term from typical survivors of glioblastoma. Cancer research *62*, 6205-6210.

Cai, J., Weiss, M.L., and Rao, M.S. (2004a). In search of "stemness". Experimental hematology *32*, 585-598.

Cai, X., Hagedorn, C.H., and Cullen, B.R. (2004b). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. RNA (New York, NY *10*, 1957-1966.

Cairnie, A.B., Lala, P.K., Leblond, C.P., Osmond, D.G., and McGill University. (1976). Stem cells of renewing cell populations : proceedings of a symposium held in

October 1975 at McGill University, Montreal, in tribute to C. P. Leblond on the occasion of his sixty-fifth birthday (New York, Academic Press).

Calin, G.A., Dumitru, C.D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., Aldler, H., Rattan, S., Keating, M., Rai, K., *et al.* (2002). Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proceedings of the National Academy of Sciences of the United States of America *99*, 15524-15529.

Calvano, S., Memeo, E., Piemontese, M.R., Melchionda, S., Bisceglia, L., Gasparini, P., and Zelante, L. (1997). Detection of dystrophin deletion carriers using FISH analysis. Clinical genetics *52*, 17-22.

Campbell, S.J., Henderson, C.J., Anthony, D.C., Davidson, D., Clark, A.J., and Wolf, C.R. (2005). The murine Cyp1a1 gene is expressed in a restricted spatial and temporal pattern during embryonic development. The Journal of biological chemistry *280*, 5828-5835.

Cappuccio, I., Verani, R., Spinsanti, P., Niccolini, C., Gradini, R., Costantino, S., Nicoletti, F., and Melchiorri, D. (2006). Context-dependent regulation of embryonic stem cell differentiation by mGlu4 metabotropic glutamate receptors. Neuropharmacology *51*, 606-611.

Card, D.A., Hebbar, P.B., Li, L., Trotter, K.W., Komatsu, Y., Mishina, Y., and Archer, T.K. (2008). Oct4/Sox2-regulated miR-302 targets cyclin D1 in human embryonic stem cells. Molecular and cellular biology *28*, 6426-6438.

Cardinali, B., Castellani, L., Fasanaro, P., Basso, A., Alema, S., Martelli, F., and Falcone, G. (2009). Microrna-221 and microrna-222 modulate differentiation and maturation of skeletal muscle cells. PloS one *4*, e7607.

Caretti, G., Di Padova, M., Micales, B., Lyons, G.E., and Sartorelli, V. (2004). The Polycomb Ezh2 methyltransferase regulates muscle gene expression and skeletal muscle differentiation. Genes & development *18*, 2627-2638.

Castellano, L., Giamas, G., Jacob, J., Coombes, R.C., Lucchesi, W., Thiruchelvam, P., Barton, G., Jiao, L.R., Wait, R., Waxman, J., *et al.* (2009). The estrogen receptor-alpha-induced microRNA signature regulates itself and its transcriptional response. Proceedings of the National Academy of Sciences of the United States of America *106*, 15732-15737.

Challen, G.A., Boles, N., Lin, K.K., and Goodell, M.A. (2009). Mouse hematopoietic stem cell identification and analysis. Cytometry A *75*, 14-24.

Chan, J.A., Krichevsky, A.M., and Kosik, K.S. (2005). MicroRNA-21 is an antiapoptotic factor in human glioblastoma cells. Cancer research *65*, 6029-6033.

Charge, S.B., and Rudnicki, M.A. (2004). Cellular and molecular regulation of muscle regeneration. Physiol Rev *84*, 209-238.

Chen, J.F., Mandel, E.M., Thomson, J.M., Wu, Q., Callis, T.E., Hammond, S.M., Conlon, F.L., and Wang, D.Z. (2006). The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. Nature genetics *38*, 228-233.

Chen, P.Y., Manninga, H., Slanchev, K., Chien, M., Russo, J.J., Ju, J., Sheridan, R., John, B., Marks, D.S., Gaidatzis, D., *et al.* (2005). The developmental miRNA profiles of zebrafish as determined by small RNA cloning. Genes & development *19*, 1288-1293.

Chidlow, G., Harnett, G., Williams, S., Levy, A., Speers, D., and Smith, D.W. (2010). Duplex real-time RT-PCR assays for the rapid detection and identification of pandemic (H1N1) 2009 and seasonal influenza viruses A/H1, A/H3 and B. Journal of clinical microbiology *48*, 862-866.

Christ, B., Jacob, H.J., and Jacob, M. (1977). Experimental analysis of the origin of the wing musculature in avian embryos. Anatomy and embryology *150*, 171-186.

Christ, B., and Ordahl, C.P. (1995). Early stages of chick somite development. Anatomy and embryology *191*, 381-396.

Christoffels, A., Koh, E.G., Chia, J.M., Brenner, S., Aparicio, S., and Venkatesh, B. (2004). Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Mol Biol Evol *21*, 1146-1151.

Coate, L.E., John, T., Tsao, M.S., and Shepherd, F.A. (2009). Molecular predictive and prognostic markers in non-small-cell lung cancer. The lancet oncology *10*, 1001-1010.

Conaco, C., Otto, S., Han, J.J., and Mandel, G. (2006). Reciprocal actions of REST and a microRNA promote neuronal identity. Proceedings of the National Academy of Sciences of the United States of America *103*, 2422-2427.

Crist, C.G., Montarras, D., Pallafacchina, G., Rocancourt, D., Cumano, A., Conway, S.J., and Buckingham, M. (2009). Muscle stem cell behavior is modified by microRNA-27 regulation of Pax3 expression. Proceedings of the National Academy of Sciences of the United States of America *106*, 13383-13387.

Croce, J.C., Wu, S.Y., Byrum, C., Xu, R., Duloquin, L., Wikramanayake, A.H., Gache, C., and McClay, D.R. (2006). A genome-wide survey of the evolutionarily conserved Wnt pathways in the sea urchin Strongylocentrotus purpuratus. Developmental biology *300*, 121-131.

Crone, T.M., Schalles, S.L., Benedict, C.M., Pan, W., Ren, L., Loy, S.E., Isom, H., and Clawson, G.A. (1999). Growth inhibition by a triple ribozyme targeted to repetitive B2 transcripts. Hepatology (Baltimore, Md *29*, 1114-1123.

Daniels, G.R., and Deininger, P.L. (1985). Repeat sequence families derived from mammalian tRNA genes. Nature *317*, 819-822.

Daubas, P., Crist, C.G., Bajard, L., Relaix, F., Pecnard, E., Rocancourt, D., and Buckingham, M. (2009). The regulatory mechanisms that underlie inappropriate transcription of the myogenic determination gene Myf5 in the central nervous system. Developmental biology *327*, 71-82.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., and Steinmetz, L.M. (2006). A high-resolution map of transcription in the yeast genome. Proceedings of the National Academy of Sciences of the United States of America *103*, 5320-5325.

Day, R.C., McNoe, L., and Macknight, R.C. (2007). Evaluation of global RNA amplification and its use for high-throughput transcript analysis of laser-microdissected endosperm. International journal of plant genomics, 61028.

DiAntonio, A., Burgess, R.W., Chin, A.C., Deitcher, D.L., Scheller, R.H., and Schwarz, T.L. (1993). Identification and characterization of Drosophila genes for synaptic vesicle proteins. J Neurosci *13*, 4924-4935.

Doherty, J.M., Geske, M.J., Stappenbeck, T.S., and Mills, J.C. (2008). Diverse adult stem cells share specific higher-order patterns of gene expression. Stem cells (Dayton, Ohio) *26*, 2124-2130.

Drew, E., Merkens, H., Chelliah, S., Doyonnas, R., and McNagny, K.M. (2002). CD34 is a specific marker of mature murine mast cells. Experimental hematology *30*, 1211.

Echenique, V., Stamova, B., Wolters, P., Lazo, G., Carollo, L., and Dubcovsky, J. (2002). Frequencies of Ty1- copia and Ty3- gypsy retroelements within the Triticeae EST databases. TAG Theoretical and applied genetics *104*, 840-844.

Edge, R.E., Falls, T.J., Brown, C.W., Lichty, B.D., Atkins, H., and Bell, J.C. (2008). A let-7 MicroRNA-sensitive vesicular stomatitis virus demonstrates tumor-specific replication. Mol Ther *16*, 1437-1443.

Eis, P.S., Tam, W., Sun, L., Chadburn, A., Li, Z., Gomez, M.F., Lund, E., and Dahlberg, J.E. (2005). Accumulation of miR-155 and BIC RNA in human B cell lymphomas. Proceedings of the National Academy of Sciences of the United States of

America *102*, 3627-3632.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America *95*, 14863-14868.

Espinoza, C.A., Allen, T.A., Hieb, A.R., Kugel, J.F., and Goodrich, J.A. (2004). B2 RNA binds directly to RNA polymerase II to repress transcript synthesis. Nature structural & molecular biology *11*, 822-829.

Espinoza, C.A., Goodrich, J.A., and Kugel, J.F. (2007). Characterization of the structure, function, and mechanism of B2 RNA, an ncRNA repressor of RNA polymerase II transcription. RNA (New York, NY *13*, 583-596.

Evsikov, A.V., and Solter, D. (2003). Comment on " 'Stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature". Science (New York, NY *302*, 393; author reply 393.

Ezhkova, E., Pasolli, H.A., Parker, J.S., Stokes, N., Su, I.H., Hannon, G., Tarakhovsky, A., and Fuchs, E. (2009). Ezh2 orchestrates gene expression for the stepwise differentiation of tissue-specific stem cells. Cell *136*, 1122-1135.

Fernandez-Guerra, A., Aze, A., Morales, J., Mulner-Lorillon, O., Cosson, B., Cormier, P., Bradham, C., Adams, N., Robertson, A.J., Marzluff, W.F., *et al.* (2006). The genomic repertoire for cell cycle control and DNA metabolism in S. purpuratus. Developmental biology *300*, 238-251.

Ferrigno, O., Virolle, T., Djabari, Z., Ortonne, J.P., White, R.J., and Aberdam, D. (2001). Transposable B2 SINE elements can provide mobile RNA polymerase II promoters. Nature genetics *28*, 77-81.

Freyhult, E., Gardner, P.P., and Moulton, V. (2005). A comparison of RNA folding measures. BMC bioinformatics *6*, 241.

Friedman, J.M., Liang, G., Liu, C.C., Wolff, E.M., Tsai, Y.C., Ye, W., Zhou, X., and Jones, P.A. (2009). The putative tumor suppressor microRNA-101 modulates the cancer epigenome by repressing the polycomb group protein EZH2. Cancer research *69*, 2623-2629.

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

Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., Lindgreen, S., Wilkinson, A.C., Finn, R.D., Griffiths-Jones, S., Eddy, S.R., *et al.* (2009). Rfam: updates to the RNA families database. Nucleic acids research *37*, D136-140.

Gayraud-Morel, B., Chretien, F., Flamant, P., Gomes, D., Zammit, P.S., and Tajbakhsh, S. (2007). A role for the myogenic determination gene Myf5 in adult regenerative myogenesis. Developmental biology *312*, 13-28.

Ge, X., Wu, Q., and Wang, S.M. (2006). SAGE detects microRNA precursors. BMC genomics *7*, 285.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., *et al.* (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol *5*, R80.

Georgantas, R.W., 3rd, Hildreth, R., Morisot, S., Alder, J., Liu, C.G., Heimfeld, S., Calin, G.A., Croce, C.M., and Civin, C.I. (2007). CD34+ hematopoietic stem-progenitor cell microRNA expression and function: a circuit diagram of differentiation control. Proceedings of the National Academy of Sciences of the United States of America *104*, 2750-2755.

Ghosh, T., Soni, K., Scaria, V., Halimani, M., Bhattacharjee, C., and Pillai, B. (2008). MicroRNA-mediated up-regulation of an alternatively polyadenylated variant of the mouse cytoplasmic {beta}-actin gene. Nucleic acids research *36*, 6318-6332.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., *et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science (New York, NY *286*, 531-537.

Grenier, G., Scime, A., Le Grand, F., Asakura, A., Perez-Iratxeta, C., Andrade-Navarro, M.A., Labosky, P.A., and Rudnicki, M.A. (2007). Resident endothelial precursors in muscle, adipose, and dermis contribute to postnatal vasculogenesis. Stem cells (Dayton, Ohio) *25*, 3101-3110.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., and Enright, A.J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. Nucleic acids

research *34*, D140-144.

Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P., and Bartel, D.P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Molecular cell *27*, 91-105.

Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B.J., Chiang, H.R., King, N., Degnan, B.M., Rokhsar, D.S., and Bartel, D.P. (2008). Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature *455*, 1193-1197.

Growney, J.D., Shigematsu, H., Li, Z., Lee, B.H., Adelsperger, J., Rowan, R., Curley, D.P., Kutok, J.L., Akashi, K., Williams, I.R., *et al.* (2005). Loss of Runx1 perturbs adult hematopoiesis and is associated with a myeloproliferative phenotype. Blood *106*, 494-504.

Gu, J., He, T., Pei, Y., Li, F., Wang, X., Zhang, J., Zhang, X., and Li, Y. (2006). Primary transcripts and expressions of mammal intergenic microRNAs detected by mapping ESTs to their flanking sequences. Mamm Genome *17*, 1033-1041.

Hammond, S.M. (2006). RNAi, microRNAs, and human disease. Cancer chemotherapy and pharmacology *58 Suppl 1*, s63-68.

Han, J., Lee, Y., Yeom, K.H., Nam, J.W., Heo, I., Rhee, J.K., Sohn, S.Y., Cho, Y., Zhang, B.T., and Kim, V.N. (2006). Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. Cell *125*, 887-901.

Hannemann, F., Bichet, A., Ewen, K.M., and Bernhardt, R. (2007). Cytochrome

P450 systems--biological variations of electron transport chains. Biochim Biophys Acta *1770*, 330-344.

Hayes, G.D., Frand, A.R., and Ruvkun, G. (2006). The mir-84 and let-7 paralogous microRNA genes of Caenorhabditis elegans direct the cessation of molting via the conserved nuclear hormone receptors NHR-23 and NHR-25. Development (Cambridge, England) *133*, 4631-4641.

Heikkinen, L., Asikainen, S., and Wong, G. (2008). Identification of phylogenetically conserved sequence motifs in microRNA 5' flanking sites from C. elegans and C. briggsae. BMC Mol Biol *9*, 105.

Heller, M.J. (2002). DNA microarray technology: devices, systems, and applications. Annu Rev Biomed Eng *4*, 129-153.

Helvik, S.A., Snove, O., Jr., and Saetrom, P. (2007). Reliable prediction of Drosha processing sites improves microRNA gene prediction. Bioinformatics (Oxford, England) *23*, 142-149.

Hillier, L.D., Lennon, G., Becker, M., Bonaldo, M.F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., *et al.* (1996). Generation and analysis of 280,000 human expressed sequence tags. Genome research *6*, 807-828.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., and Schuster, P. (1994). Fast Folding and Comparison of RNA Secondary Structures. Monatshefte f Chemie *125*, 167-188.

Hofacker, I.L., Priwitzer, B., and Stadler, P.F. (2004). Prediction of locally stable RNA secondary structures for genome-wide surveys. Bioinformatics (Oxford, England) *20*, 186-190.

Houbaviy, H.B., Dennis, L., Jaenisch, R., and Sharp, P.A. (2005). Characterization of a highly variable eutherian microRNA gene. RNA (New York, NY *11*, 1245-1257.

Houzelstein, D., Auda-Boucher, G., Cheraud, Y., Rouaud, T., Blanc, I., Tajbakhsh, S., Buckingham, M.E., Fontaine-Perus, J., and Robert, B. (1999). The homeobox gene Msx1 is expressed in a subset of somites, and in muscle progenitor cells migrating into the forelimb. Development (Cambridge, England) *126*, 2689-2701.

Howard-Ashby, M., Materna, S.C., Brown, C.T., Chen, L., Cameron, R.A., and Davidson, E.H. (2006). Gene families encoding transcription factors expressed in early development of Strongylocentrotus purpuratus. Developmental biology *300*, 90-107.

Hu, G., Kim, J., Xu, Q., Leng, Y., Orkin, S.H., and Elledge, S.J. (2009). A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. Genes & development *23*, 837-848.

Huang, K.M., Dentchev, T., and Stambolian, D. (2008). MiRNA expression in the eye. Mamm Genome *19*, 510-516.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics (Oxford, England) *18 Suppl 1*,

S96-104.

Hutvagner, G., McLachlan, J., Pasquinelli, A.E., Balint, E., Tuschl, T., and Zamore, P.D. (2001). A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. Science (New York, NY *293*, 834-838.

Ichikawa, M., Asai, T., Saito, T., Seo, S., Yamazaki, I., Yamagata, T., Mitani, K., Chiba, S., Ogawa, S., Kurokawa, M., *et al.* (2004). AML-1 is required for megakaryocytic maturation and lymphocytic differentiation, but not for maintenance of hematopoietic stem cells in adult hematopoiesis. Nature medicine *10*, 299-304.

Inacio, J., and Fonseca, A. (2004). Reinstatement of Rhodotorula colostri (Castelli) Lodder and Rhodotorula crocea Shifrine & Phaff, former synonyms of Rhodotorula aurantiaca (Saito) Lodder. FEMS Yeast Res *4*, 557-561.

Inoue, T., Kagawa, T., Fukushima, M., Shimizu, T., Yoshinaga, Y., Takada, S., Tanihara, H., and Taga, T. (2006). Activation of canonical Wnt pathway promotes proliferation of retinal stem cells derived from adult mouse ciliary margin. Stem cells (Dayton, Ohio) *24*, 95-104.

Iorio, M.V., Ferracin, M., Liu, C.G., Veronese, A., Spizzo, R., Sabbioni, S., Magri, E., Pedriali, M., Fabbri, M., Campiglio, M., *et al.* (2005). MicroRNA gene expression deregulation in human breast cancer. Cancer research *65*, 7065-7070.

Irizarry, R.A., Warren, D., Spencer, F., Kim, I.F., Biswal, S., Frank, B.C., Gabrielson, E., Garcia, J.G., Geoghegan, J., Germino, G., *et al.* (2005). Multiple-laboratory comparison of microarray platforms. Nat Methods *2*, 345-350.

Iscove, N.N., Barbara, M., Gu, M., Gibson, M., Modi, C., and Winegarden, N. (2002). Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA. Nature biotechnology *20*, 940-943.

Ishida, Y., Kubota, H., Yamamoto, A., Kitamura, A., Bachinger, H.P., and Nagata, K. (2006). Type I collagen in Hsp47-null cells is aggregated in endoplasmic reticulum and deficient in N-propeptide processing and fibrillogenesis. Mol Biol Cell *17*, 2346-2355.

Ivanovska, I., and Cleary, M.A. (2008). Combinatorial microRNAs: working together to make a difference. Cell cycle (Georgetown, Tex *7*, 3137-3142.

Jimenez, M.A., Akerblad, P., Sigvardsson, M., and Rosen, E.D. (2007). Critical role for Ebf1 and Ebf2 in the adipogenic transcriptional cascade. Molecular and cellular biology *27*, 743-757.

Joglekar, M.V., Parekh, V.S., and Hardikar, A.A. (2007). New pancreas from old: microregulators of pancreas regeneration. Trends in endocrinology and metabolism: TEM *18*, 393-400.

Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., and Shoemaker, D.D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science (New York, NY *302*, 2141-2144.

Jordan, C.T., and Lemischka, I.R. (1990). Clonal and systemic analysis of long-term hematopoiesis in the mouse. Genes & development *4*, 220-232.

Kalirai, H., and Clarke, R.B. (2006). Human breast epithelial stem cells and their regulation. J Pathol *208*, 7-16.

Kanai, M.I., Okabe, M., and Hiromi, Y. (2005). seven-up Controls switching of transcription factors that specify temporal identities of Drosophila neuroblasts. Developmental cell *8*, 203-213.

Kanellopoulou, C., Muljo, S.A., Kung, A.L., Ganesan, S., Drapkin, R., Jenuwein, T., Livingston, D.M., and Rajewsky, K. (2005). Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. Genes & development *19*, 489-501.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., *et al.* (2003). The UCSC Genome Browser Database. Nucleic acids research *31*, 51-54.

Kavanaugh, L.A., and Dietrich, F.S. (2009). Non-coding RNA prediction and verification in Saccharomyces cerevisiae. PLoS genetics *5*, e1000321.

Kee, N., Sivalingam, S., Boonstra, R., and Wojtowicz, J.M. (2002). The utility of Ki-67 and BrdU as proliferative markers of adult neurogenesis. J Neurosci Methods *115*, 97-105.

Kennedy, S.A., van Diepen, A.C., van den Hurk, C.M., Coates, L.C., Lee, T.W., Ostrovsky, L.L., Miranda, E., Perez, J., Davies, M.J., Lomas, D.A., *et al.* (2007). Expression of the serine protease inhibitor neuroserpin in cells of the human myeloid lineage. Thromb Haemost *97*, 394-399.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome research *12*, 996-1006.

Kim, H.K., Lee, Y.S., Sivaprasad, U., Malhotra, A., and Dutta, A. (2006). Muscle-specific microRNA miR-206 promotes muscle differentiation. The Journal of cell biology *174*, 677-687.

Kim, Y.K., and Kim, V.N. (2007). Processing of intronic microRNAs. The EMBO journal *26*, 775-783.

Kitzmann, M., Carnac, G., Vandromme, M., Primig, M., Lamb, N.J., and Fernandez, A. (1998). The muscle regulatory factors MyoD and myf-5 undergo distinct cell cycle-specific expression in muscle cells. The Journal of cell biology *142*, 1447-1459.

Klamut, H.J., Zubrzycka-Gaarn, E.E., Bulman, D.E., Malhotra, S.B., Bodrug, S.E., Worton, R.G., and Ray, P.N. (1989). Myogenic regulation of dystrophin gene expression. British medical bulletin *45*, 681-702.

Kohno, R.I., Ikeda, Y., Yonemitsu, Y., Hisatomi, T., Yamaguchi, M., Miyazaki, M., Takeshita, H., Ishibashi, T., and Sueishi, K. (2006). Sphere formation of ocular epithelial cells in the ciliary body is a reprogramming system for neural differentiation. Brain Res *1093*, 54-70.

Kojima, A., Goto, K., Morioka, S., Naito, T., Akema, T., Fujiya, H., Sugiura, T., Ohira, Y., Beppu, M., Aoki, H., *et al.* (2007). Heat stress facilitates the regeneration of

injured skeletal muscle in rats. J Orthop Sci *12*, 74-82.

Kondo, M., Wagers, A.J., Manz, M.G., Prohaska, S.S., Scherer, D.C., Beilhack, G.F., Shizuru, J.A., and Weissman, I.L. (2003). Biology of hematopoietic stem cells and progenitors: implications for clinical application. Annu Rev Immunol *21*, 759-806.

Krayev, A.S., Markusheva, T.V., Kramerov, D.A., Ryskov, A.P., Skryabin, K.G., Bayev, A.A., and Georgiev, G.P. (1982). Ubiquitous transposon-like repeats B1 and B2 of the mouse genome: B2 sequencing. Nucleic acids research *10*, 7461-7475.

Krek, A., Grun, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M., *et al.* (2005). Combinatorial microRNA target predictions. Nature genetics *37*, 495-500.

Krieg, S.A., Krieg, A.J., and Shapiro, D.J. (2001). A unique downstream estrogen responsive unit mediates estrogen induction of proteinase inhibitor-9, a cellular inhibitor of IL-1beta- converting enzyme (caspase 1). Mol Endocrinol *15*, 1971-1982.

Krzemien, J., Dubois, L., Makki, R., Meister, M., Vincent, A., and Crozatier, M. (2007). Control of blood cell homeostasis in Drosophila larvae by the posterior signalling centre. Nature *446*, 325-328.

Krzyzanowski, P.M., and Andrade-Navarro, M.A. (2007). Identification of novel stem cell markers using gap analysis of gene expression data. Genome Biol *8*, R193.

Kuhn, R.M., Karolchik, D., Zweig, A.S., Wang, T., Smith, K.E., Rosenbloom, K.R., Rhead, B., Raney, B.J., Pohl, A., Pheasant, M., *et al.* (2009). The UCSC Genome

Browser Database: update 2009. Nucleic acids research *37*, D755-761.

Lal, A., Pan, Y., Navarro, F., Dykxhoorn, D.M., Moreau, L., Meire, E., Bentwich, Z., Lieberman, J., and Chowdhury, D. (2009). miR-24-mediated downregulation of H2AX suppresses DNA repair in terminally differentiated blood cells. Nature structural & molecular biology *16*, 492-498.

Landais, S., Landry, S., Legault, P., and Rassart, E. (2007). Oncogenic potential of the miR-106-363 cluster and its implication in human T-cell leukemia. Cancer research *67*, 5699-5707.

Laneve, P., Di Marcotullio, L., Gioia, U., Fiori, M.E., Ferretti, E., Gulino, A., Bozzoni, I., and Caffarelli, E. (2007). The interplay between microRNAs and the neurotrophin receptor tropomyosin-related kinase C controls proliferation of human neuroblastoma cells. Proceedings of the National Academy of Sciences of the United States of America *104*, 7957-7962.

Law, R.H., Zhang, Q., McGowan, S., Buckle, A.M., Silverman, G.A., Wong, W., Rosado, C.J., Langendorf, C.G., Pike, R.N., Bird, P.I., *et al.* (2006). An overview of the serpin superfamily. Genome Biol *7*, 216.

Le Grand, F., Jones, A.E., Seale, V., Scime, A., and Rudnicki, M.A. (2009). Wnt7a activates the planar cell polarity pathway to drive the symmetric expansion of satellite stem cells. Cell stem cell *4*, 535-547.

Le Grand, F., and Rudnicki, M.A. (2007). Skeletal muscle satellite cells and adult myogenesis. Curr Opin Cell Biol *19*, 628-633.

Lee, C.T., Li, L., Takamoto, N., Martin, J.F., Demayo, F.J., Tsai, M.J., and Tsai, S.Y. (2004). The nuclear orphan receptor COUP-TFII is required for limb and skeletal muscle development. Molecular and cellular biology *24*, 10835-10843.

Lee, N.S., Kim, J.S., Cho, W.J., Lee, M.R., Steiner, R., Gompers, A., Ling, D., Zhang, J., Strom, P., Behlke, M., *et al.* (2008). miR-302b maintains "stemness" of human embryonal carcinoma cells by post-transcriptional regulation of Cyclin D2 expression. Biochemical and biophysical research communications *377*, 434-440.

Legendre, M., Ritchie, W., Lopez, F., and Gautheret, D. (2006). Differential repression of alternative transcripts: a screen for miRNA targets. PLoS computational biology *2*, e43.

Letunic, I., Doerks, T., and Bork, P. (2009). SMART 6: recent updates and new developments. Nucleic acids research *37*, D229-232.

Lewis, B.P., Burge, C., and Bartel, D.P. (2007). TargetScan: Prediction of microRNA targets (Whitehead Institute for Biomedical Research).

Li, L., and Xie, T. (2005). Stem cell niche: structure and function. Annu Rev Cell Dev Biol *21*, 605-631.

Li, S.C., Pan, C.Y., and Lin, W.C. (2006). Bioinformatic discovery of microRNA precursors from human ESTs and introns. BMC genomics *7*, 164.

Li, T., Spearow, J., Rubin, C.M., and Schmid, C.W. (1999). Physiological stresses increase mouse short interspersed element (SINE) RNA expression in vivo. Gene *239*,

367-372.

Liang, H.C., Li, H., McKinnon, R.A., Duffy, J.J., Potter, S.S., Puga, A., and Nebert, D.W. (1996). Cyp1a2(-/-) null mutant mice develop normally but show deficient drug metabolism. Proceedings of the National Academy of Sciences of the United States of America *93*, 1671-1676.

Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B., and Bartel, D.P. (2003). The microRNAs of Caenorhabditis elegans. Genes & development *17*, 991-1008.

Lin, S.L., Chang, D.C., Chang-Lin, S., Lin, C.H., Wu, D.T., Chen, D.T., and Ying, S.Y. (2008). Mir-302 reprograms human skin cancer cells into a pluripotent ES-cell-like state. RNA (New York, NY *14*, 2115-2124.

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

Liu, W.M., Chu, W.M., Choudary, P.V., and Schmid, C.W. (1995). Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. Nucleic acids research *23*, 1758-1765.

Lodish, H.F. (2000). Molecular cell biology, 4th edn (New York, W.H. Freeman).

Long, J.R., Egan, K.M., Dunning, L., Shu, X.O., Cai, Q., Cai, H., Dai, Q., Holtzman, J., Gao, Y.T., and Zheng, W. (2006). Population-based case-control study of

AhR (aryl hydrocarbon receptor) and CYP1A2 polymorphisms and breast cancer risk. Pharmacogenet Genomics *16*, 237-243.

Lund, E., Guttinger, S., Calado, A., Dahlberg, J.E., and Kutay, U. (2004). Nuclear export of microRNA precursors. Science (New York, NY *303*, 95-98.

Lutfiyya, L.L., Xu, N., D'Ordine, R.L., Morrell, J.A., Miller, P.W., and Duff, S.M. (2007). Phylogenetic and expression analysis of sucrose phosphate synthase isozymes in plants. Journal of plant physiology *164*, 923-933.

Macrae, I.J., Zhou, K., Li, F., Repic, A., Brooks, A.N., Cande, W.Z., Adams, P.D., and Doudna, J.A. (2006). Structural basis for double-stranded RNA processing by Dicer. Science (New York, NY *311*, 195-198.

Majoros, W.H., Pertea, M., and Salzberg, S.L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics (Oxford, England) *20*, 2878-2879.

Mangelsdorf, D.J., Thummel, C., Beato, M., Herrlich, P., Schutz, G., Umesono, K., Blumberg, B., Kastner, P., Mark, M., Chambon, P., *et al.* (1995). The nuclear receptor superfamily: the second decade. Cell *83*, 835-839.

Mariner, P.D., Walters, R.D., Espinoza, C.A., Drullinger, L.F., Wagner, S.D., Kugel, J.F., and Goodrich, J.A. (2008). Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. Molecular cell *29*, 499-509.

Martick, M., Horan, L.H., Noller, H.F., and Scott, W.G. (2008). A discontinuous

hammerhead ribozyme embedded in a mammalian messenger RNA. Nature *454*, 899-902.

Masui, S., Ohtsuka, S., Yagi, R., Takahashi, K., Ko, M.S., and Niwa, H. (2008). Rex1/Zfp42 is dispensable for pluripotency in mouse ES cells. BMC developmental biology *8*, 45.

Matsuoka, Y., Kubota, H., Adachi, E., Nagai, N., Marutani, T., Hosokawa, N., and Nagata, K. (2004). Insufficient folding of type IV collagen and formation of abnormal basement membrane-like structure in embryoid bodies derived from Hsp47-null embryonic stem cells. Mol Biol Cell *15*, 4467-4475.

Mayr, C., and Bartel, D.P. (2009). Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell *138*, 673-684.

McArdle, A., Broome, C.S., Kayani, A.C., Tully, M.D., Close, G.L., Vasilaki, A., and Jackson, M.J. (2006). HSF expression in skeletal muscle during myogenesis: implications for failed regeneration in old mice. Experimental gerontology *41*, 497-500.

McCarthy, J.J. (2008). MicroRNA-206: the skeletal muscle-specific myomiR. Biochim Biophys Acta *1779*, 682-691.

McDaneld, T.G., Smith, T.P., Doumit, M.E., Miles, J.R., Coutinho, L.L., Sonstegard, T.S., Matukumalli, L.K., Nonneman, D.J., and Wiedmann, R.T. (2009). MicroRNA transcriptome profiles during swine skeletal muscle development. BMC genomics *10*, 77.

McKinnell, I.W., Parise, G., and Rudnicki, M.A. (2005). Muscle stem cells and regenerative myogenesis. Current topics in developmental biology *71*, 113-130.

Meng, F., Henson, R., Wehbe-Janek, H., Ghoshal, K., Jacob, S.T., and Patel, T. (2007). MicroRNA-21 regulates expression of the PTEN tumor suppressor gene in human hepatocellular cancer. Gastroenterology *133*, 647-658.

Michael, L., Swantek, J., and Robinson, M.J. (2006). Cloning and expression of human mitogen-activated protein kinase kinase 7gamma1. Biochemical and biophysical research communications *341*, 679-683.

Missen, M.A., Haylock, D., Whitty, G., Medcalf, R.L., and Coughlin, P.B. (2006). Stage specific gene expression of serpins and their cognate proteases during myeloid differentiation. Br J Haematol *135*, 715-724.

Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M., and Yamanaka, S. (2003). The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. Cell *113*, 631-642.

Mizukami, T., Kuramitsu, M., Takizawa, K., Momose, H., Masumi, A., Naito, S., Iwama, A., Ogawa, T., Noce, T., Hamaguchi, I., *et al.* (2008). Identification of transcripts commonly expressed in both hematopoietic and germ-line stem cells. Stem cells and development *17*, 67-80.

Montcouquiol, M., Rachel, R.A., Lanford, P.J., Copeland, N.G., Jenkins, N.A., and Kelley, M.W. (2003). Identification of Vangl2 and Scrb1 as planar polarity genes in

mammals. Nature *423*, 173-177.

Montsant, A., Maheswari, U., Bowler, C., and Lopez, P.J. (2005). Diatomics: toward diatom functional genomics. Journal of nanoscience and nanotechnology *5*, 5-14.

Morin, R.D., O'Connor, M.D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M., *et al.* (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome research *18*, 610-621.

Morrison, S.J., and Kimble, J. (2006). Asymmetric and symmetric stem-cell divisions in development and cancer. Nature *441*, 1068-1074.

Morrison, S.J., and Weissman, I.L. (1994). The long-term repopulating subset of hematopoietic stem cells is deterministic and isolatable by phenotype. Immunity *1*, 661-673.

Murchison, E.P., Partridge, J.F., Tam, O.H., Cheloufi, S., and Hannon, G.J. (2005). Characterization of Dicer-deficient murine embryonic stem cells. Proceedings of the National Academy of Sciences of the United States of America *102*, 12135-12140.

Nakajima, N., Takahashi, T., Kitamura, R., Isodono, K., Asada, S., Ueyama, T., Matsubara, H., and Oh, H. (2006). MicroRNA-1 facilitates skeletal myogenic differentiation without affecting osteoblastic and adipogenic differentiation. Biochemical and biophysical research communications *350*, 1006-1012.

Nakasa, T., Ishikawa, M., Shi, M., Shibuya, H., Adachi, N., and Ochi, M. (2009).

Acceleration of muscle regeneration by local injection of muscle-specific microRNAs in rat skeletal muscle injury model. Journal of cellular and molecular medicine.

Nawrocki, E.P., Kolbe, D.L., and Eddy, S.R. (2009). Infernal 1.0: inference of RNA alignments. Bioinformatics (Oxford, England) *25*, 1335-1337.

Neely, L.A., Rieger-Christ, K.M., Neto, B.S., Eroshkin, A., Garver, J., Patel, S., Phung, N.A., McLaughlin, S., Libertino, J.A., Whitney, D., *et al.* (2008). A microRNA expression ratio defining the invasive phenotype in bladder tumors. Urologic oncology.

Nguyen, H.T., and Frasch, M. (2006). MicroRNAs in muscle differentiation: lessons from Drosophila and beyond. Current opinion in genetics & development *16*, 533-539.

Novocraft (2009). Novoalign (Kuala Lumpur, Malaysia, Novocraft Technologies Sdn Bhd).

Nudel, U., Robzyk, K., and Yaffe, D. (1988). Expression of the putative Duchenne muscular dystrophy gene in differentiated myogenic cell cultures and in the brain. Nature *331*, 635-638.

O'Toole, A.S., Miller, S., Haines, N., Zink, M.C., and Serra, M.J. (2006). Comprehensive thermodynamic analysis of 3' double-nucleotide overhangs neighboring Watson-Crick terminal base pairs. Nucleic acids research *34*, 3338-3344.

Oberhauser, A.F., Balan, V., Fernandez-Badilla, C.L., and Fernandez, J.M. (1994). RT-PCR cloning of Rab3 isoforms expressed in peritoneal mast cells. FEBS

letters *339*, 171-174.

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

Okita, K., Ichisaka, T., and Yamanaka, S. (2007). Generation of germline-competent induced pluripotent stem cells. Nature *448*, 313-317.

Okumura, A.J., Peterson, L.F., Lo, M.C., and Zhang, D.E. (2007). Expression of AML/Runx and ETO/MTG family members during hematopoietic differentiation of embryonic stem cells. Experimental hematology *35*, 978-988.

Olsson, P.A., Korhonen, L., Mercer, E.A., and Lindholm, D. (1999). MIR is a novel ERM-like protein that interacts with myosin regulatory light chain and inhibits neurite outgrowth. The Journal of biological chemistry *274*, 36288-36292.

Pabbaraju, K., Wong, S., Wong, A.A., Appleyard, G.D., Chui, L., Pang, X.L., Yanow, S.K., Fonseca, K., Lee, B.E., Fox, J.D., *et al.* (2009). Design and validation of real-time reverse transcription-PCR assays for detection of pandemic (H1N1) 2009 virus. Journal of clinical microbiology *47*, 3454-3460.

Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A.L., Mohammad, N., Babak, T., Siu, H., Hughes, T.R., Morris, Q.D., *et al.* (2004). Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Molecular cell *16*, 929-941.

Panning, M., Eickmann, M., Landt, O., Monazahian, M., Olschlager, S.,

Baumgarte, S., Reischl, U., Wenzel, J.J., Niller, H.H., Gunther, S., *et al.* (2009). Detection of influenza A(H1N1)v virus by real-time RT-PCR. Euro Surveill *14*.

Patel, S., Malde, K., Lanzen, A., Olsen, R.H., and Nerland, A.H. (2009). Identification of immune related genes in Atlantic halibut (Hippoglossus hippoglossus L.) following in vivo antigenic and in vitro mitogenic stimulation. Fish & shellfish immunology.

Pawlicki, J.M., and Steitz, J.A. (2009). Subnuclear compartmentalization of transiently expressed polyadenylated pri-microRNAs: processing at transcription sites or accumulation in SC35 foci. Cell cycle (Georgetown, Tex *8*, 345-356.

Pedersen, I.M., Cheng, G., Wieland, S., Volinia, S., Croce, C.M., Chisari, F.V., and David, M. (2007). Interferon modulation of cellular microRNAs as an antiviral mechanism. Nature *449*, 919-922.

Pepe, M.S., Longton, G., Anderson, G.L., and Schummer, M. (2003). Selecting differentially expressed genes from microarray experiments. Biometrics *59*, 133-142.

Perez-Iratxeta, C., and Andrade, M.A. (2005). Inconsistencies over time in 5% of NetAffx probe-to-gene annotations. BMC bioinformatics *6*, 183.

Perez-Iratxeta, C., Palidwor, G., Porter, C.J., Sanche, N.A., Huska, M.R., Suomela, B.P., Muro, E.M., Krzyzanowski, P.M., Hughes, E., Campbell, P.A., *et al.* (2005a). Study of stem cell function using microarray experiments. FEBS letters *579*, 1795-1801.

Perez-Iratxeta, C., Palidwor, G.A., Porter, C.J., Sanche, N.A., Huska, M.R., Suomela, B.P., Muro, E.M., Krzyzanowski, P.M., Hughes, E., Campbell, P.C., *et al.* (2005b). StemBase (Ottawa, Ontario, Ontario Genomics Innovation Centre).

Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., *et al.* (2000). Molecular portraits of human breast tumours. Nature *406*, 747-752.

Pevny, L., and Rao, M.S. (2003). The stem-cell menagerie. Trends Neurosci *26*, 351-359.

Plass, C., Hellwig, T., Traut, W., and Winking, H. (1992). Evolution of a B2 tagged sequence from a long-range repeat family in the genus Mus. Mamm Genome *3*, 197-201.

Pornprasertsuk, S., Duarte, W.R., Mochida, Y., and Yamauchi, M. (2004). Lysyl hydroxylase-2b directs collagen cross-linking pathways in MC3T3-E1 cells. J Bone Miner Res *19*, 1349-1355.

Quackenbush, J. (2006). Microarray analysis and tumor classification. N Engl J Med *354*, 2463-2472.

Quentin, Y. (1994). A master sequence related to a free left Alu monomer (FLAM) at the origin of the B1 family in rodent genomes. Nucleic acids research *22*, 2222-2227.

Ramalho-Santos, M., Yoon, S., Matsuzaki, Y., Mulligan, R.C., and Melton, D.A.

(2002). "Stemness": transcriptional profiling of embryonic and adult stem cells. Science (New York, NY *298*, 597-600.

Rao, M. (2004). Stem and precursor cells in the nervous system. J Neurotrauma *21*, 415-427.

Rao, P.K., Kumar, R.M., Farkhondeh, M., Baskerville, S., and Lodish, H.F. (2006). Myogenic factors that regulate expression of muscle-specific microRNAs. Proceedings of the National Academy of Sciences of the United States of America *103*, 8721-8726.

Rebscher, N., Zelada-Gonzalez, F., Banisch, T.U., Raible, F., and Arendt, D. (2007). Vasa unveils a common origin of germ cells and of somatic stem cells from the posterior growth zone in the polychaete Platynereis dumerilii. Developmental biology *306*, 599-611.

Ried, T., Baldini, A., Rand, T.C., and Ward, D.C. (1992). Simultaneous visualization of seven different DNA probes by in situ hybridization using combinatorial fluorescence and digital imaging microscopy. Proceedings of the National Academy of Sciences of the United States of America *89*, 1388-1392.

Riedel, D., Antonin, W., Fernandez-Chacon, R., Alvarez de Toledo, G., Jo, T., Geppert, M., Valentijn, J.A., Valentijn, K., Jamieson, J.D., Sudhof, T.C., *et al.* (2002). Rab3D is not required for exocrine exocytosis but for maintenance of normally sized secretory granules. Molecular and cellular biology *22*, 6487-6497.

Rivas, E., and Eddy, S.R. (2001). Noncoding RNA gene detection using

comparative sequence analysis. BMC bioinformatics *2*, 8.

Ro, S., Park, C., Song, R., Nguyen, D., Jin, J., Sanders, K.M., McCarrey, J.R., and Yan, W. (2007). Cloning and expression profiling of testis-expressed piRNA-like RNAs. RNA (New York, NY *13*, 1693-1702.

Robertson, A.J., Dickey, C.E., McCarthy, J.J., and Coffman, J.A. (2002). The expression of SpRunt during sea urchin embryogenesis. Mechanisms of development *117*, 327-330.

Rogler, C.E., Levoci, L., Ader, T., Massimi, A., Tchaikovskaya, T., Norel, R., and Rogler, L.E. (2009). MicroRNA-23b cluster microRNAs regulate transforming growth factor-beta/bone morphogenetic protein signaling and liver stem cell differentiation by targeting Smads. Hepatology (Baltimore, Md *50*, 575-584.

Rosenberg, M.I., Georges, S.A., Asawachaicharn, A., Analau, E., and Tapscott, S.J. (2006). MyoD inhibits Fstl1 and Utrn expression by inducing transcription of miR-206. The Journal of cell biology *175*, 77-85.

Rozen, S., and Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. Methods in molecular biology (Clifton, NJ *132*, 365-386.

Rudin, C.M., and Thompson, C.B. (2001). Transcriptional activation of short interspersed elements by DNA-damaging agents. Genes, chromosomes & cancer *30*, 64-71.

Rudnicki, M.A., Braun, T., Hinuma, S., and Jaenisch, R. (1992). Inactivation of

MyoD in mice leads to up-regulation of the myogenic HLH gene Myf-5 and results in apparently normal muscle development. Cell *71*, 383-390.

Rudnicki, M.A., Le Grand, F., McKinnell, I., and Kuang, S. (2008). The molecular regulation of muscle stem cell function. Cold Spring Harb Symp Quant Biol *73*, 323-331.

Rybak, A., Fuchs, H., Smirnova, L., Brandt, C., Pohl, E.E., Nitsch, R., and Wulczyn, F.G. (2008). A feedback loop comprising lin-28 and let-7 controls pre-let-7 maturation during neural stem-cell commitment. Nature cell biology *10*, 987-993.

Saetrom, P., Snove, O., Nedland, M., Grunfeld, T.B., Lin, Y., Bass, M.B., and Canon, J.R. (2006). Conserved microRNA characteristics in mammals. Oligonucleotides *16*, 115-144.

Saini, H.K., Griffiths-Jones, S., and Enright, A.J. (2007). Genomic analysis of human microRNA transcripts. Proceedings of the National Academy of Sciences of the United States of America *104*, 17719-17724.

Sakamoto, K., and Okada, N. (1985). Rodent type 2 Alu family, rat identifier sequence, rabbit C family, and bovine or goat 73-bp repeat may have evolved from tRNA genes. Journal of molecular evolution *22*, 134-140.

Salo, A.M., Sipila, L., Sormunen, R., Ruotsalainen, H., Vainio, S., and Myllyla, R. (2006). The lysyl hydroxylase isoforms are widely expressed during mouse embryogenesis, but obtain tissue- and cell-specific patterns in the adult. Matrix Biol *25*, 475-483.

Salzberg, S.L., Delcher, A.L., Kasif, S., and White, O. (1998). Microbial gene identification using interpolated Markov models. Nucleic acids research *26*, 544-548.

Sasaki, M., Kaneuchi, M., Fujimoto, S., Tanaka, Y., and Dahiya, R. (2003). CYP1B1 gene in endometrial cancer. Mol Cell Endocrinol *202*, 171-176.

Satoh, N., Kawashima, T., Shoguchi, E., and Satou, Y. (2006). Urochordate genomes. Genome dynamics *2*, 198-212.

Sayers, E.W., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., *et al.* (2009). Database resources of the National Center for Biotechnology Information. Nucleic acids research *37*, D5-15.

Schluter, O.M., Schmitz, F., Jahn, R., Rosenmund, C., and Sudhof, T.C. (2004). A complete genetic analysis of neuronal Rab3 function. J Neurosci *24*, 6629-6637.

Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. (2003). Human-mouse alignments with BLASTZ. Genome research *13*, 103-107.

Seale, P., Ishibashi, J., Scime, A., and Rudnicki, M.A. (2004). Pax7 is necessary and sufficient for the myogenic specification of CD45+:Sca1+ stem cells from injured muscle. PLoS biology *2*, E130.

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

Sewer, A., Paul, N., Landgraf, P., Aravin, A., Pfeffer, S., Brownstein, M.J., Tuschl, T., van Nimwegen, E., and Zavolan, M. (2005). Identification of clustered microRNAs using an ab initio prediction method. BMC bioinformatics *6*, 267.

Shackleton, M., Vaillant, F., Simpson, K.J., Stingl, J., Smyth, G.K., Asselin-Labat, M.L., Wu, L., Lindeman, G.J., and Visvader, J.E. (2006). Generation of a functional mammary gland from a single stem cell. Nature *439*, 84-88.

Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J.C., Dwight, S.S., Kaloper, M., Weng, S., Jin, H., Ball, C.A., *et al.* (2001). The Stanford Microarray Database. Nucleic acids research *29*, 152-155.

Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engele, P., McDonagh, P.D., Loerch, P.M., Leonardson, A., Lum, P.Y., Cavet, G., *et al.* (2001). Experimental annotation of the human genome using microarray technology. Nature *409*, 922-927.

Siddiqui, R.A., and Exton, J.H. (1992). Oleate stimulation of diacylglycerol formation from phosphatidylcholine through effects on phospholipase D and phosphatidate phosphohydrolase. Eur J Biochem *210*, 601-607.

Sieuwerts, A.M., Look, M.P., Meijer-van Gelder, M.E., Timmermans, M., Trapman, A.M., Garcia, R.R., Arnold, M., Goedheer, A.J., de Weerd, V., Portengen, H., *et al.* (2006). Which cyclin E prevails as prognostic marker for breast cancer? Results from a retrospective study involving 635 lymph node-negative breast cancer patients. Clin Cancer Res *12*, 3319-3328.

Singh, S.K., Clarke, I.D., Hide, T., and Dirks, P.B. (2004). Cancer stem cells in nervous system tumors. Oncogene *23*, 7267-7273.

Sipione, S., Simmen, K.C., Lord, S.J., Motyka, B., Ewen, C., Shostak, I., Rayat, G.R., Dufour, J.M., Korbutt, G.S., Rajotte, R.V., *et al.* (2006). Identification of a novel human granzyme B inhibitor secreted by cultured sertoli cells. J Immunol *177*, 5051-5058.

Sironi, M., Cagliani, R., Pozzoli, U., Bardoni, A., Comi, G.P., Giorda, R., and Bresolin, N. (2002). The dystrophin gene is alternatively spliced throughout its coding sequence. FEBS letters *517*, 163-166.

Skotheim, R.I., Abeler, V.M., Nesland, J.M., Fossa, S.D., Holm, R., Wagner, U., Florenes, V.A., Aass, N., Kallioniemi, O.P., and Lothe, R.A. (2003). Candidate genes for testicular cancer evaluated by in situ protein expression analyses on tissue microarrays. Neoplasia (New York, NY *5*, 397-404.

Skvorak, A.B., Weng, Z., Yee, A.J., Robertson, N.G., and Morton, C.C. (1999). Human cochlear expressed sequence tags provide insight into cochlear gene expression and identify candidate genes for deafness. Human molecular genetics *8*, 439-452.

Smalheiser, N.R., and Torvik, V.I. (2005). Mammalian microRNAs derived from genomic repeats. Trends Genet *21*, 322-326.

Smalheiser, N.R., and Torvik, V.I. (2006). Alu elements within human mRNAs are probable microRNA targets. Trends Genet *22*, 532-536.

Sodergren, E., Weinstock, G.M., Davidson, E.H., Cameron, R.A., Gibbs, R.A., Angerer, R.C., Angerer, L.M., Arnone, M.I., Burgess, D.R., Burke, R.D., *et al.* (2006). The genome of the sea urchin Strongylocentrotus purpuratus. Science (New York, NY *314*, 941-952.

Sohn, S.Y., Bae, W.J., Kim, J.J., Yeom, K.H., Kim, V.N., and Cho, Y. (2007). Crystal structure of human DGCR8 core. Nature structural & molecular biology *14*, 847-853.

Sokol, N.S., and Ambros, V. (2005). Mesodermally expressed Drosophila microRNA-1 is regulated by Twist and is required in muscles during larval growth. Genes & development *19*, 2343-2354.

Song, L., Han, M.H., Lesicka, J., and Fedoroff, N. (2007). Arabidopsis primary microRNA processing proteins HYL1 and DCL1 define a nuclear body distinct from the Cajal body. Proceedings of the National Academy of Sciences of the United States of America *104*, 5437-5442.

Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., *et al.* (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences of the United States of America *98*, 10869-10874.

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J.S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., *et al.* (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. Proceedings of the National Academy

of Sciences of the United States of America *100*, 8418-8423.

Steemers, F.J., and Gunderson, K.L. (2005). Illumina, Inc. Pharmacogenomics *6*, 777-782.

Stenvang, J., Silahtaroglu, A.N., Lindow, M., Elmen, J., and Kauppinen, S. (2008). The utility of LNA in microRNA-based cancer diagnostics and therapeutics. Seminars in cancer biology *18*, 89-102.

Stingl, J., Eirew, P., Ricketson, I., Shackleton, M., Vaillant, F., Choi, D., Li, H.I., and Eaves, C.J. (2006). Purification and unique properties of mammary epithelial stem cells. Nature *439*, 993-997.

Stoilov, I., Rezaie, T., Jansson, I., Schenkman, J.B., and Sarfarazi, M. (2004). Expression of cytochrome P4501b1 (Cyp1b1) during early murine development. Mol Vis *10*, 629-636.

Stoughton, R.B. (2005). Applications of DNA microarrays in biology. Annu Rev Biochem *74*, 53-82.

Strem, B.M., Hicok, K.C., Zhu, M., Wulur, I., Alfonso, Z., Schreiber, R.E., Fraser, J.K., and Hedrick, M.H. (2005). Multipotential differentiation of adipose tissue-derived stem cells. Keio J Med *54*, 132-141.

Styhler, S., Nakamura, A., Swan, A., Suter, B., and Lasko, P. (1998). vasa is required for GURKEN accumulation in the oocyte, and is involved in oocyte differentiation and germline cyst development. Development (Cambridge, England) *125*,

1569-1578.

Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A., *et al.* (2002). Large-scale analysis of the human and mouse transcriptomes. Proceedings of the National Academy of Sciences of the United States of America *99*, 4465-4470.

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., *et al.* (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proceedings of the National Academy of Sciences of the United States of America *101*, 6062-6067.

Sun, Q., Zhang, Y., Yang, G., Chen, X., Zhang, Y., Cao, G., Wang, J., Sun, Y., Zhang, P., Fan, M., *et al.* (2008). Transforming growth factor-beta-regulated miR-24 promotes skeletal muscle differentiation. Nucleic acids research *36*, 2690-2699.

Svoboda, P., and Di Cara, A. (2006). Hairpin RNA: a secondary structure of primary importance. Cell Mol Life Sci *63*, 901-908.

Sweetman, D., Goljanek, K., Rathjen, T., Oustanina, S., Braun, T., Dalmay, T., and Munsterberg, A. (2008). Specific requirements of MRFs for the expression of muscle specific microRNAs, miR-1, miR-206 and miR-133. Developmental biology *321*, 491-499.

Sweetman, D., Rathjen, T., Jefferson, M., Wheeler, G., Smith, T.G., Wheeler, G.N., Munsterberg, A., and Dalmay, T. (2006). FGF-4 signaling is involved in mir-206 expression in developing somites of chicken embryos. Dev Dyn *235*, 2185-2191.

Tajbakhsh, S., and Buckingham, M.E. (1994). Mouse limb muscle is determined in the absence of the earliest myogenic factor myf-5. Proceedings of the National Academy of Sciences of the United States of America *91*, 747-751.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell *131*, 861-872.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663-676.

Tanaka, S.S., Toyooka, Y., Akasu, R., Katoh-Fukui, Y., Nakahara, Y., Suzuki, R., Yokoyama, M., and Noce, T. (2000). The mouse homolog of Drosophila Vasa is required for the development of male germ cells. Genes & development *14*, 841-853.

Tanzer, A., and Stadler, P.F. (2004). Molecular evolution of a microRNA cluster. Journal of molecular biology *339*, 327-335.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., and Church, G.M. (1999). Systematic determination of genetic network architecture. Nature genetics *22*, 281-285.

Terskikh, A.V., Easterday, M.C., Li, L., Hood, L., Kornblum, H.I., Geschwind, D.H., and Weissman, I.L. (2001). From hematopoiesis to neuropoiesis: evidence of overlapping genetic programs. Proceedings of the National Academy of Sciences of the United States of America *98*, 7934-7939.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of

multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences of the United States of America *99*, 6567-6572.

Toma, J.G., McKenzie, I.A., Bagli, D., and Miller, F.D. (2005). Isolation and characterization of multipotent skin-derived precursors from human skin. Stem cells (Dayton, Ohio) *23*, 727-737.

Tropepe, V., Coles, B.L., Chiasson, B.J., Horsford, D.J., Elia, A.J., McInnes, R.R., and van der Kooy, D. (2000). Retinal stem cells in the adult mammalian eye. Science (New York, NY *287*, 2032-2036.

Tsutsui, Y. (2009). Effects of cytomegalovirus infection on embryogenesis and brain development. Congenital anomalies *49*, 47-55.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America *98*, 5116-5121.

Uchida, N., Buck, D.W., He, D., Reitsma, M.J., Masek, M., Phan, T.V., Tsukamoto, A.S., Gage, F.H., and Weissman, I.L. (2000). Direct isolation of human central nervous system stem cells. Proceedings of the National Academy of Sciences of the United States of America *97*, 14720-14725.

Ueda, R., Iketaki, H., Nagata, K., Kimura, S., Gonzalez, F.J., Kusano, K., Yoshimura, T., and Yamazoe, Y. (2006). A common regulatory region functions bidirectionally in transcriptional activation of the human CYP1A1 and CYP1A2 genes. Mol Pharmacol *69*, 1924-1930.

Valoczi, A., Hornyik, C., Varga, N., Burgyan, J., Kauppinen, S., and Havelda, Z. (2004). Sensitive and specific detection of microRNAs by northern blot analysis using LNA-modified oligonucleotide probes. Nucleic acids research *32*, e175.

van Rooij, E., Quiat, D., Johnson, B.A., Sutherland, L.B., Qi, X., Richardson, J.A., Kelm, R.J., Jr., and Olson, E.N. (2009). A family of microRNAs encoded by myosin genes governs myosin expression and muscle performance. Developmental cell *17*, 662-673.

Vervoort, M., Crozatier, M., Valle, D., and Vincent, A. (1999). The COE transcription factor Collier is a mediator of short-range Hedgehog-induced patterning of the Drosophila wing. Curr Biol *9*, 632-639.

Vienne, A., and Pontarotti, P. (2006). Metaphylogeny of 82 gene families sheds a new light on chordate evolution. Int J Biol Sci *2*, 32-37.

Vizoso, P., Meisel, L.A., Tittarelli, A., Latorre, M., Saba, J., Caroca, R., Maldonado, J., Cambiazo, V., Campos-Vargas, R., Gonzalez, M., *et al.* (2009). Comparative EST transcript profiling of peach fruits under different post-harvest conditions reveals candidate genes associated with peach fruit quality. BMC genomics *10*, 423.

Vogel, G. (2003). Stem cells. 'Stemness' genes still elusive. Science (New York, NY *302*, 371.

Voronina, E., Lopez, M., Juliano, C.E., Gustafson, E., Song, J.L., Extavour, C., George, S., Oliveri, P., McClay, D., and Wessel, G. (2008). Vasa protein expression is

restricted to the small micromeres of the sea urchin, but is inducible in other lineages early in development. Developmental biology *314*, 276-286.

Voskova-Goldman, A., Peier, A., Caskey, C.T., Richards, C.S., and Shaffer, L.G. (1997). DMD-specific FISH probes are diagnostically useful in the detection of female carriers of DMD gene deletions. Neurology *48*, 1633-1638.

Waggoner, S.N., Hall, C.H., and Hahn, Y.S. (2007). HCV core protein interaction with gC1q receptor inhibits Th1 differentiation of CD4+ T cells via suppression of dendritic cell IL-12 production. Journal of leukocyte biology *82*, 1407-1419.

Wang, H., Ach, R.A., and Curry, B. (2007a). Direct and sensitive miRNA profiling from low-input total RNA. RNA (New York, NY *13*, 151-159.

Wang, S.S., Betz, A.G., and Reed, R.R. (2002). Cloning of a novel Olf-1/EBF-like gene, O/E-4, by degenerate oligo-based direct selection. Mol Cell Neurosci *20*, 404-414.

Wang, Y., Jatkoe, T., Zhang, Y., Mutch, M.G., Talantov, D., Jiang, J., McLeod, H.L., and Atkins, D. (2004). Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer. J Clin Oncol *22*, 1564-1571.

Wang, Y., Klijn, J.G., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E., Yu, J., *et al.* (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet *365*, 671-679.

Wang, Y., Medvid, R., Melton, C., Jaenisch, R., and Blelloch, R. (2007b). DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. Nature genetics *39*, 380-385.

Ward, C.R., Faundes, D., and Foster, J.A. (1999). The monomeric GTP binding protein, rab3a, is associated with the acrosome in mouse sperm. Mol Reprod Dev *53*, 413-421.

Washietl, S., Hofacker, I.L., Lukasser, M., Huttenhofer, A., and Stadler, P.F. (2005a). Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. Nature biotechnology *23*, 1383-1390.

Washietl, S., Hofacker, I.L., and Stadler, P.F. (2005b). Fast and reliable prediction of noncoding RNAs. Proceedings of the National Academy of Sciences of the United States of America *102*, 2454-2459.

Washietl, S., Pedersen, J.S., Korbel, J.O., Stocsits, C., Gruber, A.R., Hackermuller, J., Hertel, J., Lindemeyer, M., Reiche, K., Tanzer, A., *et al.* (2007). Structured RNAs in the ENCODE selected regions of the human genome. Genome research *17*, 852-864.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Watson, E.D., Mattar, P., Schuurmans, C., and Cross, J.C. (2009). Neural stem cell self-renewal requires the Mrj co-chaperone. Dev Dyn *238*, 2564-2574.

Wheeler, G., Valoczi, A., Havelda, Z., and Dalmay, T. (2007). In situ detection of animal and plant microRNAs. DNA and cell biology *26*, 251-255.

White, R.J., Stott, D., and Rigby, P.W. (1990). Regulation of RNA polymerase III transcription in response to Simian virus 40 transformation. The EMBO journal *9*, 3713-3721.

Wickiser, J.K., Winkler, W.C., Breaker, R.R., and Crothers, D.M. (2005). The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch. Molecular cell *18*, 49-60.

Wienholds, E., Kloosterman, W.P., Miska, E., Alvarez-Saavedra, E., Berezikov, E., de Bruijn, E., Horvitz, H.R., Kauppinen, S., and Plasterk, R.H. (2005). MicroRNA expression in zebrafish embryonic development. Science (New York, NY *309*, 310-311.

Williams, C., Helguero, L., Edvardsson, K., Haldosen, L.A., and Gustafsson, J.A. (2009). Gene expression in murine mammary epithelial stem cell-like cells shows similarities to human breast cancer gene expression. Breast Cancer Res *11*, R26.

Williams, W.P., Tamburic, L., and Astell, C.R. (2004). Increased levels of B1 and B2 SINE transcripts in mouse fibroblast cells due to minute virus of mice infection. Virology *327*, 233-241.

Wolpert, L. (2002). Principles of development, 2nd edn (Oxford, Oxford University Press).

Wong, C.F., and Tellam, R.L. (2008). MicroRNA-26a targets the histone

methyltransferase Enhancer of Zeste homolog 2 during myogenesis. The Journal of biological chemistry *283*, 9836-9843.

Wu, Z., and Irizarry, R.A. (2005). Stochastic models inspired by hybridization theory for short oligonucleotide arrays. J Comput Biol *12*, 882-893.

Xia, Y., Yamagata, K., and Krukoff, T.L. (2006). Differential expression of the CD14/TLR4 complex and inflammatory signaling molecules following i.c.v. administration of LPS. Brain Res *1095*, 85-95.

Xiong, H., Qian, J., He, T., and Li, F. (2009). Independent transcription of miR-281 in the intron of ODA in Drosophila melanogaster. Biochemical and biophysical research communications *378*, 883-889.

Xu, C., Lu, Y., Pan, Z., Chu, W., Luo, X., Lin, H., Xiao, J., Shan, H., Wang, Z., and Yang, B. (2007). The muscle-specific microRNAs miR-1 and miR-133 produce opposing effects on apoptosis by targeting HSP60, HSP70 and caspase-9 in cardiomyocytes. J Cell Sci *120*, 3045-3052.

Xu, N., Papagiannakopoulos, T., Pan, G., Thomson, J.A., and Kosik, K.S. (2009). MicroRNA-145 regulates OCT4, SOX2, and KLF4 and represses pluripotency in human embryonic stem cells. Cell *137*, 647-658.

Yanaihara, N., Caplen, N., Bowman, E., Seike, M., Kumamoto, K., Yi, M., Stephens, R.M., Okamoto, A., Yokota, J., Tanaka, T., *et al.* (2006). Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. Cancer cell *9*, 189-198.

Yano, S., Ito, Y., Fujimoto, M., Hamazaki, T.S., Tamaki, K., and Okochi, H. (2005). Characterization and localization of side population cells in mouse skin. Stem cells (Dayton, Ohio) *23*, 834-841.

Yeakley, J.M., Fan, J.B., Doucet, D., Luo, L., Wickham, E., Ye, Z., Chee, M.S., and Fu, X.D. (2002). Profiling alternative splicing on fiber-optic arrays. Nature biotechnology *20*, 353-358.

Yeom, K.H., Lee, Y., Han, J., Suh, M.R., and Kim, V.N. (2006). Characterization of DGCR8/Pasha, the essential cofactor for Drosha in primary miRNA processing. Nucleic acids research *34*, 4622-4629.

You, L.R., Lin, F.J., Lee, C.T., DeMayo, F.J., Tsai, M.J., and Tsai, S.Y. (2005). Suppression of Notch signalling by the COUP-TFII transcription factor regulates vein identity. Nature *435*, 98-104.

Yuasa, K., Hagiwara, Y., Ando, M., Nakamura, A., Takeda, S., and Hijikata, T. (2008). MicroRNA-206 is highly expressed in newly formed muscle fibers: implications regarding potential for muscle regeneration and maturation in muscular dystrophy. Cell structure and function *33*, 163-169.

Zaghloul, N.A., and Moody, S.A. (2007). Alterations of rx1 and pax6 expression levels at neural plate stages differentially affect the production of retinal cell types and maintenance of retinal stem cell qualities. Developmental biology *306*, 222-240.

Zeng, Y., and Cullen, B.R. (2005). Efficient processing of primary microRNA hairpins by Drosha requires flanking nonstructured RNA sequences. The Journal of

biological chemistry *280*, 27595-27603.

Zerial, M., and McBride, H. (2001). Rab proteins as membrane organizers. Nat Rev Mol Cell Biol *2*, 107-117.

Zhang, B.H., Pan, X.P., Wang, Q.L., Cobb, G.P., and Anderson, T.A. (2005). Identification and characterization of new plant microRNAs using EST analysis. Cell Res *15*, 336-360.

Zhang, J., Niu, C., Ye, L., Huang, H., He, X., Tong, W.G., Ross, J., Haug, J., Johnson, T., Feng, J.Q., *et al.* (2003). Identification of the haematopoietic stem cell niche and control of the niche size. Nature *425*, 836-841.

Zhang, W., Morris, Q.D., Chang, R., Shai, O., Bakowski, M.A., Mitsakakis, N., Mohammad, N., Robinson, M.D., Zirngibl, R., Somogyi, E., *et al.* (2004). The functional landscape of mouse gene expression. J Biol *3*, 21.

Zohn, I.E., and Brivanlou, A.H. (2001). Expression cloning of Xenopus Os4, an evolutionarily conserved gene, which induces mesoderm and dorsal axis. Developmental biology *239*, 118-131.

Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule. Science (New York, NY *244*, 48-52.

Chapter 7 Appendix

7.1 Miscellaneous Figures

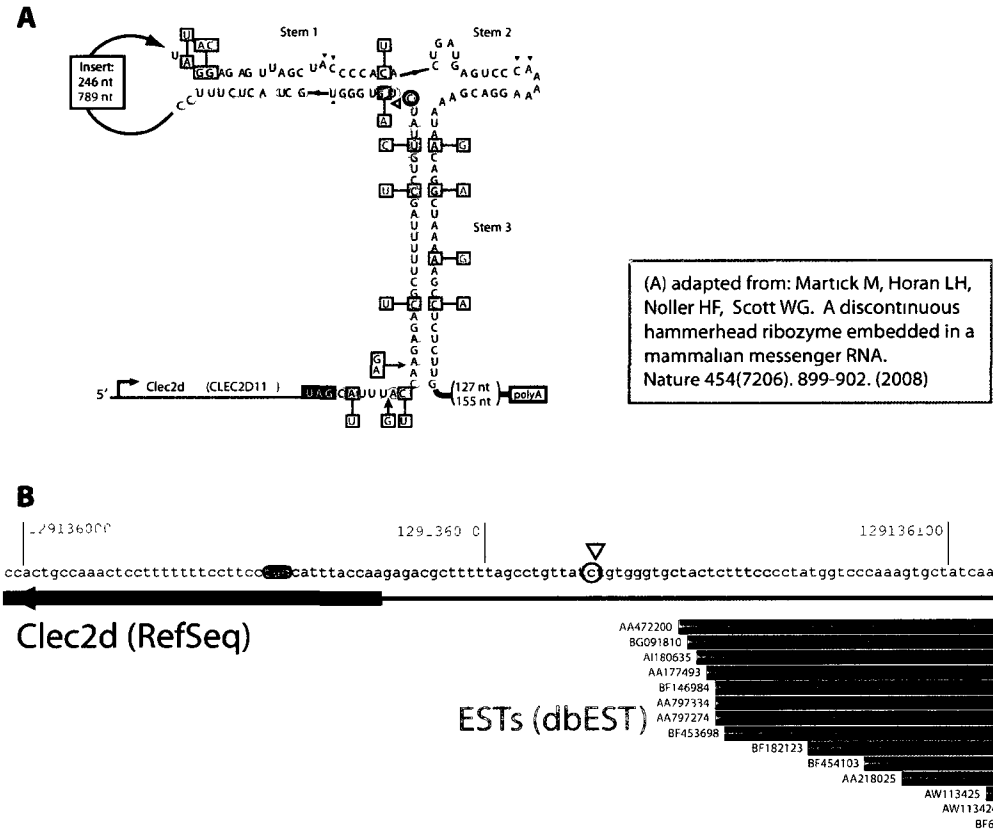

Figure 7-1: Clec2d is an example of a ribozyme with supporting EST evidence
A: Figure of a ribozyme embedded in the Clec2d 3'UTR as published in Martick et al., 2008 (Martick et al.). B: Layout of the Clec2d 3'UTR region showing a cluster of ESTs terminating in proximity of the ribozyme cleavage site (Indicated by the black triangle in both panels).