**Title**
Defining the Chromatin Signatures at Regulatory Regions of Tissue Specific Genes



**Author**
Edwards, Miguel

**Publication Date**
2014

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

# Defining the Chromatin Signatures at Regulatory Regions of Tissue Specific Genes

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Molecular Biology

by

Miguel Edwards

2014

ABSTRACT OF THE DISSERTATION

Defining the Chromatin Signatures at Regulatory Regions of Tissue Specific Genes

by

Miguel Edwards

Doctor of Philosophy in Molecular Biology

University of California, Los Angeles, 2014

Professor Stephen T. Smale, Chair

The activation of tissue specific genes relies upon the precise orchestration of a number of events that result in the initiation of transcription upon lineage specification. This process is heavily dictated by the chromatin environment both at the promoter and distal sequences, as well as by the availability of transcription factors necessary for activation. A critical role is played at distal regulatory sequences, which often are the first sites to be engaged by key regulatory proteins. This interaction often promotes a chromatin environment that is necessary for the activation of the gene and results in the recruitment of additional sequence specific factors and a direct interaction with the promoter to initiate transcription. Understanding the properties of enhancer elements for tissue specific genes is important for a clear understanding of the mechanisms of activation. A number of studies have shown that enhancers are marked long before the activation of the gene takes place, in some cases as early as the embryonic stem cell stage. A detailed study described an unmethylated window within the enhancer of the Ptcra

locus. Further analysis showed the enhancer mark to be regulated by sequence specific binding factors. These studies lacked the appropriate chromatin environment, which we know to be important. Here we use a bacterial artificial chromosome containing the Ptcra locus to demonstrate that the enhancer mark persists in a chromatin context but is not regulated in the manner described in a non-native chromatin context.

We then expand our studies to global tissue specific gene expression in order to understand more broadly the regulatory properties the define tissue specific genes. Parsing the mechanisms that drive tissue specific gene expression is critical for an understanding of pluripotency and tissue specificity. Here we use deep chromatin RNA-sequencing to accurately quantify the transcriptome of pluripotent stem cells and four primary differentiated cell types – E14.5 cortical neurons, $CD4^+$ $CD8^+$ thymocytes, bone marrow-derived macrophages and hepatocytes, in mouse. We define tissue specific genes with the broadest dynamic range in expression and define the chromatin properties at their promoters. Separating tissue specific genes with the largest dynamic range in expression allowed us to uncover cell type specific differences in the fundamental promoter properties.

The dissertation of Miguel Edwards is approved by:

Arnold J. Berk

Guoping Fan

William Lowry

Kathrin Plath

Stephen T. Smale, Committee Chair

University of California, Los Angeles

2014

In dedication to my wife, Janel

## TABLE OF CONTENTS

Acknowledgements

I would like to thank my advisor, Stephen Smale, for his continued support and guidance throughout my time as a graduate student. I would like to thank the members of my thesis committee Arnold J. Berk, Guoping Fan, William Lowry, and Kathrin Plath for their help, guidance and support. I would also like to thank all past and present members who have been a part of my graduate journey, particularly Jian Xu, Scott Pope, Abe Chang, Kevin Doty, Justin Langerman and Prabhat Purbey. I am fortunate to have made lasting friendships with so many of you.

I would like to thank my wife and family for their continued love and support

Vita

| | |
|---|---|
| 2007 | Bachelor of Arts, Biology<br>Hunter College - City University New York |
| 2005-2007 | MARC Fellowship<br>Hunter College - City University New York |
| 2005-2007 | Undergraduate Research Internship<br>Hunter College - City University New York |
| 2006 | HHMI Exrop Undergraduate Research Scholar<br>University of California, Berkeley |
| 2007-2008 | ACCESS Program<br>University of California, Los Angeles |
| 2008-present | Molecular Biology Institute<br>University of California, Los Angeles |
| 2009-2010 | Teaching Assistant<br>Department of Microbiology, Immunology and Molecular Genetics<br>University of California, Los Angeles |
| 2008-2011 | Cellular and Molecular Training Grant<br>University of California, Los Angeles |

**Presentations**

**Miguel Edwards,** Constantinos Chronis, Jessica Grindheim, Kenneth Zaret, Matteo Pellegrini, Kathrin Plath, and Stephen Smale. **2014**. Quantitative Analysis of Transcription Dynamics in Pluripotent and Differentiated Cells. Maryland NIH

# Chapter 1

# Introduction

## Pluripotency, Tissue Specificity and the Transcriptome

## A. Transcriptome Profiling

One of the most fundamental questions in biology is based on our understanding that in complex multicellular organisms, all cells contain identical DNA content, yet, they have hundreds of different cell types, with distinct functions. The variations in cell types are based on differences in gene expression. The full complement of transcribed RNA in a given cell, i.e. the transcriptome, determines the proteins to be translated that will collectively impart both the form and function of a cell. To understand what drives functional outcomes and phenotypic differences is one of the most fundamental aspects of biology. Interrogating not only what genes are expressed but also the level of their expression is critical in order to continue making strides in our understanding of cellular identity. The importance of defining and quantifying transcriptomes is clear. We can study a variety of fundamental biological phenomena from development to disease, and infer the roles of single genes or groups of genes in these processes.

The earliest attempts to characterize mammalian transcriptomes began with the publication of the human expressed sequence tags (EST) database. In this study 609 human brain complementary DNA (cDNA) clones were used to generate ESTs. Sanger sequencing of those clones resulted in the identification of 337 novel genes (Mark D. Adams, 1991). The utility of ESTs at this early stage was in discovery. It was not until the serial analysis of gene expression (SAGE) method was developed, that transcriptomes began to be interrogated in a somewhat global manner. SAGE allowed for the analysis of thousands of transcripts with relative abundancies, which simply could not be achieved cost effectively by sequencing of ESTs (Victor E. Velculescu, 1995). Almost

concurrently a new and powerful high-throughput method to quantify gene expression arose, microarrays.

**A 1. Microarray Technology**

The expansion of cDNA databases with sequences from a number of organisms precipitated the opportunity for microarray technology. Knowing some or many of the sequences expressed in a cell type or organism, one could use this information to quantify the expression of those corresponding genes. By amplifying cDNA sequences from *Arabidopsis thaliana* and depositing individual clones into single wells of a 96well plate, Schena et. al., had created a library which could be used to quantify the expression of those *Arabidopsis* genes (Schena M., 1995). Using robotics, samples from the plate were printed on glass slides. Chemical and heat treatment attached the DNA to the slide and denatured the DNA making it accessible for complementary hybridization. The mRNA to be interrogated was then reverse transcribed with the incorporation of fluorescent dyes. After hybridization of the labeled cDNA with the slide and subsequent washing, a laser scanned the slide. The intensity of the light was then used as a measure of expression. Over the next decade and through many iterations of the microarray, a vast amount of valuable information has been gathered. There are over 520,000 individual microarray experiments archived in the Gene Expression Omnibus (GEO) repository (John H Malone, 2011). Despite this success there are still a number of limitations associated with microarrays, some have been overcome but many remain. One of the most glaring deficiencies of microarrays is the reliance upon the knowledge of existing sequences. With the invention of tiling arrays, the reliance on existing knowledge of genome

structure was mitigated. However, limitations due to cross-hybridization, normalization and saturation have remained. These issues cause high background, difficulties in comparing across experiments and a limited dynamic range, respectively.

**A 2. RNA Sequencing**

The arrival of Next-Generation Sequencing (NGS) impacted the study of the transcriptome in an unprecedented manner. With this novel method of high-throughput DNA sequencing, it became possible to sequence millions of short sequences simultaneously i.e. in a massively parallel fashion. Out of this innovation sprung the method RNA sequencing (RNA-Seq), which now provides us with an unparalleled opportunity to identify and accurately quantify transcripts. One of the first demonstrations of the utility of RNA-Seq was the analysis of the LNCaP transcriptome, a prostate cancer cell line (Bainbridge et al., 2006). In this study expression was detected from over 10,000 gene loci, 25 novel splicing events were detected, 1,500 single nucleotide polymorphisms were detected, and thousands of sequences were mapped to regions of the genome where transcription was not previously predicted. Two years later more evidence for the utility of RNA-Seq emerged with the sequencing of tens of millions of ~30bp reads, compared with only 200,000 110bp reads in the LNCaP study. Three papers using RNA-Seq and Illumina sequencing technology demonstrated a high reproducibility of technical replicates, r= 0.96, higher sensitivity than microarrays, a broader dynamic range than microarrays, the ability to identify splice junctions and the capability to identify novel transcripts without prior knowledge of sequence. (Mortazavi

et al., 2008; Sultan et al., 2008; Wilhelm et al., 2008). RNA-Seq was poised to supplant microarray technology and accelerate our understanding of complex eukaryotic genomes.

Although RNA-Seq overcame many of the challenges associated with microarray technologies, it brought a number of new challenges of its own, namely bioinformatic analysis. It took almost ten years for some type of consensus to be reached for microarray analysis methods (David B. Allison, 2006). Eight years on from the invention of RNA-Seq new analysis methods continue to arise, with no clear consensus in sight. Moreover, the technology continues to evolve rapidly with increased read lengths, paired-end sequencing, strand-specific sequencing and single cell sequencing, compounding the ability to reach a consensus (Ozsolak and Milos, 2011; Parkhomchuk et al., 2009; Tang et al., 2009). The most significant issues revolve around the massive amounts of data; strategies must be devised to process, manage and store that data. Mapping millions of sequences back to a reference genome is no small task. As the number of reads has continued to increase the fastest and most efficient methods for alignment have risen to the top e.g. Bowtie (Langmead et al., 2009). Those methods continue to be revised and updated.

Despite the addition of new challenges, RNA-Seq still provides the best opportunity to date for the scientific community to fully grasp the complexity of the eukaryotic genome. Ultimately the goal of defining transcriptomes is to understand phenotype and phenotypic differences. An accurate and quantifiable assessment of transcriptomes in different cell types, developmental stages and disease states, will be indispensable in our efforts to further our understanding of gene regulation. One cell type of great interest is the embryonic stem (ES) cell.

**B. Embryonic Stem Cells**

Embryonic Stem (ES) cells, derived from the inner cell mass (ICM) of the pre-implantation blastocyst, are characterized by their ability to self-renew while retaining the capacity to differentiate into a multitude of distinct cell types. The establishment of ES cell lines demonstrated, that under the appropriate conditions, ES cells have the capacity to proliferate indefinitely, i.e. self-renewal. Germline transmission and *in vitro* differentiation demonstrated the ability of mouse and human ES cells respectively, to differentiate into cells from all three germ layers, i.e. pluripotency (M. J. Evans, 1981; Martin, 1981; Thomson, 1998). These fundamental properties are the reason ES cells hold such promise as a model for developmental regulation, different ion and therapeutic application.

Fusion of ES cells with somatic cells results in the conversion of the somatic cells to a pluripotent state (Cowan et al., 2005; Jaenisch et al., 2004). Overexpression of defined transcription factors (Oct3/4, Sox2, Klf4, and c-Myc) in both mouse and human fibroblasts also results in a conversion to an induced pluripotent state (Takahashi et al., 2007; Takahashi and Yamanaka, 2006). Defining the genes that represent the ES cell phenotype will undoubtedly facilitate our understanding of the pluripotent state and reprogramming. Moreover, a quantitative understanding of the ES cell transcriptome will bring further insight. This perspective will allow for genes to be classified based on their levels of expression. A careful quantitative delineation of gene classes will provide an opportunity to examine their regulatory properties, including the prevalence of different

promoter classes, their DNA methylation and histone modification signatures, and their interactions with key transcription factors.

**C. Pluripotency**

Substantial progress has been made in understanding the key determinants of pluripotency. These molecular mechanisms consist of complex signaling pathways, genetics and epigenetics all intertwined. The aforementioned are regulated during development allowing for the appropriate transitions between cellular states (Chambers and Tomlinson, 2009; Chen et al., 2008b; Marks et al., 2012; Marks and Stunnenberg, 2014; Ying et al., 2008). The transcription factors that regulate gene expression in ES cells are critical to ES cell identity. Oct4, Sox2 and Nanog have been termed the "core pluripotency network" for their roles in both establishment and maintenance of pluripotency, with null mutations in each unable to maintain pluripotency (Avilion et al., 2003; Dejosez and Zwaka, 2012; Jennifer Nichols et al., 1998; Kaoru Mitsui et al., 2003). These transcription factors bind to regulatory DNA sequences and control the transcription of many downstream genes necessary for the maintenance of pluripotency. These transcription factors also bind their own regulatory sequences and regulating their own expression (Boyer et al., 2005; Chen et al., 2008a; Chen et al., 2008b; Loh et al., 2006). This in effect creates a complex regulatory network sensitive to a number of stimuli.

Oct4, a homeodomain transcription factor encoded for by the gene *Pou5f1*, has an expression pattern restricted to totipotent, pluripotent and germ cells (Dejosez and Zwaka, 2012). Oct4 has been shown to be required for pluripotency both *in vivo* (Hitoshi

Niwa, 2000; Jennifer Nichols et al., 1998), and *in vitro* (Thomson et al., 2011). Silencing of Oct4 leads to the differentiation of ES cells and the ICM into trophoblast-like cells (Jennifer Nichols et al., 1998), while overexpression results in differentiation of ES cells into primitive endoderm and mesoderm (Hitoshi Niwa, 2000).

Sox2, the high mobility group (HMG) box containing transcription factor, appears to play a more nuanced role based on the evidence that inducible Sox2 null ES cells resemble the loss of Oct4 phenotype, differentiation into trophectoderm like cells (Masui et al., 2007). Furthermore, overexpression of Oct4 can rescue Sox2 null ES cells from differentiation but interestingly, Nanog cannot (Masui et al., 2007). The loss of Sox2 leads to the aberrant expression of both positive and negative regulators of Oct4, resulting in the eventual decrease in Oct4 expression and differentiation to trophectoderm (Masui et al., 2007). Sox2 has also been shown to act cooperatively with Oct4, co-occupying numerous promoters (Avilion et al., 2003). Thus, it appears the role of Sox2 is in maintaining the expression of transcription factors that are necessary for the precise expression of Oct4.

Nanog, a homeodomain containing transcription factor, has a similar expression pattern to Oct4 (Ian Chambers et al., 2003; Kaoru Mitsui et al., 2003). However, Nanog null ES cells differentiate into parietal endoderm-like cells (Ian Chambers et al., 2003; Kaoru Mitsui et al., 2003). Overexpression of Nanog bypasses the requirement of the signaling pathway downstream effector Stat3, for maintenance of ES cells *in vitro* (Kaoru Mitsui et al., 2003). These phenotypes define a distinct role of Nanog in the prevention of differentiation into extraembryonic endodermal lineages. The precise role of Nanog has become more complicated with the finding that Nanog $^{-/-}$ cells are able to self-renew and

also contribute to lineages outside of endoderm. (Chambers et al., 2007). Chambers et. al., posit Nanog may function to stabilize pluripotency of ES cells, instead of being absolutely required for maintenance. Oct4, Sox2 and Nanog lay the transcriptional foundation for the autoregulatory feedback mechanism necessary to maintain pluripotency.

**C 1. Genetic Control of Pluripotency**

Aside from these key transcription factors, a plethora of interacting partners have been described and our knowledge of the roles of these interacting partners continues to expand (Dejosez and Zwaka, 2012; Marks and Stunnenberg, 2014). Employing various genome-wide profiling techniques, from ChIP-chip to ChIP-Seq, a number of groups defined the binding sites of the core factors in both mouse and human ES cell genomes (Apostolou et al., 2013; Boyer et al., 2005; Hammachi et al., 2012; Jerabek et al., 2014; Kim et al., 2010; Loh et al., 2006; Ng and Surani, 2011; Orkin et al., 2008; van den Berg et al., 2010; Wang et al., 2006). From these studies hundreds of target genes were revealed and a number of important insights arose. Firstly, Oct4, Sox2 and Nanog appear to act co-operatively based on the high degree of overlap in their binding sites, particularly in human, 90% (Boyer et al., 2005; Loh et al., 2006). Secondly, the targets include actively transcribed genes known to be necessary for the prevention of differentiation. This list also includes genes that require repression in order to maintain an undifferentiated state. Other key targets include chromatin modifying enzymes and signal transduction pathway components (Boyer et al., 2005; Kim et al., 2008; Loh et al., 2006). Thirdly, there is a correlation between the presence of the core pluripotency transcription

factors and gene activity. In the studies performed Kim et. al., the binding of nine transcription factors (Nanog, Oct4, Sox2, Klf4 and MycDax1, Nac1, Zfp281 and Rex1) was assessed using biotinylation mediated ChIP-chip, in mouse ES cells. Two major categories emerged from their studies: (1) genes bound by greater than four factors that were largely active, and (2) those bound by only a few factors that were largely inactive. Furthermore those active genes undergo repression upon differentiation (Kim et al., 2008). Taken together, Oct4, Sox2 and Nanog bind and activate genes necessary for the maintenance of pluripotency and self-renewal, while repressing genes involved in lineage specification.

After these initial findings, and with the development of induced pluripotency, a number of other key transcriptional regulators established their position within the pluripotency hierarchy, including the Myc transcription factors (Smith et al., 2011). c-Myc was implicated in cell cycle control, DNA replication/repair and metabolism based on genome-wide binding profiles (Kim et al., 2008). However, the roles of N and c-Myc in pluripotency remained unclear due to their phenotypes only becoming apparent at later stages in development (Charron et al., 1992; Stanton et al., 1992). In addition N or c-Myc deficient ES cells remain pluripotent and replicate indefinitely (Malynn et al., 2000). Using a double knockout strategy (dKO) to circumvent the redundancies, Varlakhanova et. al, showed N and c-Myc to be critical for pluripotency and self-renewal (Varlakhanova et al., 2010). These findings were corroborated in a separate study with the addition of a mechanism for the maintenance of pluripotency. Smith et. al, showed the dKO resulted in differentiation to primitive endoderm due to the upregulation of Gata6 (Smith et al., 2010). It also became apparent that Myc has the potential to regulate

a completely different group of genes from that of the core pluripotency factors, based on its binding profile (Chen et al., 2008b; Kim et al., 2010; Smith et al., 2011). In conjunction with the finding that the exclusion of c-Myc drastically reduces the efficiency of reprogramming somatic cells to induced pluripotent stem cells, c-Myc has planted itself firmly within the pluripotency network (Wernig et al., 2008).

**C 2. Epigenetic Control of Pluripotency**

There is an ever-expanding role for epigenetic mechanisms in pluripotency. These mechanisms have continued to be refined as our understanding of the complex transcriptional networks involved in pluripotency deepens. Chromatin structure, histone modifications and DNA methylation are among these mechanisms.

One of the early implications of epigenetic mechanisms was the finding that ES cell chromatin is fundamentally different from that of differentiated cells (Meshorer et al., 2006). ES cells were determined to have an “open” chromatin configuration with a more dynamic exchange of chromatin components compared with that of differentiated cells. Heterochromatin protein HP1-α, involved in chromatin compaction, shows fewer and more diffuse foci in ES cells compared to lineage committed cells. This so called “hyperdynamic” state is considered important for the maintenance of plasticity, and thus critical to the undifferentiated ES cell state (Meshorer and Misteli, 2006; Meshorer et al., 2006).

Further support for epigenetic mechanisms arose with the finding that genes encoding developmental regulators appeared to be poised for activation upon lineage commitment based on modifications at the histone level. (reviwed by Voigt et al., 2013a).

Genome-wide ChIP-chip analyses in ES cells identified a class of genes whose promoters were marked by histone modifications associated with both activation and repression, simultaneously. These "bivalent" domains consist of histone H3 Lysine 4 trimethylation (H3K4me3), and histone H3 Lysine 27 trimethylation (H3K27me3). Upon differentiation many of these genes resolve to either the active or repressive mark. (Bernstein et al., 2006; Boyer et al., 2006a). The sum of these findings led to the hypothesis that this poised state allows for the appropriate expression of these developmental regulators in a timely manner. Much has been learnt about the mechanisms involved in the establishment of these marks but the lack of functional evidence for their role in development leaves this subject in somewhat of a controversy.

The polycomb group proteins (PcG) were identified in *Drosophila* as mediators of repression for the developmental Hox genes (Lewis, 1978). These highly conserved proteins act in two multi-subunit complexes termed polycomb repressive complexes (PRC) 1 and 2, modifying histone tails (Di Croce and Helin, 2013). PRC2 is comprised of the core PcG proteins Ezh1/2, Eed and Suz12. Ezh1/2 is the catalytic component responsible for mono-, di- and trimethylation of H3K27. H3K27me3 serves at the binding surface for CBX, a chromodomain containing subunit of PRC1 and binds along with family members of PCGF, HPH and Ring1A/1B. Ring1A/1B catalyzes the monoubiquitination of histone 2A at Lysine 119 (H2AK119Ub). Together these proteins mediate repression in a multiplayer fashion including chromatin compaction and blocking of polymerase II (Pol II) elongation (Di Croce and Helin, 2013).

Defining the PcG proteins specific role in maintenance of pluripotency has been somewhat hampered by the existence of many homologus proteins with overlapping

function. Currently one of the few pieces of evidence for the direct implication of PcG proteins in pluripotency is the Ring1A/1B dKO phenotype. The dKO results in loss of ES cell morphology and the inability to proliferate indefinitely due to the derepression of a many PRC 1 targets (Endoh et al., 2008). Although the direct evidence for the requirement of PRC1 and 2 in pluripotency and self-renewal is lacking, it is evident that PcG proteins play a key role in the ES cell differentiation program.

**D. Reprogramming of Somatic Cells to an Embryonic like State**

The landmark paper from Takahashi and Yamanaka in 2006 fundamentally changed our understanding of the pluripotent state. Using the viral overexpression of four transcription factors, Oct4, Sox2, Klf4 and c-Myc (OSKM), Takahashi and Yamanaka were able to convert mouse embryonic fibroblasts (MEFs) and adult human fibroblasts into induced pluripotent stem (iPS) cells (Takahashi et al., 2007; Takahashi and Yamanaka, 2006). Independently the Thompson group generated human iPS cells using OS and two other key contributors to ES cell identity, NANOG and LIN28 (Yu et al., 2007). Since that time researchers have devised many methods to achieve reprogramming, including the substitution of each of the original factors by other proteins, micro-RNAs, chromatin modifiers and small molecules (Anokye-Danso et al., 2011; Hussein and Nagy, 2012; Ma et al., 2013; Papp and Plath, 2013). Most recently, Oct4, in the OSKM cocktail, was replaced by mesodermal lineage specifier GATA3 in both mouse and human (G3SKM) (Montserrat et al., 2013; Shu et al., 2013). Not only are we capable of converting multiple somatic cell types into iPS cells, but also converting

fibroblasts into numerous other lineages i.e. transdifferentiation (Hussein and Nagy, 2012; Vierbuchen and Wernig, 2012).

By most accounts iPS cells are similar to their embryonic equivalent in terms of functionality, gene expression, DNA methylation and the distribution of chromatin modifications (Lowry, 2012; Plath and Lowry, 2011). Much of the focus remains on understanding and overcoming the inefficiency of the process in an effort to gain insights into the molecular mechanisms that drive reprogramming. The widely accepted model is that the factors activate the core pluripotency transcriptional network. This autoregulatory network establishes the pluripotent state, which is then maintained by endogenous factors. Simultaneously the overexpressed factors modify the expression of the genes controlling the differentiated cell type in order to facilitate and solidify the pluripotency network (Chou and Cheng, 2013). This remarkable finding presents a wonderful opportunity for the modeling of diseases and the generation patient specific iPS cells for personalized regenerative medicine.

**E. Epigenetic Regulation of Gene Expression**

The regulation of gene expression is predominantly controlled by the intrinsic properties of the DNA and the transcription factors that bind regulatory sequences. There are additional layers of heritable regulation that can occur without any changes in DNA sequence. These covalent modifications, added to both DNA and histones, have become widely accepted as a major contributing factor in gene regulation and now encompass the field of epigenetics.

## E 1. DNA Methylation

DNA methylation is one of the most well studied epigenetic modifications. It is a critical modification involved in a wide range of cellular processes including development, transcription, X chromosome inactivation, chromosome stability, suppression of mobile genetic elements and genomic imprinting (Jones, 2012; Smith and Meissner, 2013). In mammals DNA methylation, the addition of a methyl group, most often occurs on the fifth carbon of cytosines (5mC), in the context of CpG dinucleotides. Genome-wide methylation patterns remain mostly static once established in early embryogenesis. The establishment of this covalent modification is initiated by *de novo* methyltransferases and then propagated throughout somatic differentiation by maintenance methyltransferases. In mammals approximately 70-80% of CpGs are methylated (Kohli and Zhang, 2013; Smith and Meissner, 2013; Wu and Zhang, 2014) While the majority of the genome is methylated there are specific regions which tend to be refractory to DNA methylation, CpG Islands (CGI). These regions contain a high density of CG dinucleotides and are often found at transcriptional start sites. This bimodal distribution shapes of our understanding of DNA methylation (Kohli and Zhang, 2013; Smith and Meissner, 2013).

### E 1.1 CpG Islands

Methylated cytosines spontaneously deaminate to thymine. This mutagenic property has resulted in the underrepresentation of CG dinucleotides throughout

mammalian genomes (Deaton and Bird, 2011; Illingworth and Bird, 2009). This property suggests remaining CpGs have been preserved because they are important for function and or these CGI are rarely methylated and avoid mutagenesis. CGI's were first identified in mouse genomic DNA using a methyl specific restriction enzyme. This cleavage resulted in a fraction of the genome that was highly fragmented. Those fragments were dense clusters of unmethylated CpG dinucleotides (Bird et al., 1985; Cooper et al., 1983). Quantification and sequence analysis determined these fragments to encompass ~26,000 distinct CGIs (Antequera and Bird, 1993). Characterization of these sequences resulted in the original definition of a CGI; regions greater than 200bp with a G + C content greater than 50% and a ratio of observed CpG frequency to expected CpG frequency of 0.6 (Gardiner-Garden and Frommer, 1987). This definition continues to be refined with advances in computational analysis methods.

CGIs are the most common promoter type with more ~70% of annotated promoters associated with a CGI (Blackledge et al., 2013; Deaton and Bird, 2011). CGIs are most often associated with housekeeping genes and to a lesser extent tissue-specific and developmental regulator genes (Larsen et al., 1992; Zhu et al., 2008). There are CGIs located in intergenic regions separate from annotated genes but many also have the ability to initiate transcription (Illingworth et al., 2010; Maunakea et al., 2010). Many studies provide evidence supporting the relationship between CGIs and the initiation of transcription.

## E 1.2 DNA Methylation Enzymes

Three conserved enzymes catalyze DNA methylation, DNA methyltransferase (DNMT) 1, DNMT3A and DNMT3B. DNMT1, which is expressed ubiquitously, is associated with replication foci in replicating cells (Bestor et al., 1988; Leonhardt et al., 1992). DNMT1 knock out ES cells result in wide spread demethylation. The preference of DNMT1 for hemi-methylated CpGs in combination with the previous study indicated DNMT1 functioned as a maintenance methyltransferase (Lei et al., 1996; Smith and Meissner, 2013). DNMT3a and DNMT3b, however function as *de novo* methyltransferases, i.e. the introduction of 5mC at unmethylated and not hemi-methylated sites. Evidence supporting this notion included knock out studies in mice showing both 3a and 3b are required for *de novo* methylation in ES cells and imprinting in germ cells (Hata et al., 2002; Okano et al., 1999). More recently evidence suggests DNMT1 alone is not sufficient for the long-term maintenance of methylation complicating our initial understanding (Jones and Liang, 2009).

The three essential enzymes mentioned above are not the only methyltransferases. DNMT2 and DNMT3L exist in mammalian cells. DNMT2 knock out ES cells have no effect on global methylation levels suggesting no involvement in propagation of DNA methylation patterns *in vivo* (Li et al., 1992; Okano et al., 1999). DNMT3L is related but lacks DNA methyltransferase activity. It does however form complexes with DNMT3a and 3b and has been shown to modulate their activity (Goll and Bestor, 2005). A number of additional roles have begun to emerge for DNMT3L and the extent of its involvement in the maintenance and establishment of DNA methylation patterns continues to be explored (Hata et al., 2002).

## E 1.3 Functions of DNA Methylation

DNA methylation is most commonly known for its functions in long term silencing of gene transcription. Advances in genome wide mapping of this epigenetic feature have added a much broader view of the contexts in which DNA methylation acts. We are beginning to understand functional differences based on observations in these different contexts, from transcriptional start sites, gene bodies, regulatory sequences to repeat elements. The role of DNA methylation is expanding and subtle differences based on context must be integrated into our existing knowledge (Jones, 2012).

Methylated cytocines can prevent certain transcription factors from accessing their target sequences thus preventing key steps necessary for the initiation of transcription. Specific chromatin modifying enzymes can recognize methylated CpGs, and recruit co-repressors resulting in silencing (Han et al., 2011; Jones, 2012; Jones and Liang, 2009). For example, the methyl-CpG-binding protein 2 (MeCP2) recognizes symmetrically methylated CpGs and recruits a co-repressor complex, Sin3a, to cause chromatin compaction (Klose and Bird, 2006).

## E 1.4. DNA Demethylation

Although DNA methylation is considered a relatively stable epigenetic modification there are specific times in early mammalian development when these patterns are very dynamic. Genome wide demethylation of both maternal and paternal genomes takes place in early embryogenesis (Howlett and Reik, 1991; Mayer et al., 2000; Monk et al., 1987). Genome wide demethylation also occurs in primordial germ cells as they migrate (Hajkova et al., 2002). There are also instances of site-specific DNA

demethylation of tissue specific genes resulting in activation upon lineage specification (Frank et al., 1991; Warnecke and Clark, 1999; Xu et al., 2007).

There are two mechanisms of demethylation: passive and active demethylation. Passive demethylation occurs through a lack of maintenance methylation after replication. This replication dependent mechanism results in the gradual loss of methylation after multiple rounds of division (Howlett and Reik, 1991). The exclusion of the maternal DNMT1 from the nucleus, in conjunction with the timing of the demethylation of the maternal genome points to a passive mechanism (Carlson et al., 1992). In contrast, both the paternal and primordial genomes are demethylated in a rapid manner. In the case of the paternal genome global demethylation takes place without undergoing one cell division making passive demethylation highly unlikely (Mayer et al., 2000; Oswald et al., 2000). Loci specific active demethylation has also been reported in somatic cells in the absence of replication (Bruniquel and Schwartz, 2003; Martinowich et al., 2003). This evidence suggests an active mechanism for the removal of 5mC in the absence of cell division.

Although a specific demethylase has not been identified there are a number of proposed multistep mechanisms for demethylation, of which a few have gained traction (reviewed by Kohli and Zhang, 2013; Wu and Zhang, 2010). These mechanisms include: (1) Direct enzymatic removal of the methyl group by methyl-CpG-binding proteins such as MDB2. (2) Removal of 5mC by the base excision repair (BER) pathway mediated by DNA glycosylases such as T DNA glycosylase. (3) Deamination of 5mC to thymine by cytidine deaminases such as activation-induced deaminase (AID), followed by BER of the mismatch. (4) Direct removal of the 5mC by the nucleotide excision repair (NER)

pathway mediated by GADD45A. (5) The conversion of 5mC to 5-hydroxymethyl cytosine (5hmC) by the ten-eleven translocation (TET) family proteins followed by BER.

**E 2. Histone Modifications**

The nucleosome is the core structural unit into which eukaryotic genomes are packaged. The nucleosome is an octamer of four core histone proteins, H2A, H2B, H3 and H4 (two copies of each). Approximately 147bp of DNA are wrapped around each nucleosome. Histones contain flexible N- and C-terminal domains, which are subject to a bevy of dynamic posttranslational modifications from methylation, acetylation, ubiquitination, sumolyation and ADP-ribosylation of lysine (K) residues, to methylation and citrullination of arginine (R) residues, and phosphorylation, acetylation and glycosylation of serine (S), threonine (T), and tyrosine (Y) residues (Campos and Reinberg, 2009; Patel and Wang, 2013; Rothbart and Strahl, 2014). The ability of these posttranslational modifications to impact chromatin structure has added a new layer of complexity to the regulation of key processes from transcription to DNA repair. Great strides have been taken in our understanding of how these modifications affect chromatin, exert specific changes in response to stimuli, and interact with each other as well as other epigenetic modifications, such as DNA methylation. Posttranslational modifications are deposited and erased in a histone and sequence specific manner. Enzymes recognize these modifications, which serve as a platform for further recruitment of proteins with chromatin remodeling activity. The ordered recruitment of chromatin modifying proteins results in dynamic changes and thus affects chromatin dependent processes, such as the accessibility or inaccessibility to a transcriptional start site

(reviewed by Patel and Wang, 2013; Rose and Klose, 2014; Suganuma and Workman, 2011; Swygert and Peterson, 2014).

ChIP-seq has enabled us to determine the genome-wide distribution of a number of these marks. One of the most well characterized histone modifications is lysine methylation (Wozniak and Strahl, 2014). Histone lysine methylation predominantly occurs on histone H3 at K4, K9, K14, K18, K23, K27, K36 and K79 and on histone H4 K20. Histone methylation of lysine results in different effects depending on which residue is modified. Direct links to transcriptional regulation have been observed for a number of these residues including transcriptional activation, H3K4, K36 and K79, and transcriptional repression, H3K9 and H3K27 (Wozniak and Strahl, 2014). Lysine methylation does not change the charge of histones like acetylation does, so, regulation of chromatin state is based on recruitment of additional proteins to facilitate the modulation of chromatin. To add to the complexity of gene regulation, lysine residues can be modified with up to three methyl groups, mono- di- and trimethyl. Moreover, mono-, di-, and trimethylation can be specifically recognized by different modulatory proteins and thus instruct distinct outcomes (Rose and Klose, 2014; Swygert and Peterson, 2014; Wozniak and Strahl, 2014).

Lysine acetylation is highly correlated with chromatin accessibility and transcriptional activation. The addition of an acetyl group can occur on a plethora of lysine residues and has been reported on the tails of each of the core histones H2A, H2B, H3 and H4 (Kimura et al., 2005; Shahbazian and Grunstein, 2007). The acetylation of lysine neutralizes its positive charge. This fact provided a simple explanation for the alteration of chromatin properties. Upon neutralization of the lysine, the electrostatic

interactions between the histone and DNA would become weaker, allowing access to the underlying DNA. Although attractive more evidence has accumulated for another method to regulate chromatin properties, the recognition of acetylated lysines by chromatin remodeling enzymes containing domains that “read” acetyl marks (Rothbart and Strahl, 2014; Swygert and Peterson, 2014).

In the case of both methylation and acetylation a number of protein domains have been discovered that can specifically recognize these modifications. Bromodomain containing proteins can recognize and bind acetylated lysines. Chromodomain containing proteins can recognize and bind methylated lysines. To complicate matters further, some domains are capable of reading multiple modifications as well as unmodified residues e.g. Plant homeodomain (PHD) fingers, which can bind unmodified, methylated and acetylated lysines. To complicate matters even further, a number of individual proteins and large protein complexes contain multiple distinct domains (Rothbart and Strahl, 2014). With the plethora of existing modifications and the multitude of domains that can read them, the potential combinations, which could lead to different functional outcomes, are staggering.

**E 3. Chromatin Remodeling**

The nucleosome has a high affinity for DNA and so presents a substantial barrier for enzymes that require accessibility to the DNA. Higher order chromatin compaction can also serve as a further impediment for such processes. Several mechanisms exist to alter the position, stability and compaction of chromatin. Posttranslational modifications as previously discussed are one such way. Another method is the incorporation of histone

variants, such as H2AZ and H3.3, which can alter stability and provide additional residues for modification. A third mechanism is the use of ATP-dependent chromatin remodeling enzymes. Using the energy garnered from ATP hydrolysis, chromatin remodelers can position, evict or exchange nucleosomes (Swygert and Peterson, 2014).

Much has been learned about the interplay between posttranslational modifications and the action of ATP-dependent chromatin remodelers. Remodelers fall into four main families which all contain the characteristic ATPase domain: SWI/SNF, INO80, ISWI and CHD. Collectively these remodelers can cause transcriptional activation and repression. Many of these enzymes can also directly modulate histone modifications and are often part of larger protein complexes. SWI/SNF contains a bromodomain that recognizes acetylated lysines residues. It has been shown in both yeast and humans that acetylation is necessary for the recruitment of SWI/SNF (Swygert and Peterson, 2014). CHD3/4 (Mi-2 β) contains tandem chromodomains as well as multiple PHD fingers that interact with H3K9 of K36. CHD3/4 is a core component of the nucleosome remodeling and deacetylation (NuRD) complex, which also contains proteins that remove acetyl groups, HDACs. Also belonging to this complex are the methyl-CpG-binding proteins MB2 and 3. The NuRD complex exerts a multifaceted approach, it can recognize acetylated lysines and deacetylate them, it can slide and reposition nucleosomes, and it can bind methylated CpG dinucleotides, resulting in the rapid formation of heterochromatin and the silencing of transcription (Allen et al., 2013; Swygert and Peterson, 2014). Complex mechanisms have evolved to interpret multiple epigenetic inputs. How these different signals are coordinated to result in a specific output continues to be a focus of intense research.

**F. Tissue Specificity**

The development of multicellular organisms requires the orchestration of a multitude of molecular processes. During differentiation cells progressively lose their developmental potential. This continued restriction in lineage potential results in the specification of cell types that carry out distinct cellular functions. These developmental transitions depend on the correct timing and appropriate expression of a cohort of genes controlled by both genetic and epigenetic mechanisms previously mentioned. Ultimately, mature cell types exhibit a characteristic gene expression profile. Within any give cell the expression profile can be divided into two main categories; genes required for basic cellular functions, and genes required for specialized cellular functions. The former are generally expressed in a wide variety of cell types, if not all, while the latter are expressed in a limited number of cell types or tissues. Genes that are restricted in such a manner are described as tissue specific. Understanding tissue specific gene expression patterns is critical for our understanding of the molecular mechanisms underlying cellular states, development, and disease (Meister et al., 2010; Ong and Corces, 2011; Song et al., 2013).

## REFERENCES


Allen, H.F., Wade, P.A., and Kutateladze, T.G. (2013). The NuRD architecture. Cellular and molecular life sciences : CMLS *70*, 3513-3524.

Anokye-Danso, F., Trivedi, C.M., Juhr, D., Gupta, M., Cui, Z., Tian, Y., Zhang, Y., Yang, W., Gruber, P.J., Epstein, J.A., *et al.* (2011). Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency. Cell stem cell *8*, 376-388.

Antequera, F., and Bird, A. (1993). Number of CpG islands and genes in human and mouse. Proceedings of the National Academy of Sciences of the United States of America *90*, 11995-11999.

Apostolou, E., Ferrari, F., Walsh, R.M., Bar-Nur, O., Stadtfeld, M., Cheloufi, S., Stuart, H.T., Polo, J.M., Ohsumi, T.K., Borowsky, M.L., *et al.* (2013). Genome-wide chromatin interactions of the Nanog locus in pluripotency, differentiation, and reprogramming. Cell stem cell *12*, 699-712.

Avilion, A.A., Nicolis, S.K., Pevny, L.H., Perez, L., Vivian, N., and Lovell-Badge, R. (2003). Multipotent cell lineages in early mouse development depend on SOX2 function. Genes & development *17*, 126-140.

Bainbridge, M.N., Warren, R.L., Hirst, M., Romanuik, T., Zeng, T., Go, A., Delaney, A., Griffith, M., Hickenbotham, M., Magrini, V., *et al.* (2006). Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC genomics *7*, 246.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., *et al.* (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315-326.

Bestor, T., Laudano, A., Mattaliano, R., and Ingram, V. (1988). Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. Journal of molecular biology *203*, 971-983.

Bird, A., Taggart, M., Frommer, M., Miller, O.J., and Macleod, D. (1985). A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell *40*, 91-99.

Blackledge, N.P., Thomson, J.P., and Skene, P.J. (2013). CpG island chromatin is shaped by recruitment of ZF-CxxC proteins. Cold Spring Harbor perspectives in biology *5*, a018648.

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.

Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K., *et al.* (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature *441*, 349-353.

Bruniquel, D., and Schwartz, R.H. (2003). Selective, stable demethylation of the interleukin-2 gene enhances transcription by an active process. Nature immunology *4*, 235-240.

Campos, E.I., and Reinberg, D. (2009). Histones: annotating chromatin. Annual review of genetics *43*, 559-599.

Carlson, L.L., Page, A.W., and Bestor, T.H. (1992). Properties and localization of DNA methyltransferase in preimplantation mouse embryos: implications for genomic imprinting. Genes & development *6*, 2536-2541.

Chambers, I., Silva, J., Colby, D., Nichols, J., Nijmeijer, B., Robertson, M., Vrana, J., Jones, K., Grotewold, L., and Smith, A. (2007). Nanog safeguards pluripotency and mediates germline development. Nature *450*, 1230-1234.

Chambers, I., and Tomlinson, S.R. (2009). The transcriptional foundation of pluripotency. Development *136*, 2311-2322.

Charron, J., Malynn, B.A., Fisher, P., Stewart, V., Jeannotte, L., Goff, S.P., Robertson, E.J., and Alt, F.W. (1992). Embryonic lethality in mice homozygous for a targeted disruption of the N-myc gene. Genes & development *6*, 2248-2257.

Chen, X., Vega, V.B., and Ng, H.H. (2008a). Transcriptional regulatory networks in embryonic stem cells. Cold Spring Harbor symposia on quantitative biology *73*, 203-209.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J., *et al.* (2008b). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell *133*, 1106-1117.

Chou, B.K., and Cheng, L. (2013). And then there were none: no need for pluripotency factors to induce reprogramming. Cell stem cell *13*, 261-262.

Cooper, D.N., Taggart, M.H., and Bird, A.P. (1983). Unmethylated domains in vertebrate DNA. Nucleic acids research *11*, 647-658.

Cowan, C.A., Atienza, J., Melton, D.A., and Eggan, K. (2005). Nuclear reprogramming of somatic cells after fusion with human embryonic stem cells. Science *309*, 1369-1373.

David B. Allison, X.C., Grier P. Page, Mahyar Sabripour (2006). Microarray data analysis- from disarray to consolidation and consensus. Nature Reviews Genetics *7*.

Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes & development *25*, 1010-1022.

Dejosez, M., and Zwaka, T.P. (2012). Pluripotency and nuclear reprogramming. Annual review of biochemistry *81*, 737-765.

Di Croce, L., and Helin, K. (2013). Transcriptional regulation by Polycomb group proteins. Nature structural & molecular biology *20*, 1147-1155.

Endoh, M., Endo, T.A., Endoh, T., Fujimura, Y., Ohara, O., Toyoda, T., Otte, A.P., Okano, M., Brockdorff, N., Vidal, M., *et al.* (2008). Polycomb group proteins Ring1A/B are functionally linked to the core transcriptional regulatory circuitry to maintain ES cell identity. Development *135*, 1513-1524.

Frank, D., Keshet, I., Shani, M., Levine, A., Razin, A., and Cedar, H. (1991). Demethylation of CpG islands in embryonic cells. Nature *351*, 239-241.

Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. Journal of molecular biology *196*, 261-282.

Goll, M.G., and Bestor, T.H. (2005). Eukaryotic cytosine methyltransferases. Annual review of biochemistry *74*, 481-514.

Hajkova, P., Erhardt, S., Lane, N., Haaf, T., El-Maarri, O., Reik, W., Walter, J., and Surani, M.A. (2002). Epigenetic reprogramming in mouse primordial germ cells. Mechanisms of development *117*, 15-23.

Hammachi, F., Morrison, G.M., Sharov, A.A., Livigni, A., Narayan, S., Papapetrou, E.P., O'Malley, J., Kaji, K., Ko, M.S., Ptashne, M., *et al.* (2012). Transcriptional activation by Oct4 is sufficient for the maintenance and induction of pluripotency. Cell reports *1*, 99-109.

Han, H., Cortez, C.C., Yang, X., Nichols, P.W., Jones, P.A., and Liang, G. (2011). DNA methylation directly silences genes with non-CpG island promoters and establishes a nucleosome occupied promoter. Human molecular genetics *20*, 4299-4310.

Hata, K., Okano, M., Lei, H., and Li, E. (2002). Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. Development *129*, 1983-1993.

Hitoshi Niwa, J.M., & Austin G. Smith (2000). Quantitative expression of Oct4/3 defines differentiation, dedifferentiation or self-renewal of ES cells. Nature Genetics *24*, 372-376.

Howlett, S.K., and Reik, W. (1991). Methylation levels of maternal and paternal genomes during preimplantation development. Development *113*, 119-127.

Hussein, S.M., and Nagy, A.A. (2012). Progress made in the reprogramming field: new factors, new strategies and a new outlook. Current opinion in genetics & development *22*, 435-443.

Ian Chambers, D.C., Morag Robertson, J.N., Sonia Lee,, and Susan Tweedie, a.A.S. (2003). Functional Expression Cloning of Nanog, a Pluripotency Sustaining Factor in Embryonic Stem Cells. Cell *113*, 643-655.

Illingworth, R.S., and Bird, A.P. (2009). CpG islands--'a rough guide'. FEBS letters *583*, 1713-1720.

Illingworth, R.S., Gruenewald-Schneider, U., Webb, S., Kerr, A.R., James, K.D., Turner, D.J., Smith, C., Harrison, D.J., Andrews, R., and Bird, A.P. (2010). Orphan CpG islands identify numerous conserved promoters in the mammalian genome. PLoS genetics *6*, e1001134.

Jaenisch, R., Hochedlinger, K., Blelloch, R., Yamada, Y., Baldwin, K., and Eggan, K. (2004). Nuclear cloning, epigenetic reprogramming, and cellular differentiation. Cold Spring Harbor symposia on quantitative biology *69*, 19-27.

Jennifer Nichols, B.Z., Konstantinos Anastassiadis, Hitoshi Niwa, Daniela Klewe-Nebenius, I.C., Hans Scholer, and Smith, a.A. (1998). Formation of Pluripotent Stem Cells in the Mammalian Embryo Depends on the POU Transcription Factor Oct4 1998. Cell *95*, 379-391.

Jerabek, S., Merino, F., Scholer, H.R., and Cojocaru, V. (2014). OCT4: dynamic DNA binding pioneers stem cell pluripotency. Biochimica et biophysica acta *1839*, 138-154.

John H Malone, B.O. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biology.

Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature reviews Genetics *13*, 484-492.

Jones, P.A., and Liang, G. (2009). Rethinking how DNA methylation patterns are maintained. Nature reviews Genetics *10*, 805-811.

Kaoru Mitsui, Y.T., Hiroaki Itoh,, Kohichi Segawa, M.M., Kazutoshi Takahashi, M.M., and Mitsuyo Maeda, a.S.Y. (2003). The Homeoprotein Nanog Is Required for Maintenance of Pluripotency in Mouse Epiblast and ES Cells. Cell *113*, 631-642.

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S.H. (2008). An extended transcriptional network for pluripotency of embryonic stem cells. Cell *132*, 1049-1061.

Kim, J., Woo, A.J., Chu, J., Snow, J.W., Fujiwara, Y., Kim, C.G., Cantor, A.B., and Orkin, S.H. (2010). A Myc network accounts for similarities between embryonic stem and cancer cell transcription programs. Cell *143*, 313-324.

Kimura, A., Matsubara, K., and Horikoshi, M. (2005). A Decade of Histone Acetylation: Marking Eukaryotic Chromosomes with Specific Codes. Journal of Biochemistry *138*, 647-662.

Klose, R.J., and Bird, A.P. (2006). Genomic DNA methylation: the mark and its mediators. Trends in biochemical sciences *31*, 89-97.

Kohli, R.M., and Zhang, Y. (2013). TET enzymes, TDG and the dynamics of DNA demethylation. Nature *502*, 472-479.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology *10*, R25.

Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. (1992). CpG islands as gene markers in the human genome. Genomics *13*, 1095-1107.

Lei, H., Oh, S.P., Okano, M., Juttermann, R., Goss, K.A., Jaenisch, R., and Li, E. (1996). De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. Development *122*, 3195-3205.

Leonhardt, H., Page, A.W., Weier, H.U., and Bestor, T.H. (1992). A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. Cell *71*, 865-873.

Lewis, E.B. (1978). A gene complex controlling segmentation in Drosophila. Nature *276*, 565-570.

Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell *69*, 915-926.

Loh, Y.-H., Wu, Q., Chew, J.-L., Vega, V.B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., *et al.* (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat Genet *38*, 431-440.

Lowry, W.E. (2012). Does transcription factor induced pluripotency accurately mimic embryo derived pluripotency? Current opinion in genetics & development *22*, 429-434.

M. J. Evans, M.H.K. (1981). Establishment in culture of pluripotential cells from mouse embryos. Nature *292*.

Ma, T., Xie, M., Laurent, T., and Ding, S. (2013). Progress in the reprogramming of somatic cells. Circulation research *112*, 562-574.

Malynn, B.A., de Alboran, I.M., O'Hagan, R.C., Bronson, R., Davidson, L., DePinho, R.A., and Alt, F.W. (2000). N-myc can functionally replace c-myc in murine development, cellular growth, and differentiation. Genes & development *14*, 1390-1399.

Mark D. Adams, J.M.K., Jeannine D. Gocayne, J. Craig Venter (1991). Complementary DNA Sequencing- Expressed Sequence Tags and Human Genome Project. Science *252*, 1651-1656.

Marks, H., Kalkan, T., Menafra, R., Denissov, S., Jones, K., Hofemeister, H., Nichols, J., Kranz, A., Stewart, A.F., Smith, A., *et al.* (2012). The transcriptional and epigenomic foundations of ground state pluripotency. Cell *149*, 590-604.

Marks, H., and Stunnenberg, H.G. (2014). Transcription regulation and chromatin structure in the pluripotent ground state. Biochimica et biophysica acta *1839*, 129-137.

Martin, G.R. (1981). Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. PNAS.

Martinowich, K., Hattori, D., Wu, H., Fouse, S., He, F., Hu, Y., Fan, G., and Sun, Y.E. (2003). DNA methylation-related chromatin remodeling in activity-dependent BDNF gene regulation. Science *302*, 890-893.

Masui, S., Nakatake, Y., Toyooka, Y., Shimosato, D., Yagi, R., Takahashi, K., Okochi, H., Okuda, A., Matoba, R., Sharov, A.A., *et al.* (2007). Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. Nature cell biology *9*, 625-635.

Maunakea, A.K., Nagarajan, R.P., Bilenky, M., Ballinger, T.J., D'Souza, C., Fouse, S.D., Johnson, B.E., Hong, C., Nielsen, C., Zhao, Y., *et al.* (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature *466*, 253-257.

Mayer, W., Niveleau, A., Walter, J., Fundele, R., and Haaf, T. (2000). Demethylation of the zygotic paternal genome. Nature *403*, 501-502.

Meister, P., Towbin, B.D., Pike, B.L., Ponti, A., and Gasser, S.M. (2010). The spatial dynamics of tissue-specific promoters during C. elegans development. Genes & development *24*, 766-782.

Meshorer, E., and Misteli, T. (2006). Chromatin in pluripotent embryonic stem cells and differentiation. Nature reviews Molecular cell biology *7*, 540-546.

Meshorer, E., Yellajoshula, D., George, E., Scambler, P.J., Brown, D.T., and Misteli, T. (2006). Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. Developmental cell *10*, 105-116.

Monk, M., Boubelik, M., and Lehnert, S. (1987). Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. Development *99*, 371-382.

Montserrat, N., Nivet, E., Sancho-Martinez, I., Hishida, T., Kumar, S., Miquel, L., Cortina, C., Hishida, Y., Xia, Y., Esteban, C.R., *et al.* (2013). Reprogramming of human fibroblasts to pluripotency with lineage specifiers. Cell stem cell *13*, 341-350.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods *5*, 621-628.

Ng, H.-H., and Surani, M.A. (2011). The transcriptional and signalling networks of pluripotency. Nature cell biology *13*, 490-496.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247-257.

Ong, C.T., and Corces, V.G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. Nature reviews Genetics *12*, 283-293.

Orkin, S.H., Wang, J., Kim, J., Chu, J., Rao, S., Theunissen, T.W., Shen, X., and Levasseur, D.N. (2008). The transcriptional network controlling pluripotency in ES cells. Cold Spring Harbor symposia on quantitative biology *73*, 195-202.

Oswald, J., Engemann, S., Lane, N., Mayer, W., Olek, A., Fundele, R., Dean, W., Reik, W., and Walter, J. (2000). Active demethylation of the paternal genome in the mouse zygote. Current biology : CB *10*, 475-478.

Ozsolak, F., and Milos, P.M. (2011). RNA sequencing: advances, challenges and opportunities. Nature reviews Genetics *12*, 87-98.

Papp, B., and Plath, K. (2013). Epigenetics of reprogramming to induced pluripotency. Cell *152*, 1324-1343.

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., and Soldatov, A. (2009). Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic acids research *37*, e123.

Patel, D.J., and Wang, Z. (2013). Readout of epigenetic modifications. Annual review of biochemistry *82*, 81-118.

Plath, K., and Lowry, W.E. (2011). Progress in understanding reprogramming to the induced pluripotent state. Nature reviews Genetics *12*, 253-265.

Reizis, B., and Leder, P. (1999). Expression of the mouse pre-T cell receptor alpha gene is controlled by an upstream region containing a transcriptional enhancer. The Journal of experimental medicine *189*, 1669-1678.

Rose, N.R., and Klose, R.J. (2014). Understanding the relationship between DNA methylation and histone lysine methylation. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms.

Rothbart, S.B., and Strahl, B.D. (2014). Interpreting the language of histone and DNA modifications. Biochimica et biophysica acta.

Rothenberg, E.V. (2014). The chromatin landscape and transcription factors in T cell programming. Trends in immunology *35*, 195-204.

Schena M., S.D., Davis R. W., Brown P. O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science *270*.

Shahbazian, M.D., and Grunstein, M. (2007). Functions of site-specific histone acetylation and deacetylation. Annual review of biochemistry *76*, 75-100.

Shu, J., Wu, C., Wu, Y., Li, Z., Shao, S., Zhao, W., Tang, X., Yang, H., Shen, L., Zuo, X., *et al.* (2013). Induction of pluripotency in mouse somatic cells with lineage specifiers. Cell *153*, 963-975.

Smith, K.N., Lim, J.-M., Wells, L., and Dalton, S. (2011). Myc orchestrates a regulatory network required for the establishment and maintenance of pluripotency. Cell Cycle *10*, 592-597.

Smith, K.N., Singh, A.M., and Dalton, S. (2010). Myc represses primitive endoderm differentiation in pluripotent stem cells. Cell stem cell *7*, 343-354.

Smith, Z.D., and Meissner, A. (2013). DNA methylation: roles in mammalian development. Nature reviews Genetics *14*, 204-220.

Song, Y., Ahn, J., Suh, Y., Davis, M.E., and Lee, K. (2013). Identification of novel tissue-specific genes by analysis of microarray databases: a human and mouse model. PloS one *8*, e64483.

Stanton, B.R., Perkins, A.S., Tessarollo, L., Sassoon, D.A., and Parada, L.F. (1992). Loss of N-myc function results in embryonic lethality and failure of the epithelial component of the embryo to develop. Genes & development *6*, 2235-2247.

Suganuma, T., and Workman, J.L. (2011). Signals and combinatorial functions of histone modifications. Annual review of biochemistry *80*, 473-499.

Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., *et al.* (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science *321*, 956-960.

Swygert, S.G., and Peterson, C.L. (2014). Chromatin dynamics: Interplay between remodeling enzymes and histone modifications. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms.

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.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., *et al.* (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nature methods *6*, 377-382.

Thomson, J.A. (1998). Embryonic Stem Cell Lines Derived from Human Blastocysts. Science *282*, 1145-1147.

Thomson, M., Liu, S.J., Zou, L.N., Smith, Z., Meissner, A., and Ramanathan, S. (2011). Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. Cell *145*, 875-889.

van den Berg, D.L., Snoek, T., Mullin, N.P., Yates, A., Bezstarosti, K., Demmers, J., Chambers, I., and Poot, R.A. (2010). An Oct4-centered protein interaction network in embryonic stem cells. Cell stem cell *6*, 369-381.

Varlakhanova, N.V., Cotterman, R.F., deVries, W.N., Morgan, J., Donahue, L.R., Murray, S., Knowles, B.B., and Knoepfler, P.S. (2010). myc maintains embryonic stem cell pluripotency and self-renewal. Differentiation; research in biological diversity *80*, 9-19.

Victor E. Velculescu, L.Z., Bert Volgelstein, Kenneth W. Kinzler (1995). Serial Analysis of Gene Expression. Science *270*, 484-487.

Vierbuchen, T., and Wernig, M. (2012). Molecular roadblocks for cellular reprogramming. Molecular cell *47*, 827-838.

Voigt, P., Tee, W.-W., and Reinberg, D. (2013). A double take on bivalent promoters. Genes & development *27*, 1318-1338.

Wang, J., Rao, S., Chu, J., Shen, X., Levasseur, D.N., Theunissen, T.W., and Orkin, S.H. (2006). A protein interaction network for pluripotency of embryonic stem cells. Nature *444*, 364-368.

Warnecke, P.M., and Clark, S.J. (1999). DNA methylation profile of the mouse skeletal alpha-actin promoter during development and differentiation. Molecular and cellular biology *19*, 164-172.

Wernig, M., Meissner, A., Cassady, J.P., and Jaenisch, R. (2008). c-Myc is dispensable for direct reprogramming of mouse fibroblasts. Cell stem cell *2*, 10-12.

Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J., and Bahler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature *453*, 1239-1243.

Wozniak, G.G., and Strahl, B.D. (2014). Hitting the 'mark': Interpreting lysine methylation in the context of active transcription. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms.

Wu, H., and Zhang, Y. (2014). Reversing DNA Methylation: Mechanisms, Genomics, and Biological Functions. Cell *156*, 45-68.

Wu, S.C., and Zhang, Y. (2010). Active DNA demethylation: many roads lead to Rome. Nature reviews Molecular cell biology *11*, 607-620.

Xu, J., Pope, S.D., Jazirehi, A.R., Attema, J.L., Papathanasiou, P., Watts, J.A., Zaret, K.S., Weissman, I.L., and Smale, S.T. (2007). Pioneer factor interactions and unmethylated CpG dinucleotides mark silent tissue-specific enhancers in embryonic stem cells. Proceedings of the National Academy of Sciences of the United States of America *104*, 12377-12382.

Ying, Q.L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. Nature *453*, 519-523.

Yu, J., Vodyanik, M.A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J.L., Tian, S., Nie, J., Jonsdottir, G.A., Ruotti, V., Stewart, R., *et al.* (2007). Induced pluripotent stem cell lines derived from human somatic cells. Science *318*, 1917-1920.

Zhu, J., He, F., Hu, S., and Yu, J. (2008). On the nature of human housekeeping genes. Trends in genetics : TIG *24*, 481-484.

# Chapter 2

# The Tissue Specific Enhancer of pTCRα Persists in an Unmethylated State in a Chromatinized Context in Embryonic Stem Cells

## ABSTRACT

Tissue specific gene activation is the result of a number of highly coordinated events that culminate in the initiation of transcription upon lineage commitment. The chromatin environment and availability of transcription factors dictate this process. There is clear evidence for the role of tissue specific enhancers in these highly orchestrated events. Often distal regulatory sequences provide a platform for the recruitment of sequence specific activators to the promoters of their genes. It is often these distal elements that first engage with sequence specific factors and provide a permissible chromatin context for activation. It has come to light the some tissue specific enhancers are marked at the embryonic stem cell stage. In this study we interrogate the finding that the Ptcra enhancer is marked in ES cells and is subject to regulation by both positive and negative mechanisms. We employ the use of a 200kb bacterial artificial chromosome encompassing the pTa locus in order to represent the endogenous chromatin context. We monitor the methylation status of the enhancer by bisulfite sequencing, giving us single nucleotide resolution. Here we report that the pTa enhancer mark consistently reappears in the context of a BAC after premethylation and stable integration into ES cells. However mutations that caused the disappearance or spreading of the unmethylated window in a plasmid reporter are not recapitulated in the BAC. The observation that the enhancer mark consistently reappears suggests that ES cells posses the necessary factors to gain access to and cause demethylation either by an active or passive mechanism. This marking may represent an additional property of pluripotent ES cells, the ability to potentiate tissue specific genes for the appropriate and timely activation upon lineage specification.

## INTRODUCTION

The recent advances in technology have lead to waves of genome-wide studies to identify chromatin signatures and patterns associated with different gene classes, to further our understanding of the regulatory mechanisms that contribute to gene expression. Key findings arose in embryonic stem cells, in which, developmental regulators and lineage determinants were simultaneously marked, at their promoters, by positive and negative histone modifications (Bernstein et al., 2006; Sachs et al., 2013; Vastenhouw and Schier, 2012; Voigt et al., 2013a). Although important much focus and effort continues to be placed genome-wide characterizations of gene classes. While critical, there is still a need for detailed mechanistic studies in specific gene classes. Tissue specific genes are thought to be primarily regulated by the availability of lineage specifying transcription factors, which are often expressed in a temporal manner (Heinz and Glass, 2012). Tissue specific genes are not typically marked by histone modifications at their promoters in ES cells, long before they will be expressed. However, recent findings suggest that tissue specific enhancers are marked as early as the ES cell stage and these marks may be functionally relevant (Liber et al., 2010; Szutorisz et al., 2005; Xu et al., 2009). In light of this new evidence, further characterization of tissue specific enhancer marks in ES cells is of high priority.

A number of tissue specific genes have been used as a model to study gene expression programs during development. The thymocyte specific pre T-cell receptor alpha (*Ptcra*) is one of them. T lymphocyte development involves four key processes, which mirror that of many development transitions: (1) The activation of lineage specific genes. (2) The silencing of non T lymphocyte lineage genes. (3) The use of prior and

newly expressed regulatory proteins. (4) Changes in chromatin structure at cis regulatory sequences for the binding of transcription factors (Rothenberg, 2014). Ptcra encodes for the pre-T cell receptor α chain (pTα) whose expression is restricted to immature thymocytes (Reizis and Leder, 1999; Xu et al., 2007).

The paradigm for our understanding of the regulation of tissue specific genes during early development comes from DNA methylation studies. After the widespread demethylation in early embryogenesis, there is a wave of *de novo* methylation, which persists through somatic cell divisions (Hajkova et al., 2002; Monk et al., 1987; Wu and Zhang, 2010). Early evidence showed many tissue-specific genes become heavily methylated during this *de novo* methylation (Kafri et al., 1992). The resulting model was that tissue-specific genes are methylated during this time and assembled into silent chromatin until the appropriate developmental cues are received (Jones and Takai, 2001; Kafri et al., 1992). Further evidence that non-CpG island tissue specific promoters remain methylated in non-expressing lineages adds to this foundation (Miranda and Jones, 2007).

Early and emerging evidence indicates critical roles for enhancer elements regulating tissue specific expression (Ong and Corces, 2011). This has been demonstrated with pTα whose enhancer, located 4kb upstream, is necessary for the stage specific expression of pTα in transgenic mice (Reizis and Leder, 1999, 2001, 2002). The core 149bp enhancer, which is DNase hypersensitive in a thymocyte specific manner, possesses highly conserved transcription factor binding sites (Reizis and Leder, 1999, 2001; Takeuchi et al., 2001). These transcription factors have been shown to be essential for thymocyte development; c-Myb, E2A, HEB and CSL (Reizis and Leder, 1999; Takeuchi et al., 2001; Tremblay et al., 2003).

Evidence has also emerged that tissue specific genes are epigenetically marked prior to activation, in some cases as early as the embryonic stem cell stage (Dillon, 2012; Liber et al., 2010; Smale, 2010; Szutorisz et al., 2005; Xu et al., 2007). This priming, so to speak, appears to occur through site-specific transcription factor binding and in some cases histone acetylation (Szutorisz et al., 2005; Xu et al., 2007; Xu et al., 2009). The pTα core enhancer was shown to be bound by key transcription factors and marked by histone acetylation. Furthermore, this enhancer mark was reestablished in ES cell clones after a premethylated reporter construct was stably integrated (Xu et al., 2009).

A classical example of this type of regulation comes from the liver specific albumin (*Alb1*) gene, whose enhancer, 10kb upstream, contributes to the liver restricted expression of *Alb1* (Liu et al., 1991; Pinkert et al., 1987). In a similar fashion to the pTα enhancer, the albumin enhancer contains binding sites for key transcription factors involved in liver development such as GATA-4, C/EBP, FoxA1 and NF1 (Liu et al., 1991; McPherson et al., 1993). Careful analysis of the enhancer showed sequential binding and suggested priming of the enhancer in cells prior to which activation of albumin occurs (Bossard and Zaret, 1998; Gualdi et al., 1996). Detailed studies performed by Xu et. al., showed the presence of an unmethylated CpG dinucleotide within the albumin enhancer in ES cells, indicating the albumin enhancer may be marked even earlier than previously reported. This CpG dinucleotide lies directly within the FoxA1 binding site. Further categorization determined FoxD3, a close family member, to be responsible for the unmethylated state at the CpG (Xu et al., 2007; Xu et al., 2009). These data lead to the hypothesis that initial binding events by 'pioneer' factors

potentiate the locus in such a manner that it becomes permissible to activation upon lineage specification. This can be thought of as epigenetic priming.

Xu et. al., also showed the regulation of the pTα unmethylated enhancer mark by both positive and negative factors. Deletion of specific transcription factor binding sites altered the methylation state. For example, the deletion of Sp1 and E-box sites caused the reporter plasmid to be refractory to demethylation in ES cells after premethylation (Xu et al., 2009). Although intriguing these studies were performed in the context of a reporter plasmid. In order to fully understand the mechanisms in play, the roles of these transcription factor binding sites must be addressed in an endogenous context. This will provide a fuller understanding of the establishment of enhancer marks in ES cells and contribute to the knowledge of the properties of ES cells and their mechanisms to achieve appropriate epigenetic profiles at lineage specific genes.

In this study we characterize the contributions of the transcription factor binding sites in the establishment of an unmethylated state in embryonic stem cells using bacterial artificial chromosomes (BAC) and bisulfite sequencing. At 200 kilobases, the BAC has more than 50kb flanking both the 5' and 3' prime end of the pTα locus. Thus, we likely include all cis-regulatory elements necessary for the appropriate chromatin context of pTα. Consequently we provide a near ideal context, aside from directly modifying the endogenous locus, to perform these studies.

## RESULTS

### *BAC Modification*

In order to study the methylation status of the pTa enhancer in an endogenous chromatin context we took the approach of using a bacterial artificial chromosome (BAC). We used a 203,589bp BAC (RP23-288F21) containing the *Ptcra* mouse locus. The pTa locus was flanked by approximately 60kb of upstream genomic sequences from transcriptional start site and approximately 140kb of downstream genomic sequences from the transcriptional termination site (Figure 2-1). A BAC this large we hypothesize would likely contain all the necessary sequences for the formation of an endogenous chromatin environment at the pTa locus. We also hypothesize that the flanking sequences of the BAC will provide insulation of the pTa locus from the flanking endogenous chromatin at the integration site.

The Ptcra BAC was modified using homologus recombination in a two-step selection/counterselection process i.e. Recombineering (recombination-mediated genetic engineering). The system revolves around λ Red-encoded genes exo, bet and gam, whose protein products execute the recombination reactions. These lambda genes are present in a specialized E. coli strain SW102. In order for the recombination reaction to proceed, exo, a 5'-3' exonuclease, creates single stranded 3' overhangs in the double stranded linear DNA-targeting cassette introduced into the bacteria. Bet binds the 3' overhangs and anneals with the complementary sequences within the BAC to complete the homologous recombination. While the reaction proceeds, gam, a protein that inhibits the E. coli exonuclease, RecBCD, protects the linear targeting cassette from degradation. These proteins are expressed from defective lambda prophage that is stably integrated

into the E. coli strain. All three proteins are controlled by a temperature sensitive promoter, which is induced at 42$^{o}$C, resulting in rapid expression and the accumulation of high levels of each protein necessary for recombination (Warming et al., 2005).

In order to introduce the mutations the BAC was electroporated into SW102s, followed by a two-step process of recombination. The first selection step takes place as a result of a non-functional galactokinase (galK) gene in the SW102 strain. In this first round of recombination, a targeting cassette containing galK was generated with 50bp of homology to the Ptcra enhancer region. After the induction of the expression of the lambda proteins at 42$^{o}$C, the galK + Ptcra homology arms cassette was electroporated into the SW102 cells. The successful integration and expression of galK permits the growth of bacteria on minimal media plates containing galactose. Colonies were selected after plating and incubation at 32$^{o}$C for 3-8 days.

In the second stage of recombination, a targeting cassette was generated using overlapping HPLC purified oligos containing each mutation, followed by amplification to include up to 500bp in homology surrounding the enhancer. The core pTa enhancer is only 149bp, which allowed us to generate mutations directly from synthesized oligos. The cassette was then electroporated into the SW102 strain with the galK containing BAC, after induction of the recombination proteins at 42$^{o}$C. The targeting construct recombines with the BAC forcing out the galK cassette allowing for counterselection on a galK negative media. Correct clones were selected by PCR screening for the integrated cassette followed by BAC fingerprinting with selected restriction enzymes (EcorI 55 cuts). To allow for selection and stable integration in ES cells, the modified BAC was retrofitted to contain a Neomycin selection cassette (Wang et al., 2001).

Due to sequence similarity we devised a method to distinguish between the endogenous pTa locus and the stably integrated BAC pTa locus. We inserted an 8bp PmeI restriction site, GTTTAAAC, 110bp upstream of the core enhancer. By designing primers upstream of the tag and downstream of the enhancer we were able to assess both the endogenous and BAC modified pTa locus simultaneously.

***DNA Methylation Analysis***

We performed bisulfite-sequencing analysis to determine whether or not the Ptcra enhancer mark is reestablished in an endogenous chromatin context in ES cells. BACs were premethylated *in vitro* with the SssI CpG methylase and were consistently methylated at the enhancer locus (Figure 2-2). We confirmed the efficiency of methylation upstream and downstream of the enhancer, as well as the enhancer and promoter. In all cases the BAC was heavily methylated (Figure 2-3). After *in vitro* methylation the BAC was linearized with the unique restriction enzyme PI-SceI and the integrity checked by pulsed-field gel electrophoresis. We confirmed that the inclusion of the tag upstream of the enhancer did not alter the methylation state of the pTa locus. Upon premethylation and integration of the pTa BAC in to mouse ES cells, the enhancer mark faithfully reappeared in multiple clones (Figure 2-4). Moreover, the modified enhancer mirrored the endogenous enhancer, again showing clear demethylation particularly at the core enhancer, -4080bp to -3965bp, in which the CpG dinucleotides directly overlap with key transcription factor binding sites (Figure 2-5, 2-6). Interestingly, the demethylation is broader in the endogenous context compared to the reporter plasmid

context, where the only consistently unmethylated CpG coincides with the Myb site located at -4,080bp.

### *Individual Mutations in the Ebox2 or Sp1 site Do Not Cause Resistance to Demethylation*

In prior studies two key mutations resulted in the pTa enhancer remaining methylated after stable transfection in ES cells. Individual mutations in the upstream Sp1 site and the E-box 2 site prevented the reappearance of the enhancer mark (Xu et al., 2009). We recapitulated the exact mutations in the context of the BAC and assessed the methylation status of the enhancer in the endogenous chromatin context (Figure 2-7). After stable integration of the BAC into ES cells, the enhancer mark reappeared consistently in multiple clones with a methylation status similar to the WT pTa Tag BAC (compare to the premethylated controls) (Figure 2-2, 2-8). It does appear in some instances that the Sp1 mutation causes the reduction in the unmethylated window (Figure 2-8). However, the breadth of the unmethylated window appears to be variable even in the WT Tag BAC limiting any interpretations (Figure 2-8). In the chromatin context the CpG dinucleotide within the Myb site reliably remains unmethylated, consistent with the studies in the reporter plasmid (Figure 2-8).

Another key mutation that prevented the premethylated enhancer-promoter reporter plasmid from becoming demethylated was a deletion encompassing the Myb, Ebox2 and upstream Sp1 site. We generated the same deletion and stably transfected the mutant BAC into ES cells. We found that the deletion did not recapitulate the results seen

in the plasmid context. The enhancer deletion was still able to undergo demethylation resulting in an unmethylated window spanning the core enhancer (Figure 2-8C). The control was efficiently methylated and upon transfection the deletion mutant and pTa TAG BAC show similar patterns.

### *Individual Mutations in the Myb, Ebox4 or CSL site Do Not Cause Widespread Demethylation*

Again in previous studies performed by Xu et. al., the pTa enhancer mark was shown to be regulated in both positive and negative manner depending upon the mutation. In the context of the enhancer reporter plasmid, mutations in the Myb, Ebox4 or CSL site caused the spreading of the unmethylated window. This indicated the possibility that transcription factors bound to those sequences restricted the spread of methylation into the surrounding regions (Xu et al., 2009). To determine if any negative regulatory mechanism exists in the endogenous context we constructed the same mutations in the pTa BAC (Figure 2-7).

All of the mutations mirror that of the wild type enhancer with the core enhancer consistently showing the lowest levels of methylation particularly at the -4,080bp Myb site (Figure 2-9A). In order to assess the spreading of the unmethylated window we analyzed both upstream and downstream regions of the enhancer. Mutations in the CSL and E-box4 site closely match the methylation status of the pTa TAG BAC. In one of the clones of the Myb mutation there is a trend towards lower levels of methylation in the upstream but not downstream region, however the spreading of the unmethylated window

into this region can be seen in one of the wild type clones complicating any potential interpretations (Figure 2-9B).

### *The pTa Enhancer Mark is Reestablished in an Sp1 Double Mutant BAC*

As none of the individual mutations led to a consistent change in the methylation status of the pTa BAC enhancer, we moved to double mutations. Sp1 has been shown to play a key role in the prevention of de novo methylation at the mouse APRT gene (Brandeis et al., 1994; Macleod et al., 1994). In these studies the Sp1 site flanked a CpG island. The pTa enhancer contains two Sp1 sites, one of which was previously shown to be crucial for the unmethylated window in ES cells. Given the aforementioned and the fact the upstream Sp1 mutation did not prevent the pTa BAC enhancer from becoming unmethylated, we chose to mutate both upstream and downstream Sp1 sites simultaneously.

Double mutations in the Sp1 sites in the endogenous context of the pTa BAC still resulted in the reappearance of the unmethylated window (Figure 2-10). The double mutant BAC was efficiently premethylated prior to transfection (Figure 2-10). We simultaneously compared the Sp1 BAC double mutant to the endogenous pTa enhancer locus. The mutant and endogenous enhancer show very consistent methylation patterns with the core enhancer showing the lowest levels of methylation (Figure 2-10B).

Although the single upstream mutation and the double mutant showed methylation patterns similar to wild type we moved forward with a single downstream

Sp1 mutant. Not surprisingly the single downstream mutant showed the reappearance of the unmethylated window after premethylation and stable transfection in ES cells (Figure 2-11).

## DISCUSSION

A number of studies, using ChIP-seq analysis, have focused on developmentally regulated genes that are marked by active and repressive histone marks simultaneously. Here we focus on non-regulatory tissue specific genes and the mechanisms by which they are regulated. This class of gene does not always appear to be marked by histone modifications in ES cells and thus escapes identification in genome wide ChIP-seq studies (Smale, 2010). Evidence is mounting that this class of genes if often marked by unmethylated CpG dinucleotides at their enhancers and possibly potentiated prior to activation (Bossard and Zaret, 1998; Dillon, 2012; Liber et al., 2010; McPherson et al., 1993; Smale, 2010; Szutorisz et al., 2005; Xu et al., 2007; Xu et al., 2009). Understanding how these marks are maintained and whether or not they are functionally significanct is an important and unanswered question.

In this study we explored the previous findings that an unmethylated CpG dinucleotide window marks the Ptcra enhancer in ES cells, a long time prior to activation. The loss of methylation at non-regulatory tissue specific enhancers was an interesting finding, adding to the potential mechanisms that may regulate tissue specific genes. If this is in fact of functional relevance, it adds another level of complexity to the embryonic stem cell state. That is, the need for the correct potentiation of tissue specific genes, in order for the appropriate expression upon lineage specification.

Here we interrogated the methods by which the unmethylated window is established at the Ptcra enhancer in ES cells. Previous studies defined a potential role for transcription factors, which by binding to sites containing CpG dinucleotides, occlude *de novo* methylation (Xu et al., 2009). Those prior studies directly implicated sequence

specific binding by c-Myb, Sp1, CSL and E-box protein E47, as the cause of the unmethylated window. Furthermore each transcription factor binding site played either a positive or negative role in the establishment of the unmethylated window. Specifically mutations in the upstream Sp1 site, Ebox2 site, or the deletion of these sites with the inclusion of Myb, prevented the reappearance of the unmethylated window at the pTa enhancer after premethylation and stable integration into ES cells. These factors are thus thought to be critical for the establishment of the unmethylated window. Mutations in the Myb, Ebox4 or CSL site resulted in the opposite effect, the spreading of the unmethylated window (Xu et al., 2009). These results indicated that factors present in ES cells can gain access to the methylated enhancer and cause demethylation either by the prevention of maintenance methylation or by active demethylation.

Although intriguing these studies lack one critical component, chromatin context. These experiments in an enhancer-promoter reporter plasmid while important are highly subject to the surrounding chromatin of the integration site. We sought to circumvent this issue with the use of bacterial artificial chromosomes. Using a BAC that contained more than 50kb of flanking sequences gave us confidence that the pTa locus would assemble into an endogenous state. We confirmed that the BAC could be premethylated efficiently and then repeated the mutations made in the plasmid setting to ascertain the effect of chromatin context. The insertion of a PmeI site allowed us to differentiate between the BAC and endogenous loci.

Our first findings were somewhat surprising but confirmed that context is important and must be taken into consideration for further studies. The same mutations that remained methylated at the enhancer in the plasmid context resulted in a clear

unmethylated window at the enhancer in the BAC context. This raises a number of possibilities; the BAC really does recapitulate the endogenous chromatin context of the locus and thus encapsulates the true requirements for the existence of the unmethylated window. If that is true the binding of transcription factors to the Myb, Ebox2 and upstream Sp1 sites are not an absolute requirement for the unmethylated window. This prompts the question of whether additional factors outside of direct binding influence the enhancer environment. It is well known that histone modification of nucleosomes can alter the local chromatin environment. Interestingly the endogenous pTa enhancer is clearly marked by H3K4me1, H3K27 and K9Ac in deep sequencing ChIP-seq studies in ES cells (Chronis unpublished). K27Ac is thought to distinguish active enhancers from poised enhancers, marked by H3K4me1 alone, yet no transcription is detectable at the pTa locus in ES cells (Creyghton et al., 2010).

Our secondary findings determined that mutations causing the spreading of the unmethylated window in the plasmid context did not do so in the BAC context. These mutations in sites for Myb, Ebox4 and CSL do not appear to regulate the size of the unmethylated window as they did in the plasmid context. It is somewhat intriguing that the single mutation in the Myb site resulted in widespread demethylation while deletion of the Myb site in conjunction with Ebox2 and Sp1 resulted in resistance to demethylation. It is of note that the Myb binding site differs from the consensus and although it was shown to bind did so at a lower affinity (Reizis and Leder, 2001). It is therefore quite possible that a different transcription factor binds in this region and the mutational analysis although abrogating the binding of Myb may not have prevented the binding of a different factor. Redundancy is a clear consideration for the differences we

see in the status to the pTa enhancer. As it was determined early on which factors bind and promote expression of pTa, focus remained on those specific transcription factors. As we are now addressing the binding events at much earlier stages in development, biases based on previous knowledge must be eliminated.

With the understanding that Sp1 can protect from de novo methylation we went after both Sp1 sites simultaneously. Again the unmethylated window reappeared in the context of the BAC. Even a sizeable deletion resulted in the reappearance of the unmethylated window. Again this adds to the credence that the BAC more likely forms an endogenous like chromatin environment with added cues that shape and develop the pTa enhancer landscape. We must also consider the possibility that the BAC is also subject to integration specific effects, issues that we have not addressed here. In unpublished work it was demonstrated the BACs often did not integrate in their full entirety, leaving the possibility for variability between integrations.

One thing is clear; the unmethylated window at the pTa enhancer is present at the endogenous locus in ES cells and consistently reappears in the context of a stably integrated BAC that is premethylated prior to transfection. Although we believe the BAC is more suitable than a plasmid, it is not optimal. The most poignant issue is functional relevance. Although we describe a consistent unmethylated state we have no evidence of the functional relevance of this mark at the endogenous locus. To what degree the histone modification and nucleosomal context influence the enhancer state in ES cells is unclear. Advances in methodologies to manipulate endogenous loci will be critical in order to understand of the relevance of these marks in ES cells.

## MATERIALS AND METHODS

### Cell Culture

CCE ES cells were maintained in standard ES growth media. Cells were maintained in Knockout DMEM plus 1% L-glutamine, 1% pen/strep, 1% non-essential amino acids, 15% ES certified FBS (Omega Scientific), and 1,000 units/ml ESGRO (Lif, Millipore). The CCE ES cell line was maintained on gelatin-coated tissue culture flasks.

### Generation of Mutations

pTa mutations were generated using HPLC purified oliogos followed by SOEing PCR and the addition of homology arms.

### BAC Premethylation

BACs were methylated after linearization with PI-SceI using excess units of M.SssI methylase and 2ul 32nM SAM as a substrate then phenol chloroform extracted for electroporation.

### ES Cell Electroporation

BAC DNA was isolated using the Large Construct Kit (Qiagen) and linearized with PI-SceI then ethanol precipitated. 25-40ug of BAC DNA was electroporation into CCE ES cells in a 0.4cm cuvette. Electroporation was done using Bio-Rad GenePulserII at 500uFD, 0.24kV with an optimal time constant of 7.2. Electroporation was done when ES cells reached 80% confluency. ES cells were given fresh media 4 hours prior to the

electroporation. CCEs recover on ice before dilution. G418 selection media was added after 24 hours without antibiotic selection 350-150ug/mL. Media was changed everyday until individual colonies appear (10-14 days)

**Bisulfite Sequencing and PCR Amplification**

1-2ug of isolated DNA was diluted in 50ul TE. DNA was denatured with 5ul of freshly prepared 3M NaOH at 37$^{o}$C for 15-30 minutes. Denatured DNA was added to 510ul of 40.5% sodium bisulfite, 30ul 10mM hydroquinone and the total volume brought up to 610ul. Mixture was incubated protected from light overnight at 55$^{o}$C. DNA was purified using a PCR purification kit (Qiagen) and eluted in 50ul TE. To prepare for PCR amplification Sample is again denatured with 5ul 3M NaOH. Sample is neutralized with 32ul 8M ammonium acetate and precipitated. 2ul of precipitated DNA was used for nested PCR for amplification. PCR amplification products were TA cloned and selected on Xgal Amp plates. Individual colonies were picked, grown overnight, mini-prepped and the resulting DNA sequenced. Sequencing results were aligned to the original sequence and ratio of C/T was calculated to generate methylation percentages.

## FIGURE LEGENDS

**Figure 2-1. Bacterial Artificial Chromosome Genomic Context**

UCSC browser depiction of the bacterial artificial chromosome used containing the Ptcra locus. BAC clone RP23-288F21

**Figure 2-2. Premethylated BAC Controls**

DNA methylation levels throughout the enhancer in the premethylated BACs prior to transfection. Genomic location listed is relative to the transcriptional start site of pTa. Methylation percentage is determined by the ratio of C/T in bisulfite treated clones. Ratio and number of clones sequenced shown. Methylation levels are colored based on the following scale. Green 0-20%, dark green 21-40%, yellow 41-60%, orange 61-80, red 81-100.

**Figure 2-3. Premethylation of the Ptcra BAC is Efficient Throughout the Locus**

DNA methylation levels ascertained after bisulfite sequencing. Location of the CpG is relative to the pTa start site. Regions assessed include both upstream and downstream of the enhancer as well as the promoter. Methylation levels categorized as in figure 2-2.

**Figure 2-4. The Ptcra Enhancer Mark is Reestablished in the Context of a BAC in Embryonic Stem Cells**

DNA methylation levels of the pTa BAC with an enhancer TAG for four independent clones determined by bisulfite sequencing.

**Figure 2-5. The Ptcra BAC Locus Recapitulates the Methylation Status of the Endogenous Locus**

DNA methylation levels of the premethylated control, pTa TAG BAC and the endogenous pTa locus.

**Figure 2-6. Ptcra Core Enhancer Transcription Factor Binding Sites Relative to CpG Dinucleotides**

Depiction of the pTa core enhancer with transcription factor binding sites relative to the CpG dinucleotides.

**Figure 2-7. Ptcra Enhancer Mutations**

Shows the transcription factor binding sites in which mutations were made with the pTa core enhancer in the BAC.

**Figure 2-8. Individual Mutations in the Ebox2 or Sp1 site Do Not Cause Resistance to Demethylation**

DNA methylation levels at the enhancer of pTa in clones after premethylation and transfection into ES cells (A) BAC clones TAG BAC, mEbox2, mSp1 enhancer methylation. Both ration and percentages shown (B) DNA methylation analysis extended to span the entire locus BAC clones TAG BAC, mEbox2, mSp1. Percentages shown (C)

Methylation levels in a pTa deletion spanning the Myb through Sp1 site. Premethylated control and endogenous levels included

**Figure 2-9. Mutations in the Myb, Ebox4 or CSL site Do Not Cause Consistent Widespread Demethylation**

DNA methylation levels of pTa enhancer mutations after premethylation and stable integration in ES cells. (A) Mutations Myb, Ebox4 and CSL. Percentages and ratios shown (B) DNA methylation analysis extended to span the entire locus BAC clones Myb, Ebox4 and CSL. Percentages shown.

**Figure 2-10. The pTa Enhancer Mark is Reestablished in an Sp1 Double Mutant BAC**

DNA methylation levels of pTa BAC determined by bisulfite sequencing after premethylation and stable integration in ES cells (A) Premethylated control – CH3, TAG BAC – WT, and three double Sp1 mutant ES cell clones. (B) DNA methylation levels of double Sp1 mutations spanning the enhancer. Premethylated control – CH3, TAG BAC – WT.

**Figure 2-11. The pTa Enhancer Mark is Reestablished in an Sp1 Single Downstream Mutant BAC**

DNA methylation levels of pTa BAC determined by bisulfite sequencing after premethylation and stable integration in ES cells. Premethylated control – CH3, TAG

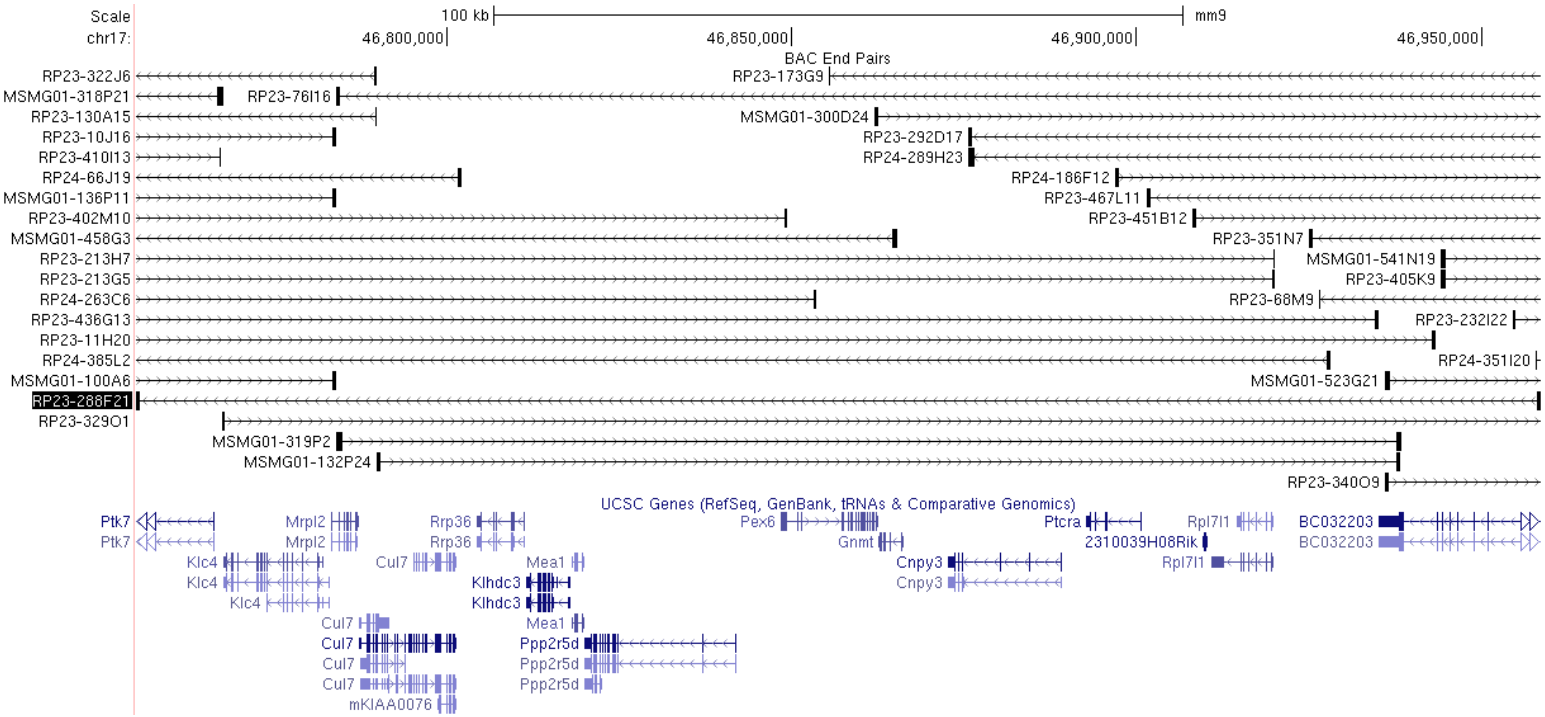
BAC – WT, and one ES cell clone. Percentages shown for all clones. Percentages and ratios shown for mutant

**Figure 2-12. pTa BAC Mutants Methylation Data Summarized**

DNA methylation levels of pTa BAC determined by bisulfite sequencing after premethylation and stable integration in ES cells. Summary of mutant clones. TAG, Myb, CSL, Ebox4, Ebox2, Sp1, Dsp1, Sp1DO

# FIGURES

## Figure 2-1 Bacterial Artificial Chromosome Genomic Context



## Figure 2-2

## Premethylated BAC Controls

| | | Premethylated BACs | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Location | | TAG | | mMyb | | mEbox2 | | mSp1 | | mCSL | | mEbox4 | |
| | | % | Ratio | % | Ratio | % | Ratio | % | Ratio | % | Ratio | % | Ratio |
| Ptcra ENH | -4130 | 100 | 17/17 | 96 | 51/53 | 97 | 30/31 | 96 | 31/32 | 100 | 13/13 | 90 | 38/43 |
| | -4080 | 76 | 13/17 | 92 | 49/53 | 100 | 31/31 | 93 | 28/30 | 92 | 12/13 | 95 | 41/43 |
| | -4042 | 100 | 16/17 | 84 | 45/53 | 97 | 30/31 | 84 | 27/32 | 100 | 13/13 | 86 | 37/43 |
| | -3997 | 100 | 17/17 | 98 | 52/53 | 94 | 29/31 | 100 | 31/31 | 85 | 11/13 | 95 | 41/43 |
| | -3965 | 90 | 17/17 | 92 | 49/53 | 84 | 26/31 | 77 | 24/31 | 100 | 13/13 | 98 | 42/43 |
| | -3947 | 100 | 17/17 | 96 | 51/53 | 94 | 29/31 | 90 | 28/31 | 100 | 13/13 | 90 | 38/42 |
| | -3900 | 82 | 14/17 | 96 | 51/53 | 93 | 25/27 | 91 | 21/23 | 100 | 13/13 | 93 | 39/42 |
| | -3818 | 70 | 12/17 | 98 | 52/53 | 96 | 26/27 | 96 | 25/26 | 100 | 13/13 | 98 | 41/42 |
| | -3806 | 100 | 17/17 | 100 | 53/53 | 100 | 27/27 | 96 | 24/25 | 92 | 12/13 | 96 | 40/42 |
| | -3800 | 100 | 17/17 | 98 | 52/53 | 92 | 24/26 | 96 | 24/25 | 92 | 12/13 | 98 | 41/42 |

**% of Methylated CpG**

- 0-20
- 21-40
- 41-60
- 61-80
- 81-100

**Figure 2-3**

**Premethylation of the Ptcra BAC is Efficient Throughout the Locus**

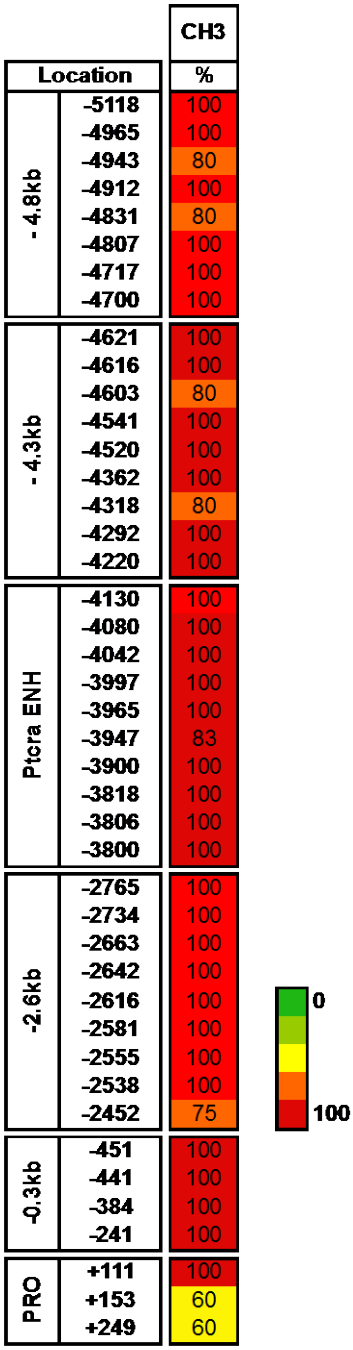
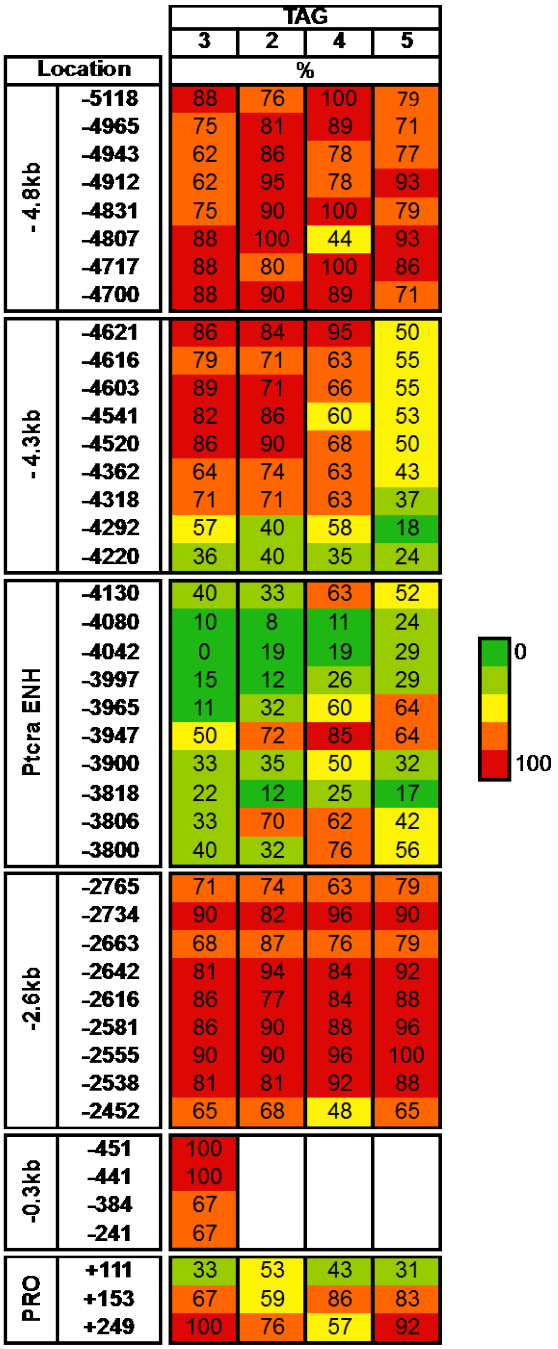


**Figure 2-4**

**The Ptcra Enhancer Mark is Reestablished in the Context of a BAC in Embryonic Stem Cells**



| | Location | TAG 3 | TAG 2 | TAG 4 | TAG 5 |
|---|---|---|---|---|---|
| | | % | | | |
| - 4.8kb | -5118 | 88 | 76 | 100 | 79 |
| | -4965 | 75 | 81 | 89 | 71 |
| | -4943 | 62 | 86 | 78 | 77 |
| | -4912 | 62 | 95 | 78 | 93 |
| | -4831 | 75 | 90 | 100 | 79 |
| | -4807 | 88 | 100 | 44 | 93 |
| | -4717 | 88 | 80 | 100 | 86 |
| | -4700 | 88 | 90 | 89 | 71 |
| - 4.3kb | -4621 | 86 | 84 | 95 | 50 |
| | -4616 | 79 | 71 | 63 | 55 |
| | -4603 | 89 | 71 | 66 | 55 |
| | -4541 | 82 | 86 | 60 | 53 |
| | -4520 | 86 | 90 | 68 | 50 |
| | -4362 | 64 | 74 | 63 | 43 |
| | -4318 | 71 | 71 | 63 | 37 |
| | -4292 | 57 | 40 | 58 | 18 |
| | -4220 | 36 | 40 | 35 | 24 |
| Ptcra ENH | -4130 | 40 | 33 | 63 | 52 |
| | -4080 | 10 | 8 | 11 | 24 |
| | -4042 | 0 | 19 | 19 | 29 |
| | -3997 | 15 | 12 | 26 | 29 |
| | -3965 | 11 | 32 | 60 | 64 |
| | -3947 | 50 | 72 | 85 | 64 |
| | -3900 | 33 | 35 | 50 | 32 |
| | -3818 | 22 | 12 | 25 | 17 |
| | -3806 | 33 | 70 | 62 | 42 |
| | -3800 | 40 | 32 | 76 | 56 |
| -2.6kb | -2765 | 71 | 74 | 63 | 79 |
| | -2734 | 90 | 82 | 96 | 90 |
| | -2663 | 68 | 87 | 76 | 79 |
| | -2642 | 81 | 94 | 84 | 92 |
| | -2616 | 86 | 77 | 84 | 88 |
| | -2581 | 86 | 90 | 88 | 96 |
| | -2555 | 90 | 90 | 96 | 100 |
| | -2538 | 81 | 81 | 92 | 88 |
| | -2452 | 65 | 68 | 48 | 65 |
| -0.3kb | -451 | 100 | | | |
| | -441 | 100 | | | |
| | -384 | 67 | | | |
| | -241 | 67 | | | |
| PRO | +111 | 33 | 53 | 43 | 31 |
| | +153 | 67 | 59 | 86 | 83 |
| | +249 | 100 | 76 | 57 | 92 |

**Figure 2-5**

**The Ptcra BAC Locus Recapitulates the Methylation Status of the Endogenous Locus**

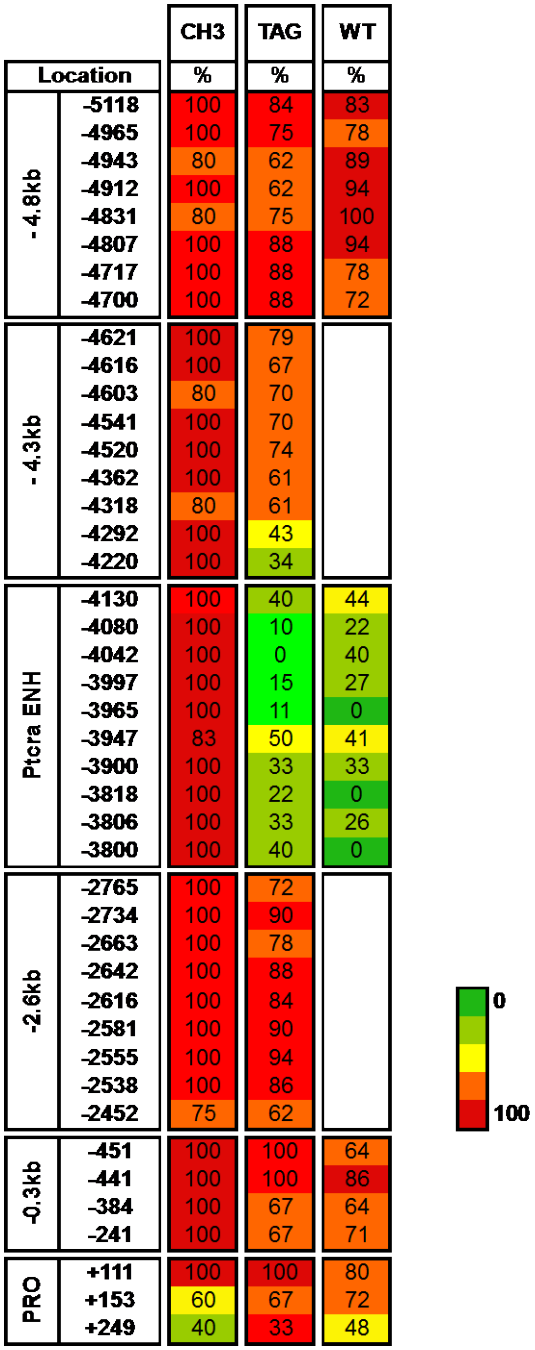


**Figure 2-6**

**Ptcra Core Enhancer Transcription Factor Binding Sites Relative to CpG Dinucleotides**

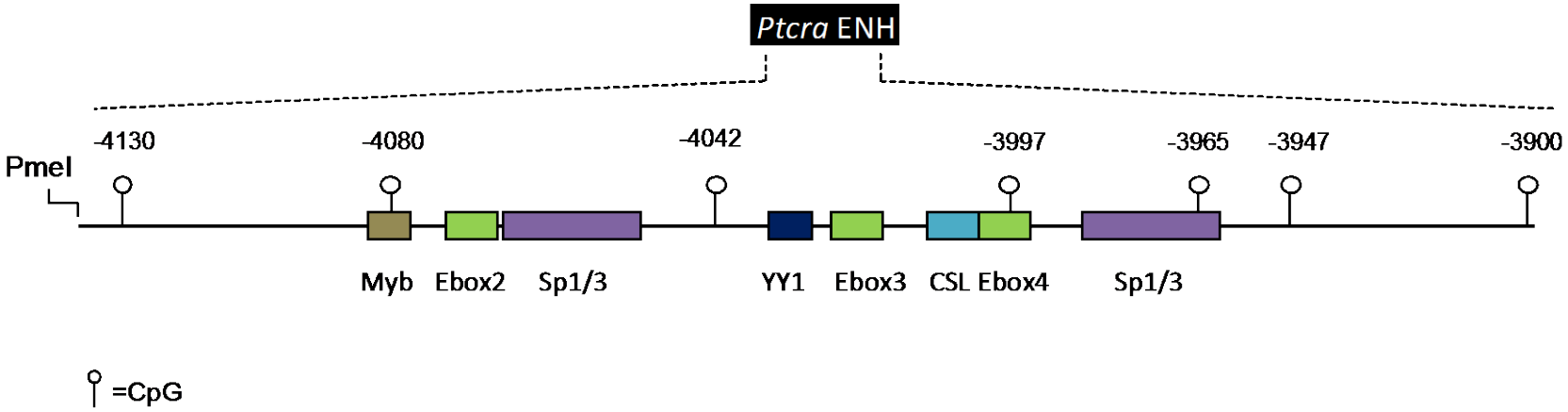
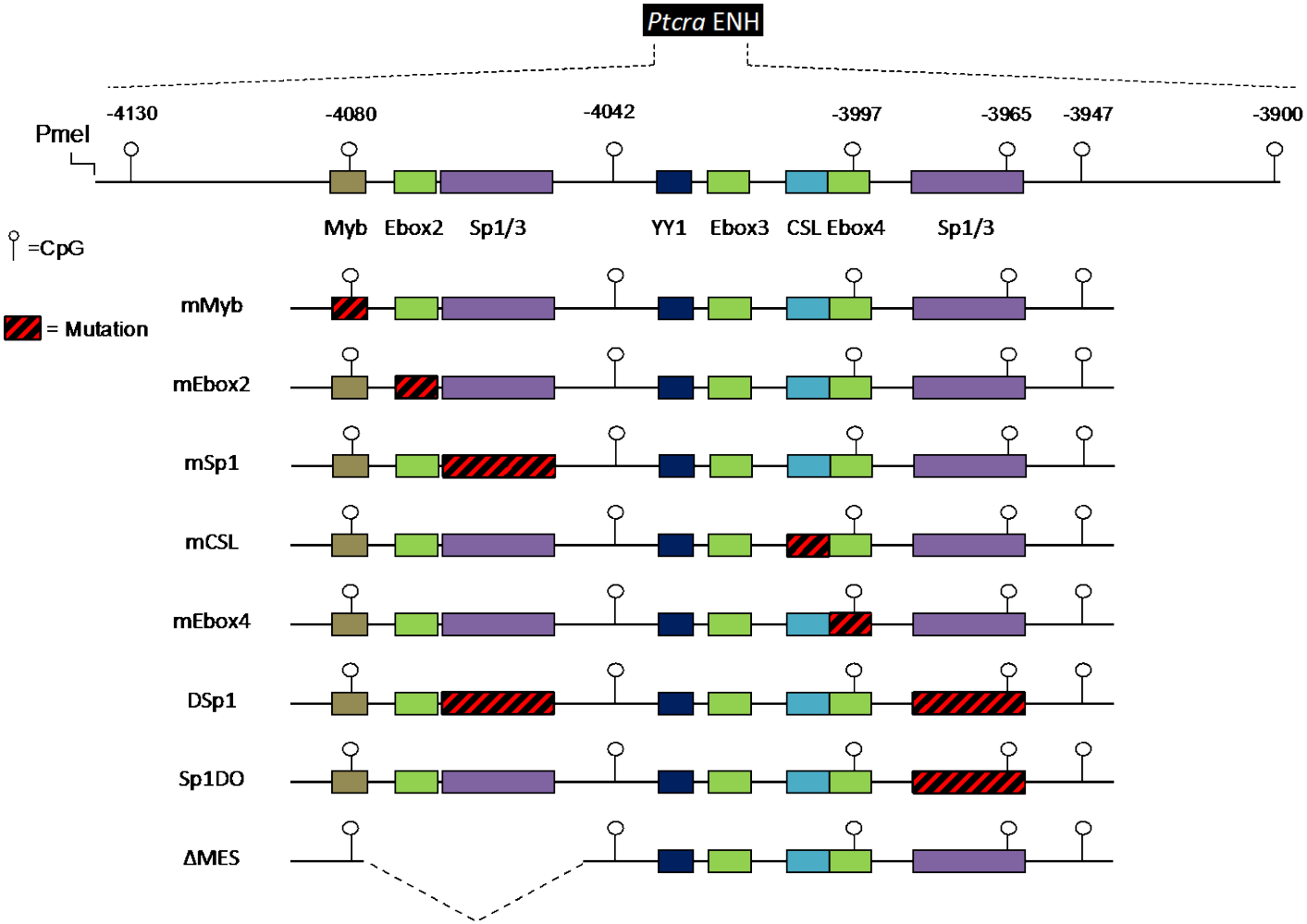


**Figure 2-7**

**Ptcra Enhancer Mutations**



## Figure 2-8

## Individual Mutations in the Ebox2 or Sp1 site Do Not Cause Resistance to Demethylation

### A

| | | WT TAG clones | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Location | | 1 | | 2 | | 3 | | 4 | | 5 | |
| | | % | Ratio | % | Ratio | % | Ratio | % | Ratio | % | Ratio |
| Ptcra ENH | -4130 | 27 | 7/26 | 33 | 5/15 | 63 | 12/19 | 63 | 12/19 | 52 | 14/27 |
| | -4080 | 13 | 4/31 | 8 | 1/13 | 11 | 2/19 | 11 | 2/19 | 24 | 5/21 |
| | -4042 | 3 | 1/30 | 19 | 3/16 | 19 | 4/21 | 19 | 4/21 | 29 | 7/24 |
| | -3997 | 10 | 3/29 | 12 | 2/17 | 21 | 4/18 | 26 | 5/19 | 29 | 7/24 |
| | -3965 | 36 | 11/31 | 32 | 6/19 | 52 | 11/21 | 60 | 12/20 | 64 | 14/22 |
| | -3947 | 53 | 16/30 | 72 | 13/18 | 86 | 18/21 | 85 | 17/20 | 64 | 14/22 |
| | -3900 | 29 | 8/28 | 35 | 6/17 | 52 | 11/21 | 50 | 10/20 | 32 | 7/22 |
| | -3818 | 10 | 3/29 | 12 | 2/17 | 24 | 5/21 | 25 | 5/20 | 17 | 4/24 |
| | -3806 | 44 | 14/32 | 70 | 14/20 | 64 | 14/22 | 62 | 13/21 | 42 | 10/24 |
| | -3800 | 44 | 14/32 | 32 | 6/19 | 77 | 17/22 | 76 | 16/21 | 56 | 13/23 |

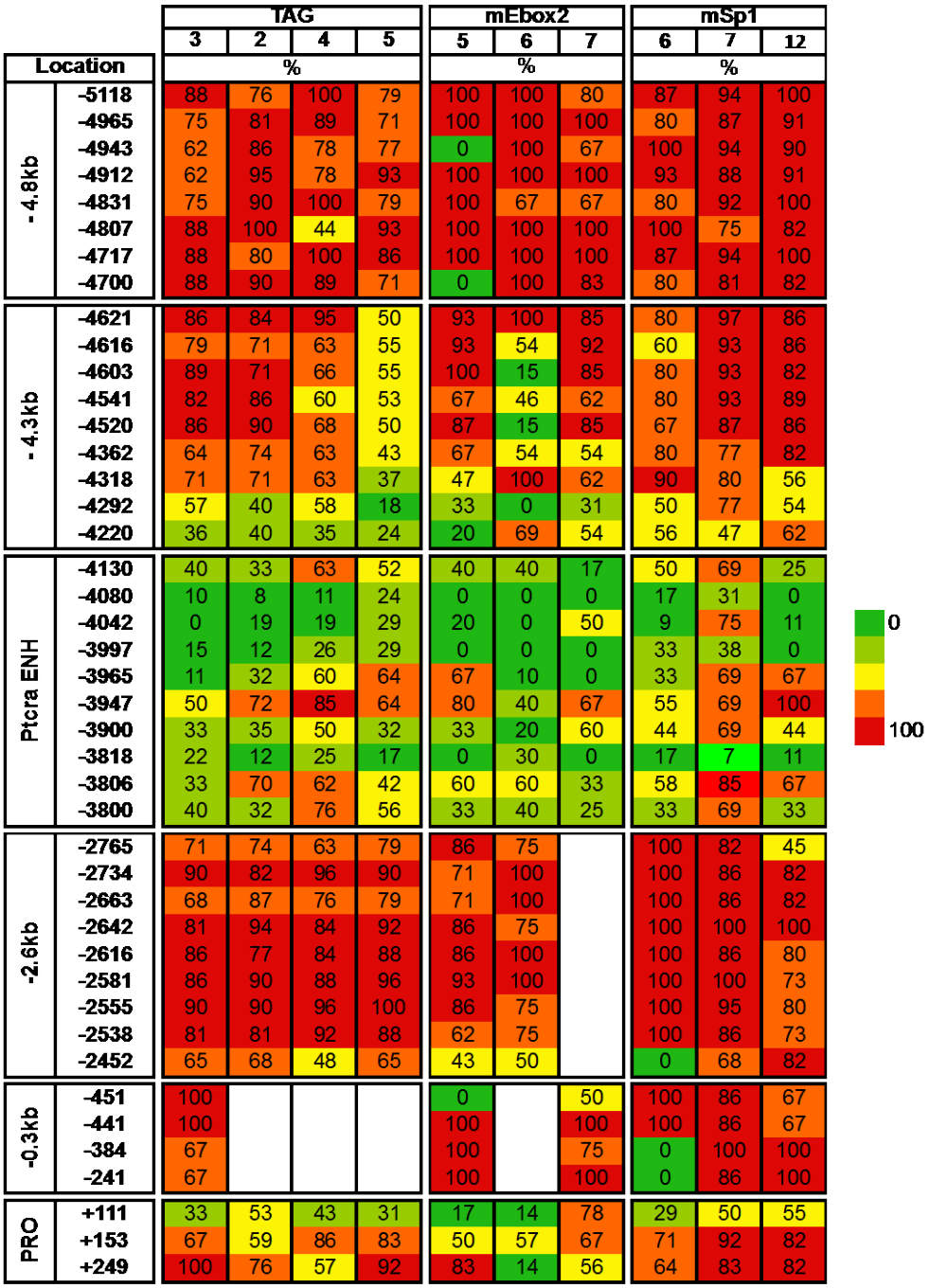
| | | mEbox2 Clones | | | | | |
|---|---|---|---|---|---|---|---|
| Location | | 5 | | 6 | | 7 | |
| | | % | Ratio | % | Ratio | % | Ratio |
| Ptcra ENH | -4130 | 7 | 1/15 | 22 | 4/18 | 0 | 0/31 |
| | -4080 | 0 | 0/15 | 6 | 1/18 | 0 | 0/31 |
| | -4042 | 13 | 2/15 | 22 | 4/18 | 0 | 0/31 |
| | -3997 | 27 | 4/15 | 0 | 0/18 | 3 | 1/29 |
| | -3965 | 7 | 1/15 | 33 | 6/18 | 55 | 17/31 |
| | -3947 | 53 | 8/15 | 59 | 10/17 | 52 | 16/31 |
| | -3900 | 40 | 6/15 | 7 | 1/15 | 30 | 9/30 |
| | -3818 | 0 | 0/15 | 17 | 3/18 | 3 | 1/31 |
| | -3806 | 13 | 2/15 | 28 | 5/18 | 32 | 10/31 |
| | -3800 | 27 | 4/15 | 33 | 6/18 | 23 | 7/30 |


| | | mSp1 Clones | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Location | | 4 | | 6 | | 7 | | 9 | | 12 | |
| | | % | Ratio | % | Ratio | % | Ratio | % | Ratio | % | Ratio |
| Ptcra ENH | -4130 | 37 | 8/22 | 57 | 12/21 | 46 | 6/13 | 40 | 6/15 | 68 | 13/19 |
| | -4080 | 14 | 3/23 | 14 | 3/22 | 13 | 2/16 | 13 | 2/16 | 9 | 2/23 |
| | -4042 | 14 | 3/21 | 30 | 6/20 | 67 | 10/15 | 19 | 3/16 | 67 | 14/21 |
| | -3997 | 18 | 4/22 | 26 | 6/19 | 44 | 7/16 | 0 | 0/15 | 29 | 6/21 |
| | -3965 | 22 | 5/23 | 55 | 12/22 | 63 | 10/16 | 66 | 10/15 | 52 | 12/23 |
| | -3947 | 73 | 16/22 | 86 | 18/21 | 73 | 11/15 | 71 | 12/17 | 87 | 20/23 |
| | -3900 | 59 | 13/22 | 43 | 9/21 | 53 | 8/15 | 50 | 8/16 | 62 | 13/21 |
| | -3818 | 9 | 2/22 | 14 | 3/22 | 44 | 7/16 | 29 | 5/17 | 39 | 9/23 |
| | -3806 | 36 | 8/22 | 41 | 9/22 | 73 | 11/15 | 59 | 10/17 | 70 | 16/23 |
| | -3800 | 36 | 8/23 | 43 | 9/21 | 73 | 11/15 | 44 | 7/16 | 73 | 16/22 |

## Figure 2-8

## Individual Mutations in the Ebox2 or Sp1 site Do Not Cause Resistance to Demethylation

**B**

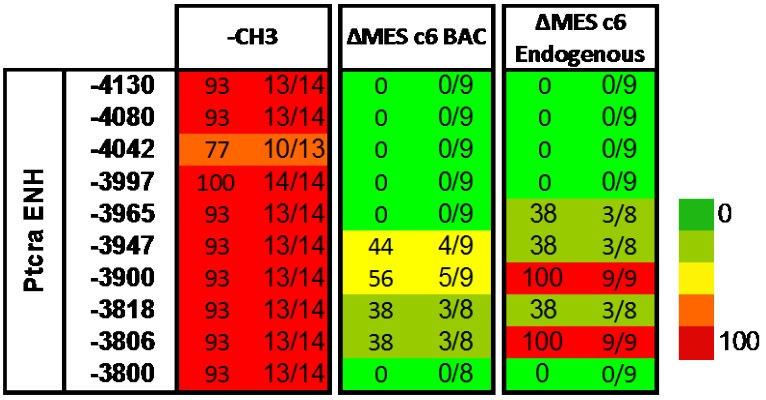


| | | TAG | | | | mEbox2 | | | mSp1 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 3 | 2 | 4 | 5 | 5 | 6 | 7 | 6 | 7 | 12 |
| Location | | % | | | | % | | | % | | |
| -4.8kb | -5118 | 88 | 76 | 100 | 79 | 100 | 100 | 80 | 87 | 94 | 100 |
| | -4965 | 75 | 81 | 89 | 71 | 100 | 100 | 100 | 80 | 87 | 91 |
| | -4943 | 62 | 86 | 78 | 77 | 0 | 100 | 67 | 100 | 94 | 90 |
| | -4912 | 62 | 95 | 78 | 93 | 100 | 100 | 100 | 93 | 88 | 91 |
| | -4831 | 75 | 90 | 100 | 79 | 100 | 67 | 67 | 80 | 92 | 100 |
| | -4807 | 88 | 100 | 44 | 93 | 100 | 100 | 100 | 100 | 75 | 82 |
| | -4717 | 88 | 80 | 100 | 86 | 100 | 100 | 100 | 87 | 94 | 100 |
| | -4700 | 88 | 90 | 89 | 71 | 0 | 100 | 83 | 80 | 81 | 82 |
| -4.3kb | -4621 | 86 | 84 | 95 | 50 | 93 | 100 | 85 | 80 | 97 | 86 |
| | -4616 | 79 | 71 | 63 | 55 | 93 | 54 | 92 | 60 | 93 | 86 |
| | -4603 | 89 | 71 | 66 | 55 | 100 | 15 | 85 | 80 | 93 | 82 |
| | -4541 | 82 | 86 | 60 | 53 | 67 | 46 | 62 | 80 | 93 | 89 |
| | -4520 | 86 | 90 | 68 | 50 | 87 | 15 | 85 | 67 | 87 | 86 |
| | -4362 | 64 | 74 | 63 | 43 | 67 | 54 | 54 | 80 | 77 | 82 |
| | -4318 | 71 | 71 | 63 | 37 | 47 | 100 | 62 | 90 | 80 | 56 |
| | -4292 | 57 | 40 | 58 | 18 | 33 | 0 | 31 | 50 | 77 | 54 |
| | -4220 | 36 | 40 | 35 | 24 | 20 | 69 | 54 | 56 | 47 | 62 |
| Ptcra ENH | -4130 | 40 | 33 | 63 | 52 | 40 | 40 | 17 | 50 | 69 | 25 |
| | -4080 | 10 | 8 | 11 | 24 | 0 | 0 | 0 | 17 | 31 | 0 |
| | -4042 | 0 | 19 | 19 | 29 | 20 | 0 | 50 | 9 | 75 | 11 |
| | -3997 | 15 | 12 | 26 | 29 | 0 | 0 | 0 | 33 | 38 | 0 |
| | -3965 | 11 | 32 | 60 | 64 | 67 | 10 | 0 | 33 | 69 | 67 |
| | -3947 | 50 | 72 | 85 | 64 | 80 | 40 | 67 | 55 | 69 | 100 |
| | -3900 | 33 | 35 | 50 | 32 | 33 | 20 | 60 | 44 | 69 | 44 |
| | -3818 | 22 | 12 | 25 | 17 | 0 | 30 | 0 | 17 | 7 | 11 |
| | -3806 | 33 | 70 | 62 | 42 | 60 | 60 | 33 | 58 | 85 | 67 |
| | -3800 | 40 | 32 | 76 | 56 | 33 | 40 | 25 | 33 | 69 | 33 |
| -2.6kb | -2765 | 71 | 74 | 63 | 79 | 86 | 75 | | 100 | 82 | 45 |
| | -2734 | 90 | 82 | 96 | 90 | 71 | 100 | | 100 | 86 | 82 |
| | -2663 | 68 | 87 | 76 | 79 | 71 | 100 | | 100 | 86 | 82 |
| | -2642 | 81 | 94 | 84 | 92 | 86 | 75 | | 100 | 100 | 100 |
| | -2616 | 86 | 77 | 84 | 88 | 86 | 100 | | 100 | 86 | 80 |
| | -2581 | 86 | 90 | 88 | 96 | 93 | 100 | | 100 | 100 | 73 |
| | -2555 | 90 | 90 | 96 | 100 | 86 | 75 | | 100 | 95 | 80 |
| | -2538 | 81 | 81 | 92 | 88 | 62 | 75 | | 100 | 86 | 73 |
| | -2452 | 65 | 68 | 48 | 65 | 43 | 50 | | 0 | 68 | 82 |
| -0.3kb | -451 | 100 | | | | 0 | | 50 | 100 | 86 | 67 |
| | -441 | 100 | | | | 100 | | 100 | 100 | 86 | 67 |
| | -384 | 67 | | | | 100 | | 75 | 0 | 100 | 100 |
| | -241 | 67 | | | | 100 | | 100 | 0 | 86 | 100 |
| PRO | +111 | 33 | 53 | 43 | 31 | 17 | 14 | 78 | 29 | 50 | 55 |
| | +153 | 67 | 59 | 86 | 83 | 50 | 57 | 67 | 71 | 92 | 82 |
| | +249 | 100 | 76 | 57 | 92 | 83 | 14 | 56 | 64 | 83 | 82 |

**Figure 2-8**

**Individual Mutations in the Ebox2 or Sp1 site Do Not Cause Resistance to Demethylation**

**C**



## Figure 2-9

## Mutations in the Myb, Ebox4 or CSL site Do Not Cause Consistent Widespread Demethylation

### A

**WT TAG clones**

| Location | | 1 | | 2 | | 3 | | 4 | | 5 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | % | Ratio | % | Ratio | % | Ratio | % | Ratio | % | Ratio |
| Ptcra ENH | -4130 | 27 | 7/26 | 33 | 5/15 | 63 | 12/19 | 63 | 12/19 | 52 | 14/27 |
| | -4080 | 13 | 4/31 | 8 | 1/13 | 11 | 2/19 | 11 | 2/19 | 24 | 5/21 |
| | -4042 | 3 | 1/30 | 19 | 3/16 | 19 | 4/21 | 19 | 4/21 | 29 | 7/24 |
| | -3997 | 10 | 3/29 | 12 | 2/17 | 21 | 4/18 | 26 | 5/19 | 29 | 7/24 |
| | -3965 | 36 | 11/31 | 32 | 6/19 | 52 | 11/21 | 60 | 12/20 | 64 | 14/22 |
| | -3947 | 53 | 16/30 | 72 | 13/18 | 86 | 18/21 | 85 | 17/20 | 64 | 14/22 |
| | -3900 | 29 | 8/28 | 35 | 6/17 | 52 | 11/21 | 50 | 10/20 | 32 | 7/22 |
| | -3818 | 10 | 3/29 | 12 | 2/17 | 24 | 5/21 | 25 | 5/20 | 17 | 4/24 |
| | -3806 | 44 | 14/32 | 70 | 14/20 | 64 | 14/22 | 62 | 13/21 | 42 | 10/24 |
| | -3800 | 44 | 14/32 | 32 | 6/19 | 77 | 17/22 | 76 | 16/21 | 56 | 13/23 |

**mMyb clones**

| | | 1 | | 2 | | 3 | | 4 | | 5 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | % | Ratio | % | Ratio | % | Ratio | % | Ratio | % | Ratio |
| Ptcra ENH | -4130 | 55 | 16/29 | 36 | 5/14 | 46 | 19/41 | 33 | 7/21 | 22 | 2/9 |
| | -4080 | 7 | 1/15 | 7 | 3/44 | 12 | 5/42 | 5 | 1/20 | 10 | 1/10 |
| | -4042 | 23 | 7/30 | 19 | 8/43 | 19 | 8/43 | 11 | 2/19 | 21 | 3/14 |
| | -3997 | 21 | 6/29 | 16 | 7/43 | 10 | 2/21 | 0 | 0/21 | 8 | 1/13 |
| | -3965 | 55 | 17/31 | 28 | 12/43 | 33 | 1/3 | 23 | 5/21 | 15 | 2/13 |
| | -3947 | 87 | 13/15 | 73 | 8/11 | 61 | 25/41 | 38 | 8/21 | 38 | 5/13 |
| | -3900 | 61 | 19/31 | 55 | 6/11 | 31 | 13/42 | 38 | 7/18 | 33 | 5/15 |
| | -3818 | 40 | 2/5 | 19 | 8/43 | 10 | 2/21 | 9 | 2/21 | 13 | 2/15 |
| | -3806 | 52 | 16/31 | 52 | 23/44 | 51 | 22/43 | 23 | 5/21 | 47 | 7/25 |
| | -3800 | 63 | 19/30 | 43 | 19/44 | 31 | 13/42 | 20 | 4/20 | 20 | 3/15 |

**mEbox4 Clones**

| Location | | 2 | | 3 | | 5 | | 10 | |
|---|---|---|---|---|---|---|---|---|---|
| | | % | Ratio | % | Ratio | % | Ratio | % | Ratio |
| Ptcra ENH | -4130 | 63 | 5/8 | 33 | 3/9 | 100 | 20/20 | 27 | 3/11 |
| | -4080 | 11 | 1/9 | 0 | 0/8 | 0 | 0/20 | 0 | 0/11 |
| | -4042 | 18 | 2/11 | 40 | 4/10 | 0 | 0/20 | 18 | 2/11 |
| | -3997 | 44 | 4/9 | 20 | 2/10 | 21 | 4/19 | 27 | 3/11 |
| | -3965 | 22 | 2/9 | 22 | 2/9 | 95 | 18/19 | 72 | 8/11 |
| | -3947 | 100 | 10/10 | 88 | 7/8 | 100 | 20/20 | 81 | 9/11 |
| | -3900 | 45 | 5/11 | 70 | 7/10 | 0 | 0/20 | 27 | 3/11 |
| | -3818 | 36 | 4/11 | 30 | 3/10 | 0 | 0/20 | 18 | 2/11 |
| | -3806 | 73 | 8/11 | 60 | 6/10 | 0 | 0/20 | 72 | 8/11 |
| | -3800 | 82 | 9/11 | 80 | 8/10 | 0 | 0/20 | 81 | 9/11 |


**mCSL Clones**

| Location | | 1 | | 2 | | 3 | | 9 | |
|---|---|---|---|---|---|---|---|---|---|
| | | % | Ratio | % | Ratio | % | Ratio | % | Ratio |
| Ptcra ENH | -4130 | 33 | 5/15 | 50 | 9/18 | 36 | 8/22 | 59 | 10/17 |
| | -4080 | 7 | 1/15 | 5 | 1/19 | 5 | 1/22 | 11 | 2/18 |
| | -4042 | 13 | 2/15 | 25 | 5/20 | 40 | 8/20 | 53 | 9/17 |
| | -3997 | 0 | 0/15 | 0 | 0/20 | 0 | 0/22 | 18 | 3/17 |
| | -3965 | 43 | 6/14 | 75 | 15/20 | 36 | 8/22 | 78 | 14/18 |
| | -3947 | 80 | 12/15 | 80 | 16/20 | 55 | 12/22 | 88 | 15/17 |
| | -3900 | 47 | 7/15 | 70 | 14/20 | 50 | 11/22 | 55 | 10/18 |
| | -3818 | 7 | 1/15 | 20 | 4/20 | 36 | 8/22 | 17 | 3/18 |
| | -3806 | 13 | 2/15 | 65 | 13/20 | 45 | 10/22 | 67 | 12/18 |
| | -3800 | 20 | 3/15 | 65 | 13/20 | 73 | 16/22 | 56 | 10/18 |

## Figure 2-9

## Mutations in the Myb, Ebox4 or CSL site Do Not Cause Consistent Widespread Demethylation

## B

| | | TAG | | | | mMyb | | | mCSL | | | mEbox4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 3 | 2 | 4 | 5 | 1 | 4 | 7 | 1 | 2 | 3 | 4 |
| Location | | % | | | | % | | | % | | | % |
| -4.8kb | -5118 | 88 | 76 | 100 | 79 | 83 | 92 | 80 | 92 | 56 | 92 | 76 |
| | -4965 | 75 | 81 | 89 | 71 | 78 | 92 | 78 | 77 | 78 | 100 | 76 |
| | -4943 | 62 | 86 | 78 | 77 | 72 | 67 | 75 | 69 | 67 | 75 | 82 |
| | -4912 | 62 | 95 | 78 | 93 | 93 | 100 | 90 | 85 | 89 | 92 | 100 |
| | -4831 | 75 | 90 | 100 | 79 | 100 | 100 | 100 | 92 | 70 | 92 | 94 |
| | -4807 | 88 | 100 | 44 | 93 | 89 | 83 | 100 | 77 | 100 | 83 | 94 |
| | -4717 | 88 | 80 | 100 | 86 | 78 | 100 | 100 | 77 | 89 | 92 | 76 |
| | -4700 | 88 | 90 | 89 | 71 | 78 | 92 | 60 | 85 | 78 | 83 | 65 |
| -4.3kb | -4621 | 86 | 84 | 95 | 50 | 93 | 7 | 50 | 77 | 83 | 100 | 91 |
| | -4616 | 79 | 71 | 63 | 55 | 93 | 36 | 50 | 87 | 83 | 86 | 75 |
| | -4603 | 89 | 71 | 66 | 55 | 93 | 15 | 60 | 84 | 61 | 86 | 75 |
| | -4541 | 82 | 86 | 60 | 53 | 80 | 21 | 70 | 68 | 61 | 88 | 81 |
| | -4520 | 86 | 90 | 68 | 50 | 93 | 36 | 70 | 77 | 61 | 75 | 84 |
| | -4362 | 64 | 74 | 63 | 43 | 87 | 29 | 30 | 61 | 61 | 75 | 66 |
| | -4318 | 71 | 71 | 63 | 37 | 73 | 29 | 20 | 65 | 61 | 63 | 63 |
| | -4292 | 57 | 40 | 58 | 18 | 67 | 14 | 20 | 26 | 28 | 63 | 34 |
| | -4220 | 36 | 40 | 35 | 24 | 73 | 29 | 20 | 19 | 22 | 25 | 38 |
| Ptcra ENH | -4130 | 40 | 33 | 63 | 52 | 50 | 25 | 18 | 10 | 54 | 0 | 64 |
| | -4080 | 10 | 8 | 11 | 24 | 100 | 63 | 0 | 0 | 15 | 20 | 9 |
| | -4042 | 0 | 19 | 19 | 29 | 50 | 25 | 27 | 0 | 38 | 0 | 9 |
| | -3997 | 15 | 12 | 26 | 29 | 0 | 25 | 9 | 22 | 15 | 20 | 18 |
| | -3965 | 11 | 32 | 60 | 64 | 100 | 0 | 22 | 30 | 43 | 0 | 27 |
| | -3947 | 50 | 72 | 85 | 64 | 50 | 75 | 45 | 40 | 50 | 60 | 27 |
| | -3900 | 33 | 35 | 50 | 32 | 0 | 38 | 18 | 20 | 50 | 66 | 55 |
| | -3818 | 22 | 12 | 25 | 17 | 0 | 13 | 0 | 22 | 23 | 33 | 27 |
| | -3806 | 33 | 70 | 62 | 42 | 100 | 75 | 27 | 30 | 54 | 33 | 36 |
| | -3800 | 40 | 32 | 76 | 56 | 50 | 63 | 18 | 20 | 50 | 33 | 36 |
| -2.6kb | -2765 | 71 | 74 | 63 | 79 | 86 | 100 | 55 | 63 | | 80 | 61 |
| | -2734 | 90 | 82 | 96 | 90 | 100 | 78 | 82 | 100 | | 93 | 94 |
| | -2663 | 68 | 87 | 76 | 79 | 80 | 67 | 91 | 63 | | 93 | 61 |
| | -2642 | 81 | 94 | 84 | 92 | 100 | 100 | 82 | 88 | | 100 | 82 |
| | -2616 | 86 | 77 | 84 | 88 | 100 | 67 | 82 | 75 | | 80 | 82 |
| | -2581 | 86 | 90 | 88 | 96 | 86 | 88 | 82 | 88 | | 87 | 83 |
| | -2555 | 90 | 90 | 96 | 100 | 86 | 100 | 90 | 100 | | 100 | 89 |
| | -2538 | 81 | 81 | 92 | 88 | 86 | 67 | 78 | 88 | | 87 | 67 |
| | -2452 | 65 | 68 | 48 | 65 | 86 | 86 | 50 | 63 | | 47 | 59 |
| -0.3kb | -451 | 100 | | | | 100 | 100 | 83 | 75 | 67 | 50 | 50 |
| | -441 | 100 | | | | 100 | 100 | 100 | 50 | 100 | 75 | 50 |
| | -384 | 67 | | | | 100 | 100 | 83 | 100 | 100 | 100 | 25 |
| | -241 | 67 | | | | 0 | 100 | 100 | 100 | 100 | 75 | 75 |
| PRO | +111 | 33 | 53 | 43 | 31 | 62 | 36 | 67 | 17 | 11 | 45 | 67 |
| | +153 | 67 | 59 | 86 | 83 | 85 | 72 | 100 | 43 | 78 | 91 | 83 |
| | +249 | 100 | 76 | 57 | 92 | 85 | 82 | 87 | 100 | 89 | 82 | 83 |

## Figure 2-10

## The pTa Enhancer Mark is Reestablished in an Sp1 Double Mutant BAC

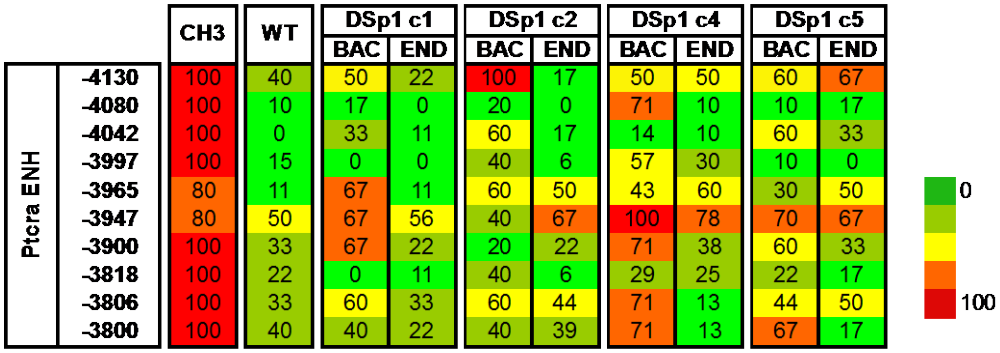
**A**

| | Location | CH3 | WT | DSp1-c1 | | DSp1 c2 | | DSp1 c4 | | DSp1 c5 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | % | % | % | Ratio | % | Ratio | % | Ratio | % | Ratio |
| -4.8kb | -5118 | 100 | 84 | 67 | 4/6 | 100 | 1/1 | 75 | 3/4 | 50 | 1/2 |
| | -4965 | 100 | 75 | 83 | 5/6 | 100 | 1/1 | 100 | 4/4 | 50 | 1/2 |
| | -4943 | 100 | 62 | 50 | 3/6 | 100 | 1/1 | 100 | 4/4 | 50 | 1/2 |
| | -4912 | 100 | 62 | 83 | 5/6 | 0 | 0/1 | 100 | 4/4 | 100 | 3/3 |
| | -4831 | 100 | 75 | 67 | 4/6 | 100 | 1/1 | 100 | 4/4 | 100 | 3/3 |
| | -4807 | 100 | 88 | 67 | 4/6 | 100 | 1/1 | 100 | 4/4 | 50 | 1/2 |
| | -4717 | 60 | 88 | 33 | 2/6 | 100 | 1/1 | 75 | 3/4 | 50 | 1/2 |
| | -4700 | 80 | 88 | 17 | 1/6 | 100 | 1/1 | 25 | 1/4 | 100 | 2/2 |
| -4.3kb | -4621 | 100 | 79 | 22 | 2/9 | 91 | 10/11 | 86 | 6/7 | 100 | 8/8 |
| | -4616 | 100 | 67 | 44 | 4/9 | 64 | 7/11 | 86 | 6/7 | 88 | 7/8 |
| | -4603 | 100 | 70 | 44 | 4/9 | 64 | 7/11 | 86 | 6/7 | 63 | 5/8 |
| | -4541 | 100 | 70 | 67 | 6/9 | 73 | 8/11 | 86 | 6/7 | 75 | 6/8 |
| | -4520 | 100 | 74 | 89 | 8/9 | 91 | 10/11 | 100 | 7/7 | 75 | 6/8 |
| | -4362 | 80 | 61 | 44 | 4/9 | 91 | 10/11 | 80 | 4/5 | 88 | 7/8 |
| | -4318 | 100 | 61 | 67 | 6/9 | 82 | 9/11 | 80 | 4/5 | 75 | 6/8 |
| | -4292 | 100 | 43 | 44 | 4/9 | 37 | 4/11 | 80 | 4/5 | 50 | 4/8 |
| | -4220 | 80 | 34 | 11 | 1/9 | 80 | 8/10 | 33 | 4/12 | 44 | 8/18 |
| Ptcra ENH | -4130 | 100 | 40 | 50 | 3/6 | 100 | 5/5 | 50 | 5/10 | 60 | 6/10 |
| | -4080 | 100 | 10 | 17 | 1/6 | 20 | 1/5 | 71 | 5/7 | 10 | 1/10 |
| | -4042 | 100 | 0 | 33 | 2/6 | 60 | 3/5 | 14 | 1/7 | 60 | 6/10 |
| | -3997 | 100 | 15 | 0 | 0/6 | 40 | 2/5 | 57 | 4/7 | 10 | 1/10 |
| | -3965 | 80 | 11 | 67 | 4/6 | 60 | 3/5 | 43 | 3/7 | 30 | 3/10 |
| | -3947 | 80 | 50 | 67 | 4/6 | 40 | 2/5 | 100 | 7/7 | 70 | 7/10 |
| | -3900 | 100 | 33 | 67 | 4/6 | 20 | 1/5 | 71 | 5/7 | 60 | 6/10 |
| | -3818 | 100 | 22 | 0 | 0/5 | 40 | 2/5 | 29 | 2/7 | 22 | 2/9 |
| | -3806 | 100 | 33 | 60 | 3/5 | 60 | 3/5 | 71 | 5/7 | 44 | 4/9 |
| | -3800 | 100 | 40 | 40 | 2/5 | 40 | 2/5 | 71 | 5/7 | 67 | 6/9 |
| -2.6kb | -2765 | 80 | 72 | 67 | 8/12 | 88 | 7/8 | 70 | 7/10 | 100 | 6/6 |
| | -2734 | 100 | 90 | 83 | 10/12 | 75 | 6/8 | 100 | 10/10 | 100 | 6/6 |
| | -2663 | 100 | 78 | 58 | 7/12 | 63 | 5/8 | 40 | 4/10 | 86 | 6/7 |
| | -2642 | 100 | 88 | 92 | 11/12 | 63 | 5/8 | 70 | 7/10 | 86 | 6/7 |
| | -2616 | 100 | 84 | 72 | 8/11 | 75 | 6/8 | 60 | 6/10 | 43 | 3/7 |
| | -2581 | 100 | 90 | 100 | 12/12 | 86 | 7/8 | 100 | 10/10 | 100 | 7/7 |
| | -2555 | 100 | 94 | 92 | 11/12 | 100 | 8/8 | 100 | 10/10 | 100 | 7/7 |
| | -2538 | 100 | 86 | 92 | 11/12 | 100 | 8/8 | 80 | 8/10 | 86 | 6/7 |
| | -2452 | 100 | 62 | 17 | 2/12 | 13 | 1/8 | 30 | 3/10 | 29 | 2/7 |
| -0.3kb | -451 | 100 | 100 | 100 | 1/1 | 75 | 3/4 | 71 | 5/7 | 40 | 2/5 |
| | -441 | 100 | 100 | 50 | 1/2 | 100 | 4/4 | 86 | 6/7 | 80 | 4/5 |
| | -384 | 100 | 67 | 100 | 1/1 | 75 | 3/4 | 100 | 7/7 | 80 | 4/5 |
| | -241 | 100 | 67 | 100 | 1/1 | 100 | 4/4 | 83 | 5/6 | 100 | 5/5 |
| PRO | +111 | 100 | 40 | 67 | 2/3 | 50 | 2/4 | 71 | 5/7 | 33 | 4/12 |
| | +153 | 100 | 74 | 100 | 3/3 | 75 | 3/4 | 86 | 6/7 | 75 | 9/12 |
| | +249 | 100 | 81 | 67 | 2/3 | 25 | 1/4 | 86 | 6/7 | 67 | 8/12 |

## Figure 2-10

## The pTa Enhancer Mark is Reestablished in an Sp1 Double Mutant BAC

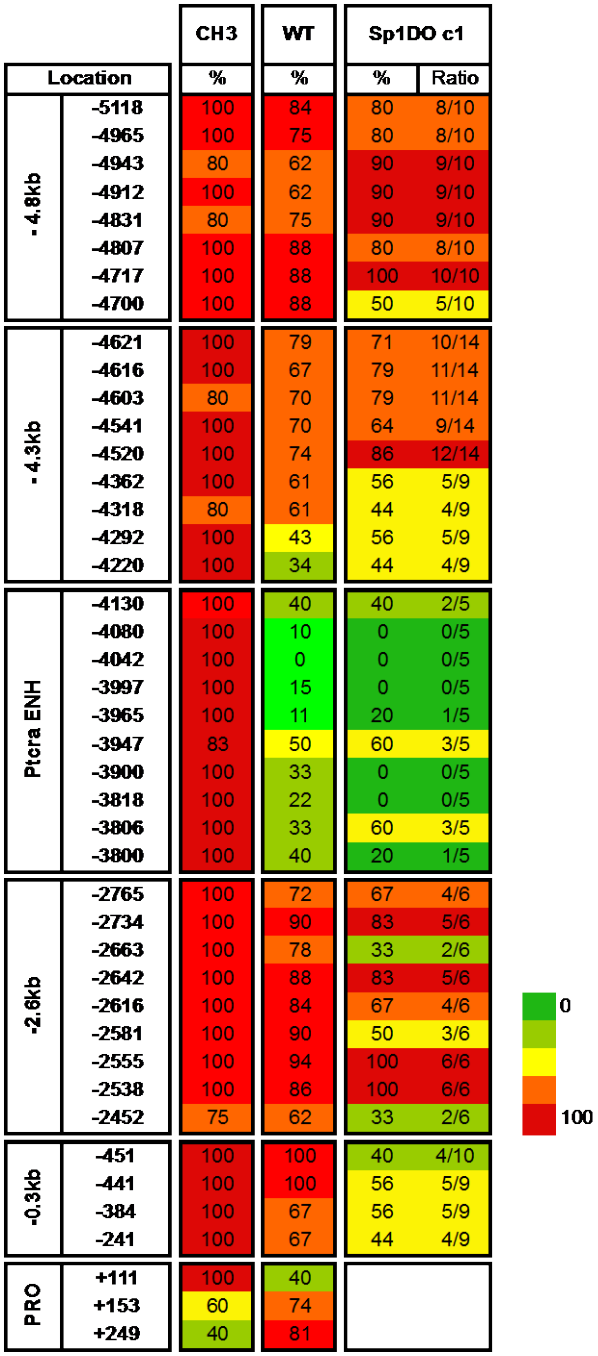
**B**



| | | CH3 | WT | DSp1 c1 | | DSp1 c2 | | DSp1 c4 | | DSp1 c5 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | BAC | END | BAC | END | BAC | END | BAC | END |
| Ptcra ENH | -4130 | 100 | 40 | 50 | 22 | 100 | 17 | 50 | 50 | 60 | 67 |
| | -4080 | 100 | 10 | 17 | 0 | 20 | 0 | 71 | 10 | 10 | 17 |
| | -4042 | 100 | 0 | 33 | 11 | 60 | 17 | 14 | 10 | 60 | 33 |
| | -3997 | 100 | 15 | 0 | 0 | 40 | 6 | 57 | 30 | 10 | 0 |
| | -3965 | 80 | 11 | 67 | 11 | 60 | 50 | 43 | 60 | 30 | 50 |
| | -3947 | 80 | 50 | 67 | 56 | 40 | 67 | 100 | 78 | 70 | 67 |
| | -3900 | 100 | 33 | 67 | 22 | 20 | 22 | 71 | 38 | 60 | 33 |
| | -3818 | 100 | 22 | 0 | 11 | 40 | 6 | 29 | 25 | 22 | 17 |
| | -3806 | 100 | 33 | 60 | 33 | 60 | 44 | 71 | 13 | 44 | 50 |
| | -3800 | 100 | 40 | 40 | 22 | 40 | 39 | 71 | 13 | 67 | 17 |

## Figure 2-11

## The pTa Enhancer Mark is Reestablished in an Sp1 Double Mutant BAC



## Figure 2-12

## pTa BAC Mutants Methylation Data Summarized

| | Location | TAG | mMyb | mCSL | mEbox4 | mEbox2 | mSp1 | DSp1 | Sp1DO |
|---|---|---|---|---|---|---|---|---|---|
| | | % | % | % | % | % | % | % | % |
| -4.8kb | -5118 | 84 | 85 | 80 | 76 | 93 | 85 | 73 | 80 |
| | -4965 | 75 | 83 | 85 | 76 | 100 | 79 | 83 | 80 |
| | -4943 | 62 | 71 | 70 | 82 | 56 | 67 | 75 | 90 |
| | -4912 | 62 | 94 | 89 | 100 | 100 | 78 | 71 | 90 |
| | -4831 | 75 | 100 | 85 | 94 | 78 | 88 | 92 | 90 |
| | -4807 | 88 | 91 | 87 | 94 | 100 | 89 | 79 | 80 |
| | -4717 | 88 | 93 | 86 | 76 | 100 | 90 | 65 | 100 |
| | -4700 | 88 | 77 | 82 | 65 | 61 | 82 | 61 | 50 |
| -4.3kb | -4621 | 79 | 50 | 87 | 91 | 93 | 64 | 75 | 71 |
| | -4616 | 67 | 60 | 85 | 75 | 80 | 63 | 71 | 79 |
| | -4603 | 70 | 56 | 77 | 75 | 67 | 63 | 64 | 79 |
| | -4541 | 70 | 57 | 72 | 81 | 58 | 64 | 75 | 64 |
| | -4520 | 74 | 66 | 71 | 84 | 62 | 70 | 89 | 86 |
| | -4362 | 61 | 49 | 66 | 66 | 58 | 55 | 76 | 56 |
| | -4318 | 61 | 41 | 63 | 63 | 70 | 51 | 76 | 44 |
| | -4292 | 43 | 34 | 39 | 34 | 21 | 38 | 53 | 56 |
| | -4220 | 34 | 41 | 22 | 38 | 48 | 37 | 42 | 44 |
| Ptcra ENH | -4130 | 40 | 31 | 21 | 64 | 32 | 36 | 65 | 40 |
| | -4080 | 10 | 54 | 12 | 9 | 0 | 32 | 30 | 0 |
| | -4042 | 0 | 34 | 13 | 9 | 23 | 17 | 42 | 0 |
| | -3997 | 15 | 11 | 19 | 18 | 0 | 13 | 27 | 0 |
| | -3965 | 11 | 41 | 24 | 27 | 26 | 26 | 50 | 20 |
| | -3947 | 50 | 57 | 50 | 27 | 62 | 53 | 69 | 60 |
| | -3900 | 33 | 19 | 45 | 55 | 38 | 26 | 55 | 0 |
| | -3818 | 22 | 4 | 26 | 27 | 10 | 13 | 23 | 0 |
| | -3806 | 33 | 67 | 39 | 36 | 51 | 50 | 59 | 60 |
| | -3800 | 40 | 44 | 34 | 36 | 33 | 42 | 55 | 20 |
| -2.6kb | -2765 | 72 | 80 | 72 | 61 | 81 | 76 | 81 | 67 |
| | -2734 | 90 | 87 | 97 | 94 | 86 | 88 | 90 | 83 |
| | -2663 | 78 | 79 | 78 | 61 | 86 | 78 | 62 | 33 |
| | -2642 | 88 | 94 | 94 | 82 | 81 | 91 | 78 | 83 |
| | -2616 | 84 | 83 | 78 | 82 | 93 | 83 | 63 | 67 |
| | -2581 | 90 | 85 | 88 | 83 | 97 | 88 | 97 | 50 |
| | -2555 | 94 | 92 | 100 | 89 | 81 | 93 | 98 | 100 |
| | -2538 | 86 | 77 | 88 | 67 | 69 | 81 | 90 | 100 |
| | -2452 | 62 | 74 | 55 | 59 | 47 | 68 | 22 | 33 |
| -0.3kb | -451 | 100 | 94 | 64 | 50 | 25 | 97 | 72 | 40 |
| | -441 | 100 | 100 | 75 | 50 | 100 | 100 | 79 | 56 |
| | -384 | 67 | 94 | 100 | 25 | 88 | 81 | 89 | 56 |
| | -241 | 67 | 67 | 92 | 75 | 100 | 67 | 96 | 44 |
| PRO | +111 | 40 | 55 | 24 | 67 | 36 | 48 | 55 | |
| | +153 | 74 | 86 | 71 | 83 | 58 | 80 | 84 | |
| | +249 | 81 | 85 | 90 | 83 | 51 | 83 | 61 | |

0

100

## REFERENCES


Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., *et al.* (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315-326.

Bossard, P., and Zaret, K.S. (1998). GATA transcription factors as potentiators of gut endoderm differentiation. Development *125*, 4909-4917.

Brandeis, M., Frank, D., Keshet, I., Siegfried, Z., Mendelsohn, M., Nemes, A., Temper, V., Razin, A., and Cedar, H. (1994). Sp1 elements protect a CpG island from de novo methylation. Nature *371*, 435-438.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., *et al.* (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences of the United States of America *107*, 21931-21936.

Dillon, N. (2012). Factor mediated gene priming in pluripotent stem cells sets the stage for lineage specification. BioEssays : news and reviews in molecular, cellular and developmental biology *34*, 194-204.

Gualdi, R., Bossard, P., Zheng, M., Hamada, Y., Coleman, J.R., and Zaret, K.S. (1996). Hepatic specification of the gut endoderm in vitro: cell signaling and transcriptional control. Genes & development *10*, 1670-1682.

Hajkova, P., Erhardt, S., Lane, N., Haaf, T., El-Maarri, O., Reik, W., Walter, J., and Surani, M.A. (2002). Epigenetic reprogramming in mouse primordial germ cells. Mechanisms of development *117*, 15-23.

Heinz, S., and Glass, C.K. (2012). Roles of lineage-determining transcription factors in establishing open chromatin: lessons from high-throughput studies. Current topics in microbiology and immunology *356*, 1-15.

Jones, P.A., and Takai, D. (2001). The role of DNA methylation in mammalian epigenetics. Science *293*, 1068-1070.

Kafri, T., Ariel, M., Brandeis, M., Shemer, R., Urven, L., McCarrey, J., Cedar, H., and Razin, A. (1992). Developmental pattern of gene-specific DNA methylation in the mouse embryo and germ line. Genes & development *6*, 705-714.

Liber, D., Domaschenz, R., Holmqvist, P.H., Mazzarella, L., Georgiou, A., Leleu, M., Fisher, A.G., Labosky, P.A., and Dillon, N. (2010). Epigenetic priming of a pre-B cell-specific enhancer through binding of Sox2 and Foxd3 at the ESC stage. Cell stem cell *7*, 114-126.

Liu, J.K., DiPersio, C.M., and Zaret, K.S. (1991). Extracellular signals that regulate liver transcription factors during hepatic differentiation in vitro. Molecular and cellular biology *11*, 773-784.

Macleod, D., Charlton, J., Mullins, J., and Bird, A.P. (1994). Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes & development *8*, 2282-2292.

McPherson, C.E., Shim, E.Y., Friedman, D.S., and Zaret, K.S. (1993). An active tissue-specific enhancer and bound transcription factors existing in a precisely positioned nucleosomal array. Cell *75*, 387-398.

Miranda, T.B., and Jones, P.A. (2007). DNA methylation: the nuts and bolts of repression. Journal of cellular physiology *213*, 384-390.

Monk, M., Boubelik, M., and Lehnert, S. (1987). Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. Development *99*, 371-382.

Ong, C.T., and Corces, V.G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. Nature reviews Genetics *12*, 283-293.

Pinkert, C.A., Ornitz, D.M., Brinster, R.L., and Palmiter, R.D. (1987). An albumin enhancer located 10 kb upstream functions along with its promoter to direct efficient, liver-specific expression in transgenic mice. Genes & development *1*, 268-276.

Reizis, B., and Leder, P. (1999). Expression of the mouse pre-T cell receptor alpha gene is controlled by an upstream region containing a transcriptional enhancer. The Journal of experimental medicine *189*, 1669-1678.

Reizis, B., and Leder, P. (2001). The upstream enhancer is necessary and sufficient for the expression of the pre-T cell receptor alpha gene in immature T lymphocytes. The Journal of experimental medicine *194*, 979-990.

Reizis, B., and Leder, P. (2002). Direct induction of T lymphocyte-specific gene expression by the mammalian Notch signaling pathway. Genes & development *16*, 295-300.

Rothenberg, E.V. (2014). The chromatin landscape and transcription factors in T cell programming. Trends in immunology *35*, 195-204.

Sachs, M., Onodera, C., Blaschke, K., Ebata, K.T., Song, J.S., and Ramalho-Santos, M. (2013). Bivalent chromatin marks developmental regulatory genes in the mouse embryonic germline in vivo. Cell reports *3*, 1777-1784.

Smale, S.T. (2010). Pioneer factors in embryonic stem cells and differentiation. Current opinion in genetics & development *20*, 519-526.

Szutorisz, H., Canzonetta, C., Georgiou, A., Chow, C.M., Tora, L., and Dillon, N. (2005). Formation of an active tissue-specific chromatin domain initiated by epigenetic marking at the embryonic stem cell stage. Molecular and cellular biology *25*, 1804-1820.

Takeuchi, A., Yamasaki, S., Takase, K., Nakatsu, F., Arase, H., Onodera, M., and Saito, T. (2001). E2A and HEB activate the pre-TCR alpha promoter during immature T cell development. Journal of immunology (Baltimore, Md : 1950) *167*, 2157-2163.

Tremblay, M., Herblot, S., Lecuyer, E., and Hoang, T. (2003). Regulation of pT alpha gene expression by a dosage of E2A, HEB, and SCL. The Journal of biological chemistry *278*, 12680-12687.

Vastenhouw, N.L., and Schier, A.F. (2012). Bivalent histone modifications in early embryogenesis. Current opinion in cell biology *24*, 374-386.

Voigt, P., Tee, W.-W., and Reinberg, D. (2013). A double take on bivalent promoters. Genes & development *27*, 1318-1338.

Wang, Z., Engler, P., Longacre, A., and Storb, U. (2001). An efficient method for high-fidelity BAC/PAC retrofitting with a selectable marker for mammalian cell transfection. Genome research *11*, 137-142.

Warming, S., Costantino, N., Court, D.L., Jenkins, N.A., and Copeland, N.G. (2005). Simple and highly efficient BAC recombineering using galK selection. Nucleic acids research *33*, e36.

Wu, S.C., and Zhang, Y. (2010). Active DNA demethylation: many roads lead to Rome. Nature reviews Molecular cell biology *11*, 607-620.

Xu, J., Pope, S.D., Jazirehi, A.R., Attema, J.L., Papathanasiou, P., Watts, J.A., Zaret, K.S., Weissman, I.L., and Smale, S.T. (2007). Pioneer factor interactions and unmethylated CpG dinucleotides mark silent tissue-specific enhancers in embryonic stem cells. Proceedings of the National Academy of Sciences of the United States of America *104*, 12377-12382.

Xu, J., Watts, J.A., Pope, S.D., Gadue, P., Kamps, M., Plath, K., Zaret, K.S., and Smale, S.T. (2009). Transcriptional competence and the active marking of tissue-specific enhancers by defined transcription factors in embryonic and induced pluripotent stem cells. Genes & development *23*, 2824-2838.

# Chapter 3

# Quantitative Analysis of Transcription Dynamics in Pluripotent and Differentiated Cells

## ABSTRACT

Cellular identity is a direct consequence of the transcriptional output for any given cell type. Understanding the regulatory properties of the genes that define each cell type is of fundamental importance for understanding pluripotency and tissue specificity. In this study, a detailed characterization of the promoter properties for tissue-specific genes in embryonic stem cells (ESC) and four terminally differentiated cell types was performed. The primary differentiated cell types, cortical neurons, double positive thymocytes, bone marrow-derived macrophages and hepatocytes are representative of all germ layers. Using deep chromatin RNA sequencing (~500 million mapped reads), allowed for accurate quantification of transcripts for genes expressed at very low levels. This provides us with the ability to distinguish genes that exhibit a broad dynamic range among cell types from those whose dynamic range of expression is more limited. We carefully analyzed the DNA and histone methylation patterns at the promoters of the most dynamically regulated genes. Our findings show striking cell type-specific differences in the fundamental promoter properties (CpG-island versus low CpG promoters) of dynamically regulated genes. Furthermore, in contrast to the mechanistic trends that have been described in studies that group the most dynamically regulated genes with those whose expression levels fluctuate only modestly, consistent properties can be attributed to specific gene classes, thereby adding clarity to our knowledge of the strategies used for the dynamic regulation of cell type-specific gene expression.

## INTRODUCTION

The full complement of transcribed RNA in a given cell, i.e. the transcriptome, determines the proteins to be translated that will collectively impart both the form and function of a cell. To understand what drives functional outcomes and phenotypic differences is one of the most fundamental aspects of biology. Interrogating not only what genes are expressed but also the level of their expression is critical in order to continue making strides in our understanding of cellular identity.

Much effort has been focused on understanding the epigenetic profiles of key developmental regulators in both mouse and human (Gifford et al., 2013; Sachs et al., 2013; Varley et al., 2013; Vastenhouw and Schier, 2012; Xie et al., 2013). The advances in sequencing technology coupled with chromatin immunoprecipitation have led to the genome wide characterization of a host of histone modifications. A major finding was the presence of H3K4me3 and H3K27me3 at the promoters of developmental regulators (Bernstein et al., 2006; Voigt et al., 2013b). These marks, active and repressive, were shown to resolve to one or the other, H3K4me3 persisting at active genes and H3K27me3 at inactive ones (Voigt et al., 2013b). This lead to the hypothesis that developmental regulators are poised in order to allow for their appropriate activation or repression upon differentiation.

Efforts have also been focused on identifying and understanding tissue specific genes and their regulation during development (Carninci et al., 2006; Efroni et al., 2008; Ernst et al., 2011; Meister et al., 2010; Song et al., 2013; Zhu et al., 2008). Studies defining tissue specific or tissue 'restricted' genes often use limited fold differences, resulting in thousands of tissue specific genes and thus grouping genes with different degrees of dynamic expression into the same

category (Carninci et al., 2006; Efroni et al., 2008; Meister et al., 2010; Song et al., 2013). Subsequent analysis performed on these gene sets does not account for possible differences in the regulatory mechanisms necessary for such dynamic regulation. Many of these tissue specific studies have been performed using microarray technology, which imposes a number of limitations (Barrera et al., 2008; Carninci et al., 2006; Okazaki et al., 2002; Song et al., 2013; Su et al., 2004; Wei et al., 2005). Microarrays are subject to cross-hybridization, normalization and saturation issues that result in high background, difficulties in comparing across experiments and a limited dynamic range of expression.

With the recent advances in sequencing technology we can revisit these fundamental questions and refine our understanding of tissue specificity. RNA sequencing not only provides an opportunity to capture the complexity of the mammalian transcriptome but also allows us to more accurately quantify transcripts (Mutz et al., 2013). A clear separation between tissue specific genes that show modest fluctuations in gene expression and genes that exhibit a much broader range of expression, will provide a more nuanced approach and a better understanding of cell type specific expression patterns and the mechanisms that govern them.

Our understanding of tissue specific gene regulation comes from studies in DNA methylation in early embryogenesis. During the wave of de novo methylation tissue specific genes are thought to be methylated and assembled into silent chromatin, where they await decondensation by lineage specific factors (Jones and Takai, 2001; Kafri et al., 1992). Identification of tissue specific genes will allow us to interrogate the associated promoter properties and provide mechanistic insights into their regulation. Although tissue specificity inversely correlates with CpG island promoter content, this has been shown to vary from cell type to cell type (Deaton and Bird, 2011; Zhu et al., 2008). Again, with the separation of tissue

specific genes that are most dynamically regulated, we can refine our understanding of these concepts.

The embryonic stem cell transcriptome is of great importance because of the fundamental properties of ES cells. They can self renew indefinitely and differentiate into all embryonic tissues, making ES cells attractive for regenerative medicine. Moreover, with the capability to convert somatic cells to an ES like state, there is a possibility for personalized stem cell therapies (Dejosez and Zwaka, 2012; Hanna et al., 2010; Jones and Takai, 2001; Kafri et al., 1992; Thomson et al., 2011). So, a clear and highly quantitative analysis of the genes necessary for the ES cell state is critical. ES cells can give rise to many distinct cells types, and thus retain developmental plasticity making them a clear choice in any tissue specific analyses. In addition, it has been reported that ES cells show basal levels of expression, even at tissue specific genes and that specification is a result of the restriction in this widespread transcription (Efroni et al., 2008). Understanding the mechanisms that regulate tissue specific genes in ES cells will contribute to our understanding of the pluripotent state.

Here we use chromatin RNA-Seq to quantify the transcriptomes of embryonic stem cells, E14.5 cortical neurons, $CD4^+$ $CD8^+$ thymocytes, bone marrow-derived macrophages and hepatocytes in order to define tissue specific genes with a broad dynamic range in expression. We assess the promoter properties of these genes in embryonic stem cells and define the chromatin signatures associated with these genes.

# RESULTS

### *Analysis of Chromatin RNA-Sequencing*

We studied tissue specific gene transcription in cells representative of all germ layers. Using the mouse embryonic stem cell line CCE and four primary differentiated mouse cell types – cortical neurons, double positive thymocytes, bone marrow-derived macrophages and hepatocytes – we defined genes that are highly specific to each cell type. In order to define the actively transcribed portions of the genome in each cell type, we employed cellular fractionation (Wuarin and Schibler, 1994) to isolate chromatin-associated transcripts, previously shown to be suitable for the analysis of nascent transcription (Bhatt et al., 2012; Pandya-Jones and Black, 2009). The purity of the chromatin fraction was assessed by Western blot analysis of Histone H3 (figure 3-1A). Strand-specific cDNA libraries were prepared from the isolated chromatin RNA and subjected to high-throughput sequencing. We performed deep sequencing, ~500 million reads, across 2-3 biological replicates for each cell type. The distribution of reads throughout the whole gene structure, introns and exons, is indicative of active transcription (figure 3-1B). In this instance (figure 3-1B), exonic peaks are consistently higher than intronic peaks, which in part, is likely due to splicing events taking place on the chromatin.

### *RPKM Distribution of Coding Genes*

We calculated RPKM values as described in (Mortazavi et al., 2008) for all Refseq coding genes in each cell type and binned expression values for comparison across cell types. All cell types show a similar distribution of expression with the exception of hepatocytes (figure 3-

2A). At the lower end of the expression spectrum, between 5000 and 8000 genes had an RPKM lower than 0.01 in all five cell types. 615 genes showed no reads within their transcriptional unit in all cell types (figure 3-2A). Of those genes expressed at lower than 0.01 RPKM, almost 3000 were common to all cell types. Interestingly, ES cells and the differentiated cell types show similar numbers of non-expressed genes (figure 3-2A), arguing against a low level of basal transcription in ES cells as previously described (Efroni et al., 2008; Lienert et al., 2011). Between 30-40% of all genes have an RPKM greater than 1 in all cell types except hepatocytes, where it is 20% (figure 3-2A). Highly expressed genes follow a similar pattern, with genes with an RPKM of at least 5 accounting for ~15% of all genes in four cell types, but only 4% in hepatocytes (figure 3-2B).

### *Defining Tissue Specificity*

To characterize the extent to which genes are specific to a cell type, we calculated minimum fold differences between one cell type and all other cell types, with a minimum RPKM of 1 or 5 in the expressing cell type. The high depth of sequencing provided us with confidence in genes expressed at low levels, an RPKM as low as 0.01. Genes which are considered 100-fold specific are expressed at a minimum of 1 or 5 RPKM in the expressing cell type, while being expressed at 0.01 or 0.05 RPKM, respectively, in all other cell types. The majority of genes expressed exhibit a limited dynamic range of expression, with greater than 50% of all genes (>1 or >5 RPKM) falling between a range of 2-5 fold in either direction for all cell types. The same holds true for greater than 75% in embryonic stem cells, cortical neurons, double positive thymocytes, and bone marrow-derived macrophages in the same range (figure 3-3A). We

employed the same strategy to assess if there are genes that are specifically repressed in one cell type. Only a handful of genes are expressed highly (>1 or >5 RPKM) in four cell types while having less than 20% expression in the fifth cell type with no clear pattern or insights emerging based on function (figure 3-3B).

### *100 Fold Tissue Specific Genes*

In order to define the promoter properties associated with tissue specific genes, we selected genes that were expressed 100 fold higher in one cell type compared to all other cell types, with a minimum RPKM of 5 in the expressing cell type. In contrast with studies that use less stringent criteria for tissue specificity, often thousands of genes per cell type, we isolated genes with a much broader dynamic range of expression (Barrera et al., 2008). Here we define 39 embryonic stem cell specific genes, 100 E14.5 cortical neuronal specific genes, 56 $CD4^+$ $CD8^+$ thymocyte specific genes, 68 bone marrow derived macrophage specific genes and 215 hepatocyte specific genes (Figure 3-4, Table 3-1 – 3-5). Despite our stringent criteria, genes critical for the function of each cell type were included in our selection such as Pou5f1, Fgf4 and Dppa4 in embryonic stem cells, Cd8a, Cd4, Rag1 and Ikzf3 in $CD4^+$ $CD8^+$ thymocytes and Alb in hepatocytes (figure 3-4). Moreover, gene ontology analysis of these limited gene sets still resulted in biological processes highly specific to each cell type: In utero embryonic development, synaptic transmission, T-cell activation, inflammatory response, and organic acid metabolic process, for embryonic stem cells, E14.5 cortical neurons, $CD4^+$ $CD8^+$ thymocytes, bone marrow derived macrophages and hepatocytes, respectively (Table 3-11 – 3-15).

100 fold was the minimum criteria for inclusion in our tissue specific genes. We analyzed the absolute fold changes to better understand the nature of their expression. The median fold differences for tissue specificity range from 200-500 fold and the maximum fold differences range from 2000 – 80,000 (Figure 3-5, Table 3-1 – 3-5). These absolute fold changes are conservative estimate as we considered 0.05 as the lower end of expression in this analysis. Tissue specificity based on our criteria can be attained either by having little to no expression in all other cell types, or, by being expressed at a much higher level in one cell type compared to all other cell types. To identify which categories our tissue specific genes fall into, we plotted the expression in the tissue specific cell type versus the average expression in all other cell types. The majority of our tissue specific genes fall into the former category with little to no expression in all other cell types (Figure 3-6). There are a limited number of genes which show elevated levels of expression in one or more of the non-expressing cell types, but of these only 1 gene has an RPKM greater than 1 (Figure 3-6).

***Embryonic Stem Cells are More Closely Correlated with Cortical Neurons***

To determine the relationship between the expression profile of the pluripotent embryonic stem cells and our differentiated cells, we calculated the fold differences between our ES cells and each individual cell type for all genes >5 RPKM. When comparing ES cells to all other cell types, the total number of 100 fold ES cell specific genes is 39 (Figure 3-7A). When comparing ES cells directly to neuronal cells, this number increases to 84, compared to thymocytes 165, to macrophages 132 and to hepatocytes 330 (Figure 3-7A). This indicates that the expression profiles of ES cells and neuronal cells are more closely related than ES cells and

the other cell types analyzed, as fewer ES cell specific genes arise when comparing directly with neuronal cells than any other cell type. This also indicates ES cells are most distantly related to hepatocytes, as this direct comparison results in many more ES cell specific genes than with any other cell type.

Although these statements hold true for a fraction of the expressed genomes, it does not address global similarities in expression profiles. To this end we calculated the correlation in expression profiles for all cell types using all refseq coding genes. Interestingly the previous observations hold true. Embryonic stem cells are most closely related to cortical neurons in their overall expression profile, while hepatocytes show the largest dissimilarity among all cell types. $CD4^+$ $CD8^+$ thymocytes and bone marrow derived macrophages cluster together but separately from ES cells and neurons, which cluster together (Figure 3-7B).

***Striking Cell Type Specific Differences Observed in Promoter Properties***

To characterize to the promoter properties of our tissue specific genes we calculated the percentage of genes that have a CpG island promoter. Chromatin RNA-Seq allowed us to accurately define the transcription start sites for each tissue specific gene (Table 3-6 – 3-10). As previously reported there is an inverse relationship between CpG island promoter content and tissue specificity (Barrera et al., 2008; Carninci et al., 2006; Saxonov et al., 2006). The vast majority of housekeeping genes have a CGI promoter (Deaton and Bird, 2011; Saxonov et al., 2006). For $CD4^+$ $CD8^+$ thymocyte, bone marrow derived macrophage, and hepatocyte specific genes we observe the same phenomenon with less than 20%, 12% and 9% of genes having a CGI promoter, respectively (Figure 3-8). In contrast E14.5 cortical neuronal specific genes, with the

broadest dynamic range in expression, show a much higher percentage of genes with a CGI promoter, 84%, which is comparable to that of housekeeping genes (Figure 3-8). Embryonic stem cell specific genes fall in between with just over 60% having a CGI promoter (Figure 3-8).

### *Tissue Specific Genes Show Two Distinct DNA Methylation Promoter Patterns in Non-Expressing Cells*

Given the bimodal distribution of CGI promoter content, we examined the DNA methylation at the promoters of tissue specific genes in embryonic stem cells, neural progenitor cells and 6-week frontal cortex neuronal cells. This allowed us to interrogate the methylation status of tissue specific gene promoters in both expressing and a non-expressing cell types. We have previously shown that the overwhelming majority of our tissue specific genes show little to no expression in all other cell types. Using promoter coordinates for $CD4^+$ $CD8^+$ thymocyte and bone marrow derived macrophage specific genes we calculated the average DNA methylation profile for each gene in the aforementioned non-expressing cell types. A clear pattern emerged: tissue specific genes that harbor a CpG island promoter, although not expressed, remain unmethylated, while those without a CGI promoter remain heavily methylated (Figure 3-9).

### *Tissue Specific CGI Genes are Marked by Both H3K4me3 and H3K27me3 in Non-Expressing Cell Types*

Given the strict rule between CGI promoter containing genes and the lack of DNA methylation for tissue specific genes in non-expressing cell types, we assessed the presence of

histone modifications at $CD4^+$ $CD8^+$ thymocyte and bone marrow derived macrophage specific genes in ES cells. The presence of an H3K4me3 or H3K27me3 ChIP-Seq peak overlapping with our tissue specific promoters was determined and again a clear rule emerged. Tissue specific genes with CGI promoters are concomitantly marked with H3K4me3 and H3K27me3 in non-expressing cell types (Figure 3-10, 3-11). We took advantage of H3K4me2 and H3K27me3 ChIP-Seq data sets for thymocytes to see if these same rules applied in another differentiated cell type. We identified bone marrow derived macrophage specific promoters with the presence of these epigenetic modifications and again a clear pattern emerged. CGI promoter containing bone marrow derived macrophage specific genes are marked concomitantly by both active and repressive histone marks in thymocytes. Although these macrophage genes show no expression in thymocytes they are clearly marked by H3K4me2 and H3K27me3 (Figure 3-13).

### *Tissue Specific Genes with a Limited Dynamic Range Lack a Clear Epigenetic Profile in Non-Expressing Cell Types*

In order to determine if these rules hold true for all tissue specific genes, we categorized these same histone modifications in $CD4^+$ $CD8^+$ thymocytes specific genes that show a more limited dynamic range of gene expression. We took the promoters of genes whose expression ranged from 10-20 fold and 5-10 and found a breakdown in the rules previously observed at genes with a much higher dynamic range in gene expression (Figure 3-12). That is, CGI promoter genes are not consistently marked with H3K4me3 and H3K27me3 in non-expressing cell types. Moreover, this rule appears to break down even further, as we reduce the dynamic range of expression from 10-20 fold to 5-10 fold (Figure 3-12).

### *Embryonic Stem Cell Specific Genes Show Unique Phases of Activation*

To assess the role of our embryonic stem cell specific genes in reprogramming, we used chromatin RNA-Seq expression data from a time course of reprogramming (Chronis unpublished). We identified which of our ES cell specific genes are induced and at what stage of reprogramming, MEFs, 48hrs post induction, preiPS and ES. The majority of our defined ES cell specific genes are not expressed until an ES cell state is reached (Figure 3-14C). One third of our ES specific genes have some level of expression at the preiPS stage (Figure 3-14B). Two genes Fgf4 and L1td1, are expressed at very early stages, 48 hours post induction (Figure 3-14A).

## DISCUSSION

Using chromatin RNA-sequencing we categorized the transcriptome of mouse embryonic stem cells, E14.5 cortical neurons, $CD4^+$ $CD8^+$ thymocytes, bone marrow-derived macrophages and hepatocytes. The four latter cell types are primary mouse tissues representing all three germ layers. Chromatin RNA-seq gave us confidence in genes expressed at lower levels of expression resulting in a broad dynamic range of gene expression for a thorough quantitation. We then defined tissue specific genes with the largest dynamic range in gene expression for each cell type. Our analysis while stringent, still resulted in tissue specific genes highly related to cellular function. Although limited in number, these tissue specific genes contribute heavily to the overall function of each cell type. We characterized 39 embryonic stem cell specific genes, 100 E14.5 cortical neuronal specific genes, 56 $CD4^+$ $CD8^+$ thymocyte specific genes, 68 bone marrow derived macrophage specific genes and 215 hepatocyte specific genes. We assessed the overall expression profiles of all refseq coding genes and found a similar distribution of expression for all cell types apart from hepatocytes. Hepatocytes as an outlier appear to be a consistent trend and likely reflect the diversity and complexity of the cell type specific functional requirements (Kmiec, 2001; Schwartz et al., 2014; Sun et al., 2013). It has previously been proposed the there is a basal level of gene expression in mouse ES cells and through differentiation there is a restriction in the expression repertoire (Efroni et al., 2008). However, even with deep sequencing, we do not observe this phenomenon in mouse ES cells, which show comparable number of genes expressed to all four differentiated cell types.

We addressed the relationship between the transcriptome of ES cells and the differentiated cell types. Analyzing genes with an expression greater than 5 RPKM we found ES cells to be most similar to cortical neurons while most dissimilar to hepatocytes. Upon

hierarchical clustering of the global transcriptomes we observed the same pattern. ES cells cluster together with and are more closely related to cortical neurons, while being most dissimilar to hepatocytes. This similarity in expression profile aligns with the 'default model' of neural induction, which states without extrinsic signals ectoderm and neuronal differentiation is the default pathway (Hemmati-Brivanlou and Melton, 1997; Munoz-Sanjuan and Brivanlou, 2002; Tropepe et al., 2001).

Chromatin RNA-Seq allowed us to accurately define transcriptional start sites and manually curate of all our tissue specific genes. After careful designation of promoters we observed striking cell type specific differences in CpG island promoter content in the most dynamically regulated genes. In line with previous studies tissue specificity is inversely correlated with CGI promoter content in most of our cell types. However, cortical neurons display a very high percentage of genes with a CGI promoter, 84%. We speculate this may be due to the need for neuronal specific genes to be both induced and shut off rapidly. Previous work has shown that CpG islands cause destabilization of nucleosomes and can provide a constitutively active chromatin environment facilitating rapid induction of transcription in response to external stimuli (Bhatt et al., 2012; Ramirez-Carrozzi et al., 2009).

Using available and unpublished genome wide DNA methylation and histone ChIP-Seq data sets we categorized the epigenetic profile of tissue specific genes. In our analysis we chose to focus on tissue specific genes with the largest dynamic range in expression. This resulted in a clear and consistent epigenetic profile for tissue specific genes in non-expressing cell types. Tissue specific genes with CpG island promoters are consistently unmethylated and marked by H3K4me3 and H3K27me3 in non-expressing cell types. This rule breaks down when considering tissue specific genes with a more narrow dynamic range in expression.

There is a vast amount of interest in bivalency and the contexts in which it exists. The marking of our tissue specific genes in ES cells with H3K4me3 and H3K27me3 is reminiscent of the bivalent domains at developmental regulators. In fact Gata3, a thymocyte specific gene in our data set, was shown to be upregulated in Eed knockout ES cells (Boyer et al., 2006b). However, the majority of our tissue specific genes are just that, highly specific to the cell type and show no enrichment for lineage determining characteristics. In this sense our tissue specific bivalent genes are not developmental regulators.

Key findings early on suggested that bivalency primarily occurred in ES cells but it has also been observed in non pluripotent cells (Voigt et al., 2013b). Here we report that the bivalent mark H3K4me2 and H3K27me3 is present at CpG island containing bone marrow macrophage specific genes in $CD4^+$ $CD8^+$ thymocytes. In this sense it is unlikely to represent a poised state, as the bone marrow macrophage gene is unlikely to be expressed in a lineage committed double positive thymocyte. For instance, Emilin2, a BMDM specific CGI promoter gene is marked by K4me3/K27me3 in ES cells and shows no expression. Emilin2 is marked by K4me2/K27me3 in thymocytes and is also silent. Although we cannot rule out that Emilin2 may be expressed at a later time in thymocyte development and thus be poised.

Consistent with existing data H3K4me3 is highly correlated with the presence of a CpG island (Voigt et al., 2013b). In almost all cases our tissue specific genes with a CpG island promoter are also marked with H3K4me3. Interestingly we see just as high a correlation with CGI promoter containing tissue specific genes and H3K27me3. Data shows the H3K27me3 marks only a subset of CGIs (Voigt et al., 2013b). It is posited this is due to the transcriptional state of a given gene. For instance, if H3K4me3 marks a CGI gene but the transcriptional activator(s) necessary for productive transcription are absent, PRC2 can be recruited to the

unmethylated CpG island and catalyze H3K27me3. Our observations provide support for this hypothesis. H3K4me3 and H3K27me3 consistently mark tissue specific CGI genes in non-expressing cell types. It is likely the lack of lineage specific transcriptional activators and the presence of the polycomb complexes the lead to silencing of these tissue specific CGI genes. Additional studies of our tissue specific genes in polycomb knockouts would yield added insight to these hypotheses.

We also assessed the timing of the activation of our ES cell specific genes during reprogramming. Interestingly there is a trend towards CGI containing genes to be activated early. Both genes, L1td1 and Fgf4, which are activated within 48hrs of induction, have a CGI promoter. 85% of genes showing expression in preiPS cells contain CGI promoters and 50% of the remaining genes that show expression in ES cells have a CpG island. The presence of a CGI, which has been previously shown to result in destabilization of nucleosomes and a permissive chromatin environment (Ramirez-Carrozzi et al., 2009), may facilitate the efficient activation of the genes thereby permitting rapid activation.

In summary, we used deep chromatin RNA-seq to provide a highly quantitative view of the transcriptome in mouse ES cells and four primary differentiated cell types. We then defined tissue specific genes with the broadest dynamic range in gene expression. With manual curation of transcriptional start sites we identified striking cell type differences in CpG island promoter content. Selecting the most dynamically regulated genes allowed us to uncover clear chromatin signatures of tissue specific genes in non-expressing cells, specifically the presence of H3K4me3 and H3K27me3. This provides additional evidence for the transcriptional state of CGI genes as a key determinant of the ability of H3K4me3 and H3K27me3 to coexist.

## MATERIALS AND METHODS

### Cell Culture

CCE ES cells were maintained in standard ES growth media. Cells were maintained in Knockout DMEM plus 1% L-glutamine, 1% pen/strep, 1% non-essential amino acids, 15% ES certified FBS (Omega Scientific), and 1,000 units/ml ESGRO (Lif, Millipore). The CCE ES cell line was maintained on gelatin-coated tissue culture flasks.

Cortical neurons were isolated from a pregnant mouse at E14.5. Embryo brains were isolated, lobes separated and meninges removed. The cortices were then dissected and trypsinized washed, counted and plated in order for projections to extend. Cells were harvested after 5 days.

CD4+ CD8+ thymocytes were isolated from the thymus of 6-8 week old mice. Thymus was isolated, minced into a single cell suspension in PBS and passed through a 70-micron filter. Cells were pelleted for 5 minutes at 1100rpm resuspended in PBS + 1% FBS and filtered again at 70 microns. Cells were counted and 1 million cells were used for each staining control. No Ab Stain, CD4-PE (1:500), CD8-APC (1:200) and CD4-PE, CD8-APC. The remaining cells were double stained for CD4-PE CD8-APC (Stain in 500uL for every $50x10^6$ cells). Cells were incubated in Ab cocktails for 20min at $4^oC$ in the dark. Cells were then washed twice with MAC buffers (PBS + 1% FBS). To sort cells were resuspended in PBS + 5% FBS. Cells were sorted on $CD4^+$ $CD8^+$ gating according to controls.

Bone marrow derived macrophages were isolated from the femur of 6-8 week old mice. Femur was flushed to push bone marrow into a petri dish with PBS. Liquid was transferred to a falcon tube and spun for 10 minutes at 1500rpm. Cells were resuspended in RBC lysis buffer for 5

minutes at room temperature. Cells were pelleted, washed with PBS and spun down 10minutes 1500rpm. Cells were then resuspended in bone marrow conditioning media. 10-25 million cells were plated. Cells were culture for 6-8 days and harvested at confluency.

Hepatocytes were isolated from 6-8week old mice by perfusion. Perfuse vena cava with 45ml Perfusion Media and 45mL Digest Media, 8ml/min transferred to 10cm dish of cold WE/PS/FBS. The liver was minced between cell scrapers and filtered through 100um cell filter into 50ml falcon. Then resuspend in 45 ml RT WE/FBS/PS, spun 5min/50g/4°C and resuspend in 10ml cold RSB+, quickly spun down 50g/3min/4°C then resuspend in 40ml cold RSB+ dounce. 25 strokes on spinning drill press and checked for nuclei.

**Cellular Fractionation and RNA Isolation**

Fractionation was performed as in (Bhatt et al., 2012). Briefly, 2.0 x $10^7$ cells were isolated, washed with 1mL cold PBS/1mM EDTA and spun for 10mins, 1K, $4^oC$. The pellet was resuspended in 200ul cold cytoplasmic lysis buffer and placed on ice for 10 mins. The lysate was then layered onto 500ul sucrose buffer and pelleted, 10mins, 14K, $4^oC$. The nuclei pellet was resuspend in 200ul glycerol buffer before adding 200ul cold nuclei lysis buffer and vortexing thoroughly. The mixture was then incubated on ice for 2mins then spun for 2mins, 14K, $4^oC$. The remaining chromatin pellet as resuspended in 50ul 1X PBS and added to TRIzol and resuspended thoroughly before purification by RNeasy column. All samples were resuspended in 30-50ul RNase-free water. Chromatin RNA was subjected to rRNA removal with the Mouse/Human Ribominus kit (Invitrogen, Catalog No. K1550-02) concentrated using glycogen and resuspended in 14ul RNase-free water. RNA quality was checked via Bioanalyzer.

**Library Preparation and Sequencing**

200ng of rRNA depleted RNA was used to generate strand specific cDNA libraries using the Illumina TruSeq RNA Sample Prep Kit v2 with the inclusion of "deoxyuridine triphosphate (dUTP)" method (Levin et al., 2010). Sequencing was performed on Illumina HiSeq 2000 as single-end 50base pair runs.

**RNA-Sequence Mapping and Analysis**

Reads were aligned to the mouse mm9 reference genome with Tophat (Trapnell et al., 2010) by using most default parameters. Alignments were restricted to uniquely mapping reads, with two possible mismatches permitted. RPKM values were calculated as described by (Mortazavi et al., 2008) for mm9 RefSeq genes (Pruitt et al., 2007). Seqmonk was used to sum reads within each transcriptional unit and Excel used to calculate final RPKM values. Genome tracks were generated with BEDTools genomeCoverageBed tool and visualized in the UCSC genome browser.

**DNA methylation ChIP-seq and Histone Modification Datasets**

Published DNA methylation data sets and ChIP-seq histone modification data sets were obtained from GEO, GSE44092 (Vincent et al., 2013), GSE30206 (Stadler et al., 2011), GSE47966 (Lister et al., 2013), and GSE31235 (Zhang et al., 2012). Average DNA methylation

profiles were calculated using BEDTools intersectBed to return DNA methylation values for individual CpGs within tissue specific promoter regions after filtering for a minimum of 3 sequencing reads for each individual CpG dinucleotide.

## FIGURE LEGENDS

**Figure 3-1. Chromatin Associated Transcripts.**

(A) A diagram of the cellular fractionation process resulting in the isolation of chromatin associated transcripts. Transcripts include a variety of species from unspliced to fully spliced. A representative western blot of ~ 20 million E14.5 cortical neurons fractionated into its cellular components displayed in the diagram. Cytoplasmic lysis was performed using 0.15% NP40. Nuclear lysis was performed with 1M Urea and 1% NP40. The nuclear lysate is separated from the precipitated chromatin pellet. Each fraction is verified by the presence of proteins known to localize to each fraction. From left to right: lane 1 – Cytoplasm β-actin, lane 2 – Nucleoplasm SNRP70, and lane 3 – Chromatin histone H3. (B) A visualization of chromatin RNA-sequencing for each cell type for the mitogen-activated protein kinase associated protein 1 gene, mapkap1. The direction of transcription is indicated by blue chevron marks throughout the transcript. Visualization was performed using bioinformatics tool genomeCoverageBed from bedtools. Reads are present along the length of the gene including introns, indicative of the chromatinized nature of the transcript.

**Figure 3-2. RPKM Distribution of All Refseq Coding Genes.**

(A) RPKMs were calculated, as described in the materials and methods above, for all refseq coding genes. The genes were then binned based on those values. (B) The RPKM distribution of the most highly expressed genes, RPKM of 5 - >50.

**Figure 3-3. The Majority of Genes Lie Within a Limited Dynamic Range of Expression.**

(A) Tissue specificity based on fold differences was determined at two thresholds, a minimum RPKM of 1 (Red) or 5 (Blue), in the expressing cell type. For each gene the minimum fold change was calculated between the expressing cell type and all other cell types. For example, embryonic stem cell genes that are in the 100-fold category must be expressed at least 100 fold more in ES cells than all other cell types. (B) Tissue specificity based on genes that are selectively repressed in one cell type compared to all other cell types. Tissue specificity was calculated based on the same RPKM thresholds but in the non-expressing cell types. For example, embryonic stem cell genes in the 100-fold category must be expressed at least 100 fold lower in ES cells compared to all other cell types i.e. selectively repressed in ES cells.

**Figure 3-4. 100 Fold Specific Genes Include Genes Known to be Important for Function.**

Diagram depicts a cluster analysis of all genes that are 100 fold specific in each of the cell types. Expression values for the tissue specific genes are shown along side the expression values for the same genes in all other cell types. Values were $\log_{10}$ transformed and clustered by cell type and descending expression values. Values are color-coded based on expression percentile. For each cell type genes known to be critical for the function were found e.g. Pou5f1 essential for pluripotency in ES cells, Rag1 critical in thymocytes and Alb key in hepatocytes.

**Figure 3-5. Absolute Fold Changes for Tissue Specific Genes.**

Box and whisker plots showing absolute fold differences for all 100-fold tissue specific genes in each cell type.

**Figure 3-6. Absolute Expression Values for 100-Fold Tissue Specific Genes.**

(A) The absolute expression values of all 100 fold embryonic specific genes. The graph shows the average value of each ES specific gene in all other cell types. ES specific RPKM is on the y-axis and an average RPKM value for the same gene in all other cell types is on the x-axis. An RPKM threshold is set at 0.06, slightly above that of the minimum believable expression value of 0.05. In red are genes with an RPKM above the minimum threshold. In black are genes that do no exceed the minimum threshold. Percent of genes above and below the threshold is shown. (B) The absolute expression values of all 100 fold E14.5 cortical neuronal genes. The graph shows the average value of each E14.5 cortical neuronal gene in all other cell types. E14.5 cortical neuronal specific RPKM is on the y-axis and an average RPKM value for the same gene in all other cell types is on the x-axis. (C) The absolute expression values of all 100 fold $CD4^+$ $CD8^+$ thymocyte specific genes. The graph shows the average value of each $CD4^+$ $CD8^+$ thymocyte specific gene in all other cell types. $CD4^+$ $CD8^+$ thymocyte specific RPKM is on the y-axis and an average RPKM value for the same gene in all other cell types is on the x-axis. (D) The absolute expression values of all 100 fold bone marrow derived macrophage specific genes. The graph shows the average value of each bone marrow derived macrophage specific gene in all other cell types. Bone marrow derived macrophage specific RPKM is on the y-axis and an average RPKM value for the same gene in all other cell types is on the x-axis. (E) The absolute

expression values of all 100 fold hepatocyte specific genes. The graph shows the average value of each hepatocyte specific gene in all other cell types. Hepatocyte specific RPKM is on the y-axis and an average RPKM value for the same gene in all other cell types is on the x-axis.

**Figure 3-7. Embryonic Stem Cell Specificity in Individual Cell Types Versus All Cell Types**

(A) Tissue specificity was calculated based on fold at a minimum expression threshold of 5 RPKM in embryonic stem cells. The graph shows a comparison of tissue specificity when comparing ES cells to all cell types and ES cells to individual cell types. For example the left most blue bar shows genes which are expressed 100 fold more in ES cells compared to all other cell types. The left most red bar shows genes that are expressed 100 fold more in ES cells compared to E14.5 cortical neurons alone. Blue: ES cells vs. all other cell types. Red: ES cells vs. E14.5 cortical neurons. Green: ES cells vs. $CD4^+$ $CD8^+$ thymocytes. Purple: ES cells vs. Bone marrow derived macrophages. Orange: ES cells vs. Hepatocytes. (B) Hierarchical clustering of all refseq coding genes shown as a dendrogram.

**Figure 3-8. Striking Difference in CpG Island Promoter Content.**

CpG island promoters are determined based on 1 base pair overlap between a CGI and the promoter (-500bp to +50bp) for all 100 fold specific genes in each cell type. Housekeeping genes are defined as genes that are expressed within 2-5 fold in all cell types.

**Figure 3-9. Tissue Specific Genes Show Two Distinct DNA Methylation Promoter Patterns in Non Expressing Cells.**

Average DNA methylation profiles were determined at promoters (-500bp to +50bp) for 100 fold $CD4^+$ $CD8^+$ thymocyte and 100 fold bone marrow macrophage specific genes. DNA methylation averages were calculated in embryonic stem cells, neural progenitor cells and 6-week frontal cortex neuronal cells (min 3 reads/CpG). Methylation levels are represented in a gradation of colors: Light green (0-20%), dark green (21-40), yellow (41-60%), orange (61-80%) and red (81-100%). Alongside the presence (blue) or absence of a CGI promoter is shown.

**Figure 3-10. Tissue Specific CGI Genes Have Active Histone Marks in Embryonic Stem Cells.**

The H3K4me3 profile in ES cells is shown for 100 fold $CD4^+$ $CD8^+$ thymocyte and 100 fold bone marrow macrophage specific gene promoters. Overlap of 1 bp required with an H3K4me3 peak to be considered positive. Average DNA methylation profile and CGI content shown alongside as in figure 3-9.

**Figure 3-11. Unmethylated Tissue Specific Genes are Associated with H3K27me3 in Embryonic Stem Cells.**

The H3K27me3 profile in ES cells is shown for 100 fold $CD4^+$ $CD8^+$ thymocyte and 100 fold bone marrow macrophage specific gene promoters. Overlap of 1 bp required with an H3K4me3 peak to be considered positive. Average DNA methylation profile and CGI content shown alongside as in figure 3-9.

**Figure 3-12. Tissue Specific Genes with a Limited Dynamic Range Lack a Clear Epigenetic Profile.**

DNA methylation in Embryonic Stem Cells, neural progenitor cells and 6-week frontal cortex neuronal cells, H3K4me3 and H3K27me3 profiles in Embryonic Stem Cells is shown for $CD4^+$ $CD8^+$ thymocyte genes 10-20 fold and 5-10 fold. Presence or absence of CGI is shown as in figure 3-9.

**Figure 3-13. Unmethylated Tissue Specific Genes are Associated with H3K4me3 and H3K27me3 in Other Non-Expressing Cell Types**

As in figure 3-11. Includes H3K4me2 and H3K27me3 profiles for 100 fold $CD4^+$ $CD8^+$ thymocyte and 100 fold bone marrow macrophage specific gene promoters in thymocytes alongside.

**Figure 3-14. Embryonic Stem Cell Specific Genes Show Unique Phases of Activation in Reprogramming.**

Chromatin RNA-Seq expression levels of 100 fold Embryonic Stem Cell specific genes through reprogramming from MEFs to iPS. (A) ES cell specific genes with expression at 48hrs after induction of reprogramming. (B) ES cell specific genes with expression at the preiPS stage. (C) ES cell specific genes with expression at the ES cell stage.

**Table 3-1. 100 Fold Embryonic Stem Cell Specific Genes.**

Gene lists containing refseq ID, gene name, RPKM value in all cell types, minimum fold and the associated p-value.

**Table 3-2. 100 Fold E 14.5 Cortical Neuron Specific Genes**

Gene lists containing refseq ID, gene name, RPKM value in all cell types, minimum fold and the associated p-value.

**Table 3-3. 100 Fold $CD4^+$ $CD8^+$ Thymocyte Specific Genes**

Gene lists containing refseq ID, gene name, RPKM value in all cell types, minimum fold and the associated p-value.

**Table 3-4. 100 Fold Bone Marrow Derived Macrophage Specific Genes**

Gene lists containing refseq ID, gene name, RPKM value in all cell types, minimum fold and the associated p-value.

**Table 3-5. 100 Fold Hepatocyte Specific Genes**

Gene lists containing refseq ID, gene name, RPKM value in all cell types, minimum fold and the associated p-value.

**Table 3.6. 100 Fold Embryonic Stem Cell Specific Promoter Coordinates**

Gene lists containing refseq ID, gene name, chromosome, genomic start, end and strand. Genomic start and end are promoter coordinates used to determine DNA methylation profiles and overlaps with histone modifications.

**Table 3-7. 100 Fold E14.5 Cortical Neuron Specific Promoter Coordinates**

Gene lists containing refseq ID, gene name, chromosome, genomic start, end and strand. Genomic start and end are promoter coordinates used to determine DNA methylation profiles and overlaps with histone modifications.

**Table 3-8. 100 Fold $CD4^+$ $CD8^+$ Thymocyte Specific Promoter Coordinates**

Gene lists containing refseq ID, gene name, chromosome, genomic start, end and strand. Genomic start and end are promoter coordinates used to determine DNA methylation profiles and overlaps with histone modifications.

**Table 3-9. 100 Fold Bone Marrow Derived Macrophage Specific Promoter Coordinates**

Gene lists containing refseq ID, gene name, chromosome, genomic start, end and strand. Genomic start and end are promoter coordinates used to determine DNA methylation profiles and overlaps with histone modifications.

**Table 3-10. 100 Fold Hepatocyte Specific Promoter Coordinates**

Gene lists containing refseq ID, gene name, chromosome, genomic start, end and strand. Genomic start and end are promoter coordinates used to determine DNA methylation profiles and overlaps with histone modifications.

**Table 3-11. 100 Fold Embryonic Stem Cell Specific Gene Ontology**

Go terms associated with Embryonic Stem Cell specific genes. Top 10 highest p-values selected.

**Table 3-12. 100 Fold E14.5 Cortical Neuron Specific Gene Ontology**

Go terms associated with E14.5 Cortical Neuron specific genes. Top 10 highest p-values selected.

**Table 3-13. 100 Fold $CD4^+$ $CD8^+$ Thymocyte Specific Gene Ontology**

Go terms associated with CD4+ CD8+ Thymocyte specific genes. Top 10 highest p-values selected.

**Table 3-14. 100 Fold Bone Marrow Derived Macrophage Specific Gene Ontology**

Go terms associated with Bone Marrow Derived Macrophage specific genes. Top 10 highest p-values selected.

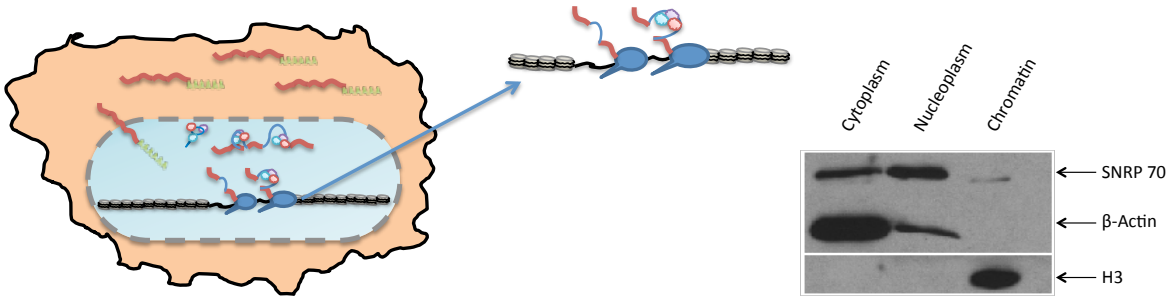
**Table 3-15. 100 Fold Hepatocyte Specific Gene Ontology**

Go terms associated with Hepatocyte specific genes. Top 10 highest p-values selected.

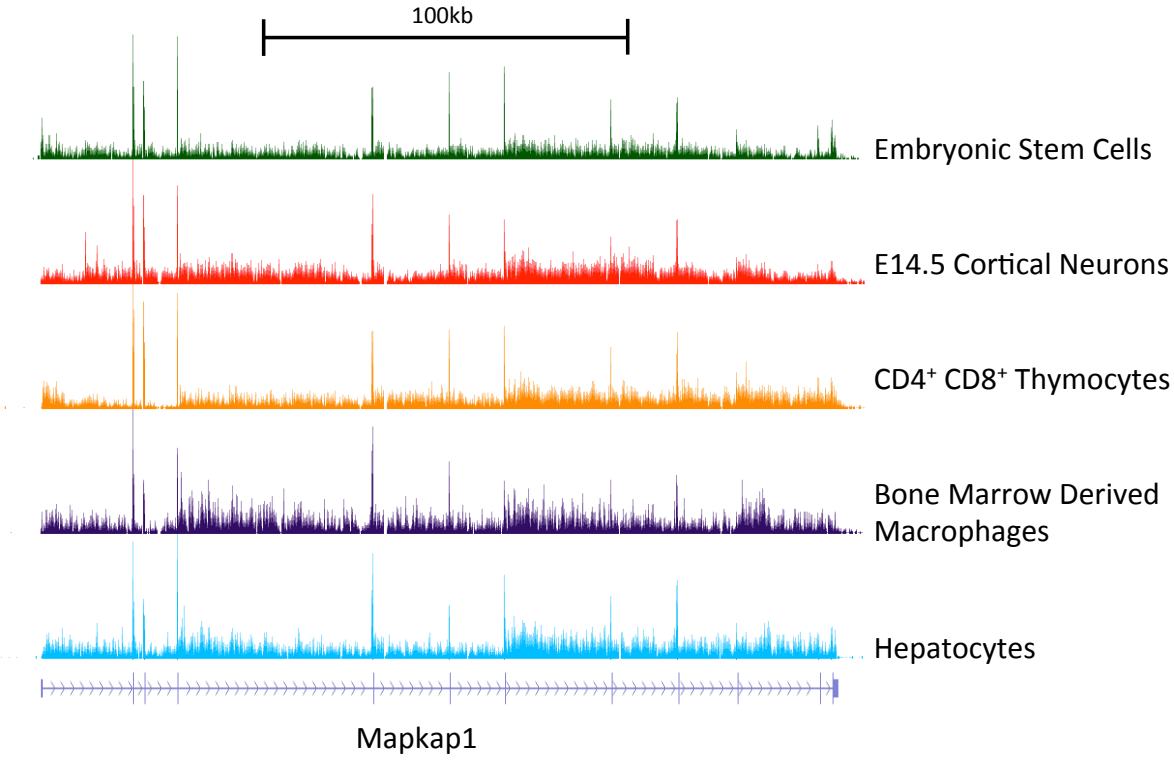
**Figure 3-1**

**Chromatin RNA-Sequencing**

**A**



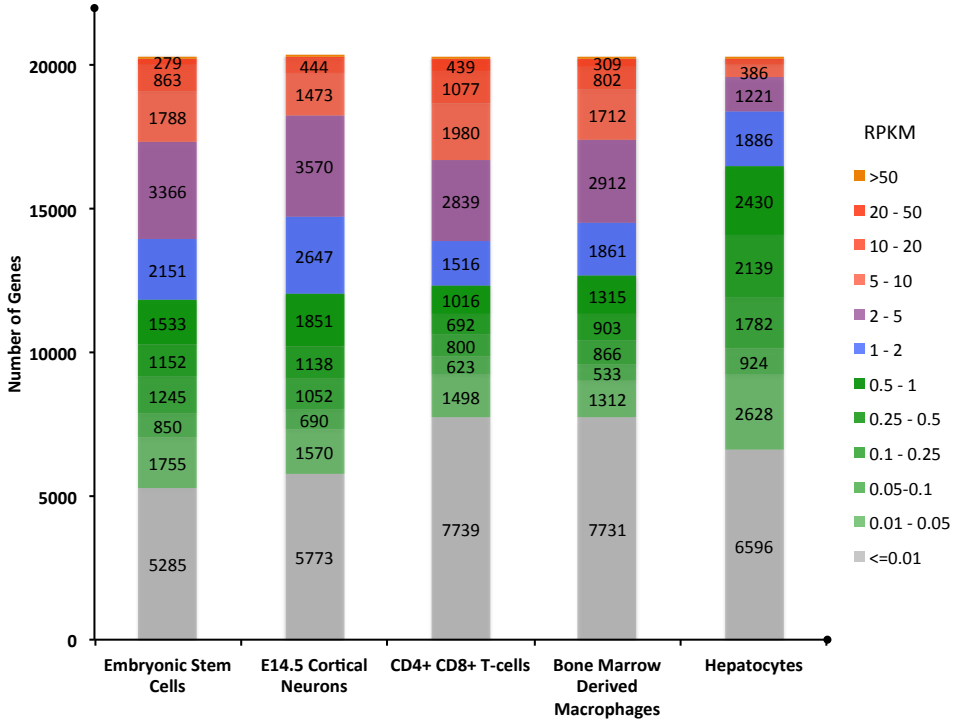

**B**



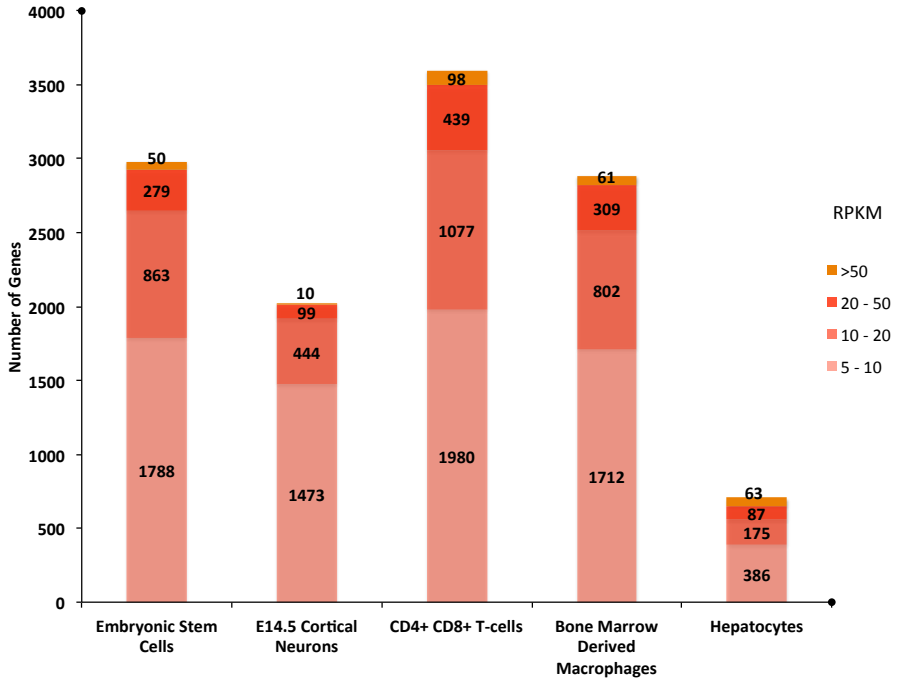
## Figure 3-2

## RPKM Distribution of All Refseq Coding Genes

**A**



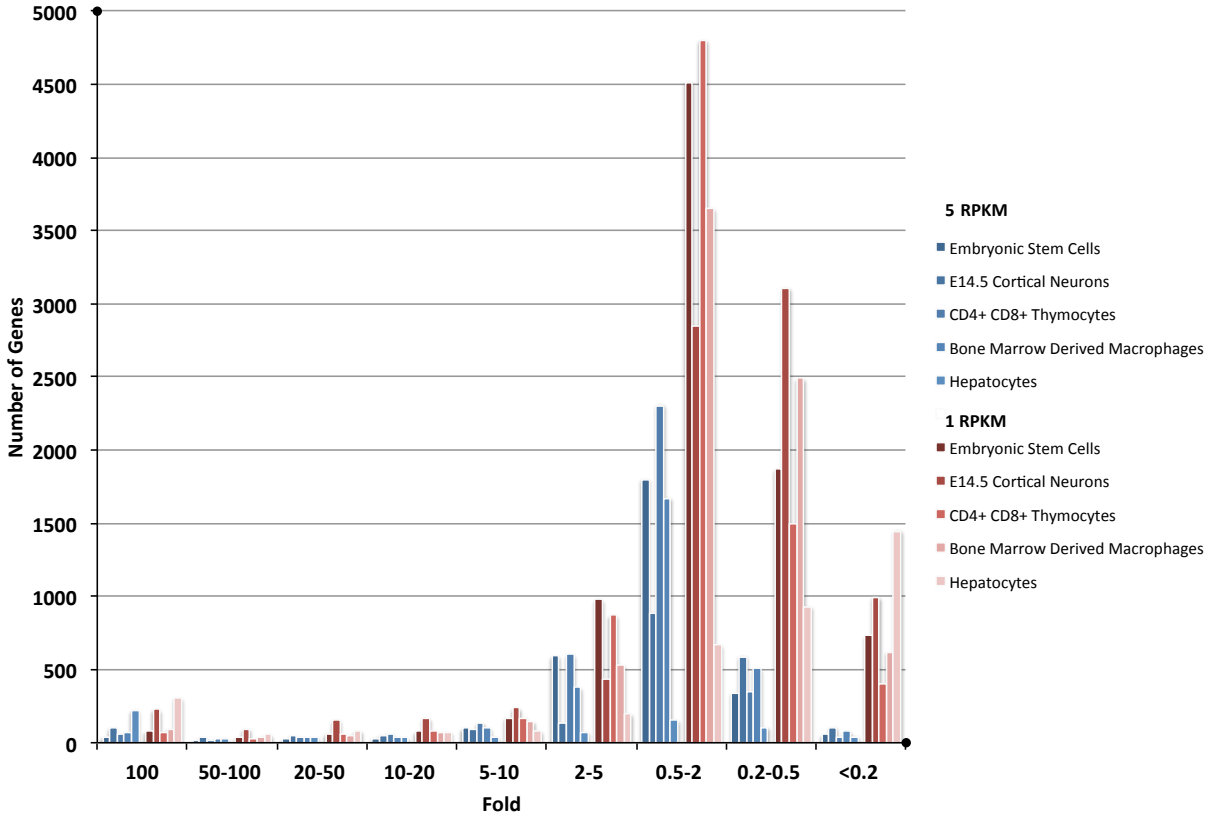

**B**



## Figure 3-3

## Tissue Specificity Based on Fold Changes

**A**



## Figure 3-3

## Tissue Specificity Based on Fold Changes

**B**

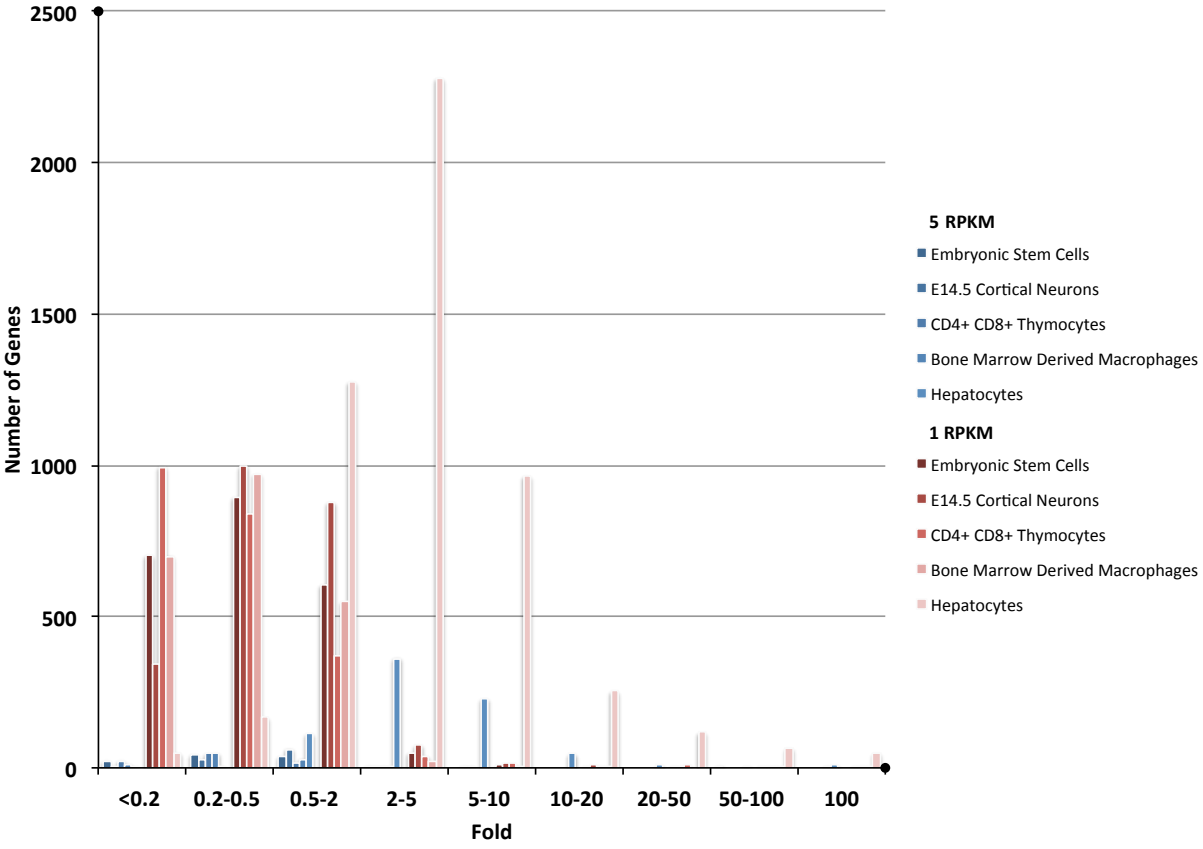
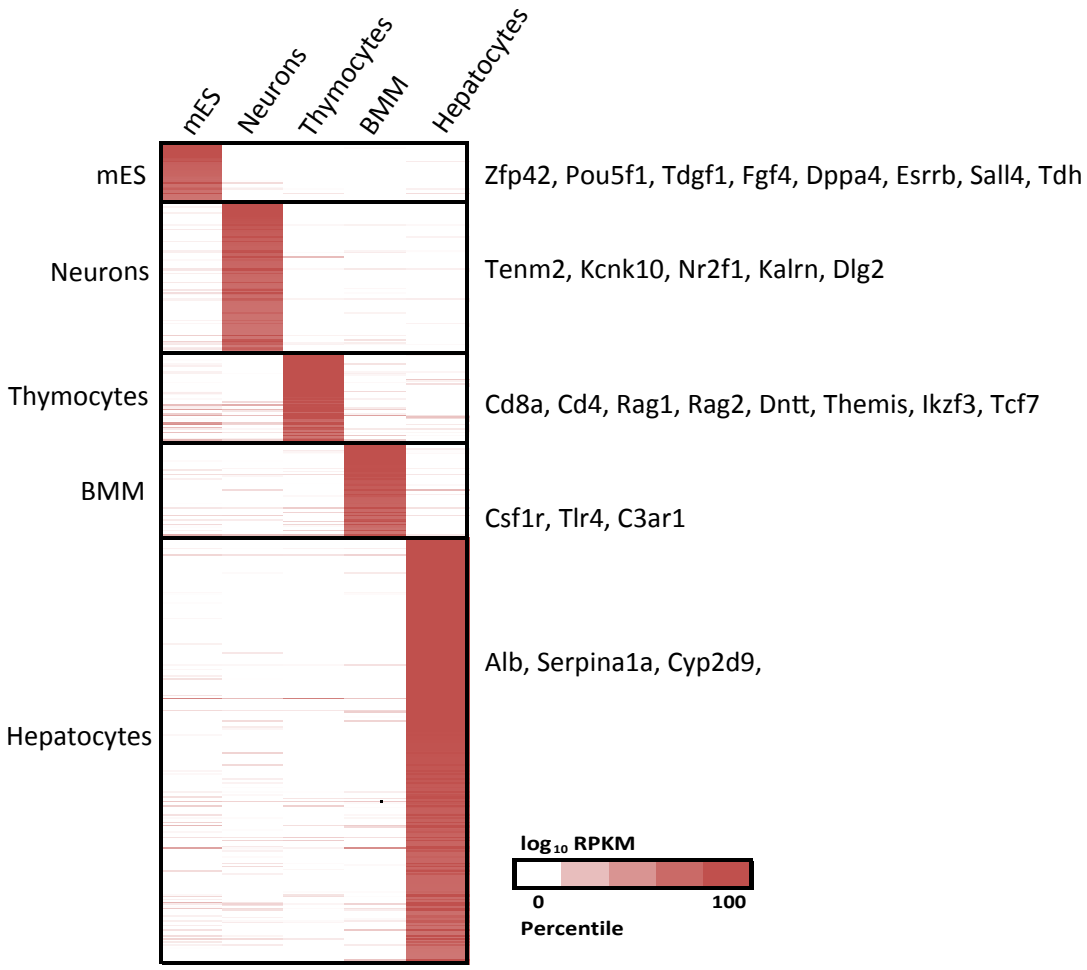


**Figure 3-4**

**100 Fold Specific Genes Include Genes Known to be Important for Function**



## Figure 3-5

## Absolute Fold Changes for Tissue Specific Genes

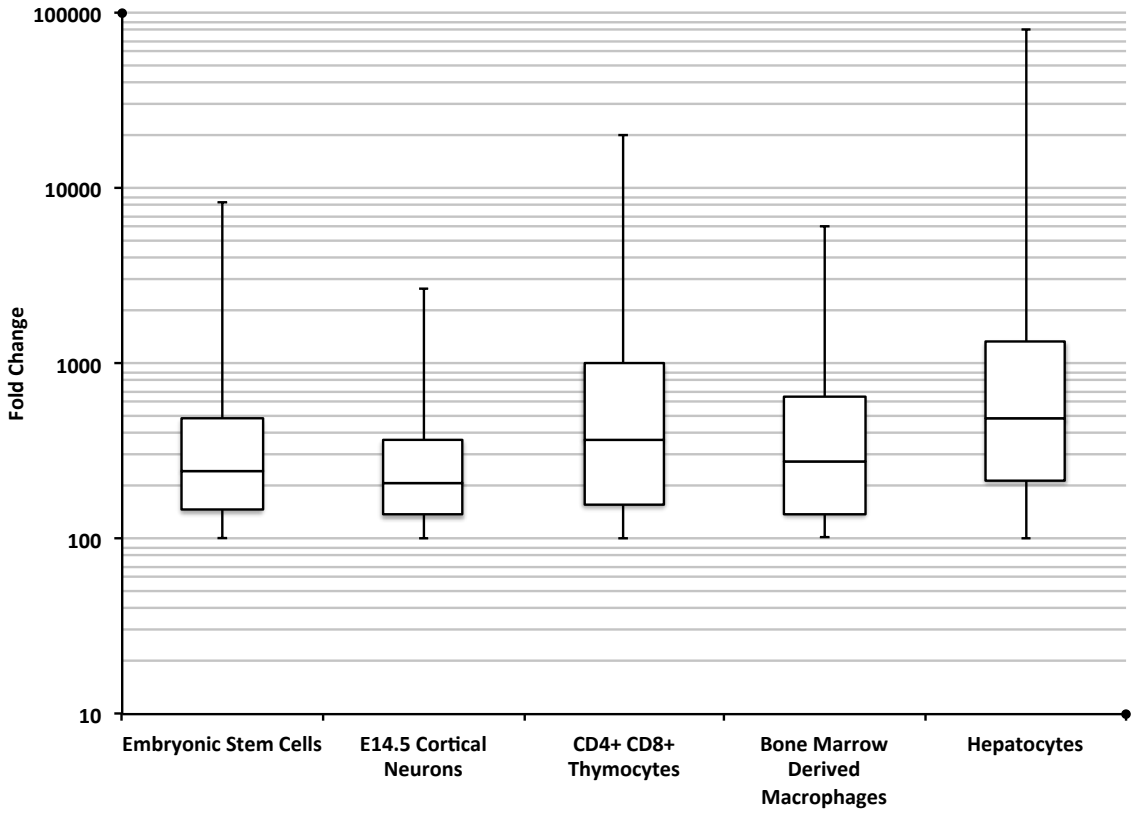
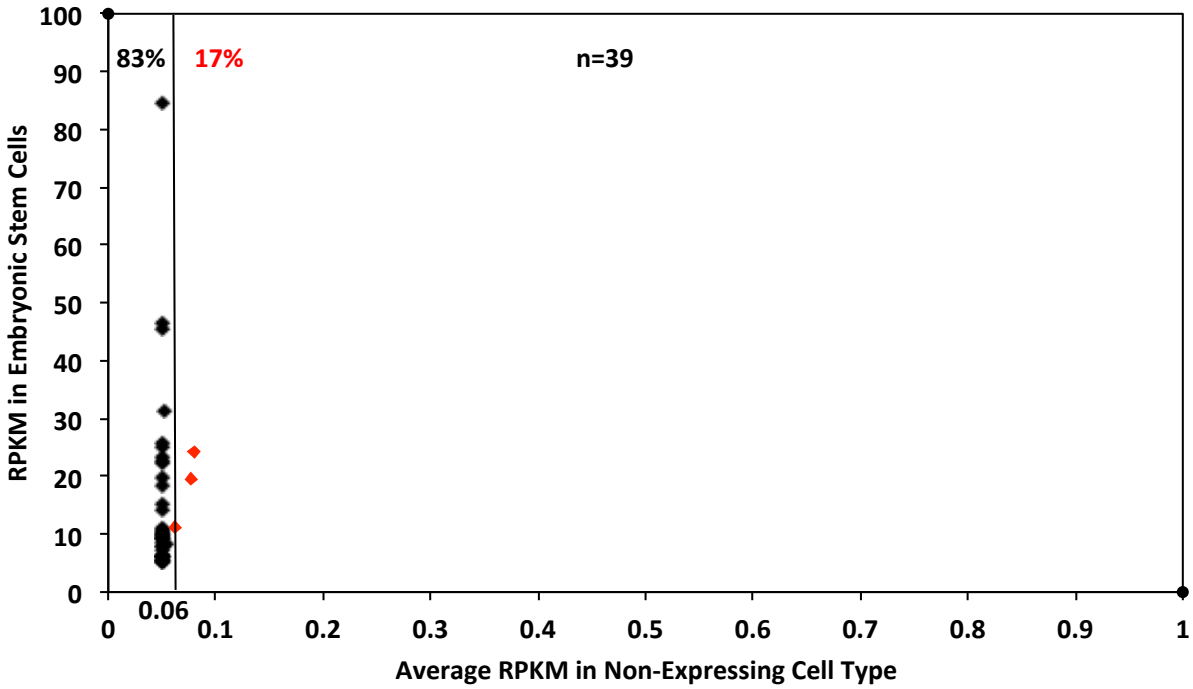


**Figure 3-6**

**Absolute Expression Values for 100-Fold Tissue Specific Genes**

**A**




**B**

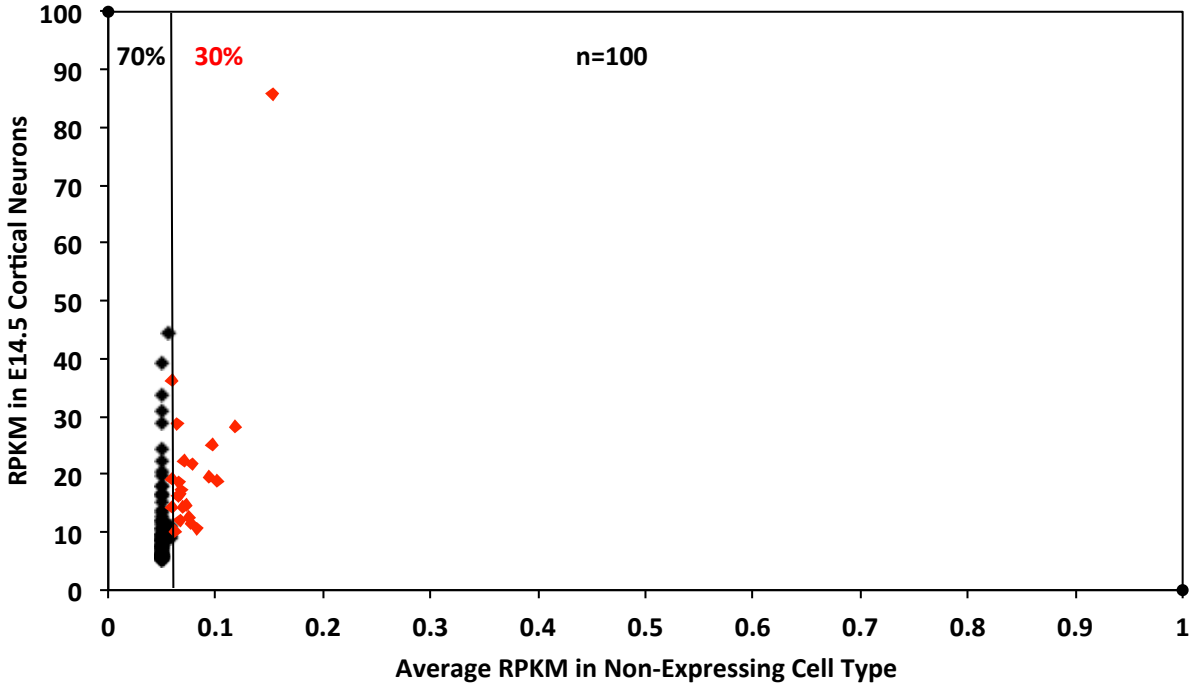
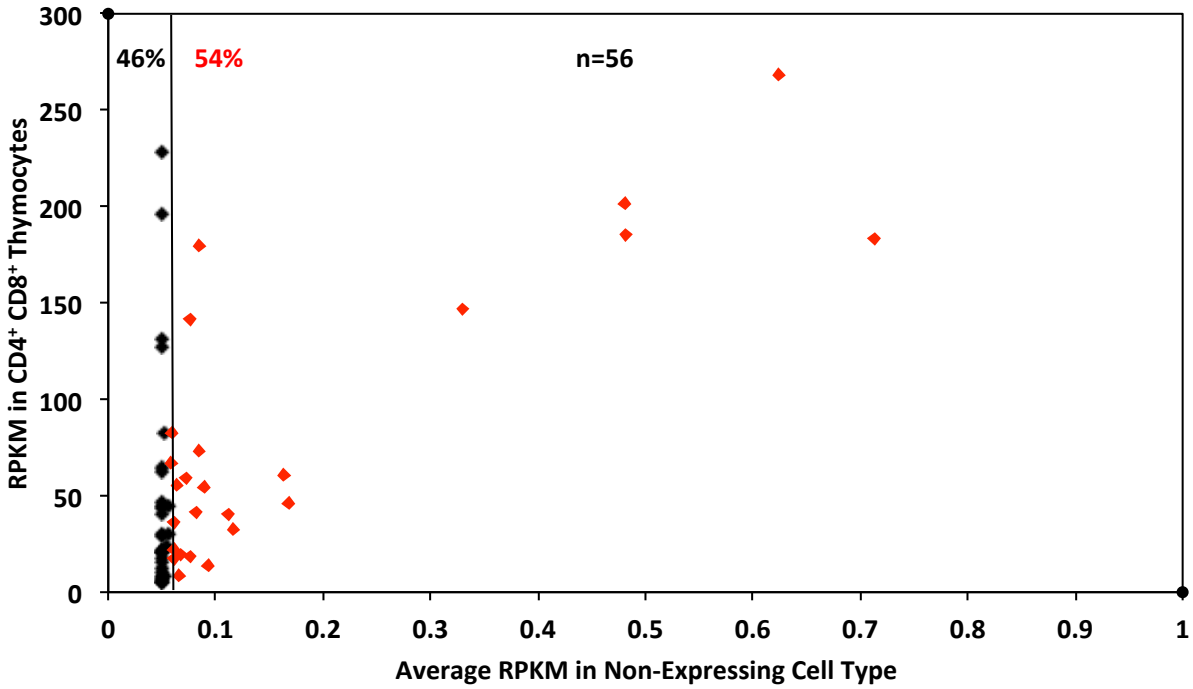


**Figure 3-6**

**Absolute Expression Values for 100-Fold Tissue Specific Genes**

**C**




**D**

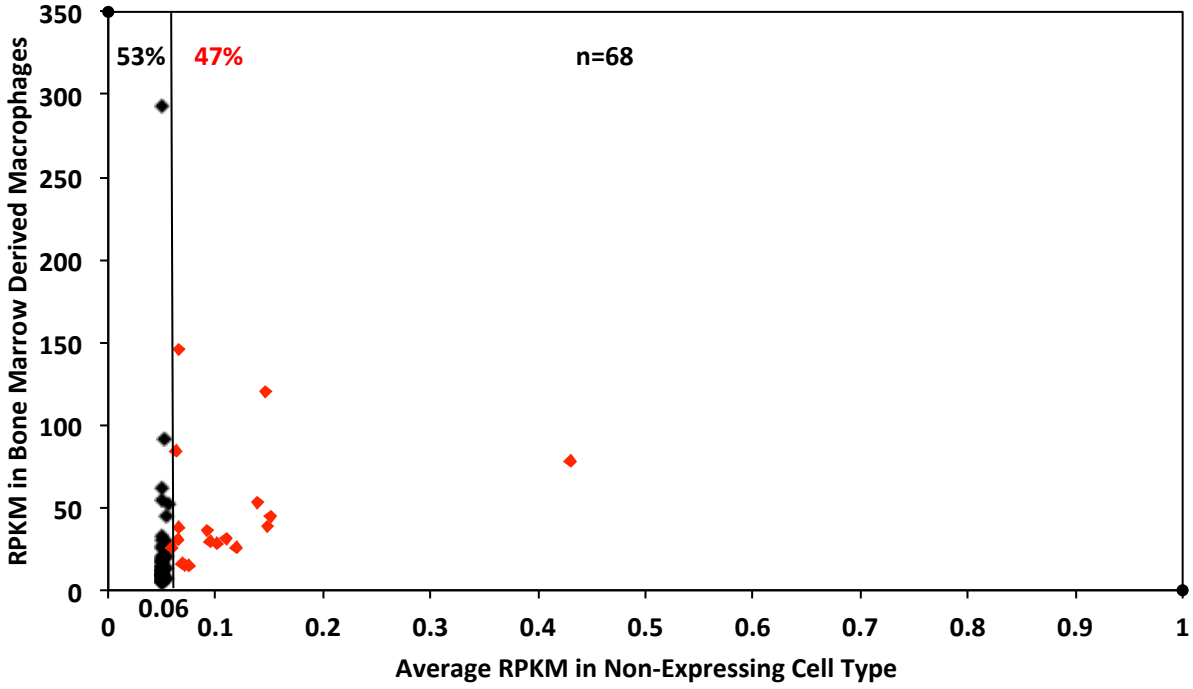
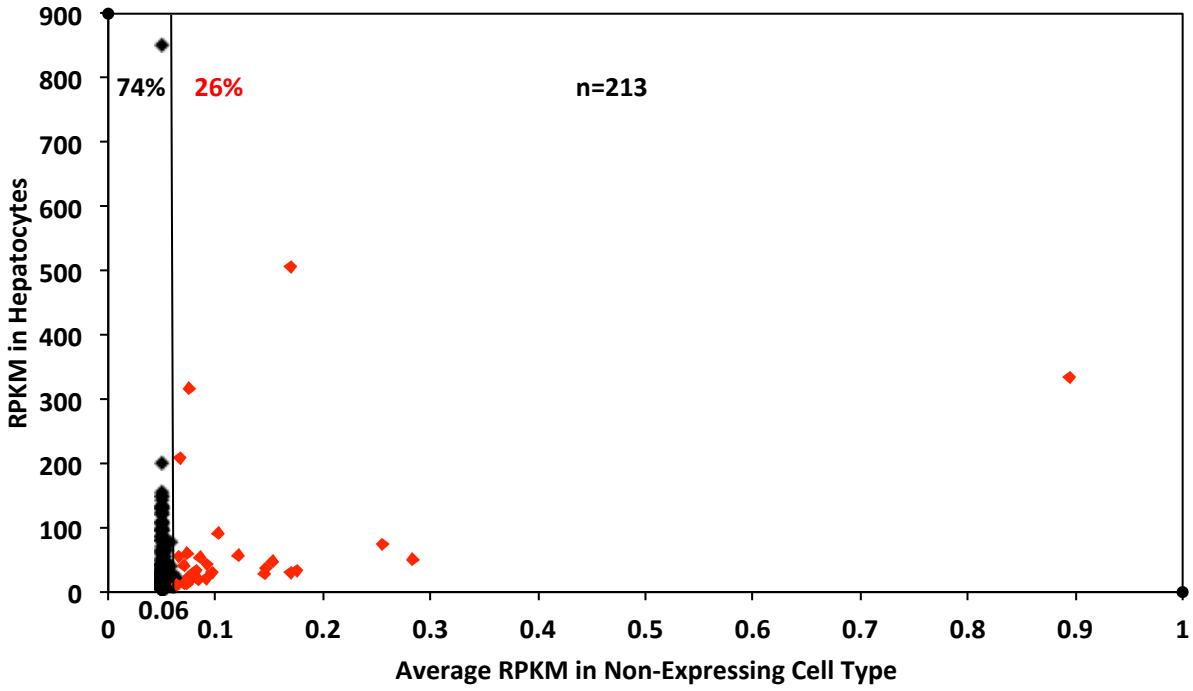


**Figure 3-6**

**Absolute Expression Values for 100-Fold Tissue Specific Genes**

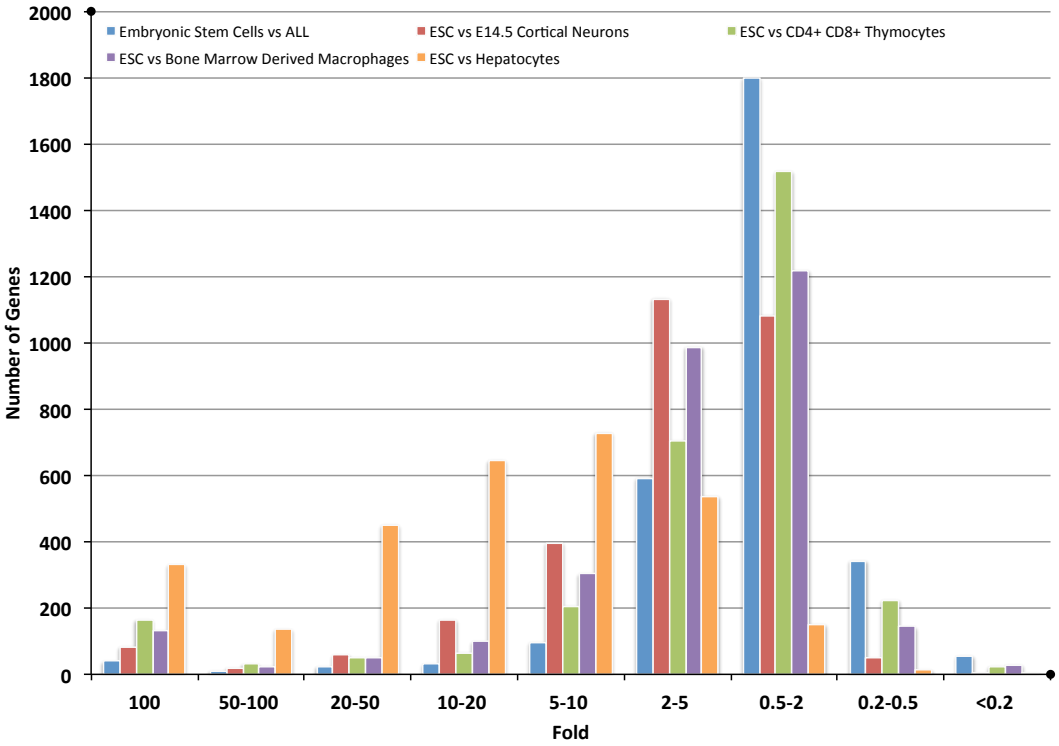
**E**



## Figure 3-7

## Embryonic Stem Cell Specificity in Individual Cell Types Versus All Cell Types

**A**




**B**

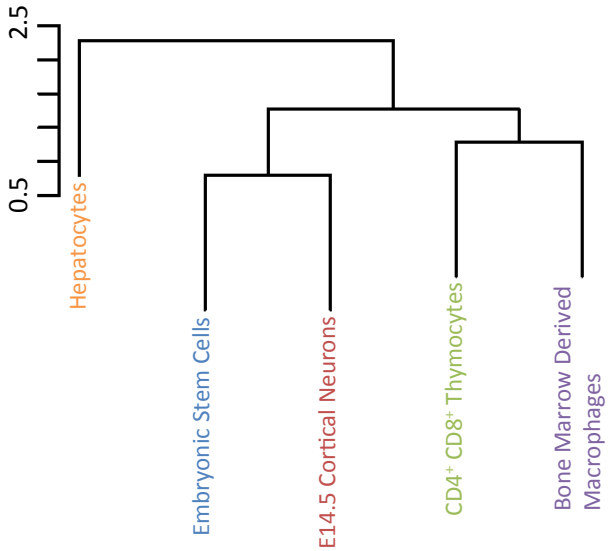


**Figure 3-8**

**Striking Difference in CpG Island Promoter Content**

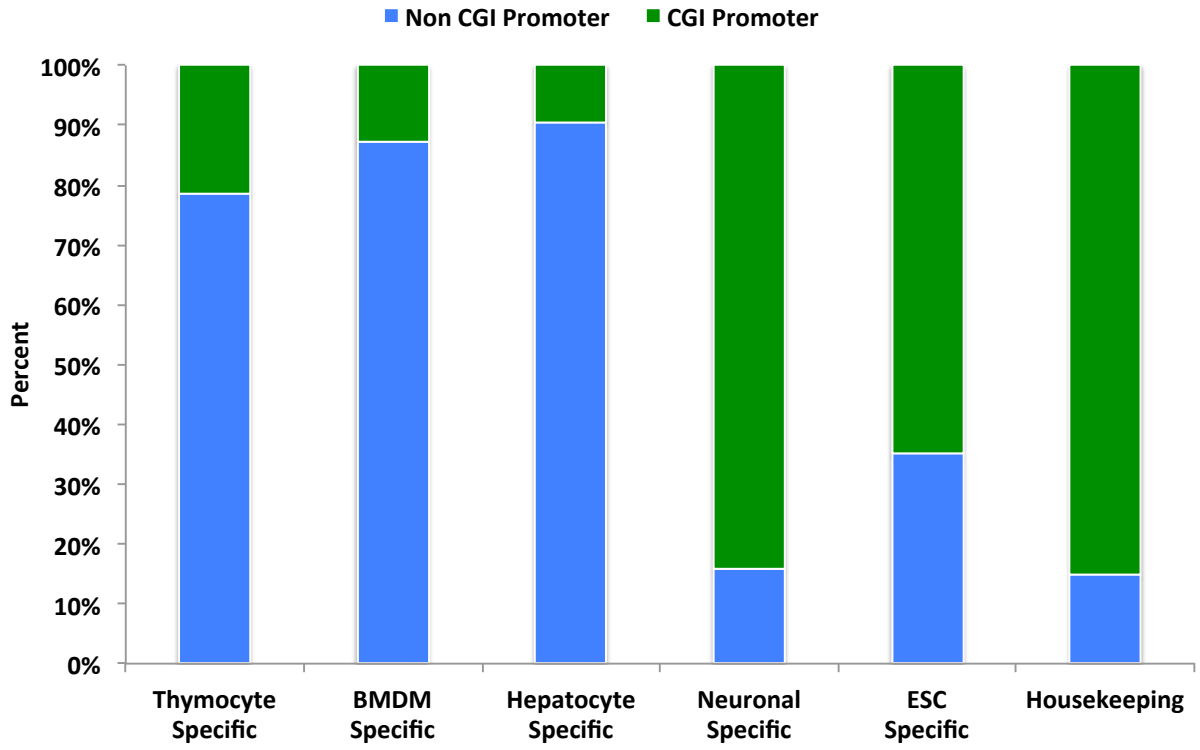


**Figure 3-9**

**Tissue Specific Genes Show Two Distinct DNA Methylation Promoter Patterns in Non Expressing Cells**

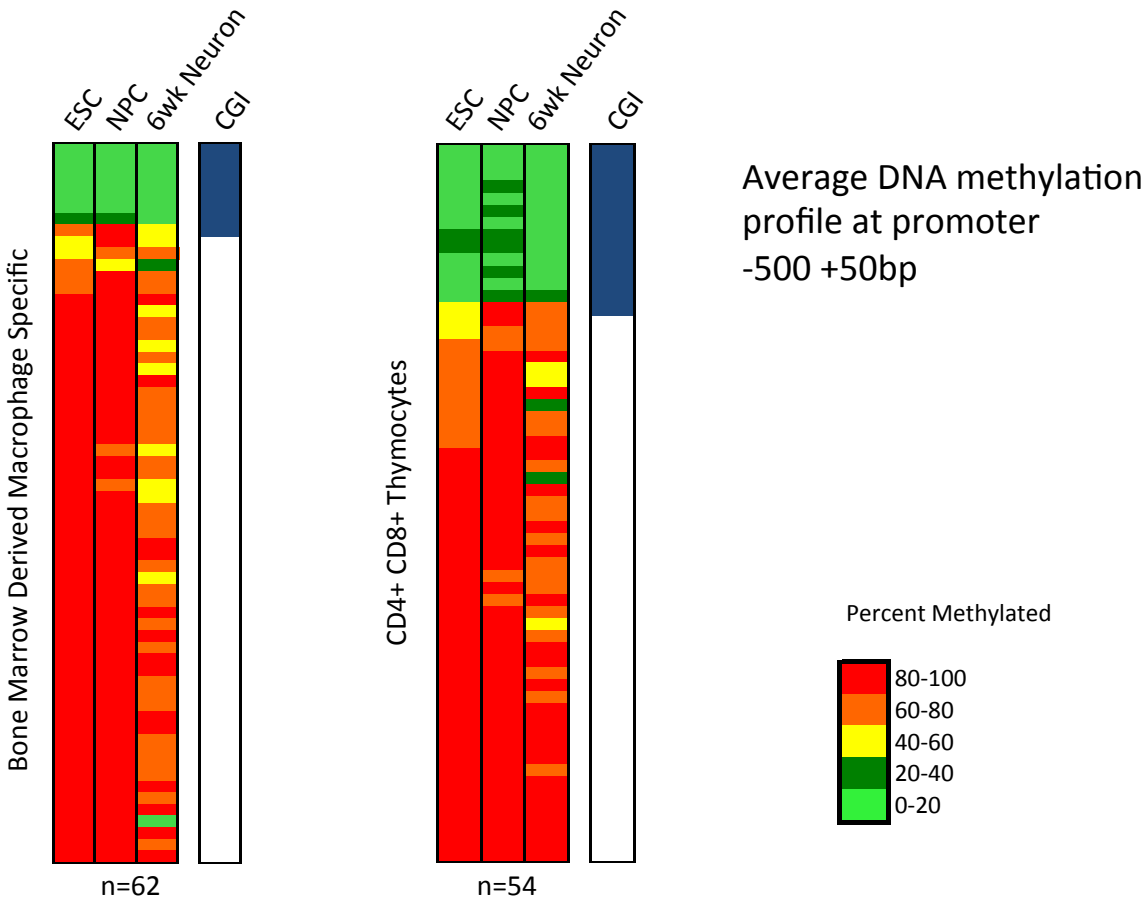


**Figure 3-10**

**Tissue Specific CGI Genes Have Active Histone Marks in Non-Expressing Tissues**

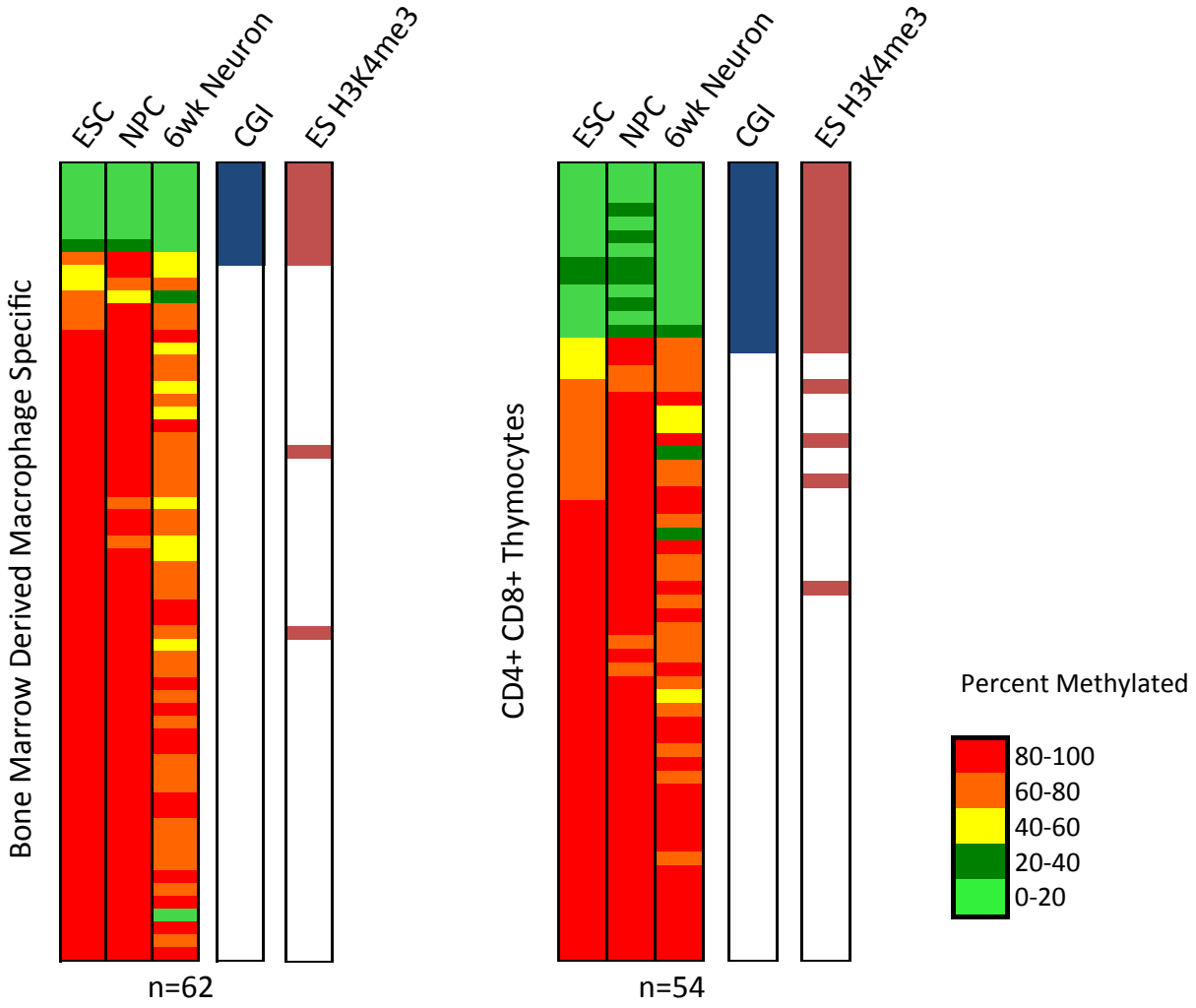


**Figure 3-11**

**Unmethylated Tissue Specific Genes are Associated with H3K27me3 in ES cells**

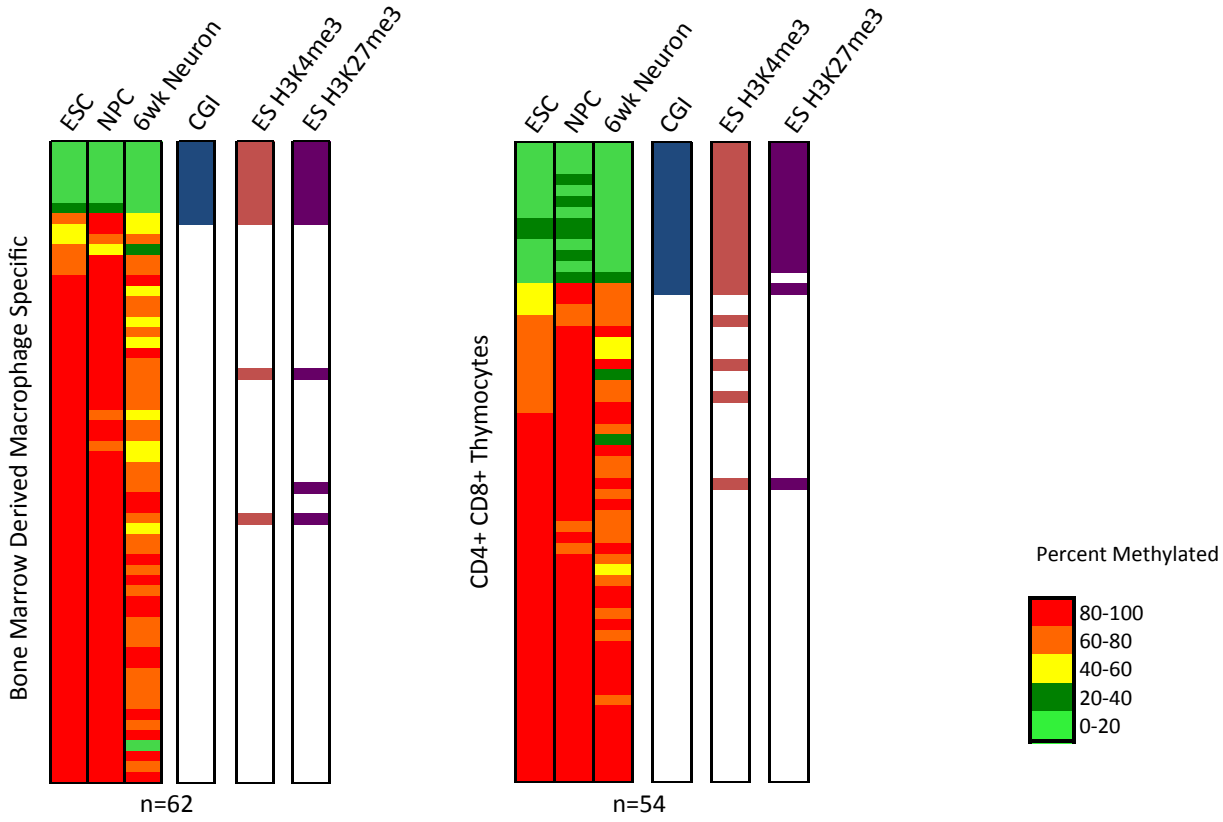


**Figure 3-12**

**Tissue Specific Genes with a Limited Dynamic Range Lack a Clear Epigenetic Profile**

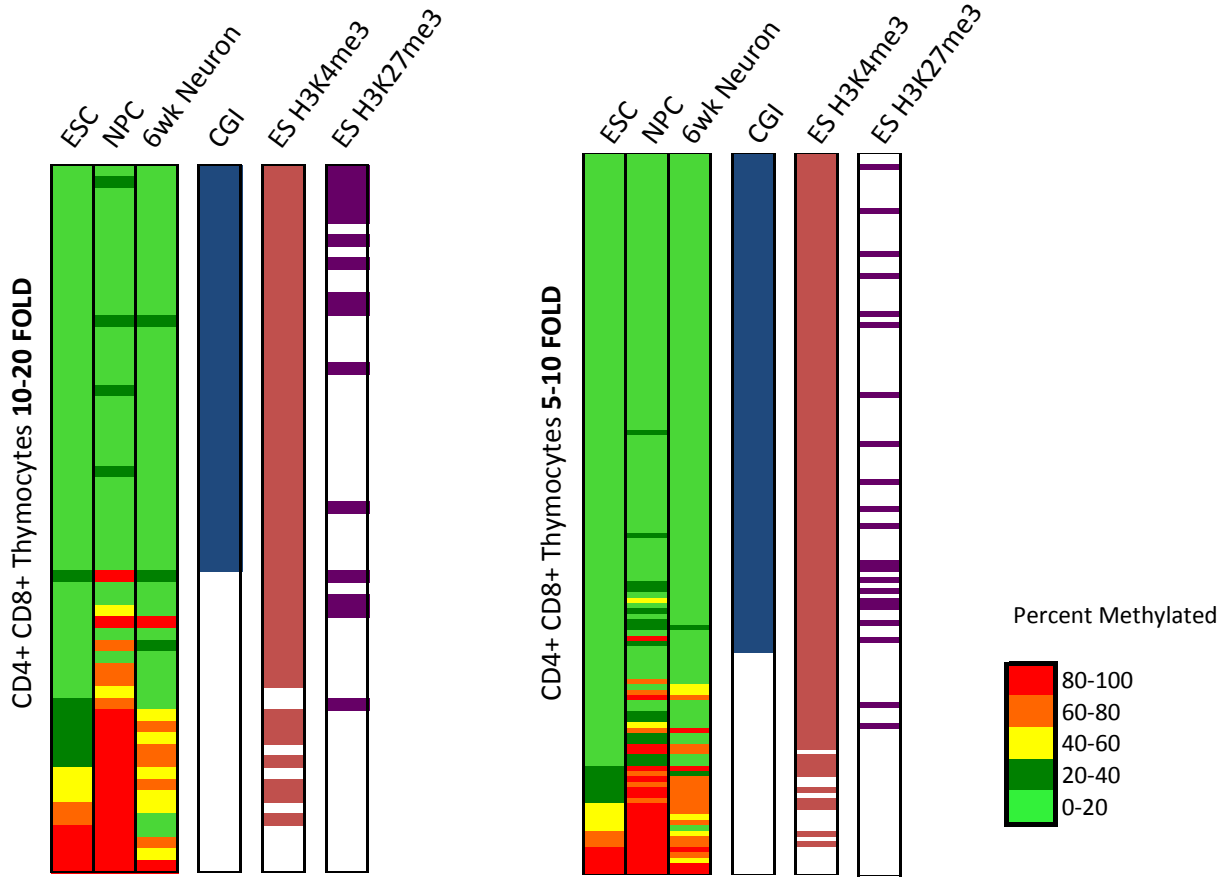
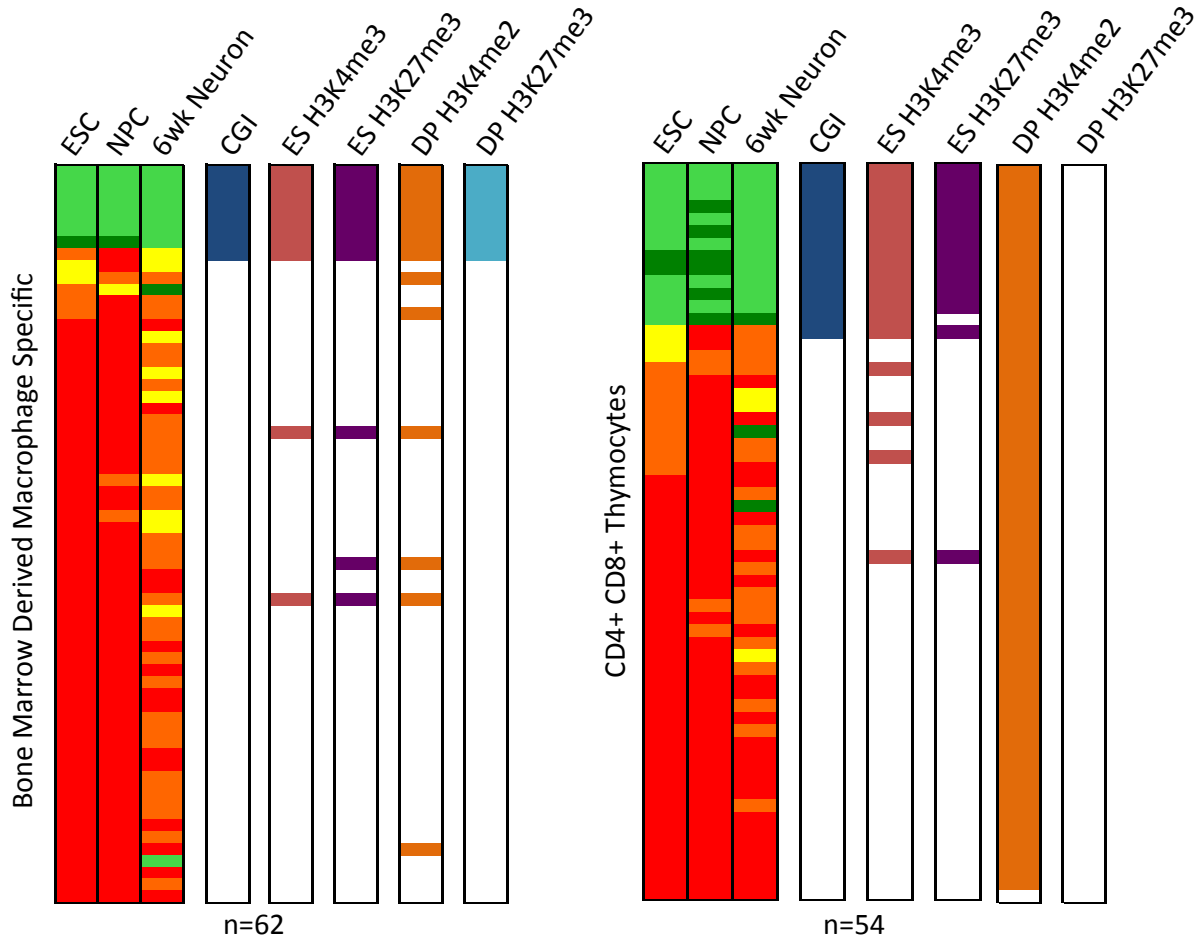


**Figure 3-13**

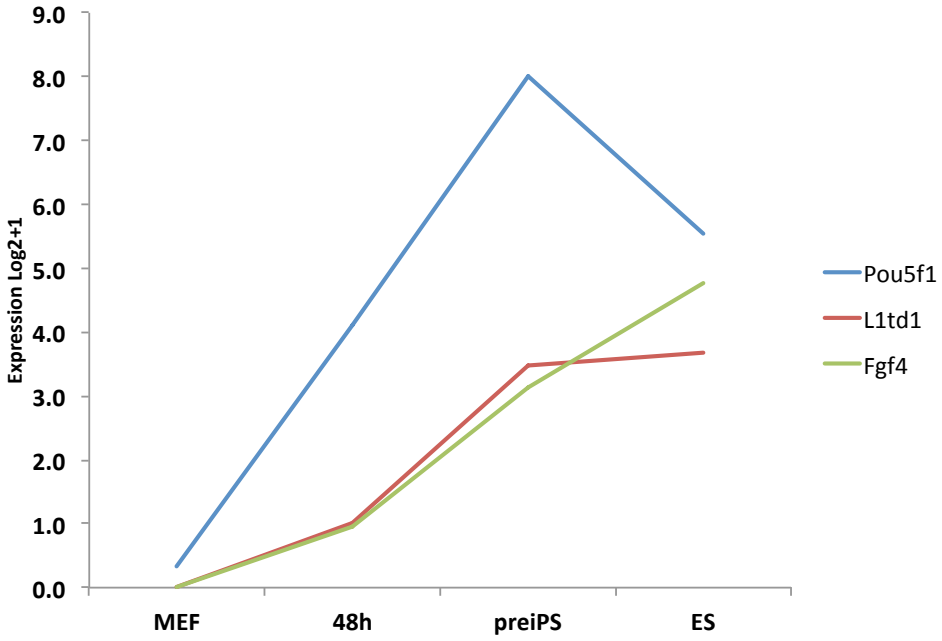
**Unmethylated Tissue Specific Genes are Associated with H3K4me3 and H3K27me3 in Other Non-Expressing Cell Types**



## Figure 3-14

## Embryonic Stem Cell Specific Genes Show Unique Phases of Activation

**A**




**B**

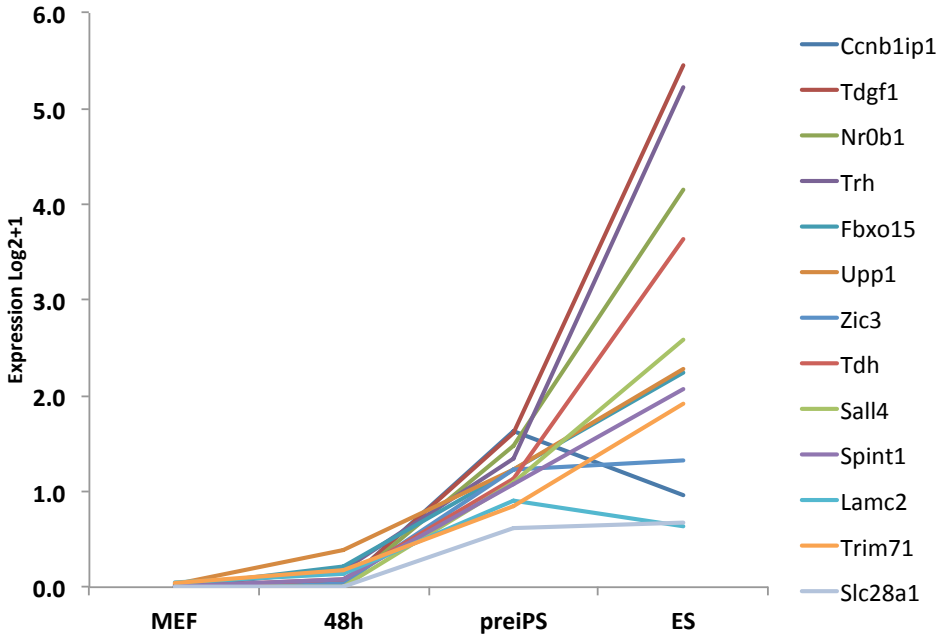
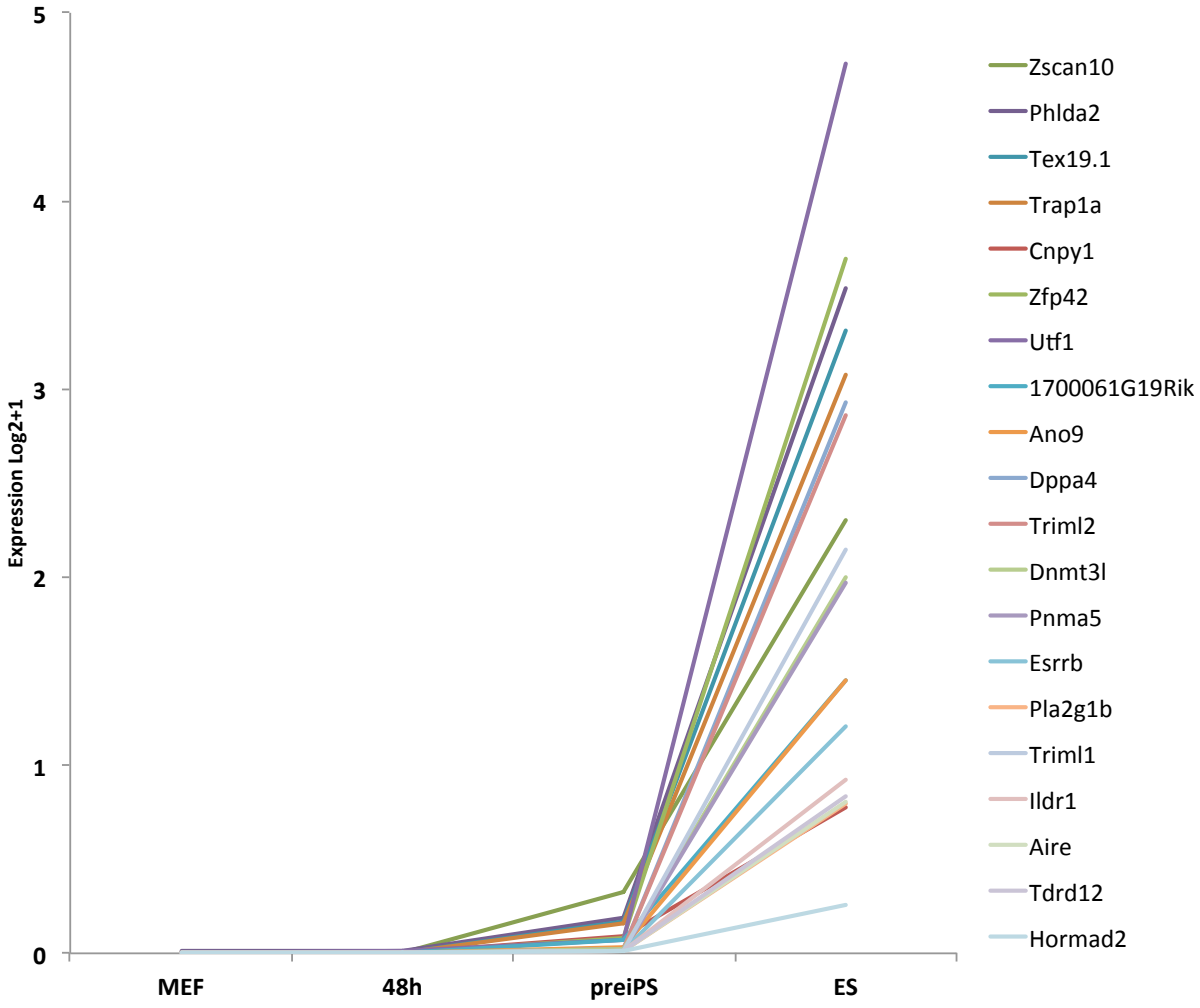


**Figure 3-14**

**Embryonic Stem Cell Specific Genes Show Unique Phases of Activation**

**C**



## Table 3-1

## 100 Fold Embryonic Stem Cell Specific Genes

| Refseq ID | Gene Name | Embryonic Stem Cells RPKM | E 14.5 Cortical Neurons RPKM | CD4+ CD8+ Thymocytes RPKM | Bone Marrow Derived Macrophages RPKM | Hepatocytes RPKM | Minumum Fold Difference | P-value |
|---|---|---|---|---|---|---|---|---|
| NM_001160412 | Triml2 | 84.5797 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 1691.5930 | 0.0031 |
| NM_177742 | Triml1 | 46.4680 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 929.3602 | 0.0130 |
| NM_009556 | Zfp42 | 45.6535 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 913.0705 | 0.0000 |
| NM_001081695 | Dnmt3l | 31.2066 | 0.0500 | 0.0500 | 0.0613 | 0.0500 | 509.1251 | 0.0196 |
| NM_013633 | Pou5f1 | 25.7053 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 514.1067 | 0.0212 |
| NM_178381 | Ano9 | 24.9591 | 0.0500 | 0.0500 | 0.0500 | 0.0519 | 480.9173 | 0.0299 |
| NM_028034 | Tdrd12 | 24.2561 | 0.1690 | 0.0500 | 0.0500 | 0.0500 | 143.5191 | 0.0339 |
| NM_011562 | Tdgf1 | 23.4477 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 468.9541 | 0.0110 |
| NM_010202 | Fgf4 | 22.7048 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 454.0961 | 0.0028 |
| NM_001081202 | L1td1 | 22.1226 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 442.4525 | 0.0187 |
| NM_001271550 | Aire | 19.6533 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 393.0661 | 0.0316 |
| NM_029458 | Hormad2 | 19.6146 | 0.0500 | 0.1225 | 0.0500 | 0.0855 | 160.0567 | 0.0339 |
| NM_015798 | Fbxo15 | 18.5008 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 370.0166 | 0.0136 |
| NM_028610 | Dppa4 | 15.2601 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 305.2023 | 0.0000 |
| NM_001159401 | Upp1 | 14.1999 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 283.9979 | 0.0302 |
| NM_020486 | Bcam | 11.2728 | 0.0917 | 0.0566 | 0.0500 | 0.0500 | 122.9946 | 0.0326 |
| NM_001033425 | Zscan10 | 11.0543 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 221.0865 | 0.0220 |
| NM_001159500 | Esrrb | 10.6368 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 212.7363 | 0.0225 |
| NM_001100461 | Pnma5 | 10.3549 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 207.0980 | 0.0289 |
| NM_009434 | Phlda2 | 10.2140 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 204.2800 | 0.0326 |
| NM_011107 | Pla2g1b | 10.1948 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 203.8953 | 0.0110 |
| NM_201395 | Sall4 | 9.6440 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 192.8802 | 0.0309 |
| NM_011635 | Trap1a | 9.3327 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 186.6530 | 0.0309 |
| NM_175651 | Cnpy1 | 9.2782 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 185.5647 | 0.0310 |
| NM_021480 | Tdh | 8.6244 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 172.4886 | 0.0208 |
| NM_009426 | Trh | 8.4980 | 0.0689 | 0.0500 | 0.0500 | 0.0500 | 123.3067 | 0.0325 |
| NM_001042503 | Trim71 | 7.8740 | 0.0550 | 0.0500 | 0.0500 | 0.0500 | 143.2189 | 0.0314 |
| NM_009482 | Utf1 | 7.8099 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 156.1978 | 0.0323 |
| NM_028946 | Slc9b1 | 7.4125 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 148.2509 | 0.0319 |
| NM_028602 | Tex19.1 | 6.3753 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 127.5056 | 0.0237 |
| NM_030141 | 1700061G19Rik | 6.3202 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 126.4037 | 0.0299 |
| NM_013611 | Nodal | 6.2391 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 124.7813 | 0.0325 |
| NM_016907 | Spint1 | 6.1974 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 123.9483 | 0.0317 |
| NM_007430 | Nr0b1 | 6.1341 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 122.6824 | 0.0156 |
| NM_001111119 | Ccnb1ip1 | 5.6051 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 112.1019 | 0.0224 |
| NM_134109 | Ildr1 | 5.5999 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 111.9980 | 0.0305 |
| NM_009575 | Zic3 | 5.5579 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 111.1570 | 0.0315 |
| NM_001004184 | Slc28a1 | 5.2316 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 104.6328 | 0.0254 |
| NM_008485 | Lamc2 | 5.1872 | 0.0500 | 0.0500 | 0.0500 | 0.0500 | 103.7445 | 0.0328 |

## Table 3-2

## 100 Fold E 14.5 Cortical Neuron Specific Genes

| Refseq ID | Gene Name | E 14.5 Cortical Neurons RPKM | Embryonic Stem Cell RPKM | CD4+ CD8+ Thymocytes RPKM | Bone Marrow Derived Macrophages RPKM | Hepatocytes RPKM | Minumum Fold Difference | P-value |
|---|---|---|---|---|---|---|---|---|
| NM_001039195 | Gria2 | 85.72376 | 0.05000 | 0.46244 | 0.05000 | 0.05000 | 185.37209 | 0.03374 |
| NM_177284 | Nrxn1 | 44.53018 | 0.07736 | 0.05000 | 0.05000 | 0.05000 | 575.63959 | 0.01022 |
| NM_011807 | Dlg2 | 39.27583 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 785.51659 | 0.00548 |
| NM_001081306 | Ptprz1 | 36.27603 | 0.08608 | 0.05000 | 0.05000 | 0.05000 | 421.40452 | 0.01604 |
| NM_001271799 | Pcdh9 | 33.61987 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 672.39750 | 0.00357 |
| NM_001111268 | Grik2 | 30.83337 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 616.66732 | 0.00939 |
| NM_001253756 | Gpm6a | 28.75782 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 575.15644 | 0.00782 |
| NM_016743 | Nell2 | 28.73517 | 0.10780 | 0.05000 | 0.05000 | 0.05000 | 266.54968 | 0.02528 |
| NM_181681 | BC005764 | 28.26208 | 0.25803 | 0.07894 | 0.08535 | 0.05000 | 109.52963 | 0.04161 |
| NM_001164268 | Kalrn | 25.05857 | 0.23700 | 0.05000 | 0.05000 | 0.05000 | 105.73265 | 0.03999 |
| NM_001109764 | Ctnna2 | 24.18948 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 483.78953 | 0.00225 |
| NM_010025 | Dcx | 22.32328 | 0.07380 | 0.06625 | 0.08149 | 0.06210 | 273.92795 | 0.01613 |
| NM_001081358 | Lrrc7 | 22.07023 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 441.40452 | 0.00103 |
| NM_007495 | Astn1 | 21.84092 | 0.15430 | 0.05000 | 0.05966 | 0.05000 | 141.54603 | 0.03926 |
| NM_011607 | Tnc | 20.34243 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 406.84862 | 0.01467 |
| NM_001170787 | Cntn5 | 19.91627 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 398.32534 | 0.00367 |
| NM_001177957 | Gpm6b | 19.48704 | 0.11686 | 0.06533 | 0.11015 | 0.08393 | 166.75987 | 0.01115 |
| NM_176930 | Nrcam | 19.22231 | 0.08886 | 0.05000 | 0.05000 | 0.05000 | 216.32597 | 0.02802 |
| NM_001271858 | Add2 | 18.91972 | 0.09169 | 0.06005 | 0.10125 | 0.15376 | 123.04507 | 0.04166 |
| NM_080285 | Cttnbp2 | 18.62222 | 0.11348 | 0.05000 | 0.05000 | 0.05000 | 164.10476 | 0.03437 |
| NM_008171 | Grin2b | 18.05726 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 361.14527 | 0.02821 |
| NM_133235 | Khdrbs2 | 17.96142 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 359.22849 | 0.01100 |
| NM_001195539 | Dclk1 | 17.35469 | 0.09000 | 0.05000 | 0.07537 | 0.05743 | 192.83840 | 0.01535 |
| NM_029792 | B3gat1 | 16.74084 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 334.81688 | 0.00714 |
| NM_001093778 | Myt1l | 16.66870 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 333.37392 | 0.00103 |
| NM_019707 | Cdh13 | 16.57510 | 0.11660 | 0.05000 | 0.05000 | 0.05000 | 142.15135 | 0.04034 |
| NM_001042617 | Cadps | 16.29513 | 0.10907 | 0.05000 | 0.05000 | 0.05000 | 149.39542 | 0.03745 |
| NM_007529 | Bcan | 16.15865 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 323.17293 | 0.02567 |
| NM_138666 | Nlgn1 | 15.26830 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 305.36610 | 0.00225 |
| NM_175750 | Plxna4 | 14.69435 | 0.05000 | 0.05000 | 0.13837 | 0.05000 | 106.19978 | 0.04161 |
| NM_175642 | Bai3 | 14.38011 | 0.08886 | 0.05000 | 0.05000 | 0.05000 | 161.82450 | 0.03745 |
| NM_027712 | Dlgap1 | 14.36652 | 0.05000 | 0.05000 | 0.05000 | 0.12685 | 113.25399 | 0.04151 |
| NM_001164316 | Ccser1 | 13.83973 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 276.79462 | 0.02875 |
| NM_007937 | Epha5 | 13.56579 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 271.31576 | 0.00636 |
| NM_172475 | Frmd4a | 12.79062 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 255.81246 | 0.00533 |
| NM_001171615 | Myt1 | 12.31853 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 246.37062 | 0.00621 |
| NM_001205341 | Ppfia2 | 12.12786 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 242.55724 | 0.00464 |
| NM_001135688 | Ly6h | 12.05067 | 0.11786 | 0.05000 | 0.05000 | 0.05000 | 102.24403 | 0.04161 |
| NM_207667 | Fgf14 | 11.68177 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 233.63543 | 0.00592 |
| NM_028627 | Psd | 11.54468 | 0.11164 | 0.08050 | 0.06496 | 0.05000 | 103.40596 | 0.04161 |

## Table 3-2 continued

## 100 Fold E 14.5 Cortical Neuron Specific Genes

| Refseq ID | Gene Name | E 14.5 Cortical Neurons RPKM | Embryonic Stem Cell RPKM | CD4+ CD8+ Thymocytes RPKM | Bone Marrow Derived Macrophages RPKM | Hepatocytes RPKM | Minumum Fold Difference | P-value |
|---|---|---|---|---|---|---|---|---|
| NM_001025074 | Ntrk2 | 11.34394 | 0.06383 | 0.05000 | 0.05000 | 0.05000 | 177.72958 | 0.03261 |
| NM_009548 | Rnf112 | 11.28899 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 225.77977 | 0.02410 |
| NM_053199 | Cadm3 | 10.97517 | 0.08334 | 0.05000 | 0.05000 | 0.05000 | 131.68391 | 0.03731 |
| NM_031404 | Actl6b | 10.89807 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 217.96146 | 0.01076 |
| NM_021286 | Sez6 | 10.83376 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 216.67524 | 0.03447 |
| NM_001285843 | Sybu | 10.59145 | 0.10248 | 0.07545 | 0.09281 | 0.05847 | 103.35014 | 0.03378 |
| NM_011856 | Tenm2 | 10.51234 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 210.24687 | 0.00763 |
| NM_001081017 | Unc79 | 10.13164 | 0.09911 | 0.05000 | 0.05000 | 0.05000 | 102.22174 | 0.04068 |
| NM_001198587 | Nrxn3 | 9.86247 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 197.24939 | 0.01061 |
| NM_001039173 | Dok6 | 9.78616 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 195.72326 | 0.02058 |
| NM_199065 | Slitrk1 | 9.77733 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 195.54654 | 0.00176 |
| NM_172290 | Ntm | 9.51210 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 190.24206 | 0.03080 |
| NM_001113325 | Gria1 | 9.49226 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 189.84513 | 0.00958 |
| NM_001253361 | Kcnma1 | 9.38737 | 0.05000 | 0.05000 | 0.06680 | 0.05000 | 140.53668 | 0.01741 |
| NM_019675 | Stmn4 | 9.37703 | 0.05000 | 0.05000 | 0.07835 | 0.05000 | 119.67828 | 0.03833 |
| NM_019724 | Mmp16 | 9.15044 | 0.07844 | 0.05000 | 0.05000 | 0.05000 | 116.65434 | 0.04004 |
| NM_053171 | Csmd1 | 9.13988 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 182.79760 | 0.00508 |
| NM_007461 | Apba2 | 9.07994 | 0.06967 | 0.05000 | 0.05000 | 0.05000 | 130.33672 | 0.03999 |
| NM_010199 | Fgf12 | 8.88513 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 177.70270 | 0.00000 |
| NM_010140 | Epha3 | 8.75578 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 175.11561 | 0.00670 |
| NM_010045 | Darc | 8.72251 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 174.45030 | 0.03012 |
| NM_011215 | Ptprn2 | 8.69861 | 0.06215 | 0.05000 | 0.05000 | 0.05000 | 139.96477 | 0.03804 |
| NM_177328 | Grm7 | 8.57727 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 171.54539 | 0.03114 |
| NM_019931 | Kcnd3 | 8.54058 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 170.81152 | 0.03217 |
| NM_133207 | Kcnh7 | 8.22942 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 164.58848 | 0.00875 |
| NM_182807 | Fam19a2 | 8.10875 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 162.17493 | 0.01037 |
| NM_001190187 | Nrg3 | 7.98912 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 159.78245 | 0.03246 |
| NM_001081348 | Hecw1 | 7.72696 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 154.53929 | 0.01027 |
| NM_001081035 | Nav3 | 7.59772 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 151.95431 | 0.01530 |
| NM_001081397 | Myo16 | 7.56557 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 151.31149 | 0.04004 |
| NM_178714 | Lrfn5 | 7.49946 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 149.98914 | 0.02254 |
| NM_183188 | Rbfox1 | 7.19904 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 143.98086 | 0.03809 |
| NM_001081414 | Grm5 | 7.18668 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 143.73368 | 0.00792 |
| NM_199024 | Nol4 | 7.11830 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 142.36594 | 0.03823 |
| NM_001199244 | Kcnip4 | 6.87500 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 137.49991 | 0.02679 |
| NM_008069 | Gabrb1 | 6.85658 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 137.13168 | 0.00548 |
| NM_177906 | Opcml | 6.61235 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 132.24709 | 0.00225 |
| NM_001081391 | Csmd3 | 6.45293 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 129.05858 | 0.00318 |
| NM_001163565 | Ptpn5 | 6.36962 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 127.39230 | 0.01237 |
| NM_172610 | Mpped1 | 6.36862 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 127.37246 | 0.03198 |

## Table 3-2 continued

## 100 Fold E 14.5 Cortical Neuron Specific Genes

| Refseq ID | Gene Name | E 14.5 Cortical Neurons RPKM | Embryonic Stem Cell RPKM | CD4+ CD8+ Thymocytes RPKM | Bone Marrow Derived Macrophages RPKM | Hepatocytes RPKM | Minumum Fold Difference | P-value |
|---|---|---|---|---|---|---|---|---|
| NM_001286388 | Trim9 | 6.26009 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 125.20184 | 0.02425 |
| NM_001039154 | Cdh8 | 6.25031 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 125.00612 | 0.01120 |
| NM_007831 | Dcc | 6.17123 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 123.42455 | 0.02963 |
| NM_001281955 | Csmd2 | 6.02276 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 120.45511 | 0.03476 |
| NM_009960 | Pcdha11 | 5.99102 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 119.82031 | 0.02758 |
| NM_010151 | Nr2f1 | 5.98832 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 119.76631 | 0.00983 |
| NM_001011874 | Xkr4 | 5.97716 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 119.54328 | 0.02552 |
| NM_001003671 | Pcdhac1 | 5.86403 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 117.28055 | 0.03418 |
| NM_001077398 | Ldb2 | 5.84576 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 116.91528 | 0.03144 |
| NM_001286013 | Dlk2 | 5.82093 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 116.41864 | 0.03789 |
| NM_009961 | Pcdha10 | 5.73365 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 114.67309 | 0.03403 |
| NM_001167748 | Egfem1 | 5.68082 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 113.61644 | 0.01193 |
| NM_198250 | Lrrc4b | 5.65458 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 113.09160 | 0.02088 |
| NM_001282102 | Lrrtm4 | 5.64930 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 112.98596 | 0.00357 |
| NM_138661 | Pcdha9 | 5.51392 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 110.27833 | 0.03031 |
| NM_008900 | Pou3f3 | 5.46738 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 109.34759 | 0.04107 |
| NM_178673 | Fstl5 | 5.43604 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 108.72074 | 0.00000 |
| NM_175549 | Robo2 | 5.17453 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 103.49059 | 0.02494 |
| NM_029911 | Kcnk10 | 5.08145 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 101.62899 | 0.00543 |

## Table 3-3

## 100 Fold $CD4^+$ $CD8^+$ Thymocyte Specific Genes

| Refseq ID | Gene Name | CD4+ CD8+ Thymocytes RPKM | Embryonic Stem Cell RPKM | E 14.5 Cortical Neurons RPKM | Bone Marrow Derived Macrophages RPKM | Hepatocytes RPKM | Minumum Fold Difference | P-value |
|---|---|---|---|---|---|---|---|---|
| NM_009331 | Tcf7 | 268.17473 | 2.16625 | 0.21040 | 0.05000 | 0.06961 | 123.79655 | 0.08603 |
| NM_009019 | Rag1 | 228.38005 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 4567.60102 | 0.00095 |
| NM_001162432 | Lck | 201.62636 | 1.57148 | 0.25186 | 0.05000 | 0.05000 | 128.30365 | 0.08603 |
| NM_001081110 | Cd8a | 196.48176 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 3929.63516 | 0.00308 |
| NM_001198914 | Myb | 185.24599 | 1.11372 | 0.51380 | 0.24920 | 0.05000 | 166.33043 | 0.08297 |
| NM_010689 | Lat | 183.36579 | 1.10864 | 0.45025 | 1.19342 | 0.09845 | 153.64674 | 0.08617 |
| NM_009858 | Cd8b1 | 179.66459 | 0.05000 | 0.05000 | 0.05000 | 0.18916 | 949.81794 | 0.04190 |
| NM_009382 | Thy1 | 146.95880 | 0.43898 | 0.77929 | 0.05000 | 0.05000 | 188.58026 | 0.07649 |
| NM_001281966 | Itk | 141.38486 | 0.06877 | 0.05000 | 0.13505 | 0.05000 | 1046.88719 | 0.04535 |
| NM_013488 | Cd4 | 131.52748 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 2630.54967 | 0.00286 |
| NM_001043228 | Dntt | 126.82478 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 2536.49556 | 0.00303 |
| NM_007648 | Cd3e | 82.85960 | 0.05710 | 0.05000 | 0.05000 | 0.05000 | 1451.09452 | 0.00526 |
| NM_178666 | Themis | 82.76471 | 0.08680 | 0.05000 | 0.05000 | 0.05000 | 953.47457 | 0.02376 |
| NM_001113391 | Cd247 | 73.07327 | 0.05496 | 0.05000 | 0.18180 | 0.05000 | 401.93346 | 0.07290 |
| NM_010742 | Ly6d | 66.98759 | 0.05000 | 0.05000 | 0.05000 | 0.08537 | 784.70122 | 0.04085 |
| NM_001166625 | Ccr9 | 64.56851 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1291.37026 | 0.00303 |
| NM_009850 | Cd3g | 62.45917 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1249.18331 | 0.00242 |
| NM_011246 | Rasgrp1 | 60.74510 | 0.50508 | 0.05000 | 0.05000 | 0.05000 | 120.26880 | 0.08603 |
| NM_001083960 | Spo11 | 59.21071 | 0.08328 | 0.05691 | 0.09196 | 0.05794 | 643.87274 | 0.00000 |
| NM_001033126 | Cd27 | 55.51463 | 0.08227 | 0.05000 | 0.07254 | 0.05000 | 674.78077 | 0.04835 |
| NM_013487 | Cd3d | 54.66024 | 0.05000 | 0.05000 | 0.05000 | 0.20872 | 261.88783 | 0.08216 |
| NM_183264 | Tespa1 | 47.33927 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 946.78532 | 0.00829 |
| NM_001276403 | Lef1 | 46.08353 | 0.42672 | 0.12212 | 0.06467 | 0.05741 | 107.99464 | 0.08595 |
| NM_007650 | Cd5 | 45.24783 | 0.05000 | 0.05000 | 0.07080 | 0.05000 | 639.09894 | 0.02481 |
| NM_001168693 | Endou | 44.87904 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 897.58087 | 0.01335 |
| NM_009937 | Colq | 43.27285 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 865.45708 | 0.00242 |
| NM_198297 | Trat1 | 41.41748 | 0.05000 | 0.17733 | 0.05000 | 0.05000 | 233.56126 | 0.07735 |
| NM_009852 | Cd6 | 40.39733 | 0.25526 | 0.05302 | 0.08567 | 0.05398 | 158.26196 | 0.08358 |
| NM_153175 | Gimap6 | 40.22641 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 804.52825 | 0.04703 |
| NM_029983 | Sla2 | 36.08206 | 0.09555 | 0.05000 | 0.05000 | 0.05000 | 377.60568 | 0.07322 |
| NM_010815 | Grap2 | 32.58080 | 0.05000 | 0.05000 | 0.31493 | 0.05000 | 103.45416 | 0.08617 |
| NM_032465 | Cd96 | 30.21901 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 604.38013 | 0.01159 |
| NM_021309 | Sh2d2a | 30.05000 | 0.05933 | 0.05000 | 0.06551 | 0.05000 | 458.67932 | 0.00000 |
| NM_011771 | Ikzf3 | 29.01566 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 580.31320 | 0.01354 |
| NM_001033186 | Skap1 | 23.46961 | 0.06394 | 0.05000 | 0.05000 | 0.05000 | 367.05457 | 0.06451 |
| NM_013698 | Txk | 22.07962 | 0.06304 | 0.05000 | 0.08295 | 0.05000 | 266.16923 | 0.01799 |
| NM_011364 | Sh2d1a | 22.04979 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 440.99570 | 0.01032 |
| NM_030710 | Slamf6 | 21.00818 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 420.16357 | 0.02308 |
| NM_019436 | Sit1 | 20.81864 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 416.37274 | 0.00362 |
| NM_001267621 | Gfi1 | 20.57217 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 411.44341 | 0.01191 |

## Table 3-3 continued

## 100 Fold $CD4^+$ $CD8^+$ Thymocyte Specific Genes

| Refseq ID | Gene Name | CD4+ CD8+ Thymocytes RPKM | Embryonic Stem Cell RPKM | E 14.5 Cortical Neurons RPKM | Bone Marrow Derived Macrophages RPKM | Hepatocytes RPKM | Minumum Fold Difference | P-value |
|---|---|---|---|---|---|---|---|---|
| NM_008859 | Prkcq | 19.54454 | 0.05000 | 0.11972 | 0.05000 | 0.05000 | 163.25090 | 0.08385 |
| NM_011346 | Sell | 18.59547 | 0.08405 | 0.05744 | 0.10424 | 0.05847 | 178.38900 | 0.01958 |
| NM_013730 | Slamf1 | 18.05000 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 361.00007 | 0.03760 |
| NM_013486 | Cd2 | 17.22083 | 0.09374 | 0.05000 | 0.05000 | 0.05000 | 183.71727 | 0.07943 |
| NM_028878 | Slc6a19 | 15.55685 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 311.13709 | 0.07740 |
| NM_175860 | Gimap1 | 13.63900 | 0.09867 | 0.06743 | 0.10895 | 0.09740 | 125.18254 | 0.07446 |
| NM_172435 | P2ry10 | 12.13710 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 242.74208 | 0.01271 |
| NM_173398 | Gpr171 | 10.83846 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 216.76917 | 0.05141 |
| NM_009824 | Cbfa2t3 | 8.61997 | 0.07721 | 0.08445 | 0.05247 | 0.05000 | 102.07608 | 0.08566 |
| NM_008091 | Gata3 | 8.06650 | 0.05698 | 0.05000 | 0.05000 | 0.05000 | 141.56663 | 0.08515 |
| NM_010165 | Eya2 | 8.05969 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 161.19388 | 0.06456 |
| NM_009020 | Rag2 | 6.80898 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 136.17965 | 0.06725 |
| NM_001038499 | Arsi | 6.07416 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 121.48313 | 0.08485 |
| NM_031395 | Sytl3 | 5.53504 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 110.70078 | 0.08595 |
| NM_001013390 | Scn4b | 5.28724 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 105.74472 | 0.07512 |
| NM_205823 | Tlr12 | 5.15783 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 103.15655 | 0.06762 |

## Table 3-4

## 100 Fold Bone Marrow Derived Macrophage Specific Genes

| Refseq ID | Gene Name | Bone Marrow Derived Macrophages RPKM | Embryonic Stem Cell RPKM | E 14.5 Cortical Neurons RPKM | CD4+ CD8+ Thymocytes RPKM | Hepatocytes RPKM | Minumum Fold Difference | P-value |
|---|---|---|---|---|---|---|---|---|
| NM_053110 | Gpnmb | 292.68255 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 5853.65092 | 0.00934 |
| NM_031254 | Trem2 | 145.85513 | 0.05748 | 0.10456 | 0.05000 | 0.05000 | 1394.96743 | 0.02462 |
| NM_138672 | Stab1 | 120.23534 | 0.05000 | 0.17103 | 0.05000 | 0.31456 | 382.23361 | 0.06090 |
| NM_001037859 | Csf1r | 91.61717 | 0.05000 | 0.05000 | 0.05000 | 0.06264 | 1462.62728 | 0.02692 |
| NM_001267695 | Ctss | 84.12816 | 0.05000 | 0.05000 | 0.10249 | 0.05000 | 820.81936 | 0.04752 |
| NM_001206390 | Hk3 | 78.08887 | 0.73943 | 0.39940 | 0.45673 | 0.12630 | 105.60637 | 0.08229 |
| NM_008147 | Gp49a | 62.05057 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1241.01146 | 0.01645 |
| NM_013532 | Lilrb4 | 54.74895 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1094.97901 | 0.02418 |
| NM_010185 | Fcer1g | 52.94950 | 0.16214 | 0.15798 | 0.17214 | 0.06359 | 307.59261 | 0.06356 |
| NM_011662 | Tyrobp | 52.27353 | 0.05000 | 0.05000 | 0.07139 | 0.05000 | 732.18121 | 0.06126 |
| NM_017372 | Lyz2 | 45.57331 | 0.05000 | 0.05000 | 0.05859 | 0.05577 | 777.86539 | 0.03951 |
| NM_001110322 | Cd72 | 44.63842 | 0.11789 | 0.10520 | 0.26383 | 0.11637 | 169.19156 | 0.08461 |
| NM_001008702 | Dab2 | 38.76748 | 0.21921 | 0.24380 | 0.05000 | 0.07858 | 159.01521 | 0.07674 |
| NM_008873 | Plau | 38.08635 | 0.11182 | 0.05000 | 0.05000 | 0.05000 | 340.60092 | 0.06361 |
| NM_015811 | Rgs1 | 36.24535 | 0.05000 | 0.05000 | 0.21881 | 0.05000 | 165.64998 | 0.07080 |
| NM_001146022 | Wdfy4 | 32.77408 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 655.48151 | 0.03188 |
| NM_178911 | Pld4 | 31.13481 | 0.08279 | 0.05000 | 0.25701 | 0.05000 | 121.14467 | 0.08258 |
| NM_001286037 | Ncf1 | 30.73459 | 0.05000 | 0.07209 | 0.08662 | 0.05133 | 354.82546 | 0.05654 |
| NM_011333 | Ccl2 | 30.05000 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 601.00000 | 0.03092 |
| NM_013652 | Ccl4 | 29.73191 | 0.05000 | 0.05000 | 0.22839 | 0.05000 | 130.17866 | 0.07390 |
| NM_010819 | Clec4d | 28.92787 | 0.05972 | 0.05329 | 0.05000 | 0.05000 | 484.38544 | 0.03797 |
| NM_010821 | Mpeg1 | 28.06373 | 0.05000 | 0.05000 | 0.05000 | 0.25492 | 110.08755 | 0.09148 |
| NM_010130 | Emr1 | 26.38304 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 527.66077 | 0.03356 |
| NM_008677 | Ncf4 | 26.15958 | 0.05000 | 0.05000 | 0.07749 | 0.05000 | 337.57613 | 0.05085 |
| NM_145227 | Oas2 | 25.63305 | 0.17608 | 0.05000 | 0.20210 | 0.05000 | 126.83474 | 0.08620 |
| NM_008533 | Cd180 | 25.51811 | 0.08797 | 0.05000 | 0.05000 | 0.05000 | 290.06453 | 0.05371 |
| NM_011311 | S100a4 | 25.31032 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 506.20636 | 0.03601 |
| NM_001163616 | 1810011H11Rik | 21.21735 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 424.34698 | 0.02359 |
| NM_011539 | Tbxas1 | 20.69167 | 0.05000 | 0.07012 | 0.05000 | 0.05000 | 295.10075 | 0.05510 |
| NM_009777 | C1qb | 19.60353 | 0.05000 | 0.05000 | 0.05000 | 0.06032 | 324.99898 | 0.05823 |
| NM_011095 | Pirb | 18.86643 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 377.32850 | 0.02479 |
| NM_145827 | Nlrp3 | 17.94324 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 358.86474 | 0.01552 |
| NM_013654 | Ccl7 | 17.86522 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 357.30442 | 0.02002 |
| NM_001113326 | Msr1 | 16.61677 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 332.33541 | 0.06050 |
| NM_001169153 | Cd300lf | 16.41884 | 0.05119 | 0.05000 | 0.05000 | 0.05000 | 320.74128 | 0.05601 |
| NM_001111058 | Cd33 | 15.95177 | 0.05109 | 0.05000 | 0.05000 | 0.12485 | 127.77178 | 0.08454 |
| NM_008625 | Mrc1 | 15.23913 | 0.05000 | 0.05000 | 0.05000 | 0.13721 | 111.06669 | 0.08488 |
| NM_001281818 | Specc1 | 14.84498 | 0.10444 | 0.09658 | 0.05000 | 0.05000 | 142.14502 | 0.08370 |
| NM_011337 | Ccl3 | 14.78896 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 295.77915 | 0.03545 |
| NM_134158 | AF251705 | 14.44968 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 288.99364 | 0.02936 |

## Table 3-4 continued

## 100 Fold Bone Marrow Derived Macrophage Specific Genes

| Refseq ID | Gene Name | Bone Marrow Derived Macrophages RPKM | Embryonic Stem Cell RPKM | E 14.5 Cortical Neurons RPKM | CD4+ CD8+ Thymocytes RPKM | Hepatocytes RPKM | Minumum Fold Difference | P-value |
|---|---|---|---|---|---|---|---|---|
| NM_023044 | Slc15a3 | 14.20659 | 0.05000 | 0.05000 | 0.05256 | 0.05000 | 270.30702 | 0.05569 |
| NM_176913 | Dpep2 | 13.71171 | 0.05000 | 0.05000 | 0.06375 | 0.05000 | 215.09560 | 0.07187 |
| NM_001040696 | Nlrp1b | 13.61658 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 272.33166 | 0.03638 |
| NM_007577 | C5ar1 | 12.66434 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 253.28689 | 0.02100 |
| NM_199221 | Cd300lb | 12.51051 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 250.21021 | 0.01601 |
| NM_001281854 | Aoah | 11.81439 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 236.28784 | 0.03914 |
| NM_009841 | Cd14 | 11.59313 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 231.86252 | 0.03066 |
| NM_145509 | 5430435G22Rik | 10.45637 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 209.12738 | 0.03770 |
| NM_001033308 | Themis2 | 10.15814 | 0.05470 | 0.05000 | 0.05000 | 0.05000 | 185.71145 | 0.07534 |
| NM_138310 | Apobr | 9.94353 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 198.87065 | 0.07082 |
| NM_177686 | Clec12a | 9.79479 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 195.89587 | 0.06640 |
| NM_153074 | Lrrc25 | 9.63310 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 192.66210 | 0.04642 |
| NM_145158 | Emilin2 | 9.31194 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 186.23873 | 0.05307 |
| NM_013482 | Btk | 9.13808 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 182.76157 | 0.06050 |
| NM_009779 | C3ar1 | 8.31513 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 166.30264 | 0.05285 |
| NM_011355 | Spi1 | 8.23520 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 164.70391 | 0.05033 |
| NM_001004435 | Pik3r6 | 8.05183 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 161.03651 | 0.05500 |
| NM_152803 | Hpse | 7.58002 | 0.05544 | 0.05000 | 0.06069 | 0.05000 | 124.90431 | 0.08754 |
| NM_001025610 | Ms4a7 | 7.49524 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 149.90470 | 0.05207 |
| NM_027763 | Treml1 | 7.40742 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 148.14841 | 0.06126 |
| NM_011426 | Siglec1 | 6.99533 | 0.05699 | 0.05000 | 0.05000 | 0.05000 | 122.73969 | 0.07767 |
| NM_205820 | Tlr13 | 6.35673 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 127.13454 | 0.02569 |
| NM_001166493 | Rasgrp3 | 6.10249 | 0.05000 | 0.05000 | 0.05000 | 0.05962 | 102.35928 | 0.09131 |
| NM_001164426 | Kcnk13 | 6.03810 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 120.76209 | 0.06554 |
| NM_009807 | Casp1 | 5.42220 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 108.44394 | 0.05432 |
| NM_021297 | Tlr4 | 5.34423 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 106.88469 | 0.03735 |
| NM_007574 | C1qc | 5.21923 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 104.38467 | 0.06676 |
| NM_144539 | Slamf7 | 5.01541 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 100.30824 | 0.06063 |

# Table 3-5

## 100 Fold Hepatocyte Specific Genes

| Refseq ID | Gene Name | Hepatocytes RPKM | Embryonic Stem Cell RPKM | E 14.5 Cortical Neurons RPKM | CD4+ CD8+ Thymocytes RPKM | Bone Marrow Derived Macrophages RPKM | Minumum Fold Difference | P-value |
|---|---|---|---|---|---|---|---|---|
| NM_009654 | Alb | 850.85912 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 17017.18239 | 0.01760 |
| NM_008645 | Mug1 | 200.53821 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 4010.76423 | 0.00000 |
| NM_013465 | Ahsg | 316.58885 | 0.08178 | 0.05589 | 0.07341 | 0.09030 | 3505.83807 | 0.07796 |
| NM_001122647 | Mup10 | 157.52651 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 3150.53013 | 0.00034 |
| NM_009692 | Apoa1 | 150.12379 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 3002.47577 | 0.08488 |
| NM_001286096 | Mup2 | 149.57525 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 2991.50505 | 0.00029 |
| NM_001252569 | Serpina1a | 142.43664 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 2848.73274 | 0.00000 |
| NM_009245 | Serpina1c | 134.17731 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 2683.54619 | 0.00000 |
| NM_001199995 | Mup12 | 133.23854 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 2664.77078 | 0.00044 |
| NM_001135127 | Mup19 | 131.82398 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 2636.47970 | 0.00029 |
| NM_001134675 | Mup7 | 130.73753 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 2614.75067 | 0.00022 |
| NM_009244 | Serpina1b | 129.78424 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 2595.68484 | 0.03156 |
| NM_001200004 | Mup15 | 126.26784 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 2525.35684 | 0.00020 |
| NM_001281979 | Mup9 | 123.16813 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 2463.36265 | 0.00022 |
| NM_001199999 | Mup14 | 121.54775 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 2430.95491 | 0.00034 |
| NM_031188 | Mup1 | 506.02931 | 0.19734 | 0.13486 | 0.12923 | 0.21791 | 2322.23819 | 0.00015 |
| NM_001039544 | Mup3 | 110.84384 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 2216.87686 | 0.00064 |
| NM_011458 | Serpina3k | 109.27784 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 2185.55683 | 0.05940 |
| NM_001199333 | LOC100048884 | 108.48562 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 2169.71233 | 0.00064 |
| NM_011044 | Pck1 | 107.22084 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 2144.41676 | 0.04950 |
| NM_001199936 | Mup16 | 100.63025 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 2012.60509 | 0.00064 |
| NM_013474 | Apoa2 | 208.20296 | 0.05000 | 0.06139 | 0.05000 | 0.10545 | 1974.39993 | 0.09172 |
| NM_021282 | Cyp2e1 | 97.75749 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1955.14973 | 0.00000 |
| NM_001164526 | Mup11 | 97.03364 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1940.67282 | 0.00064 |
| NM_001134676 | Mup8 | 96.00393 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1920.07861 | 0.00093 |
| NM_001200006 | Mup17 | 95.71986 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1914.39712 | 0.00090 |
| NM_013697 | Ttr | 88.86616 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1777.32316 | 0.08798 |
| NM_017399 | Fabp1 | 88.83222 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1776.64438 | 0.09038 |
| NM_001134674 | Mup13 | 87.57621 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1751.52427 | 0.00090 |
| NM_008096 | Gc | 81.18870 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1623.77392 | 0.08813 |
| NM_008277 | Hpd | 80.95663 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1619.13259 | 0.06153 |
| NM_001080809 | Cps1 | 80.86087 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1617.21738 | 0.09143 |
| NM_009246 | Serpina1d | 80.73597 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1614.71937 | 0.00000 |
| NM_146214 | Tat | 72.90907 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1458.18131 | 0.08969 |
| NM_019792 | Cyp3a25 | 65.11589 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1302.31782 | 0.09153 |
| NM_008646 | Mug2 | 65.02588 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1300.51764 | 0.00027 |
| NM_133862 | Fgg | 62.36437 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1247.28731 | 0.09180 |
| NM_009247 | Serpina1e | 61.22536 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1224.50710 | 0.00000 |
| NM_001190732 | Gm20594 | 4485.90508 | 3.20780 | 0.49085 | 3.66918 | 0.36733 | 1222.58915 | 0.15900 |
| NM_181849 | Fgb | 59.16627 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1183.32531 | 0.09094 |

## Table 3-5 continued

## 100 Fold Hepatocyte Specific Genes

| Refseq ID | Gene Name | Hepatocytes RPKM | Embryonic Stem Cell RPKM | E 14.5 Cortical Neurons RPKM | CD4+ CD8+ Thymocytes RPKM | Bone Marrow Derived Macrophages RPKM | Minumum Fold Difference | P-value |
|---|---|---|---|---|---|---|---|---|
| NM_198672 | Ces3a | 78.21781 | 0.06438 | 0.05000 | 0.05000 | 0.07109 | 1100.26678 | 0.00010 |
| NM_001134644 | Gm2083 | 66.02625 | 0.05535 | 0.05000 | 0.05000 | 0.06112 | 1080.27349 | 0.00147 |
| NM_010006 | Cyp2d9 | 53.25007 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1065.00149 | 0.00024 |
| NM_023114 | Apoc3 | 50.99010 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 1019.80192 | 0.08986 |
| NM_007376 | Pzp | 55.85013 | 0.05595 | 0.05000 | 0.05000 | 0.05000 | 998.19012 | 0.09282 |
| NM_017370 | Hp | 48.75876 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 975.17511 | 0.09167 |
| NM_080845 | Ftcd | 45.19420 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 903.88395 | 0.08744 |
| NM_007606 | Car3 | 45.05313 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 901.06265 | 0.09170 |
| NM_008877 | Plg | 44.94791 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 898.95822 | 0.09153 |
| NM_001111048 | Fga | 44.67690 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 893.53802 | 0.09104 |
| NM_007817 | Cyp2f2 | 43.29222 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 865.84444 | 0.06735 |
| NM_009780 | C4b | 43.08862 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 861.77245 | 0.09182 |
| NM_009693 | Apob | 40.66979 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 813.39585 | 0.09529 |
| NM_019414 | Selenbp2 | 39.12078 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 782.41562 | 0.09158 |
| NM_007443 | Ambp | 37.72946 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 754.58925 | 0.09285 |
| NM_019911 | Tdo2 | 37.56060 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 751.21201 | 0.09285 |
| NM_013478 | Azgp1 | 36.76458 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 735.29168 | 0.08500 |
| NM_009253 | Serpina3m | 35.69297 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 713.85944 | 0.07952 |
| NM_023617 | Aox3 | 46.68301 | 0.06737 | 0.05000 | 0.05000 | 0.05000 | 692.91279 | 0.12695 |
| NM_013475 | Apoh | 33.59207 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 671.84136 | 0.09165 |
| NM_010406 | Hc | 32.32595 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 646.51895 | 0.10637 |
| NM_007954 | Ces1c | 31.63888 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 632.77763 | 0.09167 |
| NM_016668 | Bhmt | 30.33683 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 606.73659 | 0.00007 |
| NM_133653 | Mat1a | 30.16094 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 603.21872 | 0.08754 |
| NM_007818 | Cyp3a11 | 29.78872 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 595.77442 | 0.09155 |
| NM_008406 | Itih1 | 28.77039 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 575.40778 | 0.08977 |
| NM_001252616 | Fut8 | 28.38661 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 567.73216 | 0.00560 |
| NM_008061 | G6pc | 28.22501 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 564.50012 | 0.09182 |
| NM_001159415 | Ces3b | 28.08855 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 561.77091 | 0.08238 |
| NM_010005 | Cyp2d10 | 28.05610 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 561.12198 | 0.08053 |
| NM_019546 | Prodh2 | 28.04150 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 560.83005 | 0.09155 |
| NM_001150749 | Rdh7 | 27.44077 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 548.81532 | 0.00007 |
| NM_007482 | Arg1 | 26.90464 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 538.09272 | 0.11742 |
| NM_080434 | Apoa5 | 26.61618 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 532.32353 | 0.04975 |
| NM_001012323 | Mup20 | 26.45991 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 529.19823 | 0.00565 |
| NM_144940 | Uroc1 | 26.34236 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 526.84725 | 0.08930 |
| NM_008101 | Gcgr | 41.47632 | 0.08299 | 0.05000 | 0.05000 | 0.05000 | 499.77008 | 0.11617 |
| NM_008407 | Itih3 | 55.37335 | 0.05000 | 0.11346 | 0.05000 | 0.05000 | 488.06354 | 0.10927 |
| NM_133946 | Nlrp6 | 24.12604 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 482.52084 | 0.11292 |
| NM_007409 | Adh1 | 23.61279 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 472.25587 | 0.09140 |

## Table 3-5 continued

## 100 Fold Hepatocyte Specific Genes

| Refseq ID | Gene Name | Hepatocytes RPKM | Embryonic Stem Cell RPKM | E 14.5 Cortical Neurons RPKM | CD4+ CD8+ Thymocytes RPKM | Bone Marrow Derived Macrophages RPKM | Minumum Fold Difference | P-value |
|---|---|---|---|---|---|---|---|---|
| NM_007815 | Cyp2c29 | 22.88506 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 457.70130 | 0.09167 |
| NM_007385 | Apoc4 | 22.01433 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 440.28655 | 0.11710 |
| NM_001013777 | Zfp488 | 21.97497 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 439.49943 | 0.10091 |
| NM_001159487 | Rbp4 | 59.57028 | 0.14257 | 0.05251 | 0.05000 | 0.05000 | 417.83679 | 0.14511 |
| NM_001009550 | Mup21 | 19.89251 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 397.85028 | 0.12098 |
| NM_032541 | Hamp | 20.07532 | 0.05000 | 0.05000 | 0.05086 | 0.05000 | 394.73763 | 0.12803 |
| NM_009474 | Uox | 19.70910 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 394.18208 | 0.09167 |
| NM_001102411 | Kng1 | 19.42512 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 388.50245 | 0.09370 |
| NM_001243063 | Nr1i3 | 54.52047 | 0.06938 | 0.06735 | 0.06456 | 0.14093 | 386.86308 | 0.10717 |
| NM_008777 | Pah | 18.80795 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 376.15890 | 0.05212 |
| NM_018795 | Abcc6 | 25.72854 | 0.06973 | 0.05000 | 0.05000 | 0.05000 | 368.95198 | 0.14824 |
| NM_011082 | Pigr | 18.23932 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 364.78633 | 0.09148 |
| NM_013786 | Hsd17b6 | 18.07698 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 361.53960 | 0.10649 |
| NM_021022 | Abcb11 | 17.60137 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 352.02732 | 0.11964 |
| NM_009778 | C3 | 90.75740 | 0.05000 | 0.05000 | 0.05000 | 0.25971 | 349.45686 | 0.15096 |
| NM_133657 | Cyp2a12 | 17.30371 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 346.07425 | 0.09092 |
| NM_144909 | Gckr | 29.06369 | 0.08694 | 0.05178 | 0.05000 | 0.05000 | 334.31251 | 0.13047 |
| NM_030611 | Akr1c6 | 16.58442 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 331.68833 | 0.00022 |
| NM_010582 | Itih2 | 16.37381 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 327.47611 | 0.09192 |
| NM_010391 | H2-Q10 | 41.85909 | 0.13046 | 0.05000 | 0.05000 | 0.05000 | 320.86103 | 0.14736 |
| NM_009252 | Serpina3n | 15.89211 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 317.84221 | 0.09285 |
| NM_026701 | Pbld1 | 15.53241 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 310.64816 | 0.11111 |
| NM_008649 | Mup5 | 15.34678 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 306.93567 | 0.09556 |
| NM_144930 | Ces1f | 15.11324 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 302.26473 | 0.09172 |
| NM_001199306 | 2810007J24Rik | 15.04765 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 300.95295 | 0.09241 |
| NM_183249 | 1100001G20Rik | 14.86022 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 297.20432 | 0.14062 |
| NM_146148 | C8a | 14.71787 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 294.35742 | 0.12668 |
| NM_013797 | Slco1a1 | 13.59993 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 271.99866 | 0.09754 |
| NM_013467 | Aldh1a1 | 13.30656 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 266.13121 | 0.09180 |
| NM_011316 | Saa4 | 13.25616 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 265.12314 | 0.09184 |
| NM_007812 | Cyp2a5 | 13.19805 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 263.96110 | 0.00017 |
| NM_053176 | Hrg | 13.15506 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 263.10118 | 0.09854 |
| NM_017371 | Hpx | 56.55486 | 0.05000 | 0.16900 | 0.05000 | 0.21597 | 261.86159 | 0.19325 |
| NM_177002 | Slc22a30 | 12.98061 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 259.61218 | 0.00046 |
| NM_053200 | Ces1d | 12.97609 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 259.52182 | 0.10461 |
| NM_013485 | C9 | 12.87222 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 257.44434 | 0.08111 |
| NM_021489 | F12 | 12.83755 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 256.75094 | 0.11448 |
| NM_001105160 | Cyp3a59 | 12.76412 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 255.28242 | 0.09412 |
| NM_001276710 | Agxt | 12.73384 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 254.67671 | 0.09121 |
| NM_145565 | Sds | 12.70550 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 254.11003 | 0.15272 |

## Table 3-5 continued

## 100 Fold Hepatocyte Specific Genes

| Refseq ID | Gene Name | Hepatocytes RPKM | Embryonic Stem Cell RPKM | E 14.5 Cortical Neurons RPKM | CD4+ CD8+ Thymocytes RPKM | Bone Marrow Derived Macrophages RPKM | Minumum Fold Difference | P-value |
|---|---|---|---|---|---|---|---|---|
| NM_001031851 | Agxt2 | 12.58547 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 251.70949 | 0.09192 |
| NM_133995 | Upb1 | 12.57255 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 251.45105 | 0.11937 |
| NM_145146 | Afm | 12.46505 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 249.30107 | 0.09184 |
| NM_007428 | Agt | 12.17997 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 243.59944 | 0.09268 |
| NM_018746 | Itih4 | 47.75628 | 0.17799 | 0.12164 | 0.11656 | 0.19654 | 242.98191 | 0.18699 |
| NM_001081408 | Agmat | 11.64528 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 232.90565 | 0.13619 |
| NM_001160303 | Gm4788 | 12.66841 | 0.05000 | 0.05000 | 0.05000 | 0.05477 | 231.28445 | 0.00183 |
| NM_028093 | Entpd8 | 11.54180 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 230.83598 | 0.14682 |
| NM_013547 | Hgd | 16.72058 | 0.05000 | 0.07252 | 0.05000 | 0.05000 | 230.55970 | 0.14761 |
| NM_027902 | Tmprss6 | 14.99184 | 0.06702 | 0.05000 | 0.05000 | 0.05000 | 223.69065 | 0.19080 |
| NM_009993 | Cyp1a2 | 11.15190 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 223.03802 | 0.09400 |
| NM_054094 | Acsm1 | 10.58340 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 211.66801 | 0.08879 |
| NM_029692 | Upp2 | 13.71767 | 0.05000 | 0.06507 | 0.05000 | 0.05000 | 210.80168 | 0.18325 |
| NM_001029867 | Ugt2b36 | 10.39757 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 207.95139 | 0.00125 |
| NM_001161667 | Acox2 | 10.35723 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 207.14454 | 0.09192 |
| NM_008768 | Orm1 | 10.30788 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 206.15753 | 0.11272 |
| NM_020495 | Slco1b2 | 20.90911 | 0.05000 | 0.10203 | 0.05000 | 0.05000 | 204.93441 | 0.16367 |
| NM_027853 | Mettl7b | 10.07579 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 201.51580 | 0.10685 |
| NM_144903 | Aldob | 43.62623 | 0.05000 | 0.21708 | 0.05000 | 0.05000 | 200.96422 | 0.17137 |
| NM_030739 | Vmn1r58 | 9.74313 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 194.86254 | 0.00171 |
| NM_031884 | Abcg5 | 9.51614 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 190.32287 | 0.09285 |
| NM_025834 | Proz | 9.47919 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 189.58379 | 0.17750 |
| NM_011707 | Vtn | 33.23511 | 0.05000 | 0.17833 | 0.05000 | 0.05000 | 186.36894 | 0.16337 |
| NM_001081372 | Ces1b | 9.28366 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 185.67315 | 0.13027 |
| NM_153193 | Hsd3b2 | 9.10834 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 182.16678 | 0.09419 |
| NM_010007 | Cyp2j5 | 9.06969 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 181.39374 | 0.09165 |
| NM_011134 | Pon1 | 9.02366 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 180.47316 | 0.09854 |
| NM_007703 | Elovl3 | 9.01511 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 180.30217 | 0.37156 |
| NM_001102409 | Kng2 | 9.00910 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 180.18205 | 0.09192 |
| NM_010775 | Mbl1 | 8.99874 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 179.97474 | 0.09092 |
| NM_027852 | Rarres2 | 12.84925 | 0.07144 | 0.05000 | 0.05000 | 0.05000 | 179.85847 | 0.19797 |
| NM_139300 | Mylk | 8.49603 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 169.92053 | 0.17726 |
| NM_009467 | Ugt2b5 | 8.47992 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 169.59833 | 0.00687 |
| NM_007576 | C4bp | 8.47590 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 169.51791 | 0.09683 |
| NM_147100 | Olfr613 | 8.43631 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 168.72628 | 0.06718 |
| NM_178758 | Acsm5 | 8.33358 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 166.67167 | 0.08879 |
| NM_145499 | Cyp2c70 | 8.26245 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 165.24897 | 0.09167 |
| NM_010012 | Cyp8b1 | 8.20916 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 164.18320 | 0.09167 |
| NM_001204333 | Cyp4f14 | 8.18182 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 163.63635 | 0.14751 |

## Table 3-5 continued

## 100 Fold Hepatocyte Specific Genes

| Refseq ID | Gene Name | Hepatocytes RPKM | Embryonic Stem Cell RPKM | E 14.5 Cortical Neurons RPKM | CD4+ CD8+ Thymocytes RPKM | Bone Marrow Derived Macrophages RPKM | Minumum Fold Difference | P-value |
|---|---|---|---|---|---|---|---|---|
| NM_019503 | Fxyd1 | 14.65518 | 0.05000 | 0.09121 | 0.05000 | 0.05000 | 160.67111 | 0.16015 |
| NM_001163486 | Hsd17b13 | 7.92915 | 0.05007 | 0.05000 | 0.05000 | 0.05000 | 158.34595 | 0.19601 |
| NM_001042767 | Proc | 23.67694 | 0.15190 | 0.05000 | 0.05000 | 0.05000 | 155.86746 | 0.19056 |
| NM_001025575 | Cfhr2 | 7.75950 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 155.18993 | 0.15641 |
| NM_008341 | Igfbp1 | 8.80248 | 0.05000 | 0.05675 | 0.05000 | 0.05000 | 155.11787 | 0.21168 |
| NM_053096 | Cml2 | 9.83357 | 0.06387 | 0.05000 | 0.05000 | 0.05000 | 153.96518 | 0.18447 |
| NM_023623 | Cyp2d40 | 7.56099 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 151.21980 | 0.13013 |
| NM_010168 | F2 | 36.84730 | 0.25137 | 0.05000 | 0.23542 | 0.05000 | 146.58838 | 0.20613 |
| NM_009997 | Cyp2a4 | 7.27598 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 145.51958 | 0.06251 |
| NM_010321 | Gnmt | 50.14684 | 0.34566 | 0.22398 | 0.24054 | 0.32417 | 145.07551 | 0.19885 |
| NM_008878 | Serpinf2 | 20.64251 | 0.05000 | 0.05000 | 0.12162 | 0.14241 | 144.94835 | 0.19349 |
| NM_133977 | Trf | 334.39893 | 1.12245 | 0.06475 | 0.07966 | 2.31252 | 144.60347 | 0.20669 |
| NM_028066 | F11 | 7.20855 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 144.17099 | 0.10825 |
| NM_177142 | Lipi | 13.04970 | 0.08252 | 0.05640 | 0.05404 | 0.09112 | 143.20771 | 0.01716 |
| NM_008124 | Gjb1 | 8.07029 | 0.05000 | 0.05649 | 0.05000 | 0.05000 | 142.85622 | 0.18183 |
| NM_001161742 | Hsd3b3 | 7.13770 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 142.75409 | 0.09221 |
| NM_007686 | Cfi | 6.98796 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 139.75915 | 0.11597 |
| NM_007494 | Ass1 | 74.90605 | 0.54195 | 0.05000 | 0.14924 | 0.28095 | 138.21532 | 0.21104 |
| NM_153168 | Lars2 | 22790.06503 | 32.37085 | 8.52373 | 94.26926 | 165.34499 | 137.83342 | 0.24617 |
| NM_001177964 | Dcdc2c | 6.87996 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 137.59926 | 0.12729 |
| NM_007468 | Apoa4 | 6.79957 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 135.99150 | 0.08977 |
| NM_009060 | Rgn | 6.71353 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 134.27066 | 0.09536 |
| NM_145365 | Creb3l3 | 18.54222 | 0.13841 | 0.05000 | 0.05000 | 0.06700 | 133.96245 | 0.21175 |
| NM_009108 | Nr1h4 | 9.46711 | 0.06438 | 0.05814 | 0.05000 | 0.07109 | 133.17108 | 0.11512 |
| NM_001101475 | F830016B08Rik | 6.60806 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 132.16128 | 0.16528 |
| NM_146230 | Acaa1b | 30.52828 | 0.23569 | 0.05000 | 0.05000 | 0.05000 | 129.52919 | 0.21178 |
| NM_007493 | Asgr2 | 6.71086 | 0.05299 | 0.05000 | 0.05000 | 0.05000 | 126.65326 | 0.19315 |
| NM_027552 | Kynu | 7.64463 | 0.05000 | 0.05000 | 0.05000 | 0.06052 | 126.31935 | 0.18359 |
| NM_152811 | Ugt2b1 | 6.28873 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 125.77468 | 0.09189 |
| NM_144834 | Serpina10 | 6.28684 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 125.73681 | 0.10935 |
| NM_001166350 | Serpina11 | 6.07745 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 121.54891 | 0.09172 |
| NM_133660 | Ces1e | 6.02437 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 120.48737 | 0.09637 |
| NM_183257 | Hamp2 | 5.94592 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 118.91833 | 0.00205 |
| NM_008455 | Klkb1 | 7.00257 | 0.05000 | 0.05000 | 0.05000 | 0.05902 | 118.64891 | 0.18733 |
| NM_009349 | Inmt | 5.91996 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 118.39921 | 0.08248 |
| NM_026180 | Abcg8 | 5.85629 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 117.12588 | 0.09930 |
| NM_080844 | Serpinc1 | 29.98274 | 0.14434 | 0.12476 | 0.25923 | 0.15027 | 115.66017 | 0.20537 |
| NM_001081318 | Gm6614 | 5.71494 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 114.29872 | 0.12859 |
| NM_001104531 | Cyp2d11 | 5.68582 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 113.71636 | 0.09698 |
| NM_001013820 | Slc22a28 | 5.64302 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 112.86042 | 0.10162 |

## Table 3-5 continued

## 100 Fold Hepatocyte Specific Genes

| Refseq ID | Gene Name | Hepatocytes RPKM | Embryonic Stem Cell RPKM | E 14.5 Cortical Neurons RPKM | CD4+ CD8+ Thymocytes RPKM | Bone Marrow Derived Macrophages RPKM | Minumum Fold Difference | P-value |
|---|---|---|---|---|---|---|---|---|
| NM_022884 | Bhmt2 | 5.50363 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 110.07251 | 0.15386 |
| NM_008223 | Serpind1 | 5.46256 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 109.25116 | 0.16110 |
| NM_029562 | Cyp2d26 | 29.19095 | 0.05000 | 0.21333 | 0.26761 | 0.05000 | 109.08118 | 0.21395 |
| NM_009202 | Slc22a1 | 5.31982 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 106.39634 | 0.20662 |
| NM_134127 | Cyp4f15 | 11.45146 | 0.05000 | 0.10774 | 0.05000 | 0.05000 | 106.28916 | 0.20259 |
| NM_009150 | Selenbp1 | 19.24069 | 0.05059 | 0.05000 | 0.05000 | 0.18316 | 105.05063 | 0.20012 |
| NM_177406 | Cyp4a12a | 5.16129 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 103.22583 | 0.29316 |
| NM_010011 | Cyp4a10 | 5.12611 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 102.52216 | 0.00232 |
| NM_008198 | Cfb | 33.16717 | 0.32410 | 0.17231 | 0.05918 | 0.14708 | 102.33557 | 0.22190 |
| NM_010936 | Nr1i2 | 9.21386 | 0.09007 | 0.05000 | 0.05000 | 0.05000 | 102.30092 | 0.21926 |
| NM_001167875 | Cyp2c50 | 5.08752 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 101.75041 | 0.12106 |
| NM_146101 | Habp2 | 5.05905 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 101.18103 | 0.09172 |
| NM_001113418 | Ppara | 6.07911 | 0.05000 | 0.06021 | 0.05000 | 0.05000 | 100.96801 | 0.20620 |
| NM_027904 | Cpn2 | 5.04735 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 100.94697 | 0.09561 |
| NM_009714 | Asgr1 | 13.61249 | 0.05000 | 0.13604 | 0.05734 | 0.05000 | 100.06381 | 0.20613 |

## Table 3.6

## 100 Fold Embryonic Stem Cell Specific Promoter Coordinates

| Refseq ID | Gene Name | Chrom | Start | End | Strand |
|---|---|---|---|---|---|
| NM_008485 | Lamc2 | 1 | 155033527 | 155034077 | - |
| NM_201395 | Sall4 | 2 | 168592651 | 168593201 | - |
| NM_016907 | Spint1 | 2 | 119062595 | 119063145 | + |
| NM_028946 | Slc9b1 | 3 | 135010500 | 135011050 | + |
| NM_001081202 | L1td1 | 4 | 98392944 | 98393494 | + |
| NM_175651 | Cnpy1 | 5 | 28536512 | 28537062 | - |
| NM_011107 | Pla2g1b | 5 | 115915774 | 115916324 | + |
| NM_009426 | Trh | 6 | 92194594 | 92195144 | - |
| NM_178381 | Ano9 | 7 | 148303655 | 148304205 | - |
| NM_028034 | Tdrd12 | 7 | 36322713 | 36323263 | - |
| NM_020486 | Bcam | 7 | 20355831 | 20356381 | - |
| NM_010202 | Fgf4 | 7 | 152046790 | 152047340 | + |
| NM_009482 | Utf1 | 7 | 147129254 | 147129804 | + |
| NM_009434 | Phlda2 | 7 | 150688379 | 150688929 | - |
| NM_001004184 | Slc28a1 | 7 | 88259184 | 88259734 | + |
| NM_177742 | Triml1 | 8 | 44235317 | 44235867 | - |
| NM_009556 | Zfp42 | 8 | 44392313 | 44392863 | - |
| NM_001160412 | Triml2 | 8 | 44265295 | 44265845 | + |
| NM_011562 | Tdgf1 | 9 | 110848612 | 110849162 | - |
| NM_001042503 | Trim71 | 9 | 114473437 | 114473987 | - |
| NM_013611 | Nodal | 10 | 60880219 | 60880769 | + |
| NM_019448 | Dnmt3l | 10 | 77512086 | 77512636 | + |
| NM_001271550 | Aire | 10 | 77506305 | 77506855 | - |
| NM_029458 | Hormad2 | 11 | 4341035 | 4341585 | - |
| NM_028602 | Tex19.1 | 11 | 121006956 | 121007506 | + |
| NM_001159401 | Upp1 | 11 | 9017510 | 9018060 | + |
| NM_0119348* | Esrrb* | 12 | 87810566 | 87811116 | + |
| NM_021480 | Tdh | 14 | 64127879 | 64128429 | - |
| NM_001111119* | Ccnb1ip1* | 14 | 51424682 | 51425232 | - |
| NM_134109 | Ildr1 | 16 | 36693563 | 36694113 | + |
| NM_028610 | Dppa4 | 16 | 48283347 | 48283897 | + |
| NM_030141 | 1700061G19Rik | 17 | 57014555 | 57015105 | + |
| NM_013633 | Pou5f1 | 17 | 35642476 | 35643026 | + |
| NM_001033425 | Zscan10 | 17 | 23737322 | 23737872 | + |
| NM_015798 | Fbxo15 | 18 | 85103916 | 85104466 | + |
| NM_011635 | Trap1a | X | 135867721 | 135868271 | + |
| NM_009575 | Zic3 | X | 55283304 | 55283854 | + |
| NM_007430 | Nr0b1 | X | 83436613 | 83437163 | + |
| NM_001100461 | Pnma5 | X | 70284885 | 70285435 | - |

*** Altered start site**

## Table 3.7

## 100 Fold E14.5 Cortical Neuron Specific Promoter Coordinates

| Refseq ID | Gene Name | Chrom | Start | End | Strand |
|---|---|---|---|---|---|
| NM_175642 | Bai3 | 1 | 25886502 | 25887052 | - |
| NM_133235 | Khdrbs2 | 1 | 32229150 | 32229700 | + |
| NM_053199 | Cadm3 | 1 | 175297776 | 175298326 | - |
| NM_010045 | Darc | 1 | 175263584 | 175264134 | - |
| NM_008900 | Pou3f3 | 1 | 42753490 | 42754040 | + |
| NM_007495 | Astn1 | 1 | 160291934 | 160292484 | + |
| NM_001011874 | Xkr4 | 1 | 3661529 | 3662079 | - |
| NM_172475 | Frmd4a* | 2 | 3938296 | 3938846 | + |
| NM_133207 | Kcnh7 | 2 | 63022294 | 63022844 | - |
| NM_001171615 | Myt1 | 2 | 181497536 | 181498086 | + |
| NM_178673 | Fstl5 | 3 | 75877982 | 75878532 | + |
| NM_138666 | Nlgn1 | 3 | 26230781 | 26231331 | - |
| NM_019978 | Dclk1 | 3 | 55045947 | 55046497 | + |
| NM_007529 | Bcan | 3 | 87804228 | 87804778 | - |
| NM_001167748 | Egfem1 | 3 | 28980998 | 28981548 | + |
| NM_001081358 | Lrrc7 | 3 | 158225135 | 158225685 | - |
| NM_001039347 | Kcnd3 | 3 | 105254747 | 105255297 | + |
| NM_001039195 | Gria2 | 3 | 80606663 | 80607213 | - |
| NM_019724 | Mmp16 | 4 | 17780128 | 17780678 | + |
| NM_011607 | Tnc | 4 | 63707999 | 63708549 | - |
| NM_001281955 | Csmd2 | 4 | 127664787 | 127665337 | + |
| NM_031404 | Actl6b | 5 | 137994282 | 137994832 | + |
| NM_008069 | Gabrb1 | 5 | 72090754 | 72091304 | + |
| NM_007937 | Epha5 | 5 | 84846357 | 84846907 | - |
| NM_001199244 | Kcnip4* | 5 | 49915890 | 49916440 | - |
| NM_001077398 | Ldb2 | 5 | 45190935 | 45191485 | - |
| NM_177328 | Grm7 | 6 | 110595091 | 110595641 | + |
| NM_175750 | Plxna4* | 6 | 32537654 | 32537704 | - |
| NM_080285 | Cttnbp2 | 6 | 18464775 | 18465325 | - |
| NM_008171 | Grin2b | 6 | 136123479 | 136124029 | - |
| NM_001282102 | Lrrtm4 | 6 | 79968370 | 79968920 | + |
| NM_001271858 | Add2 | 6 | 85978174 | 85978724 | + |
| NM_001164316 | Ccser1 | 6 | 61129818 | 61130368 | + |
| NM_001109764 | Ctnna2 | 6 | 77929611 | 77930161 | - |
| NM_001081306 | Ptprz1 | 6 | 22825001 | 22825551 | + |
| NM_198250 | Lrrc4b* | 7 | 51684327 | 51684877 | + |
| NM_013643 | Ptpn5 | 7 | 54389004 | 54389554 | - |
| NM_011807 | Dlg2* | 7 | 97625182 | 97625732 | + |
| NM_007461 | Apba2 | 7 | 71646091 | 71646641 | + |

**Table 3.7 continued**

**100 Fold E14.5 Cortical Neuron Specific Promoter Coordinates**

| Refseq ID | Gene Name | Chrom | Start | End | Strand |
|---|---|---|---|---|---|
| NM_001081414 | Grm5 | 7 | 94732177 | 94732727 | + |
| NM_053171 | Csmd1 | 8 | 17535335 | 17535885 | - |
| NM_019707 | Cdh13 | 8 | 120807154 | 120807704 | + |
| NM_001253754 | Gpm6a | 8 | 55864286 | 55864836 | + |
| NM_001081397 | Myo16 | 8 | 10153422 | 10153972 | + |
| NM_001039154 | Cdh8 | 8 | 101940321 | 101940871 | - |
| NM_177906 | Opcml | 9 | 27598353 | 27598903 | + |
| NM_172290 | Ntm | 9 | 29770664 | 29771214 | - |
| NM_029792 | B3gat1* | 9 | 26540901 | 26541450 | + |
| NM_001170787 | Cntn5 | 9 | 10904725 | 10905275 | - |
| NM_182807 | Fam19a2 | 10 | 122700631 | 122701181 | + |
| NM_181681 | BC005764 | 10 | 79337329 | 79337879 | - |
| NM_001205341 | Ppfia2 | 10 | 105906865 | 105907415 | + |
| NM_001111268 | Grik2 | 10 | 49508510 | 49509060 | - |
| NM_001081035 | Nav3 | 10 | 109893210 | 109893760 | - |
| NM_021286 | Sez6 | 11 | 77743944 | 77744494 | + |
| NM_011856 | Tenm2* | 11 | 37049719 | 37050268 | - |
| NM_009548 | Rnf112 | 11 | 61267338 | 61267888 | - |
| NM_001113325 | Gria1 | 11 | 56824619 | 56825169 | + |
| NM_178714 | Lrfn5* | 12 | 62623587 | 62624187 | + |
| NM_176930 | Nrcam | 12 | 45429371 | 45429921 | + |
| NM_029911 | Kcnk10 | 12 | 99812872 | 99813422 | - |
| NM_011215 | Ptprn2 | 12 | 117723692 | 117724242 | + |
| NM_001286388 | Trim9 | 12 | 71448551 | 71449101 | - |
| NM_001198587 | Nrxn3* | 12 | 89960820 | 89961370 | + |
| NM_001093778 | Myt1l | 12 | 30212748 | 30213298 | + |
| NM_001081017 | Unc79 | 12 | 104186568 | 104187118 | + |
| NM_010151 | Nr2f1 | 13 | 78338193 | 78338743 | - |
| NM_001282961 | Ntrk2 | 13 | 58907429 | 58907979 | + |
| NM_001081348 | Hecw1 | 13 | 14615443 | 14615993 | - |
| NM_207667 | Fgf14* | 14 | 125076668 | 125077218 | - |
| NM_199065 | Slitrk1 | 14 | 109313406 | 109313956 | - |
| NM_019675 | Stmn4 | 14 | 66962711 | 66963261 | + |
| NM_001271799 | Pcdh9 | 14 | 94289838 | 94290388 | - |
| NM_001253361 | Kcnma1 | 14 | 24823377 | 24823927 | - |
| NM_001190187 | Nrg3 | 14 | 40286326 | 40286876 | - |
| NM_001042617 | Cadps | 14 | 13655543 | 13656093 | - |
| NM_176998 | Sybu | 15 | 44619559 | 44620109 | - |
| NM_172610 | Mpped1 | 15 | 83609952 | 83610502 | + |

**Table 3.7 continued**

**100 Fold E14.5 Cortical Neuron Specific Promoter Coordinates**

| Refseq ID | Gene Name | Chrom | Start | End | Strand |
|---|---|---|---|---|---|
| NM_016743 | Nell2 | 15 | 95359087 | 95359637 | - |
| NM_001135688 | Ly6h | 15 | 75397236 | 75397786 | - |
| NM_001081391 | Csmd3* | 15 | 48624164 | 48624714 | - |
| NM_213614 | Sept5 | 16 | 18629981 | 18630531 | - |
| NM_175549 | Robo2* | 16 | 74412020 | 74412570 | - |
| NM_021477 | Rbfox1 | 16 | 5884385 | 5884935 | + |
| NM_010199 | Fgf12 | 16 | 28753279 | 28753829 | - |
| NM_010140 | Epha3 | 16 | 63863933 | 63864483 | - |
| NM_001164268 | Kalrn* | 16 | 34573568 | 34574118 | - |
| NM_027712* | Dlgap1* | 17 | 70318208 | 70318758 | + |
| NM_177284 | Nrxn1 | 17 | 91492092 | 91492642 | - |
| NM_001286013 | Dlk2 | 17 | 46434376 | 46434926 | + |
| NM_199024 | Nol4 | 18 | 23200104 | 23200654 | - |
| NM_138661 | Pcdha9 | 18 | 37157033 | 37157583 | + |
| NM_054072 | Pcdhac1 | 18 | 37089438 | 37089988 | + |
| NM_009961 | Pcdha10 | 18 | 37164473 | 37165023 | + |
| NM_009960 | Pcdha11 | 18 | 37170011 | 37170561 | + |
| NM_007831 | Dcc | 18 | 72510673 | 72511223 | - |
| NM_001039173 | Dok6 | 18 | 89938478 | 89939028 | - |
| NM_028627 | Psd | 19 | 46401596 | 46402146 | - |
| NM_010025 | Dcx | X | 140367712 | 140368262 | - |
| NM_001177961 | Gpm6b | X | 162676374 | 162676924 | + |

## Table 3.8

## 100 Fold $CD4^+$ $CD8^+$ Thymocyte Specific Promoter Coordinates

| Refseq ID | Gene Name | Chrom | Start | End | Strand |
|---|---|---|---|---|---|
| NM_030710 | Slamf6 | 1 | 173847167 | 173847717 | + |
| NM_013730 | Slamf1 | 1 | 173696762 | 173697312 | + |
| NM_011346 | Sell | 1 | 165991706 | 165992256 | + |
| NM_001113391 | Cd247 | 1 | 167718311 | 167718861 | + |
| NM_029983 | Sla2 | 2 | 156712764 | 156713314 | - |
| NM_011246 | Rasgrp1 | 2 | 117168563 | 117169113 | - |
| NM_010165 | Eya2 | 2 | 165480297 | 165480847 | + |
| NM_009020 | Rag2 | 2 | 101464404 | 101464954 | + |
| NM_009019 | Rag1 | 2 | 101489639 | 101490189 | - |
| NM_008859 | Prkcq | 2 | 11093508 | 11094058 | + |
| NM_008091 | Gata3 | 2 | 9800177 | 9800727 | - |
| NM_001083960 | Spo11* | 2 | 172802701 | 172803251 | + |
| NM_173398 | Gpr171 | 3 | 58905693 | 58906243 | - |
| NM_021309 | Sh2d2a | 3 | 87650176 | 87650726 | + |
| NM_013486 | Cd2 | 3 | 101091812 | 101092362 | - |
| NM_010703 | Lef1 | 3 | 130812714 | 130813264 | + |
| NM_205823 | Tlr12 | 4 | 128295813 | 128296363 | - |
| NM_019436 | Sit1 | 4 | 43496531 | 43497081 | - |
| NM_001162432 | Lck | 4 | 129235566 | 129236116 | - |
| NM_001267621 | Gfi1 | 5 | 108154775 | 108155325 | - |
| NM_001122754 | Txk | 5 | 73143962 | 73144512 | - |
| NM_175860 | Gimap1 | 6 | 48688545 | 48689095 | + |
| NM_153175 | Gimap6 | 6 | 48658193 | 48658743 | - |
| NM_013488 | Cd4 | 6 | 124838177 | 124838727 | - |
| NM_009858 | Cd8b1 | 6 | 71272305 | 71272855 | + |
| NM_001081110 | Cd8a | 6 | 71322920 | 71323470 | + |
| NM_001033126 | Cd27 | 6 | 125186995 | 125187545 | - |
| NM_010689 | Lat | 7 | 133512998 | 133513548 | - |
| NM_009824 | Cbfa2t3 | 8 | 125222959 | 125223509 | - |
| NM_013487 | Cd3d | 9 | 44789368 | 44789918 | + |
| NM_009850 | Cd3g | 9 | 44788464 | 44789014 | - |
| NM_009382 | Thy1 | 9 | 43850966 | 43851516 | + |
| NM_007648 | Cd3e | 9 | 44817623 | 44818173 | - |
| NM_001166625 | Ccr9 | 9 | 123675828 | 123676378 | + |
| NM_001013390 | Scn4b | 9 | 44946624 | 44947174 | + |
| NM_183264 | Tespa1 | 10 | 129759407 | 129759957 | + |
| NM_178666 | Themis | 10 | 28387700 | 28388250 | + |
| NM_001198914 | Myb | 10 | 20880740 | 20881290 | - |
| NM_011771 | Ikzf3 | 11 | 98407295 | 98407845 | - |

**Table 3.8 continued**

**100 Fold $CD4^+$ $CD8^+$ Thymocyte Specific Promoter Coordinates**

| Refseq ID | Gene Name | Chrom | Start | End | Strand |
|---|---|---|---|---|---|
| NM_009331 | Tcf7* | 11 | 52096466 | 52097016 | - |
| NM_001281966 | Itk | 11 | 46202967 | 46203517 | - |
| NM_001033186 | Skap1 | 11 | 96325404 | 96325954 | + |
| NM_028878 | Slc6a19 | 13 | 73838093 | 73838643 | - |
| NM_009937 | Colq | 14 | 32390519 | 32391069 | - |
| NM_010815 | Grap2 | 15 | 80402524 | 80403074 | + |
| NM_010742 | Ly6d | 15 | 74593947 | 74594497 | - |
| NM_001168693 | Endou | 15 | 97561786 | 97562336 | - |
| NM_198297 | Trat1 | 16 | 48772019 | 48772569 | - |
| NM_032465 | Cd96 | 16 | 46120311 | 46120861 | - |
| NM_183368 | Sytl3 | 17 | 6909678 | 6910228 | + |
| NM_001038499 | Arsi | 18 | 61071393 | 61071943 | + |
| NM_009852 | Cd6 | 19 | 10904498 | 10905048 | - |
| NM_007650 | Cd5 | 19 | 10813414 | 10813964 | - |
| NM_001043228 | Dntt | 19 | 41103264 | 41103814 | + |
| NM_172435 | P2ry10 | X | 104284173 | 104284723 | + |
| NM_011364 | Sh2d1a | X | 39855284 | 39855834 | + |

## Table 3-9

## 100 Fold Bone Marrow Derived Macrophage Specific Promoter Coordinates

| Refseq ID | Gene Name | Chrom | Start | End | Strand |
|---|---|---|---|---|---|
| NM_144539 | Slamf7 | 1 | 173579080 | 173579630 | - |
| NM_145509 | 5430435G22Rik | 1 | 133584771 | 133585321 | + |
| NM_015811 | Rgs1 | 1 | 146096184 | 146096734 | - |
| NM_010185 | Fcer1g | 1 | 173164430 | 173164980 | - |
| NM_008625 | Mrc1 | 2 | 14150540 | 14151090 | + |
| NM_011426 | Siglec1 | 2 | 130912451 | 130913001 | - |
| NM_011355 | Spi1 | 2 | 90922047 | 90922597 | + |
| NM_011311 | S100a4 | 3 | 90407191 | 90407741 | + |
| NM_001267695 | Ctss | 3 | 95330207 | 95330757 | + |
| NM_021297 | Tlr4 | 4 | 66488344 | 66488894 | + |
| NM_009777 | C1qb | 4 | 136442042 | 136442592 | - |
| NM_007574 | C1qc | 4 | 136448779 | 136449329 | - |
| NM_001110322 | Cd72 | 4 | 43467448 | 43467998 | - |
| NM_001033308 | Themis2 | 4 | 132352229 | 132352779 | - |
| NM_152803 | Hpse | 5 | 101148652 | 101149202 | - |
| NM_145227 | Oas2 | 5 | 121199807 | 121200357 | - |
| NM_001286037 | Ncf1 | 5 | 134705445 | 134705995 | - |
| NM_177686 | Clec12a | 6 | 129299761 | 129300311 | + |
| NM_053110 | Gpnmb | 6 | 48986016 | 48986566 | + |
| NM_011539 | Tbxas1 | 6 | 38868484 | 38869034 | + |
| NM_010819 | Clec4d | 6 | 123211624 | 123212174 | + |
| NM_009779 | C3ar1 | 6 | 122806125 | 122806675 | - |
| NM_001111058 | Cd33 | 7 | 50788491 | 50789041 | - |
| NM_138310 | Apobr | 7 | 133728021 | 133728571 | + |
| NM_011662 | Tyrobp | 7 | 31198306 | 31198856 | + |
| NM_011095 | Pirb | 7 | 3671934 | 3672484 | - |
| NM_007577 | C5ar1 | 7 | 16844839 | 16845389 | - |
| NM_176913 | Dpep2 | 8 | 108508907 | 108509457 | - |
| NM_153074 | Lrrc25 | 8 | 73140242 | 73140792 | + |
| NM_001113326 | Msr1 | 8 | 40727982 | 40728532 | - |
| NM_009807 | Casp1 | 9 | 5298016 | 5298566 | + |
| NM_017372 | Lyz2 | 10 | 116719278 | 116719828 | - |
| NM_013532 | Lilrb4 | 10 | 51210280 | 51210830 | + |
| NM_008147 | Gp49a | 10 | 51199984 | 51200534 | + |
| NM_199221 | Cd300lb | 11 | 114795650 | 114796200 | - |
| NM_145827 | Nlrp3 | 11 | 59354587 | 59355137 | + |
| NM_134158 | AF251705 | 11 | 114863144 | 114863694 | - |
| NM_013654 | Ccl7 | 11 | 81858713 | 81859263 | + |
| NM_013652 | Ccl4 | 11 | 83475585 | 83476135 | + |

**Table 3-9 continued**

**100 Fold Bone Marrow Derived Macrophage Specific Promoter Coordinates**

| Refseq ID | Gene Name | Chrom | Start | End | Strand |
|---|---|---|---|---|---|
| NM_011337 | Ccl3 | 11 | 83462830 | 83463380 | - |
| NM_011333 | Ccl2 | 11 | 81848578 | 81849128 | + |
| NM_001281818 | Specc1 | 11 | 61769764 | 61770314 | + |
| NM_001169153 | Cd300lf | 11 | 114995256 | 114995806 | - |
| NM_001040696 | Nlrp1b | 11 | 71044185 | 71044735 | - |
| NM_001004435 | Pik3r6 | 11 | 68316020 | 68316570 | + |
| NM_178911 | Pld4 | 12 | 113998365 | 113998915 | + |
| NM_001164426 | Kcnk13 | 12 | 101202208 | 101202758 | + |
| NM_008533 | Cd180 | 13 | 103483137 | 103483687 | + |
| NM_001281854 | Aoah | 13 | 20885481 | 20886031 | + |
| NM_001033245 | Hk3 | 13 | 55122696 | 55123246 | - |
| NM_138672 | Stab1 | 14 | 31981777 | 31982327 | - |
| NM_008873 | Plau | 14 | 21655383 | 21655933 | + |
| NM_001163616 | 1810011H11Rik | 14 | 33598648 | 33599198 | + |
| NM_001146022 | Wdfy4 | 14 | 33998202 | 33998752 | - |
| NM_023118 | Dab2 | 15 | 6336247 | 6336797 | + |
| NM_008677 | Ncf4 | 15 | 78074740 | 78075290 | + |
| NM_001166493 | Rasgrp3 | 17 | 75834744 | 75835294 | + |
| NM_145158 | Emilin2 | 17 | 71660255 | 71660805 | - |
| NM_031254 | Trem2 | 17 | 48485225 | 48485775 | + |
| NM_027763 | Treml1 | 17 | 48498740 | 48499290 | + |
| NM_010130 | Emr1 | 17 | 57497608 | 57498158 | + |
| NM_009841 | Cd14 | 18 | 36886258 | 36886808 | - |
| NM_001037859 | Csf1r | 18 | 61264725 | 61265275 | + |
| NM_010821 | Mpeg1 | 19 | 12534768 | 12535318 | + |
| NM_023044 | Slc15a3 | 19 | 10916533 | 10917083 | + |
| NM_001025610 | Ms4a7 | 19 | 11410586 | 11411136 | - |
| NM_205820 | Tlr13 | X | 103338113 | 103338663 | + |
| NM_013482 | Btk | X | 131117629 | 131118179 | - |

## Table 3-10

## 100 Fold Hepatocyte Specific Promoter Coordinates

| Refseq ID | Gene Name | Chrom | Start | End | Strand |
|---|---|---|---|---|---|
| NM_001290273 | Marc1 | 1 | 186635129 | 186635679 | - |
| NM_080844 | Serpinc1 | 1 | 162908236 | 162908786 | + |
| NM_023617 | Aox3 | 1 | 58169479 | 58170029 | + |
| NM_013474 | Apoa2 | 1 | 173154684 | 173155234 | + |
| NM_011082 | Pigr | 1 | 132722760 | 132723310 | + |
| NM_007576 | C4bp | 1 | 132558145 | 132558695 | - |
| NM_001276710 | Agxt | 1 | 95031316 | 95031866 | + |
| NM_001243063 | Nr1i3 | 1 | 173143600 | 173144150 | + |
| NM_001160303 | Gm4788 | 1 | 141677766 | 141678316 | - |
| NM_001080809 | Cps1 | 1 | 67169100 | 67169650 | + |
| NM_001025575 | Cfhr2 | 1 | 141915448 | 141915998 | - |
| NM_029692 | Upp2 | 2 | 58607094 | 58607644 | + |
| NM_028093 | Entpd8 | 2 | 24935342 | 24935892 | + |
| NM_027552 | Kynu | 2 | 43410344 | 43410894 | + |
| NM_021022 | Abcb11 | 2 | 69180623 | 69181173 | - |
| NM_011044 | Pck1 | 2 | 172978073 | 172978623 | + |
| NM_010582 | Itih2 | 2 | 10052260 | 10052810 | - |
| NM_010406 | Hc | 2 | 34916911 | 34917461 | - |
| NM_010168 | F2 | 2 | 91476521 | 91477071 | - |
| NM_007494 | Ass1 | 2 | 31325289 | 31325839 | + |
| NM_153193 | Hsd3b2 | 3 | 98528416 | 98528966 | - |
| NM_001161742 | Hsd3b3 | 3 | 98567001 | 98567551 | - |
| NM_181849 | Fgb | 3 | 82853662 | 82854212 | - |
| NM_133862 | Fgg | 3 | 82811317 | 82811867 | + |
| NM_019911 | Tdo2 | 3 | 81779600 | 81780150 | - |
| NM_019414 | Selenbp2 | 3 | 94496994 | 94497544 | + |
| NM_009474 | Uox | 3 | 146259612 | 146260162 | + |
| NM_009150 | Selenbp1 | 3 | 94736504 | 94737054 | + |
| NM_007686 | Cfi | 3 | 129539156 | 129539706 | + |
| NM_007606 | Car3 | 3 | 14863037 | 14863587 | + |
| NM_007409 | Adh1 | 3 | 137940108 | 137940658 | + |
| NM_001111048 | Fga | 3 | 82829574 | 82830124 | + |
| NM_010011 | Cyp4a10 | 4 | 115190391 | 115190941 | + |
| NM_177406 | Cyp4a12a | 4 | 114971150 | 114971700 | + |
| NM_146148 | C8a | 4 | 104548953 | 104549503 | - |
| NM_144903 | Aldob | 4 | 49562305 | 49562855 | - |
| NM_031188 | Mup1 | 4 | 60514782 | 60515332 | - |
| NM_010007 | Cyp2j5 | 4 | 96330760 | 96331310 | - |
| NM_008768 | Orm1 | 4 | 63005099 | 63005649 | + |

**Table 3-10 continued**

**100 Fold Hepatocyte Specific Promoter Coordinates**

| Refseq ID | Gene Name | Chrom | Start | End | Strand |
|---|---|---|---|---|---|
| NM_007443 | Ambp | 4 | 62815126 | 62815676 | - |
| NM_001286096 | Mup2 | 4 | 60152678 | 60153228 | - |
| NM_001281979 | Mup9 | 4 | 60434774 | 60435324 | - |
| NM_001200006 | Mup17 | 4 | 61256814 | 61257364 | - |
| NM_001200004 | Mup15 | 4 | 60152680 | 60153230 | - |
| NM_001199999 | Mup14 | 4 | 60964980 | 60965530 | - |
| NM_001199995 | Mup12 | 4 | 60736103 | 60736653 | - |
| NM_001199936 | Mup16 | 4 | 61180468 | 61181018 | - |
| NM_001199333 | LOC100048884 | 4 | 61335078 | 61335628 | - |
| NM_001164526 | Mup11 | 4 | 60675217 | 60675767 | - |
| NM_001135127 | Mup19 | 4 | 61443208 | 61443758 | - |
| NM_001134676 | Mup8 | 4 | 60235421 | 60235971 | - |
| NM_001134675 | Mup7 | 4 | 60083297 | 60083847 | - |
| NM_001134674 | Mup13 | 4 | 60889268 | 60889818 | - |
| NM_001134644 | Gm2083 | 4 | 60675187 | 60675737 | - |
| NM_001122647 | Mup10 | 4 | 60594977 | 60595527 | - |
| NM_001081408 | Agmat | 4 | 141302089 | 141302639 | + |
| NM_001039544 | Mup3 | 4 | 61748296 | 61748846 | - |
| NM_001012323 | Mup20 | 4 | 61715101 | 61715651 | - |
| NM_001009550 | Mup21 | 4 | 61811825 | 61812375 | - |
| NM_001163486 | Hsd17b13 | 5 | 104406357 | 104406907 | - |
| NM_152811 | Ugt2b1 | 5 | 87355478 | 87356028 | - |
| NM_145565 | Sds | 5 | 120926055 | 120926605 | + |
| NM_145146 | Afm | 5 | 90947474 | 90948024 | + |
| NM_144909 | Gckr | 5 | 31599453 | 31600003 | + |
| NM_019792 | Cyp3a25 | 5 | 146821143 | 146821693 | - |
| NM_013478 | Azgp1 | 5 | 138422248 | 138422798 | + |
| NM_009654 | Alb | 5 | 90889414 | 90889964 | + |
| NM_009467 | Ugt2b5 | 5 | 87569315 | 87569865 | - |
| NM_008277 | Hpd | 5 | 123632645 | 123633195 | - |
| NM_008096 | Gc | 5 | 89886873 | 89887423 | - |
| NM_007818 | Cyp3a11 | 5 | 146691380 | 146691930 | - |
| NM_001105160 | Cyp3a59 | 5 | 146890333 | 146890883 | + |
| NM_001029867 | Ugt2b36 | 5 | 87521530 | 87522080 | - |
| NM_001190732 | Gm20594 | 6 | 79767524 | 79768074 | + |
| NM_144940 | Uroc1 | 6 | 90282782 | 90283332 | + |
| NM_053096 | Cml2 | 6 | 85819081 | 85819631 | - |
| NM_027852 | Rarres2 | 6 | 48522619 | 48523169 | - |
| NM_020495 | Slco1b2 | 6 | 141577538 | 141578088 | + |

**Table 3-10 continued**

**100 Fold Hepatocyte Specific Promoter Coordinates**

| Refseq ID | Gene Name | Chrom | Start | End | Strand |
|---|---|---|---|---|---|
| NM_017399 | Fabp1 | 6 | 71149381 | 71149931 | + |
| NM_013797 | Slco1a1 | 6 | 141895433 | 141895983 | - |
| NM_011134 | Pon1 | 6 | 5143896 | 5144446 | - |
| NM_009349 | Inmt | 6 | 55124934 | 55125484 | - |
| NM_008646 | Mug2 | 6 | 121956315 | 121956865 | + |
| NM_008645 | Mug1 | 6 | 121788058 | 121788608 | + |
| NM_007376 | Pzp | 6 | 128476688 | 128477238 | - |
| NM_001081318 | Gm6614 | 6 | 141960775 | 141961325 | - |
| NM_183257 | Hamp2 | 7 | 31709150 | 31709700 | - |
| NM_178758 | Acsm5 | 7 | 126669278 | 126669828 | + |
| NM_175250 | 2810007J24Rik | 7 | 15031886 | 15032436 | - |
| NM_133946 | Nlrp6 | 7 | 148106300 | 148106850 | + |
| NM_133657 | Cyp2a12 | 7 | 27813608 | 27814158 | + |
| NM_054094 | Acsm1 | 7 | 126760841 | 126761391 | + |
| NM_032541 | Hamp | 7 | 31728986 | 31729536 | - |
| NM_021282 | Cyp2e1 | 7 | 147949230 | 147949780 | + |
| NM_019546 | Prodh2 | 7 | 31278176 | 31278726 | + |
| NM_019503 | Fxyd1 | 7 | 31840625 | 31841175 | - |
| NM_018795 | Abcc6 | 7 | 53285606 | 53286156 | - |
| NM_017371 | Hpx | 7 | 112748580 | 112749130 | - |
| NM_011316 | Saa4 | 7 | 53987863 | 53988413 | - |
| NM_009997 | Cyp2a4 | 7 | 27091710 | 27092260 | + |
| NM_007817 | Cyp2f2 | 7 | 27904473 | 27905023 | + |
| NM_007812 | Cyp2a5 | 7 | 27619857 | 27620407 | + |
| NM_007385 | Apoc4 | 7 | 20266759 | 20267309 | - |
| NM_198672 | Ces3a | 8 | 107571998 | 107572548 | + |
| NM_146214 | Tat | 8 | 112513835 | 112514385 | + |
| NM_144930 | Ces1f | 8 | 95803585 | 95804135 | - |
| NM_133660 | Ces1e | 8 | 95753468 | 95754018 | - |
| NM_053200 | Ces1d | 8 | 95721653 | 95722203 | - |
| NM_028066 | F11 | 8 | 46347335 | 46347885 | - |
| NM_025834 | Proz | 8 | 13060407 | 13060957 | + |
| NM_017370 | Hp | 8 | 112103022 | 112103572 | - |
| NM_008455 | Klkb1 | 8 | 46380139 | 46380689 | - |
| NM_007954 | Ces1c | 8 | 95655132 | 95655682 | - |
| NM_007428 | Agt | 8 | 127093557 | 127094107 | - |
| NM_001159415 | Ces3b | 8 | 107607154 | 107607704 | + |
| NM_001081372 | Ces1b | 8 | 95603866 | 95604416 | - |
| NM_146230 | Acaa1b | 9 | 119066161 | 119066711 | - |

**Table 3-10 continued**

**100 Fold Hepatocyte Specific Promoter Coordinates**

| Refseq ID | Gene Name | Chrom | Start | End | Strand |
|---|---|---|---|---|---|
| NM_133977 | Trf | 9 | 103132566 | 103133116 | - |
| NM_080434 | Apoa5 | 9 | 46076190 | 46076740 | + |
| NM_023114 | Apoc3 | 9 | 46043332 | 46043882 | - |
| NM_010012 | Cyp8b1 | 9 | 121825373 | 121825923 | - |
| NM_009993 | Cyp1a2 | 9 | 57531412 | 57531962 | - |
| NM_009692 | Apoa1 | 9 | 46036212 | 46036762 | + |
| NM_007468 | Apoa4 | 9 | 46048426 | 46048976 | + |
| NM_013786 | Hsd17b6 | 10 | 127444514 | 127445064 | - |
| NM_145365 | Creb3l3 | 10 | 80561567 | 80562117 | - |
| NM_133995 | Upb1 | 10 | 74869155 | 74869705 | + |
| NM_080845 | Ftcd | 10 | 76037892 | 76038442 | + |
| NM_027853 | Mettl7b | 10 | 128397994 | 128398544 | - |
| NM_026701 | Pbld1 | 10 | 62523864 | 62524414 | + |
| NM_008777 | Pah | 10 | 86984039 | 86984589 | + |
| NM_007482 | Arg1 | 10 | 24647226 | 24647776 | - |
| NM_001163504 | Nr1h4 | 10 | 88996317 | 88996867 | - |
| NM_001150749 | Rdh7 | 10 | 127325739 | 127326289 | - |
| NM_183249 | 1100001G20Rik | 11 | 83559941 | 83560491 | + |
| NM_013475 | Apoh | 11 | 108256110 | 108256660 | + |
| NM_011707 | Vtn | 11 | 78312121 | 78312671 | + |
| NM_009714 | Asgr1 | 11 | 69867370 | 69867920 | + |
| NM_008878 | Serpinf2 | 11 | 75252953 | 75253503 | - |
| NM_008341 | Igfbp1 | 11 | 7097289 | 7097839 | + |
| NM_008101 | Gcgr | 11 | 120383680 | 120384230 | + |
| NM_008061 | G6pc | 11 | 101228543 | 101229093 | + |
| NM_007493 | Asgr2 | 11 | 69905645 | 69906195 | + |
| NM_144834 | Serpina10 | 12 | 104869565 | 104870115 | - |
| NM_011458 | Serpina3k | 12 | 105576195 | 105576745 | + |
| NM_009693 | Apob | 12 | 7983982 | 7984532 | + |
| NM_009253 | Serpina3m | 12 | 105624873 | 105625423 | + |
| NM_009252 | Serpina3n | 12 | 105644417 | 105644967 | + |
| NM_009247 | Serpina1e | 12 | 105195057 | 105195607 | - |
| NM_009246 | Serpina1d | 12 | 105011793 | 105012343 | - |
| NM_009245 | Serpina1c | 12 | 105143110 | 105143660 | - |
| NM_009244 | Serpina1b | 12 | 104976349 | 104976899 | - |
| NM_001252569 | Serpina1a | 12 | 105101779 | 105102329 | - |
| NM_001166350 | Serpina11 | 12 | 105228117 | 105228667 | - |
| NM_030611 | Akr1c6 | 13 | 4433088 | 4433638 | + |
| NM_022884 | Bhmt2 | 13 | 94444207 | 94444757 | - |

**Table 3-10 continued**

**100 Fold Hepatocyte Specific Promoter Coordinates**

| Refseq ID | Gene Name | Chrom | Start | End | Strand |
|---|---|---|---|---|---|
| NM_021489 | F12 | 13 | 55528113 | 55528663 | - |
| NM_016668 | Bhmt | 13 | 94407663 | 94408213 | - |
| NM_133653 | Mat1a | 14 | 41917821 | 41918371 | + |
| NM_018746 | Itih4 | 14 | 31699161 | 31699711 | + |
| NM_010775 | Mbl1 | 14 | 41964244 | 41964794 | + |
| NM_008407 | Itih3 | 14 | 31736723 | 31737273 | - |
| NM_008406 | Itih1 | 14 | 31756425 | 31756975 | - |
| NM_001161667 | Acox2 | 14 | 9091303 | 9091853 | - |
| NM_029562 | Cyp2d26 | 15 | 82624625 | 82625175 | - |
| NM_027902 | Tmprss6 | 15 | 78299014 | 78299564 | - |
| NM_023623 | Cyp2d40 | 15 | 82594502 | 82595052 | - |
| NM_013485 | C9 | 15 | 6394832 | 6395382 | + |
| NM_010006 | Cyp2d9 | 15 | 82282306 | 82282856 | + |
| NM_010005 | Cyp2d10 | 15 | 82237574 | 82238124 | - |
| NM_001113418 | Ppara | 15 | 85565493 | 85566043 | + |
| NM_001104531 | Cyp2d11 | 15 | 82224402 | 82224952 | - |
| NM_001031851 | Agxt2 | 15 | 10287833 | 10288383 | + |
| NM_013465 | Ahsg | 16 | 22891587 | 22892137 | + |
| NM_139300 | Mylk | 16 | 34744796 | 34745346 | + |
| NM_053176 | Hrg | 16 | 22950644 | 22951194 | + |
| NM_027904 | Cpn2 | 16 | 30267568 | 30268118 | - |
| NM_013547 | Hgd | 16 | 37579738 | 37580288 | + |
| NM_010936 | Nr1i2 | 16 | 38294860 | 38295410 | - |
| NM_008223 | Serpind1 | 16 | 17330963 | 17331513 | + |
| NM_001102411 | Kng1 | 16 | 23057872 | 23058422 | + |
| NM_001102409 | Kng2 | 16 | 23029124 | 23029674 | - |
| NM_134127 | Cyp4f15 | 17 | 32822103 | 32822653 | + |
| NM_031884 | Abcg5 | 17 | 85082213 | 85082763 | - |
| NM_026180 | Abcg8 | 17 | 85081960 | 85082510 | + |
| NM_010391 | H2-Q10 | 17 | 35606533 | 35607083 | + |
| NM_010321 | Gnmt | 17 | 46866064 | 46866614 | - |
| NM_009780 | C4b | 17 | 34880792 | 34881342 | - |
| NM_009778 | C3 | 17 | 57367509 | 57368059 | - |
| NM_009202 | Slc22a1 | 17 | 12868654 | 12869204 | - |
| NM_008877 | Plg | 17 | 12570974 | 12571524 | + |
| NM_008198 | Cfb | 17 | 34999409 | 34999959 | - |
| NM_001204333 | Cyp4f14 | 17 | 33054224 | 33054774 | - |
| NM_013697 | Ttr | 18 | 20823250 | 20823800 | + |
| NM_008934 | Proc | 18 | 32297809 | 32298359 | - |

**Table 3-10 continued**

**100 Fold Hepatocyte Specific Promoter Coordinates**

| Refseq ID | Gene Name | Chrom | Start | End | Strand |
|---|---|---|---|---|---|
| NM_001101475 | F830016B08Rik | 18 | 60452533 | 60453083 | + |
| NM_177002 | Slc22a30 | 19 | 8479545 | 8480095 | - |
| NM_146101 | Habp2 | 19 | 56361927 | 56362477 | + |
| NM_145499 | Cyp2c70 | 19 | 40261726 | 40262276 | - |
| NM_013467 | Aldh1a1 | 19 | 20675971 | 20676521 | + |
| NM_007815 | Cyp2c29 | 19 | 39361074 | 39361624 | + |
| NM_007703 | Elovl3 | 19 | 46205888 | 46206438 | + |
| NM_001167875 | Cyp2c50 | 19 | 40163668 | 40164218 | + |
| NM_001159487 | Rbp4 | 19 | 38199761 | 38200311 | - |
| NM_001013820 | Slc22a28 | 19 | 8206422 | 8206972 | - |
| NM_009060 | Rgn | X | 20126443 | 20126993 | + |
| NM_008124 | Gjb1 | X | 98572175 | 98572725 | + |

**Table 3-11**

**100 Fold ESC Specific Gene Ontology**

| Gene Ontology Enrichment | P-value |
|---|---|
| In utero embryonic development | 3.6E-12 |
| Embryo development | 8.3E-12 |
| Chordate embryonic development | 2.3E-10 |
| Embryo development ending in birth or egg hatching | 2.9E-10 |
| Placenta development | 1.9E-08 |
| Blastocyst development | 2.7E-08 |
| Blastocyst formation | 2.9E-08 |
| Embryonic morphogenesis | 2.8E-07 |
| Stem cell maintenance | 1.7E-06 |
| Trophectodermal cellular morphogenesis | 4.3E-06 |

**Table 3-12**

**100 Fold E14.5 Cortical Neuron Specific Gene Ontology**

| Gene Ontology Enrichment | P-value |
|---|---|
| Synaptic transmission | 1.34E-30 |
| Cell-cell signaling | 1.37E-26 |
| Nervous system development | 1.67E-25 |
| Synapse organization | 1.35E-17 |
| Neuron projection development | 1.53E-16 |
| Generation of neurons | 1.94E-16 |
| System development | 1.29E-15 |
| Neurogenesis | 3.00E-15 |
| Synaptic transmission, glutamatergic | 8.86E-15 |
| Neuron differentiation | 2.33E-14 |

**Table 3-13**

**100 Fold $CD4^+$ $CD8^+$ Thymocyte Specific Gene Ontology**

| Gene Ontology Enrichment | P-value |
|---|---|
| Immune system process | 7.36E-13 |
| Lymphocyte activation | 2.86E-10 |
| Leukocyte activation | 5.26E-09 |
| T cell activation | 3.80E-08 |
| Cell activation | 5.22E-08 |
| Positive regulation of immune response | 2.20E-07 |
| T cell receptor signaling pathway | 8.93E-07 |
| Regulation of immune system process | 1.30E-06 |
| Positive regulation of immune system process | 8.00E-06 |
| Regulation of immune response | 1.22E-05 |

**Table 3-14**

**100 Fold Bone Marrow Derived Macrophage Specific Gene Ontology**

| Gene Ontology Enrichment | P-value |
|---|---|
| Immune system process | 1.87E-23 |
| Defense response | 1.93E-22 |
| Inflammatory response | 3.33E-22 |
| Response to wounding | 1.38E-21 |
| Immune response | 2.73E-18 |
| Regulation of response to external stimulus | 1.15E-13 |
| Positive regulation of response to stimulus | 4.82E-13 |
| Positive regulation of cytokine production | 2.11E-12 |
| Response to external stimulus | 2.36E-12 |
| Regulation of immune system process | 2.68E-12 |

**Table 3-15**

**100 Fold Hepatocyte Specific Gene Ontology**

| Gene Ontology Enrichment | P-value |
|---|---|
| Organic acid metabolic process | 2.23E-37 |
| Oxidation-reduction process | 9.05E-37 |
| Carboxylic acid metabolic process | 3.91E-35 |
| Oxoacid metabolic process | 8.79E-34 |
| Negative regulation of peptidase activity | 2.69E-33 |
| Single-organism metabolic process | 1.16E-31 |
| Negative regulation of hydrolase activity | 5.83E-30 |
| Negative regulation of endopeptidase activity | 4.44E-29 |
| Blood coagulation | 2.50E-24 |
| Hemostasis | 4.35E-24 |

## REFERENCES


Barrera, L.O., Li, Z., Smith, A.D., Arden, K.C., Cavenee, W.K., Zhang, M.Q., Green, R.D., and Ren, B. (2008). Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs. Genome research *18*, 46-59.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., *et al.* (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315-326.

Bhatt, D.M., Pandya-Jones, A., Tong, A.J., Barozzi, I., Lissner, M.M., Natoli, G., Black, D.L., and Smale, S.T. (2012). Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. Cell *150*, 279-290.

Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K., *et al.* (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature *441*, 349-353.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A., Taylor, M.S., Engstrom, P.G., Frith, M.C., *et al.* (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet *38*, 626-635.

Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes & development *25*, 1010-1022.

Dejosez, M., and Zwaka, T.P. (2012). Pluripotency and nuclear reprogramming. Annual review of biochemistry *81*, 737-765.

Efroni, S., Duttagupta, R., Cheng, J., Dehghani, H., Hoeppner, D.J., Dash, C., Bazett-Jones, D.P., Le Grice, S., McKay, R.D., Buetow, K.H., *et al.* (2008). Global transcription in pluripotent embryonic stem cells. Cell stem cell *2*, 437-447.

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

Gifford, C.A., Ziller, M.J., Gu, H., Trapnell, C., Donaghey, J., Tsankov, A., Shalek, A.K., Kelley, D.R., Shishkin, A.A., Issner, R., *et al.* (2013). Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. Cell *153*, 1149-1163.

Hanna, J.H., Saha, K., and Jaenisch, R. (2010). Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. Cell *143*, 508-525.

Hemmati-Brivanlou, A., and Melton, D. (1997). Vertebrate embryonic cells will become nerve cells unless told otherwise. Cell *88*, 13-17.

Jones, P.A., and Takai, D. (2001). The role of DNA methylation in mammalian epigenetics. Science *293*, 1068-1070.

Kafri, T., Ariel, M., Brandeis, M., Shemer, R., Urven, L., McCarrey, J., Cedar, H., and Razin, A. (1992). Developmental pattern of gene-specific DNA methylation in the mouse embryo and germ line. Genes & development *6*, 705-714.

Kmiec, Z. (2001). Cooperation of liver cells in health and disease. Advances in anatomy, embryology, and cell biology *161*, Iii-xiii, 1-151.

Levin, J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A., and Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nature methods *7*, 709-715.

Lienert, F., Mohn, F., Tiwari, V.K., Baubec, T., Roloff, T.C., Gaidatzis, D., Stadler, M.B., and Schubeler, D. (2011). Genomic prevalence of heterochromatic H3K9me2 and transcription do not discriminate pluripotent from terminally differentiated cells. PLoS genetics *7*, e1002090.

Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D.*, et al.* (2013). Global epigenomic reconfiguration during mammalian brain development. Science *341*, 1237905.

Meister, P., Towbin, B.D., Pike, B.L., Ponti, A., and Gasser, S.M. (2010). The spatial dynamics of tissue-specific promoters during C. elegans development. Genes & development *24*, 766-782.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods *5*, 621-628.

Munoz-Sanjuan, I., and Brivanlou, A.H. (2002). Neural induction, the default model and embryonic stem cells. Nature reviews Neuroscience *3*, 271-280.

Mutz, K.O., Heilkenbrinker, A., Lonne, M., Walter, J.G., and Stahl, F. (2013). Transcriptome analysis using next-generation sequencing. Current opinion in biotechnology *24*, 22-30.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H.*, et al.* (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature *420*, 563-573.

Pandya-Jones, A., and Black, D.L. (2009). Co-transcriptional splicing of constitutive and alternative exons. RNA (New York, NY) *15*, 1896-1908.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research *35*, D61-65.

Ramirez-Carrozzi, V.R., Braas, D., Bhatt, D.M., Cheng, C.S., Hong, C., Doty, K.R., Black, J.C., Hoffmann, A., Carey, M., and Smale, S.T. (2009). A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. Cell *138*, 114-128.

Sachs, M., Onodera, C., Blaschke, K., Ebata, K.T., Song, J.S., and Ramalho-Santos, M. (2013). Bivalent chromatin marks developmental regulatory genes in the mouse embryonic germline in vivo. Cell reports *3*, 1777-1784.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proceedings of the National Academy of Sciences of the United States of America *103*, 1412-1417.

Schwartz, R.E., Fleming, H.E., Khetani, S.R., and Bhatia, S.N. (2014). Pluripotent stem cell-derived hepatocyte-like cells. Biotechnology advances *32*, 504-513.

Song, Y., Ahn, J., Suh, Y., Davis, M.E., and Lee, K. (2013). Identification of novel tissue-specific genes by analysis of microarray databases: a human and mouse model. PloS one *8*, e64483.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Scholer, A., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., Tiwari, V.K., *et al.* (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490-495.

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, P., Zhou, X., Farnworth, S.L., Patel, A.H., and Hay, D.C. (2013). Modeling human liver biology using stem cell-derived hepatocytes. International journal of molecular sciences *14*, 22011-22021.

Thomson, M., Liu, S.J., Zou, L.N., Smith, Z., Meissner, A., and Ramanathan, S. (2011). Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. Cell *145*, 875-889.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology *28*, 511-515.

Tropepe, V., Hitoshi, S., Sirard, C., Mak, T.W., Rossant, J., and van der Kooy, D. (2001). Direct neural fate specification from embryonic stem cells: a primitive mammalian neural stem cell stage acquired through a default mechanism. Neuron *30*, 65-78.

Varley, K.E., Gertz, J., Bowling, K.M., Parker, S.L., Reddy, T.E., Pauli-Behn, F., Cross, M.K., Williams, B.A., Stamatoyannopoulos, J.A., Crawford, G.E., *et al.* (2013). Dynamic DNA methylation across diverse human cell lines and tissues. Genome research *23*, 555-567.

Vastenhouw, N.L., and Schier, A.F. (2012). Bivalent histone modifications in early embryogenesis. Current opinion in cell biology *24*, 374-386.

Vincent, J.J., Huang, Y., Chen, P.Y., Feng, S., Calvopina, J.H., Nee, K., Lee, S.A., Le, T., Yoon, A.J., Faull, K., *et al.* (2013). Stage-specific roles for tet1 and tet2 in DNA demethylation in primordial germ cells. Cell stem cell *12*, 470-478.

Voigt, P., Tee, W.W., and Reinberg, D. (2013). A double take on bivalent promoters. Genes & development *27*, 1318-1338.

Wei, C.L., Miura, T., Robson, P., Lim, S.K., Xu, X.Q., Lee, M.Y., Gupta, S., Stanton, L., Luo, Y., Schmitt, J., *et al.* (2005). Transcriptome profiling of human and murine ESCs identifies divergent paths required to maintain the stem cell state. Stem cells *23*, 166-185.

Wuarin, J., and Schibler, U. (1994). Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. Molecular and cellular biology *14*, 7219-7225.

Xie, W., Schultz, M.D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J.W., Tian, S., Hawkins, R.D., Leung, D., *et al.* (2013). Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell *153*, 1134-1148.

Zhang, J.A., Mortazavi, A., Williams, B.A., Wold, B.J., and Rothenberg, E.V. (2012). Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity. Cell *149*, 467-482.

Zhu, J., He, F., Hu, S., and Yu, J. (2008). On the nature of human housekeeping genes. Trends in genetics : TIG *24*, 481-484.

# Chapter 4

# Concluding Remarks

In the studies presented here we sought to address the fundamental properties associated with tissue specific genes and the chromatin signatures that define the regulatory elements associated with these genes. We took two approaches, one narrow, which focused closely on the well-characterized thymocyte specific gene Ptcra and its associated enhancer, the other broad, using deep chromatin RNA-sequencing to generate a highly quantitative and unbiased assessment of the transcriptome in multiple cell types.

The Ptcra enhancer is marked by unmethylated CpG dinucleotides in embryonic stem cells. We used a bacterial artificial chromosome containing the pTa gene locus to recapitulate the endogenous chromatin environment. Key transcription factor binding sites reside within the enhancer mark, which are involved in the thymocyte specification. We conducted mutational analyses in the pTa enhancer in the BAC to determine if these transcription factor binding sites contribute to the unmethylated state of the enhancer. We show that no individual mutation in the transcription factor binding sites alters the methylation status of the enhancer. We also show that deletions of in the enhancer do no alter the methylation status of the enhancer in the context of the BAC. Together these data show the pTa enhancer mark reappears robustly after stable integration of the BAC after *in vitro* premethylation. The BAC provided a native chromatin context but is not ideal. To further these studies efforts must be made towards targeting and altering endogenous loci. Recently emerging techniques, such as CRISPR-Cas, make it feasible to modify endogenous loci, rapidly, efficiently and with ease.

The reappearance of the mark, although interesting, does not indicate functional significance. In order to progress these analyses, systems must be designed to determine the relevance of the marks in relation to expression in the differentiated cell type. This would include a method to stably alter the chromatin environment at the enhancer, followed by differentiation

to the appropriate cell type. If important, differences in expression compared to wild type would be observed. If enhancer marks in ES cells are necessary for transcription upon lineage differentiation, it will add to the properties pluripotency. In order for ES cells to be truly considered pluripotent the appropriate markings must be present at enhancers.

We then broadened our scope to look genome-wide at tissue specific genes. We used chromatin RNA-sequencing, which has clear advantages over older transcriptome profiling techniques. We quantified the transcriptome in embryonic stem cells, E14.5 cortical neurons, $CD4^+$ $CD8^+$ thymocytes, bone marrow derived macrophages and hepatocytes. We used this quantification to identify tissue specific genes that show the largest dynamic range in expression. We determined the DNA methylation and histone properties found at the promoters of our tissue specific genes. Consistent with the literature we show an inverse correlation between CpG island content and tissue specificity. Interestingly, we observed a very high percentage of neuronal specific genes that possess a CpG island at their promoter. We speculate that our tissue specific neuronal genes may need to be dynamically regulated, as the presence of a CpG island correlates with a permissive chromatin environment and is thus accessible for rapid activation.

We then addressed the histone modification profile of our tissue specific promoters in expressing and non-expressing cells. As expected in expressing cells tissue specific promoters are unmethylated and associated with positive histone marks such as H3K4me3 and H3K9Ac. In non-expressing cells the epigenetic signatures fell into two categories, genes with and without a CpG island promoter. Genes that lacked a CpG island promoter followed the existing paradigm, showing heavily methylated promoters in all non-expressing cell types. Genes with a CpG island follow an almost perfect rule in non-expressing cell types; they are unmethylated and marked with H3K4me3 and H3K27me3. This provides some insights into the idea of bivalency, the

simultaneous marking of genes with H3K4me3 and H3K27me3. We show marking by K4/K27 in non-pluripotent, differentiated cell types. This lends towards the transcriptional status of the associated gene to be a direct indicator of the ability of CpG islands to be marked with K27. In this fashion is does not appear to be a mechanism for poising.

To further these studies there are a number of clear directions. Tissue specific gene expression is heavily influenced by distal regulatory elements. Identifying and locating enhancers for these tissue specific genes will likely yield insights into the mechanisms of tissue specific regulation. The use of chromatin RNA-Seq also provides an opportunity to identify non-coding transcripts and determine if tissue specific differences exist.

All Together we have described an interesting feature associated with the Ptcra enhancer in ES cells. We have performed quantitative transcriptome analysis and defined tissue specific genes with the largest dynamic range in gene expression. With those genes, we have characterized the DNA methylation and histone properties in non-expressing cells and provided added insights into bivalency. Beside these contributions the highly quantitative transcriptome analysis will be of great utility for future studies.