Genomic Analysis of Control of Cell Type

by

Garrett M. Frampton

B.S. Biological Chemistry, Biology
The University of Chicago, 2001

SUBMITTED TO THE DEPARTMENT OF BIOLOGY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN BIOLOGY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 2011

Signature of Author .....
Garrett M. Frampton
January 31, 2011

Certified by .....
Dr. Richard A. Young
Professor of Biology
Thesis Supervisor

Accepted by .....
Dr. Steve Bell
Professor of Biology
Chairperson, Biology Graduate Committee

Genomic Analysis of Control of Cell Type

by

Garrett M. Frampton

B.S. Biological Chemistry, Biology
The University of Chicago, 2001

SUBMITTED TO THE DEPARTMENT OF BIOLOGY ON JANUARY 31, 2011
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN BIOLOGY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Abstract

In mammalian development, a single fertilized egg grows into a complex organism, comprised of organs and tissues made up of hundreds of different specialized cell types. All of these cells contain the same genome, but express distinct sets of genes and proteins, which give the cells their specialized functions. Understanding how this process occurs is one of the fundamental goals of biology.

Research using new technology for high-resolution genome-wide location analysis and gene expression profiling has allowed characterization of the transcriptional regulatory circuitry of cells in unprecedented detail. From this research emerges an improved understanding of how these cells work, how they malfunction in disease, and several general principles. This thesis describes several studies designed to understand the transcriptional regulatory circuitry of three different medically important cell types; embryonic stem cells, regulatory T cells and MLL leukemia cells.

Thesis Supervisor: Dr. Richard A. Young

Title: Professor of Biology

Acknowledgments

I want to thank many people for the role that they have played in helping me get to complete this degree. My parents, John and Judith, have strongly supported me in just about everything that I have done. This has led me to believe that I could accomplish whatever I put my mind to and carried me through difficult times. My wife, Shefali, has been a loving companion though all of the ups and downs of graduate school. My sister, Stephanie, I am so glad I have gotten to spend the past 6 years living above you. Thank you all very much, I love you.

The 6+ years that I have spent at MIT have been a gift. It is truly a wonderful place which gave me the incredible opportunity to experience just about anything that I wanted so to the MIT community, thank you very much. To all of my friends at MIT, I am glad that our paths crossed, good luck to you.

I would like to thank Michael Christman, Marc Lenburg, and Norman Gerry, with whom I worked at the Boston University Department of Genetics and Genomics. Without your guidance I would not be where I am, I am truly grateful. I am also indebted to my colleagues and mentors at the University of Chicago, the Newton Massachusetts public schools, and at Cambridge Harvard Community Health Plan pharmacy.

I would especially like to thank all of my collaborators on the research that led to the stories that are presented in this document. The science that we do is a team effort.

I want to thank the members of my various thesis committees, Rudolf Jaenisch, Aviv Regev, Dave Housman, Uttam Rajbandhary, Mike Laub, and Anglea DePace for taking time out of there extremely busy schedules to help me.

Lastly, I want to thank my thesis advisor, Richard Young. Without your research foundantion and mentorship, none of this would have been possible.

Thank you all very much.

Table of Contents

Chapter 1

Introduction: Transcriptional Regulatory Circuitry and Genomic Analysis

Abstract

Understanding how genes are selectively transcribed to generate cell type specific expression programs is one of the fundamental goals of biology. New technology for high-resolution genome-wide location analysis and gene expression profiling has begun to allow the characterization of the transcriptional regulatory circuitry of several medically important cell types in unprecedented detail. From this research is emerging an improved understanding of how these cells work, how they malfunction in disease, and several general themes. In this introduction, I briefly describe how transcription factors and chromatin regulators contribute to the control of gene expression programs. I then describe the key types of high-throughput data that are being used to investigate transcriptional control and focus the remainder of the chapter on proper analysis of such genomic data.

Transcriptional Regulatory Circuitry

In mammalian development, a single fertilized egg grows into a complex organism, comprised of organs and tissues made up of hundreds of different specialized cell types. All of these cells contain the same genome, but express distinct sets of genes and proteins, which give the cells their specialized functions. Understanding how this process occurs is one of the fundamental goals of biology.

Cells use an array of biochemical mechanisms to control their gene expression programs and, of these, the action of transcription factors and chromatin regulators is especially important. These proteins physically interact with the genome and regulate transcription by modulating the activity of various components of the transcriptional apparatus. By interacting differently with different portions of the genome, they promote the expression of a subset set of genes at appropriate levels and the repression of others. Furthermore, they provide a means for cells to maintain gene expression programs over time and through cell division. We call the collection of these elements, working together to control gene expression and cell state, the transcriptional regulatory circuitry of the cell (Figure 1).

Developing new means to treat human disease is a constant beacon for biology research. Understanding the transcriptional regulatory circuitry that controls cell type will play an increasingly important role in developing new treatments for human disease. It is likely that defects in normal regulatory circuitry are the cause of a subset of common human diseases, so detailed knowledge of the circuitry of these diseased cells should help researchers to understand the molecular underpinnings of disease.

Figure 1

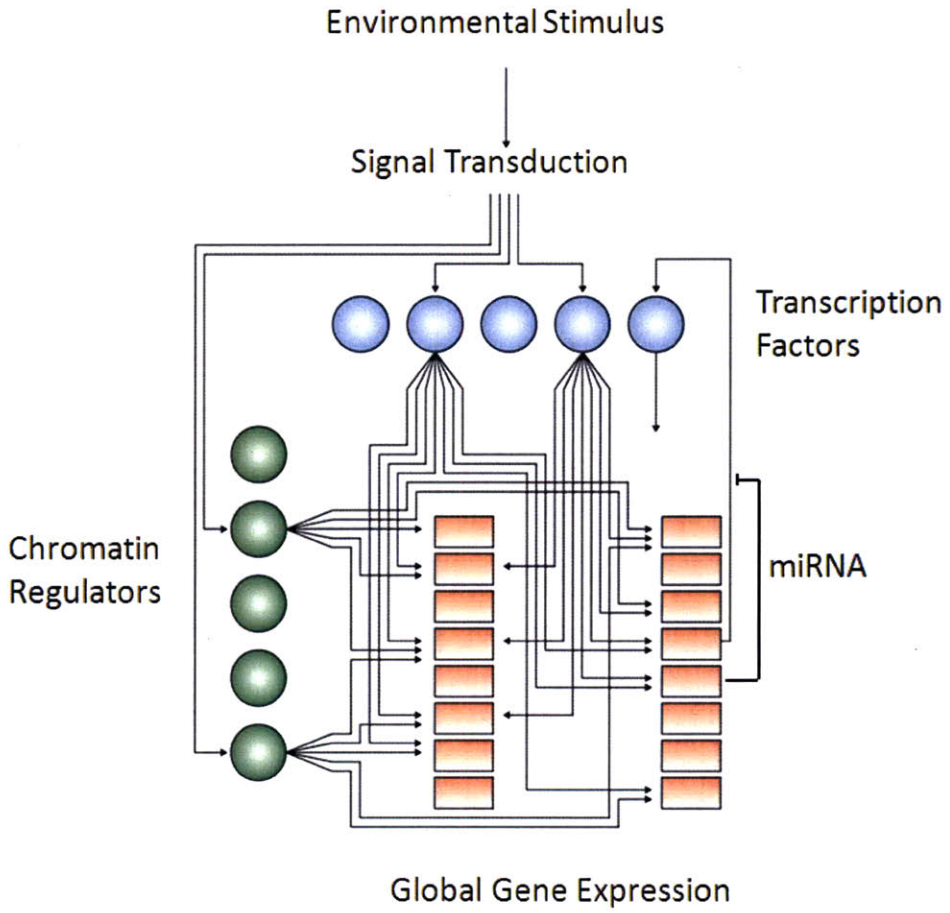


Figure 1. Model of transcriptional regulatory circuitry

A schematic showing the hypothetical circuitry that connects signal transduction pathways, transcription factors (blue circles), chromatin regulators (green circles), and their target genes (orange rectangles) to form transcriptional regulatory circuitry (Jaenisch and Young 2008).

Transcription Factors

Early studies of the *E. coli lac* operon (Jacob and Monod 1961) are fundamental to our current understanding how transcription is regulated by the action of transcription factors (TFs). In the absence of lactose molecules, transcription of the *lac* operon is turned off by the Lac repressor transcription factor. This protein binds to a specific DNA sequence that occurs close to the transcription start site of the *lac* operon. When bound to this DNA element, the Lac repressor interacts with RNA polymerase, preventing productive transcription. Lactose molecules, if present, bind to the Lac repressor protein with high affinity, altering its conformation and abolishing its ability to bind to DNA, thereby promoting transcription of the *lac* operon. Later experiments in yeast showed how, in eukaryotes, a single TF species can regulate many genes and control an entire gene expression program (Giniger et al. 1985; Harbison et al. 2004). Since TFs can regulate the expression of many genes, in multicellular organisms they provide a powerful mechanism for enabling cell type specific gene expression programs.

Work in *Drosophila melanogaster* examining the function of the bithorax gene cluster demonstrated that cells contain genes that govern cell type during development (Lewis 1963; Lewis 1978). Homeotic mutations such as the famous Antennapedia mutant, in which a fly's normal antennae are replaced by a pair of legs, are caused by defects in these genes (Figure 2). In the 1980s it was discovered that the genes within the bithorax gene cluster encode homeobox TFs and that, remarkably, the homeobox genes clusters are conserved from flies to humans (Figure 3) (McGinnis et al. 1984; Scott and Weiner 1984; McGinnis and Krumlauf 1992; Hughes and Kaufman 2002). Thus, homeobox TFs govern cell type by enacting cell type specific gene expression programs. For most well characterized cell types, a small set of TFs are key to those cells' transcriptional regulatory circuitry and define their developmental state. We call these the master regulator TFs of that cell type.

Figure 2

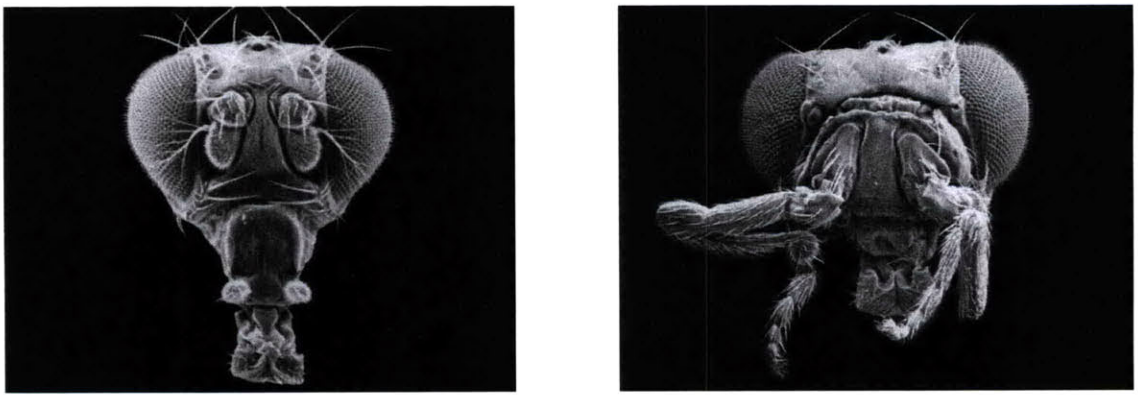


Figure 2. The phenotype of the *Drosophila* Antennapedia mutant

The head of a wild-type fly (left) is shown in comparison to the head of an Antennapedia mutant fly (right), demonstrating a homeotic transformation of antennae into ectopic legs (Turner and Mahowald 1979).

Figure 3

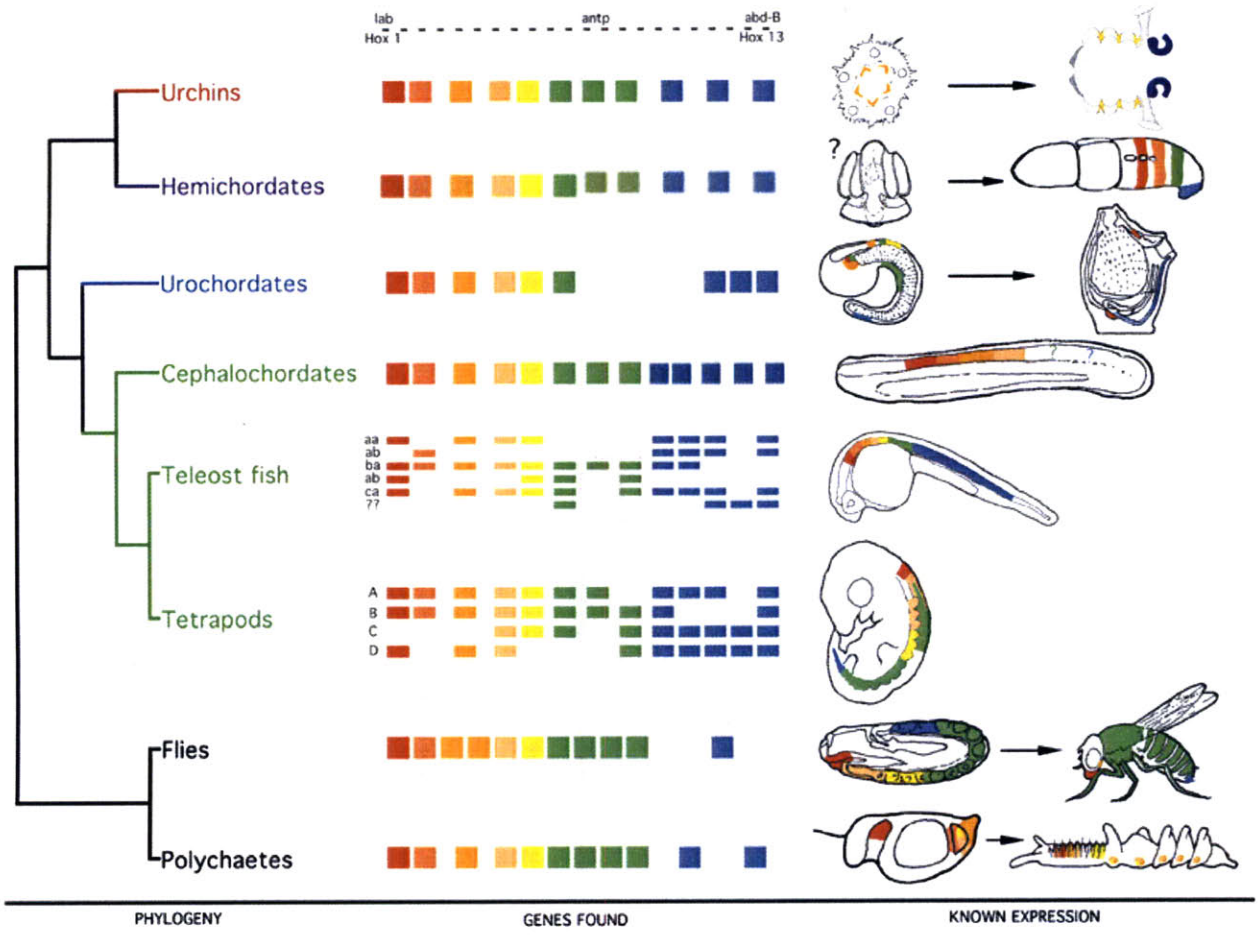


Figure 3. Hox gene clusters are conserved through deuterostomes

The evolutionary relationships between deuterostome clades, the genomic organization of the Hox cluster(s), and the anatomical expression patterns of the Hox genes are shown. There are eight Hox gene clusters in teleost fishes, four Hox gene clusters in tetrapod vertebrates, and a single cluster in invertebrates. In many species the Hox gene clusters are broken up onto two chromosomes (Swalla 2006).

Another early example of the power of a master regulator TF to control cell type was the demonstration that mouse fibroblast cells could be converted into myoblast cells by the ectopic expression of the gene encoding the transcription factor MyoD1 (Davis et al. 1987). It has since been shown that the MyoD1 TF can mediate trans-differentiation to generate myoblast cells from many, but not all, starting cell types in several species including human (Weintraub et al. 1991). The MyoD1 protein has been so highly conserved through evolution that the mouse protein can effectively substitute for the human protein (Weintraub 1993). Deficiencies in the function of MyoD1 and the partially redundant TF Myf5 result in failure to develop muscle cells (Rudnicki et al. 1993), and MyoD1 is required for several cell types that occur sequentially during muscle development (Buckingham et al. 2003).

The MyoD1 transcription factor occupies many thousands of binding sites throughout the genome in myoblast cells (Cao et al. 2010). These sites contain the specific DNA motif that is bound by MyoD1, these motifs are significantly more conserved than neighboring sequences, and are also significantly more conserved than instances of the same sequence that are not occupied by MyoD1 (Cao et al. 2010). MyoD1 tends to occupy gene promoter regions and is biased towards genes that are expressed in muscle cells. It is even more strongly biased towards genes that are specifically expressed in muscle cells, but not other cell types (Cao et al. 2010). Lastly, ectopic expression of MyoD1 in ES cells results in MyoD1 occupying a large fraction of the sites it occupied in muscle cells and promotes some parts of the muscle cell gene expression program (Nishiyama et al. 2009).

The role of MyoD1 in the transcriptional regulatory circuitry of myoblast cells demonstrates several general themes in the control of regulatory circuitry by master regulator TFs. These properties are not necessarily true of all master regulator TFs, but do serve as useful guides. Briefly, they are summarized as follows: 1) Master regulator TFs are usually necessary for the proper development of the cells that they specify and can be sufficient to induce trans-differentiation. 2) Master regulator TFs frequently specify several cell types that occur sequentially during development. 3) Master regulator TFs are highly conserved through evolution in terms of the cell types that they control, their protein structure, and the DNA motifs to which they bind. 4) Master regulator TFs bind to thousands of sites in the genome, particularly in gene promoter regions, and these sites are more likely to be conserved than surrounding sequences. 5) The genes that are targeted by master regulator TFs are enriched in

genes that are actively expressed and highly enriched in genes which are expressed specifically in that cell type.

The example of MyoD1 was the first demonstration of TF mediated trans-differentiation, but subsequently, numerous additional examples of this phenomenon have been reported. For example, adipocyte cells can be created from several cell types including fibroblast cells using the TF PPAR-gamma-2 (Tontonoz et al. 1994). Macrophage cells can be created from fibroblast cells using the TFs PU.1 and CEBP-alpha or CEBP-beta (Feng et al. 2008). Finally, the creation of induced pluripotent stem cells (iPS) cells using several different sets of TFs, including Oct4, Sox2, and Nanog demonstrate the power of TFs to control cell type and even force reprogramming through transitions between cell types that are far from each other in development (Takahashi and Yamanaka 2006; Takahashi et al. 2007; Wernig et al. 2007; Yu et al. 2007).

DNA-binding transcription factors are the largest single class of proteins encoded in the human genome, representing approximately 10% of all protein-coding genes (Lander et al. 2001; Levine and Tjian 2003; Babu et al. 2004). The DNA binding specificities for a large portion of these genes have been determined *in vitro* via protein binding microarray (Badis et al. 2009), and the precise DNA elements that are bound *in vivo* have been determined for several transcription factors with ChIP-Seq experiments (Barski et al. 2007; Robertson et al. 2007; Marson et al. 2008). Understanding the global role of transcription factors in controlling the transcriptional regulatory circuitry of every cell cell types is an important goal for biology and medicine.

Chromatin Regulators

Another key molecular mechanism that cells use to control cell type specific gene expression programs is the regulation of chromatin state. In mammalian cells, DNA exists together in physical complexes with an array of proteins that together make up chromatin. The majority of these proteins are histones, which form an octameric nucleosome core particle, wrapping 147 base pairs of DNA around their protein spool (Olins and Olins 1974; Finch and Klug 1976; McGhee and Felsenfeld 1980; Luger et al. 1997; Davey et al. 2002). Nucleosomes physically compact DNA and are also fundamentally important in the regulation of transcription (Li et al. 2007). Genomic regions that are densely populated by nucleosomes tend to be less transcriptionally active than regions that have lower nucleosome density (Knezetic and Luse 1986; Gilbert et al. 2004). Nucleosomes are generally displaced from active transcription start sites, and the presence of transcription factors and the transcriptional machinery can influence local nucleosome positioning (Bernstein et al. 2004; Yuan et al. 2005; Segal et al. 2006; Mavrich et al. 2008; Schones et al. 2008; Tirosh and Barkai 2008; Jiang and Pugh 2009).

Chromatin regulators can be subdivided into two general classes. The first, chromatin remodeling complexes, can add, move, or remove nucleosomes from DNA, which can enhance or reduce access to specific DNA sequences by transcription factors or affect mobility of the transcription apparatus with consequent effects on gene activity (Tsukiyama et al. 1994; Causton et al. 2001; Narlikar et al. 2002; de la Serna et al. 2006; Segal et al. 2006; Schones et al. 2008; Ho and Crabtree 2010). Additionally, chromatin remodeling complexes can introduce variant histones into chromatin as a means to regulate transcription (Ahmad and Henikoff 2002; Mizuguchi et al. 2004).

The second class of chromatin state regulators are the histone modifying enzymes, which catalyze the chemical modification (methylation, actelyation, phosphorylation, ubiquitinization, sumolyation) of specific residues (lysine, serine, arginine) in the tails of histones (Phillips 1963; Brownell et al. 1996; Ogryzko et al. 1996; Strahl and Allis 2000; Jenuwein and Allis 2001; Pokholok et al. 2005). There are also histone modifying enzymes that catalyze the removal of many of these modifications (Nagy et al. 1997; Jones et al. 1998; Nan et al. 1998; Agger et al. 2008). Modification of histones can create binding sites that recruit certain transcriptional regulators and abolish binding sites for others. For example, the acetylation of lysine residues

forms binding sites for proteins that contain bromodomains (Haynes et al. 1992; Dhalluin et al. 1999; Mujtaba et al. 2007) and methylation of the same lysine residues forms binding sites for proteins that contain chromodomains (Pearce et al. 1992; Ekwall et al. 1995; Fischle et al. 2003). Additionally, chemical modification of histone tails can alter their charge or steric properties and modify their interaction with DNA (Schiessel 2003; Korolev et al. 2006).

Many histone modifications are associated with local gene activation or repression and with specific portions of gene bodies. For example, H3K4me3 modified nucleosomes occur at the start site of genes that experience transcriptional initiation (Guenther et al. 2007). H3K79me2 modified nucleosomes occur in the first exon and intron of genes that are being actively transcribed (Marson et al. 2008) while H3K36me3 modified nucleosomes occur from the second exon to the end of the transcribed region (Mikkelsen et al. 2007). H3K27me3 and H2AK119ub modified nucleosomes are associated with transcriptional repression and occur in broad domains at the promoters of genes encoding key regulators of development (Ku et al. 2008). Most chromatin regulators do not have DNA sequence specific binding properties and are recruited to specific sites in the genome by transcription factors, components of the transcription apparatus, or non-coding RNA species. Consequently, understanding how key chromatin regulators are recruited to their genomic targets is an area of active research.

Genomic Experiments and Analysis

Since the advent of DNA sequencing technologies in the 1980s, the adoption of the microarray in the 1990s, and high throughput sequencing in this decade, new genomic technologies have allowed biology researchers to generate unprecedented quantities of data about their systems of interest. Unfortunately, in many cases, these data have not been fully exploited, in large part due to a lack of expertise in dealing with genomic datasets. Genomic experiments produce large and complex datasets and there are many different ways to analyze this data. The methods that are used often have a critical impact on the ability to gain biological insight from the data. Furthermore, technology in this area is rapidly changing, so developing and employing the best possible data analysis methods is a constant challenge. Indeed, the development of new data analysis techniques was crucial to the discoveries that are described in this thesis.

In the research conducted in the Young lab, there are two principle experimental techniques that we rely upon to quantify the state of cells. The first, chromatin-immunoprecipitation (ChIP) based location analysis experiments, allow the identification the genomic sites occupied by a TF, chromatin regulator, or histone modification (Ren et al. 2000; Farnham 2009). The second, RNA expression analysis experiments, allow the simultaneous quantification of the transcription level of every gene in the genome.

In ChIP experiments, cells are crosslinked to chemically attach proteins to DNA, lysed, and then sonicated to break the DNA into small fragments. Subsequently, a specific antibody is used to immunoprecipitate the protein of interest, coupled with the DNA to which it was bound. This DNA is then assayed to determine the regions of the genome that were occupied by the protein. When the work described in this thesis was begun, our laboratory used microarrays (ChIP-chip) to assay immunoprecipitated material from location analysis experiments (Lee et al. 2006). In 2007, the ChIP-Seq method was developed. ChIP-Seq uses high throughput sequencing to assay ChIP material and offers genome-wide coverage as well as improved sensitivity and specificity, as compared with ChIP-chip (Barski et al. 2007; Mardis 2007; Robertson et al. 2007; Park 2009; Hawkins et al. 2010).

The other experimental technique that we rely upon heavily is RNA expression analysis. As with location analysis, microarrays had been used as the principle means to assay RNA abundance in RNA expression analysis. In these experiments RNA is isolated from a sample of

interest, converted into cDNA, labeled, and hybridized to a microarray which contains DNA probes that are complementary to every gene. Probes that match highly transcribed genes will have a bright signal when the microarray is scanned while probes that match genes which are not transcribed will not light up. Just as in location analysis, high-throughput sequencing has supplanted microarrays as a means to assay the genomic material for RNA expression analysis as well (Mortazavi et al. 2008; Nagalakshmi et al. 2008). The RNA-Seq method allows accurate quantification of RNA species and yields much more information about RNA splicing and novel transcripts than microarray based gene expression experiments (Hawkins et al. 2010).

Both location analysis experiments and RNA expression experiments produce large and complex datasets, and the choice of analysis methods has important implications in the interpretation of the results of these experiments. In the remainder of this introductory chapter, I will highlight some of the key features of the analysis of genomic data.

First, it is important to understand the structure of a genomic dataset. The results of a single genomic experiment can usually be described as a vector. This vector contains the values obtained for a large number of measurements made in parallel on a single sample. For RNA expression experiments each row corresponds to the measured RNA expression level of a single species of RNA. For location analysis experiments each row corresponds to the measured occupancy level of a protein of interest at a single position in the genome. A typical gene expression experiment measures the expression of between 10 and 100 thousand genes, while a typical ChIP-Seq location analysis experiment measures the genomic occupancy at between 10 and 100 million positions in the genome. A genomic dataset, containing the results for multiple samples can be described as a matrix. It is important to note that since both gene expression data and location analysis data have a similar structure, many of the same analytic methods can be used for each.

A second key feature of genomic data analysis is sample normalization. This is used to make to signal values from different samples in a genomic dataset as comparable as possible. Consider the following example in which three gene expression samples are being compared. The first sample has a minimum signal of 3 units, a median signal of 213 units, and a maximum signal of 25,123 units. The second sample has a minimum of 11, a median of 872, and a maximum of 75,824. The third sample has a minimum of 52, a median of 541, and a maximum

of 60,996. In this example, if one wanted to compare data from these three samples, it would be difficult, because each sample has a different range of values.

One of the best methods for sample normalization is quantile or rank normalization which was described by Bolstad et al. (Bolstad et al. 2003). In quantile normalization, the row in each dataset with the greatest signal is identified. The average of these values is calculated and then the greatest signal in each dataset is replaced with that average. This is repeated for the second greatest signal in each dataset, then third greatest signal in each dataset, and continues to the lowest signal in each dataset. If two or more rows are tied then each is given the mean of the set of ranks that the tied genes span. In the preceding example, after quantile normalization each sample would have a minimum signal of 22 units, a median signal of 512 units, and a maximum signal of 53,981 units. The result of quantile normalization is that each sample now the same range of signal values and the samples are maximally comparable.

Two other common methods for sample normalization are linear scaling and locally weighted scatterplot smoothing (LOESS) (Cleveland 1979; Smyth and Speed 2003). When compared with linear scaling methods, quantile normalization does a much better job of normalization over the full range of signal values, while remaining computationally simple. When compared with LOESS methods, quantile normalization does a similar job of normalization, and it is computationally much simpler (Bolstad et al. 2003). It the analysis of genomic data, all things being equal, simpler is better. In my experience, it is generally a mistake to use complicated analysis methods to derive a result, when simpler methods would be adequate.

A third key feature of genomic data analysis is using an analysis of variance framework (ANOVA) for identifying statically significant differences between the values of a measurement for two or more groups of samples (Kerr et al. 2000; Wolfinger et al. 2001; Cui and Churchill 2003). ANOVA can be used to answer the question "Are two sets of measurements different from each other?" Consider the following example (Example #1) of the values of a single gene expression measurement made in two groups of samples:

Example #1:
Group 1: 2,036, 1,559, 2,124, 1,946, and 1,477
Group 2: 4,876, 3,538, 4,381, 4,315, and 3,879

In this example, it looks like the measurements here are pretty different, but how do we know? Analysis of variance allows us to answer this question.

The basic principle behind ANOVA is a signal to noise calculation. For an ANOVA test we calculate a t score which is the difference between the means of two set of measurements divided by the standard error of those measurements, which is a function of the variability occurring within each of the groups. This t score is then used to calculate a *P* value, which is the probability of observing a given t score if there were truly no difference between the two groups.

In Example #1 the mean of the measurements in group 1 is about 1,800 with a standard deviation of about 300 and in the measurements in group 2 average about 4,100 with a standard deviation of about 500. If you plug those values into the ANOVA formula, you get a t score of a little more than 10, which corresponds to a *P* value of 1 in a million. So, the probability of observing this large a difference between the two groups of measurements if there were truly no underlying difference between the two groups is about one in a million. So, for this example, ANOVA tells us that there is a statistically significant difference between these two groups.

A fourth key feature of genomic data analysis is correcting for multiple hypothesis testing. This is an important adjustment that must be made to significance levels in a genomic experiment to account for the large number of statistical tests that are being performed. Consider the following example (Example #2):

Example #2
Group 1: 1,215, 1,476, 1,187, 1,212, 1,368, and 1,092
Group 2: 1,814, 19,964, 1,735, 1,327, 1,840, and 1,537

When one performs the ANOVA calculations on this data it can be determined that the measurements of from these two groups of samples are different from each other at a significance level of .002, or 2 in 1,000. At first glance it seems like the statistics indicate that this is a significant difference. You have to take into account however, that you are doing a large number of statistical tests, so making a 2 in 1,000 observation is going to happen frequently. For example, in a gene expression experiment, you might measure the expression level of 50,000 genes. Even if there is no difference between the two groups of samples, 2 out of 1,000 genes

will have a *P* value of .002 or less at random. Thus, we would expect that 100 genes in the dataset would be false positives at a *P* value of .002.

Benjamini and Hochberg proposed a method to account for this issue and translate *P* values into more meaningful False Discovery Rates (Benjamini and Hochberg 1995). The False Discovery Rate associated with a given *P* value is the overall expected false positive rate at that *P* value across the entire dataset. Thus if 200 genes in this experiment were significant at *P* less than .002, then this would correspond to a false discovery rate of 50%. If 2,000 genes were significant at a *P* value of less than .002, then this would correspond to a false discovery rate of 5%.

The fifth and final key feature of genomic data analysis that I want to highlight is measurement error estimation. This serves to avoid false positives that are caused by spuriously small variability in small numbers of measurements. Consider the following example (Example#3):

Example #3
Group 1: 1,223, 1,212, 1,204, 1,198, 1,206, and 1,205
Group 2: 1,246, 1,250, 1,248, 1,237, 1,242, and 1,240

In this example the difference between the means of the two groups of samples is quite small, but the standard deviation of the measurements within each group even smaller. Since the difference between the means of the two groups is much larger than then the variability within the groups, the t score is quite high and the *P* value is very low. In reality we know that our measurements are not this precise and that if we made these measurements again, we would be extremely unlikely to observe the same results. This is not a gene that we believe to have a meaningful difference between the two groups of samples, but since the ANOVA calculations are based only on this small set of measurements, this gene would be called significant.

The typical error of a gene expression measurement is determined in large part by the overall signal intensity of the measurement. This is evident from a plot of the standard deviation of a gene expression measurement as a function of signal intensity (Figure 4). We see here that a measurement of approximately 1,000 units should have a standard deviation of at least 100 units.

Baldi and Long proposed a method that would incorporate our knowledge about measurement error into ANOVA calculations (Baldi and Long 2001; Long et al. 2001). A

normal ANOVA uses the stand error from the actual measurements as the denominator in the t score calculation. The method of Baldi and Long replaces this value with an estimate of measurement error. The estimate is created by combining information from the actual error of the measurements and a model of the error based on the signal intensity of the measurement. This adjustment in ANOVA calculations results in data behaving like those from Example #3 no longer being called statistically significant.

To do a statistical analysis on gene expression data using the framework that I just described, there are several online tools such as the NIA array analysis tool (http://lgsun.grc.nia.nih.gov/ANOVA/; Sharov et al. 2005), which I use, or cyber-T (http://cybert.ics.uci.edu/). To conduct this analysis with ChIP-Seq data there are no tools available. Consequently, I developed a set of code, written in the python language, to implement these statistical methods for the analysis of ChIP-Seq data.

Since the form of ChIP-Seq data is similar to gene expression data, it is useful to repurpose algorithms from gene expression analysis for ChIP-Seq. In ChIP-Seq data, each of the principles that were outlined above are applicable. 1) ChIP-Seq data takes the form of a matrix. 2) Quantile normalization does a good job at sample normalization for ChIP-Seq data, and its computational simplicity is nice since ChIP-Seq datasets are so much larger than gene expression datasets. 3) An ANOVA framework is a logical choice for asking the question, "What regions show statistically significant differences in ChIP-Seq density between two groups of samples?" 4) A multiple hypothesis testing correction is extremely important for the statistical analysis of ChIP-Seq data, since millions of statistical test are performed. 5) Measurement error estimation is extremely useful in the analysis of ChIP-Seq data because measurement error can be accurately modeled as a function of signal intensity, because current experiments typically contain only a small number of samples, and because analysis of ChIP-Seq data requires making million of statistical tests, so there are many opportunities for spuriously small measurement error.

Figure 4

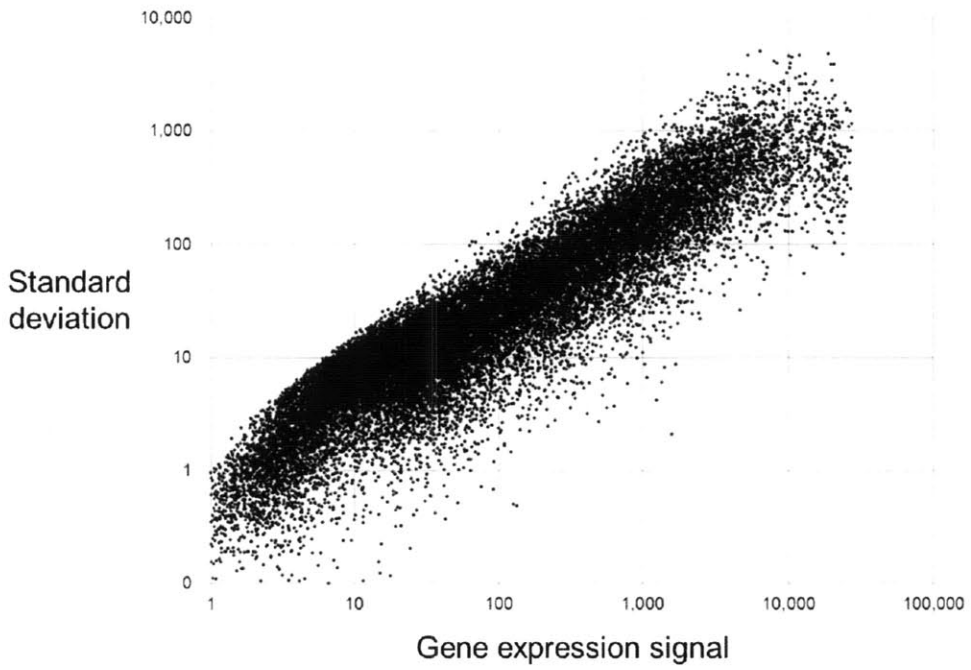


Figure 4. Measurement error in gene expression experiments is a function of signal intensity.

The standard deviation of replicate gene expression measurements was plotted against the average signal of those measurements. Data shown are from experiments performed using Affymetrix GeneChip microarrays.

When our laboratory initially began doing ChIP-Seq experiments there was extremely limited software available for analysis of the resulting data. Consequently, I developed a set of code, written in the Python language (www.python.org), to analyze our ChIP-Seq data. I have improved and expanded the code over time, as new functionalities were required, to best address the analytic needs of the lab's experimental questions. This program now has many functions including the ability to: 1) identify ChIP enriched genomic regions using background correction, 2) map enriched regions to genes, 3) output DNA sequences for motif discovery, 4) analyze evolutionary conservation of enriched regions, 5) make genome browser track files, 6) make heatmap and clustergram outputs, 7) filter input reads, 8) normalize multiple ChIP-Seq datasets and 9) identify statistically significant differences between ChIP-Seq datasets using an ANOVA framework with measurement error estimation and multiple hypothesis testing corrections. So far, the program has been used by more than 40 researchers in 15 laboratories. Although there have been many ChIP-Seq analysis software packages published (Fejes et al. 2008; Ji et al. 2008; Zhang et al. 2008; Rozowsky et al. 2009), the vast majority of researchers at the Whitehead Institute who are doing ChIP-Seq experiments are using the code I developed for analysis.

Research Summary

In graduate school, I have collaborated with wet lab researchers to best utilize the power of genomics to try to answer some of the most important questions that are currently facing biology. My principle research aim has been to develop and employ new algorithms, analytic methods, and presentation techniques to best understand the results of genomic experiments. I have also worked to create tools that will allow biologists to examine their own datasets and to teach them how to use the suite of tools that I employ. My work has focused on understanding the regulatory mechanisms that cells use to control cell type specific gene expression programs.

Chapter 2 describes research investigating whether induced pluripotent stem (iPS) cells are an equivalent cell type to embryonic stem (ES) cells. Recent studies have suggested that ES and iPS cells represent different pluripotent states with substantially different gene expression programs. We compared global chromatin structure and gene expression data for a panel of human ES and iPS cells. We found that while there may be small differences in chromatin state and gene expression between various pluripotent cell lines, there is no consistent signature separating iPS cells from ES cells. I originated this project, developed new methods for normalizing and comparing ChIP-Seq datasets in an ANOVA framework and performed all data analysis. In addition, I wrote the manuscript and was responsible for supplemental materials and figure generation.

Chapter 3 describes research integrating miRNAs into expanded models of the regulatory circuitry of ES cells. We developed a new map of the transcriptional regulatory circuitry of ES cells that incorporates both protein-coding and miRNA genes, and which is based on high-resolution ChIP-seq data, systematic identification of miRNA promoters, and quantitative sequencing of short transcripts in multiple cell types. We found that the key ES cell transcription factors are associated with promoters for most miRNAs that are preferentially expressed in ES cells and with promoters for a set of silent miRNA genes. For this work, I developed new methods and code for ChIP-Seq analysis and performed all data analysis. In addition, I wrote the manuscript and supplemental materials and generated all figures.

Chapter 4 describes a study designed to elucidate the role of the transcription factor Foxp3 in the transcriptional regulatory circuitry of regulatory T (T_{reg}) cells. This sub-population of helper T cells is essential for the prevention of autoimmunity because they can suppress the

proliferation and effector function of neighboring T cells. The forkhead transcription factor Foxp3 is required for T_{reg} development and function, and is sufficient to induce a T_{reg} phenotype in conventional helper T cells. Prior to our study the global set of genes regulated directly by Foxp3 was not known. Consequently, how this transcription factor controls the T_{reg} gene expression program was not understood. We identified the genome-wide targets of Foxp3 and report that many of these are key modulators of T cell activation and function. Remarkably, the predominant, although not exclusive, effect of Foxp3 occupancy is to suppress the activation of target genes on T cell stimulation. My contribution to this project included writing the manuscript and supplemental materials, figure generation, and data analysis.

Chapter 5 describes research into how MLL fusion proteins corrupt the regulatory circuitry of hematopoietic cells to generate leukemia. We show that the MLL-AF4 fusion protein occupies developmental regulatory genes important for hematopoietic stem cell identity and self-renewal in human leukemia cells, and these regions have grossly altered chromatin structure. Our results define the direct targets of the MLL fusion protein, reveal the global role of epigenetic misregulation in leukemia, and identify new targets for therapeutic intervention in cancer. For this project I developed new methods for ChIP-Seq data analysis and performed all data analysis. In addition, I wrote the manuscript and supplemental materials and was responsible for all figure generation.

Chapter 6 describes an examination of the relationship between CpG islands and Polycomb (H3K27me3) and Trithorax (H3K4me3) mediated chromatin modifications. The action of these chromatin modifying complexes is vitally important for the control of cell type during development, but how they are targeted to their sites of action is unknown. We report that in pluripotent cells, but not other cell types, H3K4me3 modifications and CpG islands occur at precisely the same genomic regions. Furthermore, we report that H3K27me3 modifications occur at gene promoters with multiple CpG islands, but not at promoters with zero or one CpG island. Lastly, we report that the occupancy of H3K27me3 modifications is consistent with a model in which PcG proteins are recruited in part by ncRNA species that contain a characteristic RNA stem loop structure. I originated this project, wrote the manuscript, developed new computational methods, performed data analysis, and made figures and supplemental materials.

In addition to the work described in this thesis, I have been an integral part of many other collaborations, which have resulted in several publications (Bilodeau et al. 2009; Bienvenu et al.

2010; Creyghton et al. 2010; Dejosez et al. 2010; Lengner et al. 2010; Novershtern et al. 2011) and manuscripts in review (Whyte et al. 2010). My major contribution to these projects was the development and application of genomic data analysis. However, my role also included writing text, generating figures and tables, preparing supplemental materials, and preparing data for submission to public repositories.

Acknowledgments

I wish to thank members of the Young lab, especially J. Reddy, T.I. Lee, and R.A. Young for helpful comments during the preparation of this chapter. Figure 1 was adapted from J. Zeitlinger and from Jaenisch and Young, 2008. Figures 2 and 3 were obtained from the referenced manuscripts.

References

Agger, K., J. Christensen, P. A. C. Cloos and K. Helin (2008). The emerging functions of histone demethylases. Current Opinion in Genetics & Development *18*, 159-168.

Ahmad, K. and S. Henikoff (2002). The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. Molecular Cell *9*, 1191-1200.

Babu, M. M., N. M. Luscombe, L. Aravind, M. Gerstein and S. A. Teichmann (2004). Structure and evolution of transcriptional regulatory networks. Current Opinion in Structural Biology *14*, 283-291.

Badis, G., M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Y. Chen, et al. (2009). Diversity and Complexity in DNA Recognition by Transcription Factors. Science *324*, 1720-1723.

Baldi, P. and A. D. Long (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics *17*, 509-519.

Barski, A., S. Cuddapah, K. R. Cui, T. Y. Roh, D. E. Schones, Z. B. Wang, G. Wei, I. Chepelev and K. J. Zhao (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate - A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B-Methodological *57*, 289-300.

Bernstein, B. E., C. L. Liu, E. L. Humphrey, E. O. Perlstein and S. L. Schreiber (2004). Global nucleosome occupancy in yeast. Genome Biology *5*.

Bienvenu, F., S. Jirawatnotai, J. E. Elias, C. A. Meyer, K. Mizeracka, A. Marson, G. M. Frampton, M. F. Cole, D. T. Odom, J. Odajima, et al. (2010). Transcriptional role of cyclin D1 in development revealed by a genetic-proteomic screen. Nature *463*, 374-378.

Bilodeau, S., M. H. Kagey, G. M. Frampton, P. B. Rahl and R. A. Young (2009). SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. Genes & Development *23*, 2484-2489.

Bolstad, B. M., R. A. Irizarry, M. Astrand and T. P. Speed (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185-193.

Brownell, J. E., J. X. Zhou, T. Ranalli, R. Kobayashi, D. G. Edmondson, S. Y. Roth and C. D. Allis (1996). Tetrahymena histone acetyltransferase A: A homolog to yeast Gcn5p linking histone acetylation to gene activation. Cell *84*, 843-851.

Buckingham, M., L. Bajard, T. Chang, P. Daubas, J. Hadchouel, S. Meilhac, D. Montarras, D. Rocancourt and F. Relaix (2003). The formation of skeletal muscle: from somite to limb. Journal of Anatomy *202*, 59-68.

Cao, Y., Z. Z. Yao, D. Sarkar, M. Lawrence, G. J. Sanchez, M. H. Parker, K. L. MacQuarrie, J. Davison, M. T. Morgan, W. L. Ruzzo, et al. (2010). Genome-wide MyoD Binding in Skeletal Muscle Cells: A Potential for Broad Cellular Reprogramming. Developmental Cell *18*, 662-674.

Causton, H. C., B. Ren, S. S. Koh, C. T. Harbison, E. Kanin, E. G. Jennings, T. I. Lee, H. L. True, E. S. Lander and R. A. Young (2001). Remodeling of yeast genome expression in response to environmental changes. Molecular Biology of the Cell *12*, 323-337.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association *74*, 829-836.

Creyghton, M. P., A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, et al. (2010). Histone H3K27ac Discriminates Between Active Enhancers and Those Poised for Future Developmental States. Proceedings of the National Academy of Sciences of the United States of America (in publication).

Cui, X. Q. and G. A. Churchill (2003). Statistical tests for differential expression in cDNA microarray experiments. Genome Biology *4*, 10.

Davey, C. A., D. F. Sargent, K. Luger, A. W. Maeder and T. J. Richmond (2002). Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 angstrom resolution. Journal of Molecular Biology *319*, 1097-1113.

Davis, R. L., H. Weintraub and A. B. Lassar (1987). Expression of a single transfected cDNA converts fibroblasts to myoblasts. Cell *51*, 987-1000.

de la Serna, I. L., Y. Ohkawa and A. N. Imbalzano (2006). Chromatin remodelling in mammalian differentiation: lessons from ATP-dependent remodellers. Nature Reviews Genetics *7*, 461-473.

Dejosez, M., S. S. Levine, G. M. Frampton, W. A. Whyte, S. A. Stratton, M. C. Barton, P. H. Gunaratne, R. A. Young and T. P. Zwaka (2010). Ronin/Hcf-1 binds to a hyperconserved enhancer element and regulates genes involved in the growth of embryonic stem cells. Genes & Development *24*, 1479-1484.

Dhalluin, C., J. E. Carlson, L. Zeng, C. He, A. K. Aggarwal and M. M. Zhou (1999). Structure and ligand of a histone acetyltransferase bromodomain. Nature *399*, 491-496.

Ekwall, K., J. P. Javerzat, A. Lorentz, H. Schmidt, G. Cranston and R. Allshire (1995). The chromodomain protein SWI6 - A key component at fission yeast centromeres. Science *269*, 1429-1431.

Farnham, P. J. (2009). Insights from genomic profiling of transcription factors. Nature Reviews Genetics *10*, 605-616.

Fejes, A. P., G. Robertson, M. Bilenky, R. Varhol, M. Bainbridge and S. J. M. Jones (2008). FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics *24*, 1729-1730.

Feng, R., S. C. Desbordes, H. F. Xie, E. S. Tillo, F. Pixley, E. R. Stanley and T. Graf (2008). PUA and C/EBP alpha/beta convert fibroblasts into macrophage-like cells. Proceedings of the National Academy of Sciences of the United States of America *105*, 6057-6062.

Finch, J. T. and A. Klug (1976). Solenoidal model for superstructure in chromatin. Proceedings of the National Academy of Sciences of the United States of America *73*, 1897-1901.

Fischle, W., Y. M. Wang, S. A. Jacobs, Y. C. Kim, C. D. Allis and S. Khorasanizadeh (2003). Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 bv Polvcomb and HP1 chromodomains. Genes & Development *17*, 1870-1881.

Gilbert, N., S. Boyle, H. Fiegler, K. Woodfine, N. P. Carter and W. A. Bickmore (2004). Chromatin architecture of the human genome: Gene-rich domains are enriched in open chromatin fibers. Cell *118*, 555-566.

Giniger, E., S. M. Varnum and M. Ptashne (1985). Specific DNA-binding of GAL4, a positive regulatory protein of yeast. Cell *40*, 767-774.

Guenther, M. G., S. S. Levine, L. A. Boyer, R. Jaenisch and R. A. Young (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77-88.

Harbison, C. T., D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J. B. Tagne, D. B. Reynolds, J. Yoo, et al. (2004). Transcriptional regulatory code of a eukaryotic genome. Nature *431*, 99-104.

Hawkins, R. D., G. C. Hon and B. Ren (2010). Next-generation genomics: an integrative approach. Nature Reviews Genetics *11*, 476-486.

Haynes, S. R., C. Dollard, F. Winston, S. Beck, J. Trowsdale and I. B. Dawid (1992). The bromodomain - A conserved sequence found in human, Drosophila and yeast proteins. Nucleic Acids Research *20*, 2603-2603.

Ho, L. and G. R. Crabtree (2010). Chromatin remodelling during development. Nature *463*, 474-484.

Hughes, C. L. and T. C. Kaufman (2002). Hox genes and the evolution of the arthropod body plan. Evolution & Development *4*, 459-499.

Jacob, F. and J. Monod (1961). Genetic regulatory mechanisms in synthesis of proteins. Journal of Molecular Biology *3*, 318-&.

Jaenisch, R. and R. Young (2008). Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell *132*, 567-582.

Jenuwein, T. and C. D. Allis (2001). Translating the histone code. Science *293*, 1074-1080.

Ji, H. K., H. Jiang, W. X. Ma, D. S. Johnson, R. M. Myers and W. H. Wong (2008). An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nature Biotechnology *26*, 1293-1300.

Jiang, C. Z. and B. F. Pugh (2009). Nucleosome positioning and gene regulation: advances through genomics. Nature Reviews Genetics *10*, 161-172.

Jones, P. L., G. J. C. Veenstra, P. A. Wade, D. Vermaak, S. U. Kass, N. Landsberger, J. Strouboulis and A. P. Wolffe (1998). Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nature Genetics *19*, 187-191.

Kerr, M. K., M. Martin and G. A. Churchill (2000). Analysis of variance for gene expression microarray data. Journal of Computational Biology *7*, 819-837.

Knezetic, J. A. and D. S. Luse (1986). The presence of nucleosomes on a DNA-template prevents initiation by RNA polymerase II in vitro. Cell *45*, 95-104.

Korolev, N., A. P. Lyubartsev and L. Nordenskiold (2006). Computer modeling demonstrates that electrostatic attraction of nucleosomal DNA is mediated by histone tails. Biophysical Journal *90*, 4305-4316.

Ku, M., R. P. Koche, E. Rheinbay, E. M. Mendenhall, M. Endoh, T. S. Mikkelsen, A. Presser, C. Nusbaum, X. H. Xie, A. S. Chi, et al. (2008). Genomewide Analysis of PRC1 and PRC2 Occupancy Identifies Two Classes of Bivalent Domains. Plos Genetics *4*, 14.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Lee, T. I., S. E. Johnstone and R. A. Young (2006). Chromatin immunoprecipitation and microarray-based analysis of protein location. Nature Protocols *1*, 729-748.

Lengner, C. J., A. A. Gimelbrant, J. A. Erwin, A. W. Cheng, M. G. Guenther, G. G. Welstead, R. Alagappan, G. M. Frampton, P. Xu, J. Muffat, et al. (2010). Derivation of Pre-X Inactivation Human Embryonic Stem Cells under Physiological Oxygen Concentrations. Cell *141*, 872-883.

Levine, M. and R. Tjian (2003). Transcription regulation and animal diversity. Nature *424*, 147-151.

Lewis, E. B. (1963). Genes and developmental pathways. American Zoologist *3*, 33-56.

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

Li, B., M. Carey and J. L. Workman (2007). The role of chromatin during transcription. Cell *128*, 707-719.

Long, A. D., H. J. Mangalam, B. Y. P. Chan, L. Tolleri, G. W. Hatfield and P. Baldi (2001). Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework - Analysis of global gene expression in Escherichia coli K12. Journal of Biological Chemistry *276*, 19937-19944.

Luger, K., A. W. Mader, R. K. Richmond, D. F. Sargent and T. J. Richmond (1997). Crystal structure of the nucleosome core particle at 2.8 angstrom resolution. Nature *389*, 251-260.

Mardis, E. R. (2007). ChIP-seq: welcome to the new frontier. Nature Methods *4*, 613-614.

Marson, A., S. S. Levine, M. F. Cole, G. M. Frampton, T. Brambrink, S. Johnstone, M. G. Guenther, W. K. Johnston, M. Wernig, J. Newman, et al. (2008). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell *134*, 521-533.

Mavrich, T. N., C. Z. Jiang, I. P. Ioshikhes, X. Y. Li, B. J. Venters, S. J. Zanton, L. P. Tomsho, J. Qi, R. L. Glaser, S. C. Schuster, et al. (2008). Nucleosome organization in the Drosophila genome. Nature *453*, 358-U27.

McGhee, J. D. and G. Felsenfeld (1980). Nucleosome structure. Annual Review of Biochemistry *49*, 1115-1156.

McGinnis, W. and R. Krumlauf (1992). Homoebox genes and axial patterning. Cell *68*, 283-302.

McGinnis, W., M. S. Levine, E. Hafen, A. Kuroiwa and W. J. Gehring (1984). A conserved DNA sequence in homoeotic gene of the Drosophila Antennapedia and bithorax complexes. Nature *308*, 428-433.

Mikkelsen, T. S., M. C. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T. K. Kim, R. P. Koche, et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-U2.

Mizuguchi, G., X. T. Shen, J. Landry, W. H. Wu, S. Sen and C. Wu (2004). ATP-Driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. Science *303*, 343-348.

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

Mujtaba, S., L. Zeng and M. M. Zhou (2007). Structure and acetyl-lysine recognition of the bromodomain. Oncogene *26*, 5521-5527.

Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein and M. Snyder (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320*, 1344-1349.

Nagy, L., H. Y. Kao, D. Chakravarti, R. J. Lin, C. A. Hassig, D. E. Ayer, S. L. Schreiber and R. M. Evans (1997). Nuclear receptor repression mediated by a complex containing SMRT, mSin3A, and histone deacetylase. Cell *89*, 373-380.

Nan, X. S., H. H. Ng, C. A. Johnson, C. D. Laherty, B. M. Turner, R. N. Eisenman and A. Bird (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature *393*, 386-389.

Narlikar, G. J., H. Y. Fan and R. E. Kingston (2002). Cooperation between complexes that regulate chromatin structure and transcription. Cell *108*, 475-487.

Nishiyama, A., L. Xin, A. A. Sharov, M. Thomas, G. Mowrer, E. Meyers, Y. L. Piao, S. Mehta, S. Yee, Y. Nakatake, et al. (2009). Uncovering Early Response of Gene Regulatory Networks in ESCs by Systematic Induction of Transcription Factors. Cell Stem Cell *5*, 420-433.

Novershtern, N., A. Subramanian, L. N. Lawton, R. H. Mak, W. N. Haining, M. E. McConkey, N. Habib, N. Yosef, C. Y. Chang, T. Shay, et al. (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell *144*, 296-309.

Ogryzko, V. V., R. L. Schiltz, V. Russanova, B. H. Howard and Y. Nakatani (1996). The transcriptional coactivators p300 and CBP are histone acetyltransferases. Cell *87*, 953-959.

Olins, A. L. and D. E. Olins (1974). Spheroid chromatin units (RU bodies). Science *183*, 330-332.

Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nature Reviews Genetics *10*, 669-680.

Pearce, J. J. H., P. B. Singh and S. J. Gaunt (1992). The mouse has a Polycomb-like chromobox gene. Development *114*, 921-929.

Phillips, D. M. (1963). Presence of acetyl groups in histones. Biochemical Journal *87*, 258-&.

Pokholok, D. K., C. T. Harbison, S. Levine, M. Cole, N. M. Hannett, T. I. Lee, G. W. Bell, K. Walker, P. A. Rolfe, E. Herbolsheimer, et al. (2005). Genome-wide map of nucleosorne acetylation and methylation in yeast. Cell *122*, 517-527.

Ren, B., F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, et al. (2000). Genome-wide location and function of DNA binding proteins. Science *290*, 2306-+.

Robertson, G., M. Hirst, M. Bainbridge, M. Bilenky, Y. J. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature Methods *4*, 651-657.

Rozowsky, J., G. Euskirchen, R. K. Auerbach, Z. D. D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder and M. B. Gerstein (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nature Biotechnology *27*, 66-75.

Rudnicki, M. A., P. N. J. Schnegelsberg, R. H. Stead, T. Braun, H. H. Arnold and R. Jaenisch (1993). MYOD or MYF-5 is required for the formation of skeletal-muscle. Cell *75*, 1351-1359.

Schiessel, H. (2003). The physics of chromatin. Journal of Physics-Condensed Matter *15*, R699-R774.

Schones, D. E., K. R. Cui, S. Cuddapah, T. Y. Roh, A. Barski, Z. B. Wang, G. Wei and K. J. Zhao (2008). Dynamic regulation of nucleosome positioning in the human genome. Cell *132*, 887-898.

Scott, M. P. and A. J. Weiner (1984). Structural relationships among genes that control development: sequence homology between the Antennapedia, Ultrabithorax, and fushi tarazu loci of Drosophila. Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences *81*, 4115-4119.

Segal, E., Y. Fondufe-Mittendorf, L. Y. Chen, A. Thastrom, Y. Field, I. K. Moore, J. P. Z. Wang and J. Widom (2006). A genomic code for nucleosome positioning. Nature *442*, 772-778.

Sharov, A. A., D. B. Dudekula and M. S. H. Ko (2005). A web-based tool for principal component and significance analysis of microarray data. Bioinformatics *21*, 2548-2549.

Smyth, G. K. and T. Speed (2003). Normalization of cDNA microarray data. Methods *31*, 265-273.

Strahl, B. D. and C. D. Allis (2000). The language of covalent histone modifications. Nature *403*, 41-45.

Swalla, B. J. (2006). Building divergent body plans with similar genetic pathways. Heredity *97*, 235-243.

Takahashi, K., K. Tanabe, M. Ohnuki, M. Narita, T. Ichisaka, K. Tomoda and S. Yamanaka (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell *131*, 861-872.

Takahashi, K. and S. Yamanaka (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663-676.

Tirosh, I. and N. Barkai (2008). Two strategies for gene regulation by promoter nucleosomes. Genome Research *18*, 1084-1091.

Tontonoz, P., E. D. Hu and B. M. Spiegelman (1994). Stimluation of adipogenesis in fibroblasts by PPAR-gamma-2, a lipid-activated transcription factor. Cell *79*, 1147-1156.

Tsukiyama, T., P. B. Becker and C. Wu (1994). ATP-dependent nucleosome disruption at a heat-shock promoter mediated by binding of GAGA transcription factor. Nature *367*, 525-532.

Turner, F. R. and A. P. Mahowald (1979). Scanning electron microscopy of Drosophila melanogaster embryogenesis:: 3. Formation of the head and caudal segments. Developmental Biology *68*, 96-109.

Weintraub, H. (1993). The MyoD family and myogenesis: redundancy, networks, and thresholds. Cell *75*, 1241-1244.

Weintraub, H., R. Davis, S. Tapscott, M. Thayer, M. Krause, R. Benezra, T. K. Blackwell, D. Turner, R. Rupp, S. Hollenberg, et al. (1991). The MYOD gene family - Nodal point during specification of the muscle-cell lineage. Science *251*, 761-766.

Wernig, M., A. Meissner, R. Foreman, T. Brambrink, M. C. Ku, K. Hochedlinger, B. E. Bernstein and R. Jaenisch (2007). In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. Nature *448*, 318-U2.

Whyte, W. A., S. Bilodeau, G. M. Frampton, D. A. Orlando and R. A. Young (2010). Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. Nature (in review).

Wolfinger, R. D., G. Gibson, E. D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari and R. S. Paules (2001). Assessing gene significance from cDNA microarray expression data via mixed models. Journal of Computational Biology *8*, 625-637.

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

Yuan, G. C., Y. J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler and O. J. Rando (2005). Genome-scale identification of nucleosome positions in S-cerevisiae. Science *309*, 626-630.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nussbaum, R. M. Myers, M. Brown, W. Li, et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biology *9*, 9.

Chapter 2

Chromatin Structure and Gene Expression Programs of Human Embryonic and Induced Pluripotent Stem Cells

Published as: Matthew G. Guenther*, Garrett M. Frampton*, Frank Soldner, Dirk Hockemeyer, Maya Mitalipova, Rudolf Jaenisch and Richard A. Young (2010). Chromatin Structure and Gene Expression Programs of Human Embryonic and Induced Pluripotent Stem Cells. Cell Stem Cell *7*, 249-57. * equal contribution

Abstract

Knowledge of both the global chromatin structure and the gene expression programs of human embryonic stem (ES) cells and induced pluripotent stem (iPS) cells should provide a robust means to assess whether the genomes of these cells have similar pluripotent states. Recent studies have suggested that ES and iPS cells represent different pluripotent states with substantially different gene expression profiles. We describe here a comparison of global chromatin structure and gene expression data for a panel of human ES and iPS cells. Genome-wide maps of nucleosomes with histone H3K4me3 and H3K27me3 modifications indicate that there is little difference between ES and iPS cells with respect to these marks. Gene expression profiles confirm that the transcriptional programs of ES and iPS cells show very few consistent differences. Although some variation in chromatin structure and gene expression was observed in these cell lines, these variations did not serve to distinguish ES from iPS cells.

Introduction

Mammalian cells can be directly reprogrammed into induced pluripotent stem (iPS) cells by introduction of defined sets of transcriptional regulators (Takahashi and Yamanaka, 2006; Maherali et al., 2007; Nakagawa et al., 2007; Okita et al., 2007; Takahashi et al., 2007; Wernig et al., 2007; Yu et al., 2007; Aoi et al., 2008). These iPS cells hold great potential for regenerative medicine because they are similar to pluripotent embryonic stem (ES) cells and can be derived in a patient-specific manner from adult somatic cells (Yamanaka, 2007; Saha and Jaenisch, 2009). ES and iPS cells are highly similar in a broad range of phenotypic behaviors, including cell morphology, expression of pluripotency markers, teratoma formation, ability to differentiate into germ layers, and tetraploid complementation (Okita et al., 2007; Wernig et al., 2007; Boland et al., 2009; Kang et al., 2009; Smith et al., 2009; Zhao et al., 2009). However, recent studies comparing the gene expression profiles of ES cells and iPS cells have suggested that iPS cells are a unique cellular subtype distinct from ES cells (Chin et al., 2009; Marchetto et al., 2009) and that iPS cells retain some of the expression program of their cell of origin (Ghosh et al., 2010). It is important to understand if there are genuine differences in the global chromatin structure and the gene expression programs of human ES and iPS cells, as such differences may impact the potential therapeutic use of iPS cells.

Trithorax group (TrxG) and Polycomb group (PcG) protein complexes are key regulators of chromatin structure that are required for segmental identity in the developing embryo and contribute to maintenance of the pluripotent ES cell state (Ringrose and Paro, 2004; Schuettengruber et al., 2007; Pietersen and van Lohuizen, 2008). TrxG complexes catalyze histone H3 lysine 4 trimethylation (H3K4me3) at promoters of protein-coding genes (Bernstein et al., 2002; Santos-Rosa et al., 2002; Ng et al., 2003; Schneider et al., 2004; Guenther et al., 2007; Mikkelsen et al., 2007), miRNA loci (Marson et al., 2008; Ozsolak et al., 2008), and non-coding lincRNA loci (Guttman et al., 2009). The PcG protein complex PRC2 catalyzes histone H3 lysine 27 trimethylation (H3K27me3), which contributes to repression of developmental genes (Schuettengruber et al., 2007; Schwartz and Pirrotta, 2008; Simon and Kingston, 2009). Histone H3K4me3 and H3K27me3 histone modifications are generally associated with transcriptionally active and repressed domains of the genome, respectively, although both modifications can occur at silent genes encoding developmental regulators that are poised for

future activation (Bernstein et al., 2006; Lee et al., 2006; Guenther et al., 2007). Genome-wide maps of these histone modifications, which produce robust signals in ChIP-Seq experiments, can be especially useful for comparing transcriptional and developmental states of cells, particularly when coupled with gene expression profiling.

Microarray based gene expression profiling has proven to be a powerful approach to characterize the transcriptional state of cells and to identify differences between cells of different types or states (Ebert and Golub, 2004; Ivanova et al., 2006). Comparing the gene expression profiles of ES and iPS cells could permit identification of any unique and consistent differences between these two cell types. However, comparative analysis of expression data can be challenging due to differences in the homogeneity of cell populations, cell handling, reagents, and analytical techniques. In comparing the expression profiles of ES and iPS cells, it is therefore important to use analytical methods that account for the noise in the data and require reproducible results across multiple experiments (Bammler et al., 2005).

We have investigated whether a panel of human ES cells differ consistently from a panel of human iPS cells using both genome-wide maps of histone H3K4me3 and H3K27me3 modifications and gene expression analysis. We have also re-analyzed a large collection of previously published gene expression data using different analysis methods. Our results reveal that small variations in the chromatin structure or gene expression occur among different ES and iPS cell lines, but we do not observe a consistent signature that distinguishes iPS cell lines from ES cell lines when examined after extended culture.

Results and Discussion

Genome-wide maps of chromatin modifications show human iPS cells share key features with ES cells

We used ChIP-Seq to map H3K4me3 and H3K27me3 occupancy genome-wide in six independent ES cell lines and six independent iPS cell lines grown under identical conditions (Figure 1). The ES cells included two male lines (BG01 and WIBR1) and four female lines (BG03, WIBR2, WIBR3 and WIBR7), each derived from a different donor (Lengner et al., 2010). The iPS lines were generated from human fibroblasts using a doxycycline-inducible

reprogramming system with OCT4, SOX2 and KLF4 genes (Hockemeyer et al., 2008; Soldner et al., 2009). Four of the iPS cells were derived from a female donor (iPS A1, iPS C1, iPS4, and iPS A6; described and characterized in Hockemeyer et al., 2008) and two from a male donor (iPS PDB^{2lox}-17 and iPS PDB^{2lox}-21; described and characterized in Soldner et al., 2009). All iPS cell lines contained integrated transgenes, but were doxycycline-independent for growth, indicating that transgene expression was not required for propagation of these pluripotent cells.

The maps of histone H3K4me3 and H3K27me3 were highly similar at protein coding and non-coding genes in all 12 ES and iPS cell lines when examined by enrichment profiles (Figure 1A, 1C), heat maps (Figure 1B, 1D) or inspection of gene tracks (Figure 1E, 1F). Nucleosomes with H3K4me3 occurred at the vast majority of protein-coding genes in both ES and iPS cells (~85%), with maximal enrichment occurring approximately 200 bp downstream of transcriptional start sites (Figure 1A, B, E; Table S1). H3K4me3 modified nucleosomes also occupied the start sites of known and predicted noncoding RNAs in both cell types, which include ES cell specific RNAs associated with pluripotency (Marson et al., 2008) (Table S1). H3K27me3 modified nucleosomes occurred primarily in the promoters of ~2,000 repressed genes, many of which encode key regulators of development (Figure 1C, D, F; Table S1). H3K27me3 marked small domains (1-5kb) within certain gene promoter regions, and large domains extending across >100kb of the HOX gene clusters. These results indicate that the genomes of human iPS cells possess the general features of Trx and PcG -mediated histone modifications previously described in ES cells (Bernstein et al., 2006; Boyer et al., 2006; Bracken et al., 2006; Lee et al., 2006; Guenther et al., 2007; Pan et al., 2007; Zhao et al., 2007; Mikkelsen et al., 2008).

Similarity in genes and regions occupied by modified histones in ES and iPS cells

Inspection of gene tracks revealed some variation in H3K4me3 and H3K27me3 nucleosome occupancy among these cells lines, prompting us to systematically compare the sets of genes occupied by these histone modifications in each cell line with the set occupied in all other lines (Figure 2; Supplemental Experimental Procedures). We first performed all pair-wise comparisons among the ES cell lines and found that 1.4±0.8% of genes had different H3K4me3 occupancy and 5.5±2.0% of genes had different H3K27me3 occupancy. Similarly, 0.7±0.3% genes varied for H3K4me3 and 6.0±2.6% varied for H3K27me3 among the iPS cell lines.

Pairwise comparison of ES and iPS cell lines revealed that the variation for H3K4me3 (1.3±0.7%) and H3K27me3 (6.0±2.4%) occupied genes was not significantly different than the variation observed within ES cell lines or within iPS cell lines. In contrast, comparisons of the genes enriched for H3K4me3 or H3K27me3 between pluripotent cells (ES and iPS cells) and adult fibroblast cells revealed significantly larger differences (12.3±0.4% for H3K4me3; 67.8±2.8% for H3K27me3) (Figure 2; Table S2). Thus, we observed no more variation between ES and iPS cells than was evident within the ES cell lines or within the iPS cell lines. We also examined the magnitude of the ChIP-Seq peaks associated with each gene and again found that differences between ES and iPS cells were no greater than the differences observed within ES cell lines and within iPS cell lines (Table S2). These results suggest that there were few, if any, consistent differences in the set of genes occupied by H3K4me3 or H3K27me3 between these human ES and iPS cells.

We developed a statistical method (Supplemental Experimental Procedures) to scan the entire genome and identify regions with significant differential H3K4me3 and H3K27me3 occupancy between ES and iPS cells. To confirm the sensitivity and specificity of this method, we compared all male ES and iPS lines to all female ES and iPS lines, and found a number of regions with significant differences in histone modifications, which were located almost exclusively on the sex chromosomes, as expected (Supplemental Table S3). We then compared female ES cells to female iPS cells and found that iPS cells showed increased H3K27me3 occupancy relative to ES cells at several X-linked genes including Xist. We believe these differences in occupancy are likely a consequence of exposing female iPS cells to high oxygen conditions during derivation, since high oxygen growth conditions induce X-inactivation in human pluripotent cells (Lengner et al., 2010). To further validate our ability to detect chromatin differences, we compared all pluripotent cells (ES and iPS cells) to donor fibroblasts, and observed a large number of regions with differences in histone modifications (Table S3) that were strongly associated with differences in gene expression (Figure 2B,C).

We then applied this method to identify statistically significant differences in chromatin structure between ES and iPS cells and found 50 genomic regions (29 genes) with differential H3K4me3 occupancy and 4 regions (2 genes) with differential H3K27me3 occupancy (Supplemental Table S3). These regions of differential occupancy represent a tiny fraction of the genome (0.003%), and although there was no obvious theme associated with them, we

considered several possible causes for the differential modification. First, we investigated if these differences were due to the presence of exogenous reprogramming factors in iPS cells, but there were no significant differences in these chromatin modifications between trans-gene containing and trans-gene excised iPS cells (Table S2; Soldner et al., 2009). Second, we investigated if the chromatin differences between ES and iPS cells were due to residual epigenetic signatures left from the parental fibroblast cell line, but found no evidence that iPS cells contain H3K4me3 or H3K27me3 signatures that reflect their cell of origin (Table S4). Lastly, we asked if any gene expression changes were associated with differences in histone modification between ES and iPS cells, but found that this was not the case (Figure S1). We conclude that there are a small number of regions in these human ES and iPS cells that show differences in H3K4me3 and H3K27me3 modified nucleosomes, which involves a small fraction of the genome and has little or no influence on gene expression. However, we cannot exclude the possibility that these small chromatin differences observed in undifferentiated cells may exert subtle effects on cells upon differentiation.

Limited variation in gene expression between human ES and iPS cells

Although the H3K4me3 and H3K27me3 profiles of the human ES and iPS cells were nearly identical, we investigated the possibility that there were differences in the gene expression profiles between these ES and iPS cells. All 12 ES and iPS cell lines, in addition to donor fibroblast cells, were subjected to expression profiling and the data were analyzed using a single-factor analysis of variance for testing statistical significance with a Bayesian model of measurement error and a false discovery rate correction for multiple hypothesis testing (Sharov et al., 2005; Supplemental Experimental Procedures). We found zero genes with statistically significant differential expression between ES and iPS cells using this method. To gain greater statistical power to identify small differences in transcript levels between ES and iPS cells, we included expression datasets from several additional ES and iPS cells lines that were not used in the chromatin studies and repeated the analysis. In this panel of 39 samples, we found only 4 genes with statistically significant differential expression between ES and iPS cells (Figure 3A, B; Table S5). These results are consistent with a study that shows the overall mRNA and microRNA expression patterns of isogenic mouse ES and iPS cells are nearly indistinguishable within the exception of a few transcripts on chromosome 12qF1 (Stadtfeld et al., 2010).

Previous reports have observed that ES and iPS cells exhibit considerable differences in gene expression (Chin et al., 2009; Marchetto et al., 2009). To determine if these gene expression differences were consistently observed in multiple laboratories, we re-examined a large collection of previously published expression data comparing ES and iPS cells (Table S5) (Maherali et al., 2008; Chin et al., 2009; Yu et al., 2009). If there were truly consistent gene expression differences between ES and iPS cells that were not a product of laboratory specific biases in cell culture conditions, passage number, RNA preparation methods, or data processing methods, it would be expected that similar sets of genes would be identified as differentially expressed in more than one of these studies. However, we found that overlap between the genes identified in each of these studies was extremely low (Figure 3B, D) and conclude that there are very few, if any, consistent differences in the gene expression programs of ES and iPS cell lines. In contrast, the differential expression observed between pluripotent (ES and iPS) cell lines and fibroblast lines was highly reproducible across laboratories (Figure 3C,E; Table S5).

Several studies have described a few hundred to several thousand genes that show statistically significant differential expression between ES and iPS cells (Chin et al., 2009; Marchetto et al., 2009; Ghosh et al., 2010). There are several possible reasons why this result may have been obtained even if there is not, in reality, a unique and consistent expression signature that distinguishes iPS from ES cells. Cell culture conditions, derivation method, passaging technique, reagents, amount of time in culture, microarray methods, and operator-specific microarray assay biases can affect gene expression profiles. It is likely that uncontrolled variables such as these contributed to the observation of differential gene expression between the ES and iPS samples, because we observe that ES and iPS cell expression data cluster by laboratory and not by ES/iPS identity (Figure 4). Our own data indicates that differences in gene expression do exist between various ES and iPS cell lines, but these differences do not consistently distinguish iPS cells from ES cells.

Discussion

ES and iPS cells have been shown to share key features of pluripotency, including expression of pluripotency markers, teratoma formation, cell morphology, ability to differentiate into germ

layers, and tetraploid complementation (Okita et al., 2007; Wernig et al., 2007; Boland et al., 2009; Kang et al., 2009; Smith et al., 2009; Zhao et al., 2009; Stadtfeld et al., 2010). Human iPS cells offer great promise for regenerative medicine and the establishment of patient or subgroup-specific disease models, but some reports suggest that ES and iPS cells may not be equivalent (Chin et al., 2009; Marchetto et al., 2009; Ghosh et al., 2010). We have mapped two histone modifications that are critical for cell state and development in human ES and iPS cells and find a very small number of consistent differences between ES and iPS cells for these marks. These differences in chromatin structure are not associated with differential gene expression. Furthermore, the consistent differences between ES and iPS cells are considerably smaller than the overall variability among these cell lines. In our analysis of gene expression data across several studies we find that variations in gene expression occur among different ES and iPS cell lines, but we do not observe a consistent signature that distinguishes iPS cell lines from ES cell lines.

Previous studies comparing the gene expression profiles of human ES and iPS cells argued that a recurrent gene expression signature appears in iPS cells regardless of their origin or the method by which they were generated, and suggested that iPS cells should be considered a unique subtype of pluripotent cell (Chin et al., 2009). This interpretation is inconsistent with our own. This discrepancy is likely due to four features of the analytic methods used by Chin et al. (2009) that, in our view, are key to accurate data interpretation. First, a correction for multiple hypothesis testing was not used, so the number of statistically significant differentially expressed genes was greatly over-estimated. Second, there was not a requirement that gene expression change in the same direction. Third, the same ES cell expression data was compared to both "early" and "late" passage iPS cells, negating the assumption that these sets of differentially expressed genes would be independent. Fourth, biases were introduced into the identification of differentially expressed genes that violate the assumption of independent assortment, which is relied upon to calculate the statistical significance of gene list overlaps. The use of a fold change threshold creates a bias towards identifying genes with larger measurement error, and collapsing measurements for several probesets into a single measurement for each gene could cause some genes to be more likely identified as differentially expressed across several datasets. In studies that came to the conclusion that ES and iPS cells have different gene expression signatures (Chin et al., 2009; Marchetto et al., 2009; Ghosh et al., 2010), we believe uncontrolled laboratory-

specific variables likely contributed to the observation of differential gene expression. When we repeat the analysis of published data using standard methods, we do not observe significant overlap between the genes that are differentially expressed between ES and iPS cells across several laboratories (Figure 3B,C). Instead, we find that ES and iPS cell expression data clusters more by laboratory than by ES/iPS identity (Figure 4). Similarly, in mouse, most expression differences between ES and iPS cells are not consistently observed across laboratories and are likely caused by variations in genetic background or method of iPS cell production (Chin et al., 2009; Stadtfeld et al., 2010).

In summary, our experiments and analysis do not demonstrate a significant difference between the H3K4me3 or H3K27me3 modifications or a consistent difference in the gene expression programs of ES and iPS cells. It is possible that there are small differences between ES and iPS cells that we lacked the statistical power to observe, or that differences may occur in non-coding or regulatory RNAs. Additionally, it is possible that there exist important epigenetic differences between ES and iPS cells that are not reflected in the chromatin marks that we examined, such as DNA methylation events (Meissner et al., 2008; Doi et al., 2009). These possible changes in histone modification, DNA methylation or other epigenetic marks may result in subtle functional differences that could affect differentiation or other cell processes (Feng et al.; Hu et al., 2010). Nevertheless, our results and the phenotypic similarities shared by ES and iPS cells (Smith et al., 2009; Zhao et al., 2009), support the view that ES and iPS cells are nearly identical cell types.

Experimental Procedures

Human ES and iPS cell culture

All primary fibroblast cell lines described in this paper were purchased from the Coriell Cell Repository (Camden, NJ). Fibroblasts were cultured in fibroblast medium (Dulbecco's modified Eagle's medium [DMEM] supplemented with 15% fetal bovine serum [FBS; Hyclone], 1 mM glutamine [Invitrogen], 1% nonessential amino acids [Invitrogen], and penicillin/streptomycin [Invitrogen]).

hiPS cells iPS A1, iPS C1, iPS4, iPS A6 (Hockemeyer et al. 2009); hiPS cells iPS PDB^{2lox}-17 and iPS PDB^{2lox}-21 (Soldner et al. 2009); hESC lines BG01 and BG03 (National Institutes of Health code: BG01 and BG03; BresaGen, Inc., Athens, GA); and hESC cell lines WIBR1, WIBR2, WIBR3, and WIBR7 (Whitehead Institute Center for Human Stem Cell Research) (Lengner et al., 2010) were maintained on mitomycin C (MMC)-inactivated mouse embryonic fibroblast feeder layers in hESC medium (DMEM/F12 [Invitrogen] supplemented with 15% FBS [Hyclone], 5% KnockOut Serum Replacement [Invitrogen], 1 mM glutamine [Invitrogen], 1% nonessential amino acids [Invitrogen], 0.1 mM β-mercaptoethanol [Sigma], and 4 ng/ml FGF2 [R&D Systems]). Cultures were passaged every 5 to 7 days either manually or enzymatically with collagenase type IV (Invitrogen; 1.5 mg/ml).

ChIP-Seq

Detailed descriptions of antibodies, antibody specificity, ChIP, and ChIP-Seq analysis methods used in this study have been published previously and are provided in the Supplemental Experimental Procedures. The antibodies for ChIP were specific for H3K4me3 (ab8580; Abcam) and H3K27me3 (ab6002; Abcam). Purified immunoprecipitated DNA was prepared for sequencing according to a modified version of the Solexa Genomic DNA protocol, applied to a flow-cell using the Solexa Cluster Station fluidics device, and sequenced according to Illumina's standard protocols. Images acquired from the Solexa sequencer were processed through the bundled Solexa image extraction pipeline and aligned to the March 2006 build (NCBI36.1/hg18) of the human genome using Bowtie software (Langmead et al., 2009). Complete ChIP-Seq data has been submitted to the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) database under accession number pending.

Expression analysis

For RNA analysis, hES and hiPS colonies were mechanically isolated and pooled for RNA extraction. Total RNA was isolated from ES, iPS, and fibroblast cells using RNeasy MiniKit (Qiagen). 5 microgram of total RNA was labeled according to standard Affymetrix protocols and hybridized to Affymetrix HG-U133 2.0 plus arrays. The data were analyzed by using Affymetrix Gene Chip Operating Software using default settings. Complete gene expression data has been submitted to the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo)

database under accession number pending. Expression data was quantile normalized and analyzed for differential expression using the NIA Array Analysis Tool (http://lgsun.grc.nia.nih.gov/ANOVA/). A more detailed description of the gene expression analysis methods is provided in the Supplemental Experimental Procedures.

Acknowledgments

We thank Tom Volkert, Sumeet Gupta, Jen Love, Jeong-Ah Kwon of the WICMT for help with direct sequencing and data analysis; Ping Xu and Raaji Alagappan for help with human ES cell culture, Bingbing Yuan, George Bell and Whitehead Institute BaRC for computational analysis. We acknowledge the generous support of Lilianne and Hillel Bachrack and Susan Whitehead. DH is a Merck Fellow of the Life Science Research Foundation. RJ is an advisor to Stemgen and a cofounder of Fate Therapeutics.

References

Aoi, T., Yae, K., Nakagawa, M., Ichisaka, T., Okita, K., Takahashi, K., Chiba, T., and Yamanaka, S. (2008). Generation of Pluripotent Stem Cells from Adult Mouse Liver and Stomach Cells. Science *321*, 699-702.

Bammler, T., Beyer, R.P., Bhattacharya, S., Boorman, G.A., Boyles, A., Bradford, B.U., Bumgarner, R.E., Bushel, P.R., Chaturvedi, K., Choi, D., *et al.* (2005). Standardizing global gene expression analysis between laboratories and across platforms. Nat Methods *2*, 351-356.

Bernstein, B.E., Humphrey, E.L., Erlich, R.L., Schneider, R., Bouman, P., Liu, J.S., Kouzarides, T., and Schreiber, S.L. (2002). Methylation of histone H3 Lys 4 in coding regions of active genes. Proc Natl Acad Sci *99*, 8695-8700.

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.

Boland, M.J., Hazen, J.L., Nazor, K.L., Rodriguez, A.R., Gifford, W., Martin, G., Kupriyanov, S., and Baldwin, K.K. (2009). Adult mice generated from induced pluripotent stem cells. Nature *461*, 91-94.

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.

Bracken, A.P., Dietrich, N., Pasini, D., Hansen, K.H., and Helin, K. (2006). Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. Genes Dev *20*, 1123-1136.

Chin, M.H., Mason, M.J., Xie, W., Volinia, S., Singer, M., Peterson, C., Ambartsumyan, G., Aimiuwu, O., Richter, L., Zhang, J., *et al.* (2009). Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. Cell Stem Cell *5*, 111-123.

Doi, A., Park, I.H., Wen, B., Murakami, P., Aryee, M.J., Irizarry, R., Herb, B., Ladd-Acosta, C., Rho, J., Loewer, S., *et al.* (2009). Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. Nat Genet *41*, 1350-1353.

Ebert, B.L., and Golub, T.R. (2004). Genomic approaches to hematologic malignancies. Blood *104*, 923-932.

Feng, Q., Lu, S.J., Klimanskaya, I., Gomes, I., Kim, D., Chung, Y., Honig, G.R., Kim, K.S., and Lanza, R. (2010). Hemangioblastic Derivatives from Human Induced Pluripotent Stem Cells Exhibit Limited Expansion and Early Senescence. Stem Cells.

Ghosh, Z., Wilson, K.D., Wu, Y., Hu, S., Quertermous, T., and Wu, J.C. (2010). Persistent donor cell gene expression among human induced pluripotent stem cells contributes to differences with human embryonic stem cells. PLoS ONE *5*, e8975.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77-88.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., *et al.* (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223-227.

Hockemeyer, D., Soldner, F., Cook, E.G., Gao, Q., Mitalipova, M., and Jaenisch, R. (2008). A drug-inducible system for direct reprogramming of human somatic cells to pluripotency. Cell Stem Cell *3*, 346-353.

Hu, B.Y., Weick, J.P., Yu, J., Ma, L.X., Zhang, X.Q., Thomson, J.A., and Zhang, S.C. (2010). Neural differentiation of human induced pluripotent stem cells follows developmental principles but with variable potency. Proc Natl Acad Sci U S A *107*, 4335-4340.

Ivanova, N., Dobrin, R., Lu, R., Kotenko, I., Levorse, J., DeCoste, C., Schafer, X., Lun, Y., and Lemischka, I.R. (2006). Dissecting self-renewal in stem cells with RNA interference. Nature *442*, 533-538.

Kang, L., Wang, J., Zhang, Y., Kou, Z., and Gao, S. (2009). iPS cells can support full-term development of tetraploid blastocyst-complemented embryos. Cell Stem Cell *5*, 135-138.

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 Biol *10*, R25.

Lee, T.I., Jenner, R.G., Boyer, L.A., Guenther, M.G., Levine, S.S., Kumar, R.M., Chevalier, B., Johnstone, S.E., Cole, M.F., Isono, K., *et al.* (2006). Control of developmental regulators by Polycomb in human embryonic stem cells. Cell *125*, 301-313.

Lengner, C.J., Gimelbrant, A.A., Cheung, W.A., Erwin, J.a., Guenther, M.G., Alagappan, R., Xu, P., Powers, D., Barrett, B.C., Young, R.A., *et al.* (2010). Derivation of pre-X inactivation human embryonic stem cells under physiological oxygen concentrations. Cell *141*, 872-883.

Maherali, N., Ahfeldt, T., Rigamonti, A., Utikal, J., Cowan, C., and Hochedlinger, K. (2008). A high-efficiency system for the generation and study of human induced pluripotent stem cells. Cell Stem Cell *3*, 340-345.

Maherali, N., Sridharan, R., Xie, W., Utikal, J., Eminli, S., Arnold, K., Stadtfeld, M., Yachechko, R., Tchieu, J., Jaenisch, R., *et al.* (2007). Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. Cell Stem Cell *1*, 55-70.

Marchetto, M.C., Yeo, G.W., Kainohana, O., Marsala, M., Gage, F.H., and Muotri, A.R. (2009). Transcriptional signature and memory retention of human-induced pluripotent stem cells. PLoS ONE *4*, e7076.

Marson, A., Levine, S.S., Cole, M.F., Frampton, G.M., Brambrink, T., Johnstone, S., Guenther, M.G., Johnston, W.K., Wernig, M., Newman, J., *et al.* (2008). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell *134*, 521-533.

Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B., *et al.* (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature *454*, 766-770.

Mikkelsen, T.S., Hanna, J., Zhang, X., Ku, M., Wernig, M., Schorderet, P., Bernstein, B.E., Jaenisch, R., Lander, E.S., and Meissner, A. (2008). Dissecting direct reprogramming through integrative genomic analysis. Nature *454*, 49-55.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P., *et al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-560.

Nakagawa, M., Koyanagi, M., Tanabe, K., Takahashi, K., Ichisaka, T., Aoi, T., Okita, K., Mochiduki, Y., Takizawa, N., and Yamanaka, S. (2007). Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. Nat Biotechnol.

Ng, H.H., Robert, F., Young, R.A., and Struhl, K. (2003). Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. Mol Cell *11*, 709-719.

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

Ozsolak, F., Poling, L.L., Wang, Z., Liu, H., Liu, X.S., Roeder, R.G., Zhang, X., Song, J.S., and Fisher, D.E. (2008). Chromatin structure analyses identify miRNA promoters. Genes Dev *22*, 3172-3183.

Pan, G., Tian, S., Nie, J., Yang, C., Ruotti, V., Wei, H., Jonsdottir, G.A., Stewart, R., and Thomson, J.A. (2007). Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. Cell Stem Cell *1*, 299-312.

Pietersen, A.M., and van Lohuizen, M. (2008). Stem cell regulation by polycomb repressors: postponing commitment. Curr Opin Cell Biol *20*, 201-207.

Ringrose, L., and Paro, R. (2004). Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. Annu Rev Genet *38*, 413-443.

Saha, K., and Jaenisch, R. (2009). Technical Challenges in Using Human Induced Pluripotent Stem Cells to Model Disease. Cell Stem Cell *5*, 584-595.

Santos-Rosa, H., Schneider, R., Bannister, A.J., Sherriff, J., Bernstein, B.E., Emre, N.C., Schreiber, S.L., Mellor, J., and Kouzarides, T. (2002). Active genes are tri-methylated at K4 of histone H3. Nature *419*, 407-411.

Schneider, R., Bannister, A.J., Myers, F.A., Thorne, A.W., Crane-Robinson, C., and Kouzarides, T. (2004). Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. Nat Cell Biol *6*, 73-77.

Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B., and Cavalli, G. (2007). Genome regulation by polycomb and trithorax proteins. Cell *128*, 735-745.

Schwartz, Y.B., and Pirrotta, V. (2008). Polycomb complexes and epigenetic states. Curr Opin Cell Biol *20*, 266-273.

Sharov, A.A., Dudekula, D.B., and Ko, M.S. (2005). A web-based tool for principal component and significance analysis of microarray data. Bioinformatics *21*, 2548-2549.

Simon, J.A., and Kingston, R.E. (2009). Mechanisms of polycomb gene silencing: knowns and unknowns. Nat Rev Mol Cell Biol *10*, 697-708.

Smith, K.P., Luong, M.X., and Stein, G.S. (2009). Pluripotency: toward a gold standard for human ES and iPS cells. J Cell Physiol *220*, 21-29.

Soldner, F., Hockemeyer, D., Beard, C., Gao, Q., Bell, G.W., Cook, E.G., Hargus, G., Blak, A., Cooper, O., Mitalipova, M., *et al.* (2009). Parkinson's disease patient-derived induced pluripotent stem cells free of viral reprogramming factors. Cell *136*, 964-977.

Stadtfeld, M., Apostolou, E., Akutsu, H., Fukuda, A., Follett, P., Natesan, S., Kono, T., Shioda, T., and Hochedlinger, K. (2010). Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. Nature.

Takahashi, K., Okita, K., Nakagawa, M., and Yamanaka, S. (2007). Induction of pluripotent stem cells from fibroblast cultures. Nat Protoc *2*, 3081-3089.

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.

Wernig, M., Meissner, A., Foreman, R., Brambrink, T., Ku, M., Hochedlinger, K., Bernstein, B.E., and Jaenisch, R. (2007). In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. Nature *448*, 318-324.

Yamanaka, S. (2007). Strategies and new developments in the generation of patient-specific pluripotent stem cells. Cell Stem Cell *1*, 39-49.

Yu, J., Hu, K., Smuga-Otto, K., Tian, S., Stewart, R., Slukvin, II, and Thomson, J.A. (2009). Human induced pluripotent stem cells free of vector and transgene sequences. Science *324*, 797-801.

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.

Zhao, X.D., Han, X., Chew, J.L., Liu, J., Chiu, K.P., Choo, A., Orlov, Y.L., Sung, W.K., Shahab, A., Kuznetsov, V.A., *et al.* (2007). Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. Cell Stem Cell *1*, 286-298.

Zhao, X.Y., Li, W., Lv, Z., Liu, L., Tong, M., Hai, T., Hao, J., Guo, C.L., Ma, Q.W., Wang, L., *et al.* (2009). iPS cells produce viable mice through tetraploid complementation. Nature *461*, 86-90.

Figure 1

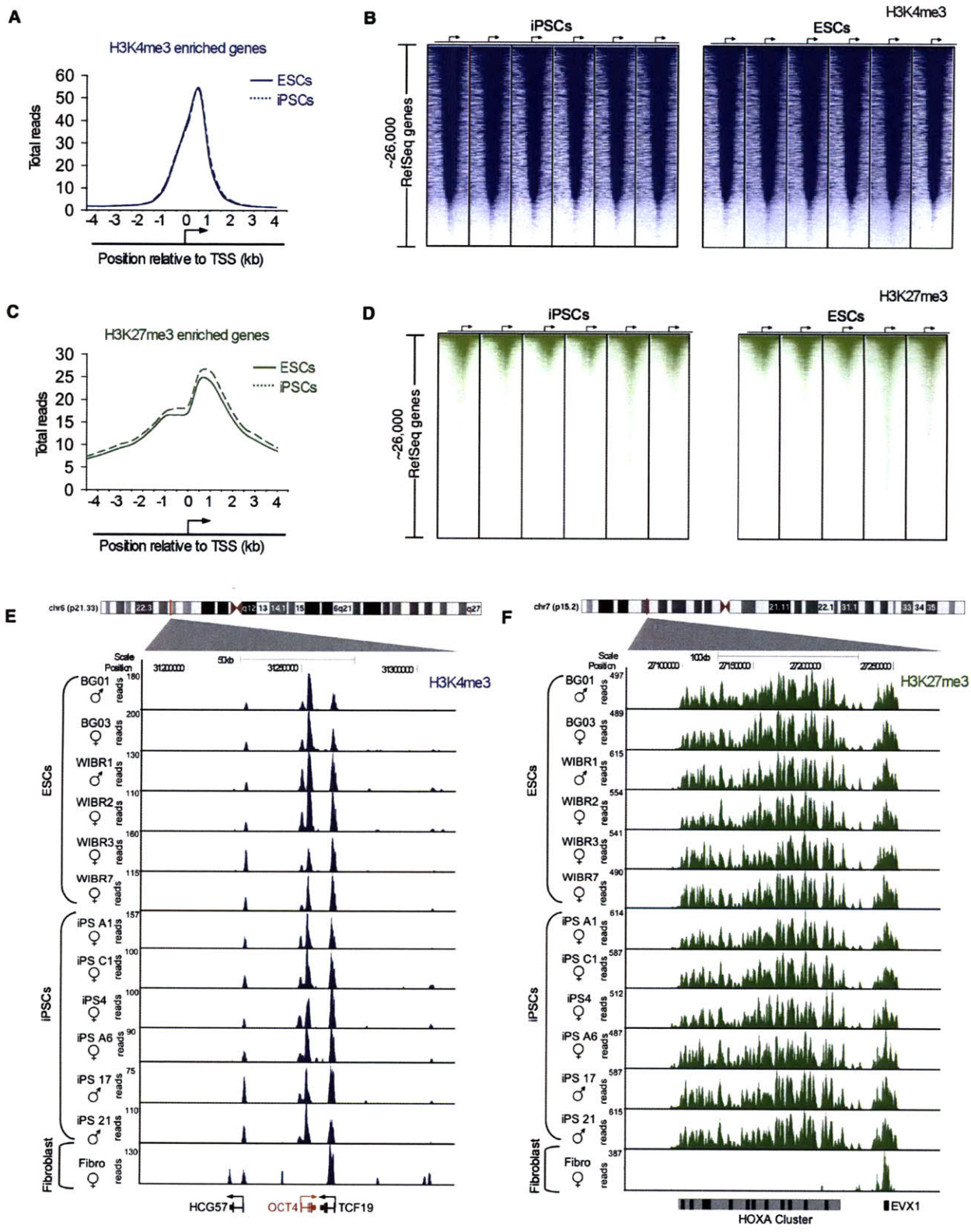

A
H3K4me3 enriched genes
Total reads
ESCs
iPSCs
Position relative to TSS (kb)
B
H3K4me3
iPSCs
ESCs
~26,000 RefSeq genes
C
H3K27me3 enriched genes
Total reads
ESCs
iPSCs
Position relative to TSS (kb)
D
H3K27me3
iPSCs
ESCs
~26,000 RefSeq genes
E
H3K4me3
ESCs
iPSCs
Fibroblast
BG01
BG03
WIBR1
WIBR2
WIBR3
WIBR7
iPS A1
iPS C1
iPS4
iPS A6
iPS 17
iPS 21
Fibro
HCG57
OCT4
TCF19
F
H3K27me3
ESCs
iPSCs
Fibroblast
HOXA Cluster
EVX1

Figure 1. Genome-wide maps of chromatin modifications show human iPS cells share key features with ES cells

A. Composite H3K4me3 enrichment profile for all RefSeq genes in ES cells (solid blue) and iPS cells (dashed blue). The transcription start site (TSS) and direction of transcription of the average gene is noted by an arrow.

B. ChIP-Seq density heat map of histone H3K4me3 (blue) for all RefSeq genes. Gene order was determined by highest average ChIP-Seq density in ES cells and arranged from highest to lowest density. The TSS and direction of transcription of genes is indicated by an arrow and the genomic region from –4.5kb to +4.5kb relative to the TSS shown.

C. Composite H3K27me3 enrichment profile for all RefSeq genes enriched for H3K27me3 in ES cells (solid green) and iPS cells (dashed green). The TSS and direction of transcription of the average gene is noted by an arrow.

D. ChIP-Seq density heat map of histone H3K27me3 (green) for all RefSeq genes. Gene order was determined by highest average ChIP-Seq density in ES cells and arranged from highest to lowest density. The TSS and direction of transcription of genes is indicated by an arrow and the genomic region from –4.5kb to +4.5kb relative to the TSS shown.

E. ChIP-Seq density for H3K4me3 (blue) at the OCT4 locus in human ES, iPS, and fibroblast cell lines. The position of the loci within chromosome 6 and the scale is shown above the gene tracks.

F. ChIP-Seq density for H3K27me3 (green) in the HOXA cluster in human ES, iPS, and fibroblast cell lines. The position of the cluster within chromosome 7 and the scale is shown above the gene tracks. See also Tables S1-S4.

Figure 2

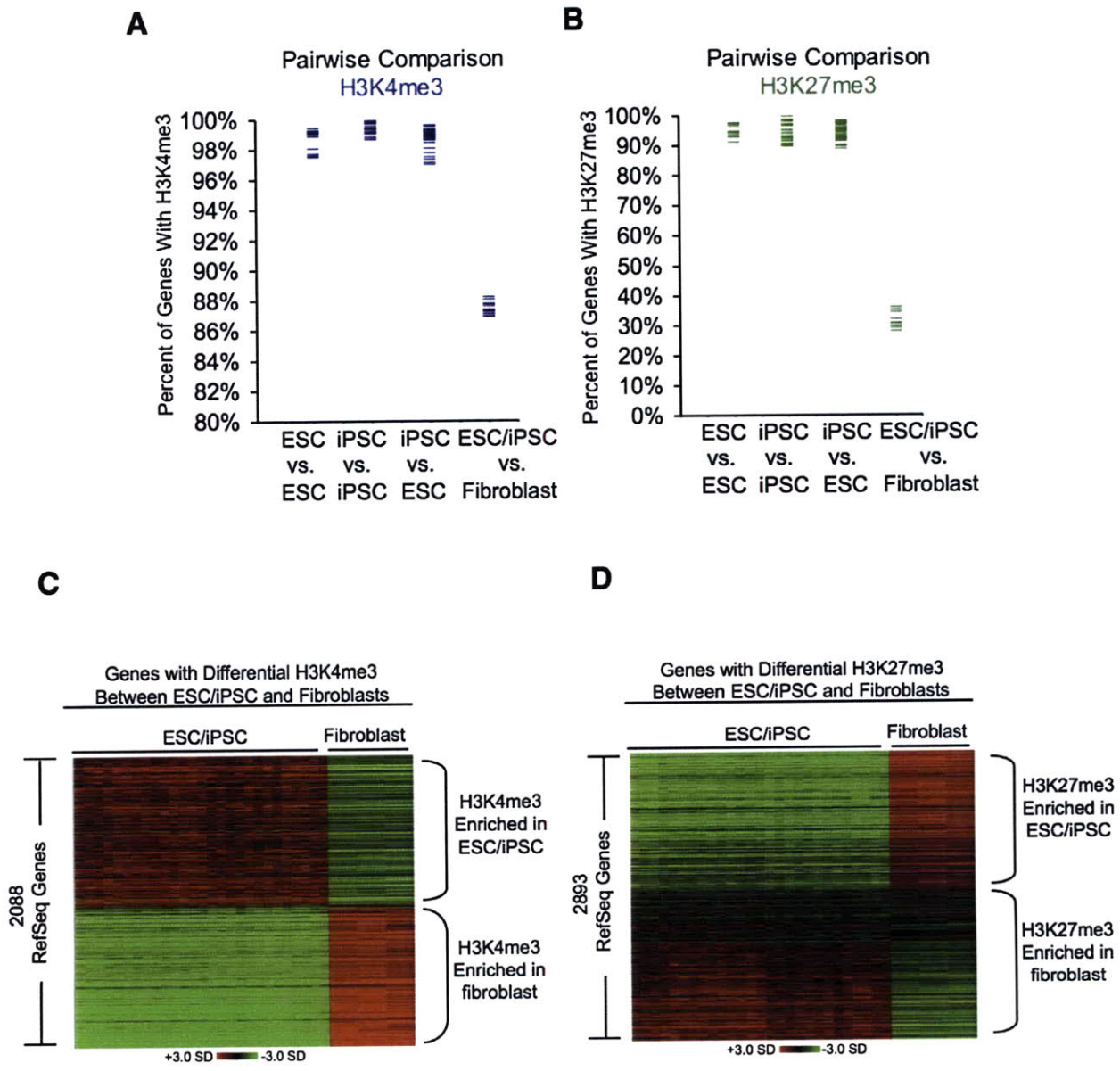

A
Pairwise Comparison
H3K4me3
Percent of Genes With H3K4me3
100%
98%
96%
94%
92%
90%
88%
86%
84%
82%
80%
ESC vs. ESC
iPSC vs. iPSC
iPSC vs. ESC
ESC/iPSC vs. Fibroblast
B
Pairwise Comparison
H3K27me3
Percent of Genes With H3K27me3
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
ESC vs. ESC
iPSC vs. iPSC
iPSC vs. ESC
ESC/iPSC vs. Fibroblast
C
Genes with Differential H3K4me3 Between ESC/iPSC and Fibroblasts
ESC/iPSC
Fibroblast
2088 RefSeq Genes
H3K4me3 Enriched in ESC/iPSC
H3K4me3 Enriched in fibroblast
+3.0 SD
-3.0 SD
D
Genes with Differential H3K27me3 Between ESC/iPSC and Fibroblasts
ESC/iPSC
Fibroblast
2893 RefSeq Genes
H3K27me3 Enriched in ESC/iPSC
H3K27me3 Enriched in fibroblast
+3.0 SD
-3.0 SD

Figure 2. Similarity in genes and regions occupied by modified histones in ES and iPS cells

A. Pairwise comparisons of genes occupied by H3K4me3 in ES, iPS, and fibroblast cells. Each blue bar represents an individual pairwise comparison of the set of genes identified as enriched in one cell line with the set enriched in a second cell line. Comparisons between two ES lines (ES vs ES), between two iPS lines (iPS vs iPS), between an ES line and an iPS line (ES vs iPS), and between an ES or iPS line and fibroblast cells (ES/iPS vs fibroblast) are shown in separate columns. Gene occupancy was determined as described in Supplemental Experimental Procedures.

B. Pairwise comparisons of genes occupied by H3K27me3 in ES, iPS, and fibroblast cells. Each green bar represents an individual pairwise comparison of the set of genes identified as enriched in one cell line with the set enriched in a second cell line as in (A).

C. Expression data for genes differentially occupied by H3K4me3 in pluripotent cells (ES and iPS) and fibroblast cells. Genes are ordered by the magnitude of differential H3K4me3 occupancy and relative gene expression is shown. Samples with higher than average expression are shown in red and samples with lower than average expression are shown in green (scale in standard deviations).

D. Expression data for genes differentially occupied by H3K27me3 in pluripotent cells (ES and iPS) and fibroblast cells. Genes are ordered by the magnitude of differential H3K27me3 occupancy and relative gene expression is shown. Samples with higher than average expression are shown in red and samples with lower than average expression are shown in green (scale in standard deviations). See also Figure S1 and Tables S1-S4.

Figure 3

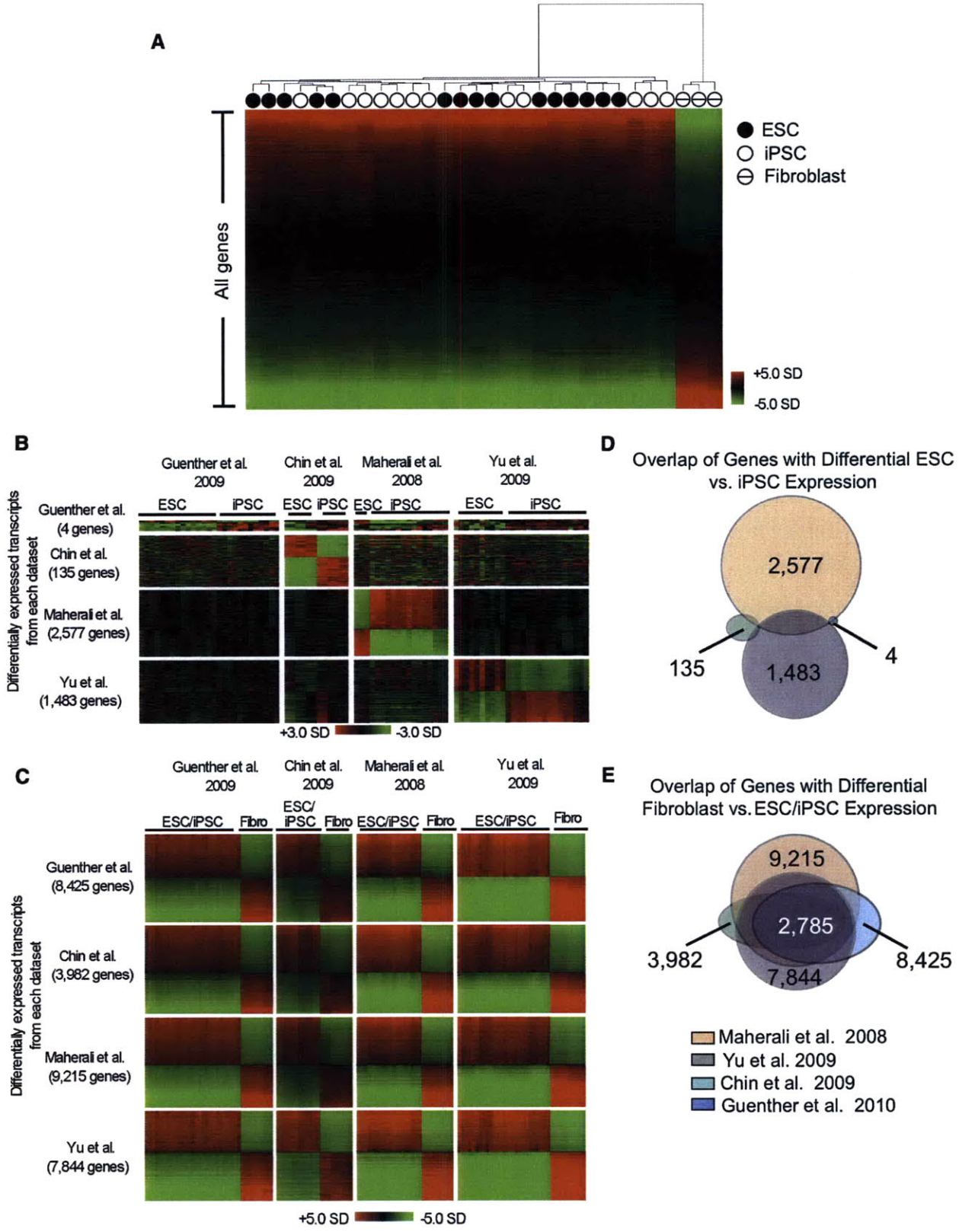

A
All genes
ESC
iPSC
Fibroblast
+5.0 SD
-5.0 SD
B
Guenther et al. 2009
Chin et al. 2009
Maherali et al. 2008
Yu et al. 2009
ESC
iPSC
Guenther et al. (4 genes)
Chin et al. (135 genes)
Maherali et al. (2,577 genes)
Yu et al. (1,483 genes)
Differentially expressed transcripts from each dataset
+3.0 SD
-3.0 SD
C
ESC/iPSC
Fibro
Guenther et al. (8,425 genes)
Chin et al. (3,982 genes)
Maherali et al. (9,215 genes)
Yu et al. (7,844 genes)
+5.0 SD
-5.0 SD
D
Overlap of Genes with Differential ESC vs. iPSC Expression
2,577
135
1,483
4
E
Overlap of Genes with Differential Fibroblast vs. ESC/iPSC Expression
9,215
2,785
3,982
7,844
8,425
Maherali et al. 2008
Yu et al. 2009
Chin et al. 2009
Guenther et al. 2010

Figure 3. Limited variation in gene expression between human ES and iPS cells

A. Unsupervised hierarchical clustering of whole genome expression data from ES cells (closed circle), iPS cells (open circle), and fibroblasts (hashed circle). Expression data were ordered by the magnitude of differential expression between pluripotent cells (ES or iPS) cells and fibroblast cells. Normalization and analysis for all expression data is described in Supplemental Experimental Procedures.

B. Differential gene expression between ES and iPS cells across multiple expression datasets. For each dataset (top to bottom) the transcripts with statistically significant differential expression between ES and iPS cells are shown. Within the set of differentially expressed transcripts from each dataset, expression data was ordered by the statistical significance of differential expression between ES and iPS cells and then aligned to all other datasets for comparison. Samples with higher than average expression are shown in red and samples with lower than average expression are shown in green (scale in standard deviations).

C. Differential gene expression between fibroblasts and pluripotent (ES and iPS) cells. Expression data were ordered by the statistical significance of differential expression between fibroblasts and ES/iPS cells in each dataset and then aligned to all other datasets for comparison. Samples with higher than average expression are shown in red and samples with lower than average expression are shown in green (scale in standard deviations).

D. Overlap of differentially expressed genes between ES and iPS cells in various expression datasets. The numbers of genes differentially expressed between ES and iPS cells are indicated in black. The total overlap of all gene sets is zero.

E. Overlap of differentially expressed genes between fibroblast and pluripotent (ES and iPS) cells in various expression datasets. The numbers of genes differentially expressed between fibroblast and pluripotent cells are indicated in black. The total overlap of all gene sets is shown in white. See also Figure S2 and Table S5.

Figure 4

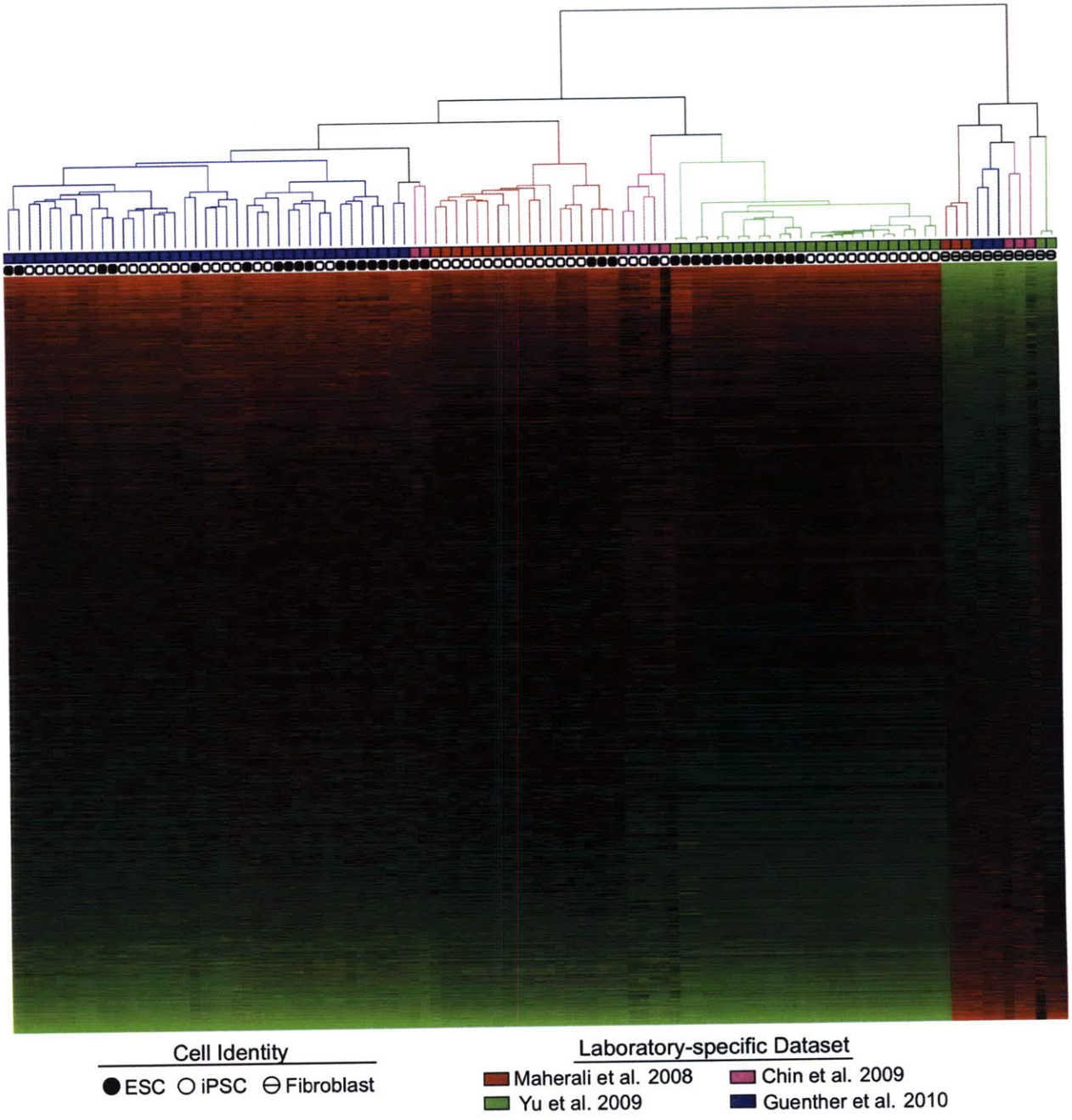

Cell Identity
ESC iPSC Fibroblast
Laboratory-specific Dataset
Maherali et al. 2008
Chin et al. 2009
Yu et al. 2009
Guenther et al. 2010

Figure 4. Human ES and iPS cell expression data clusters more by laboratory than by ES/iPS identity

Gene expression datasets for human ES, iPS, and fibroblasts cells from four laboratories (Guenther et al., present study; Maherali et al., 2008; Chin et al., 2009; Yu et al., 2009) were quantile normalized as one group. Genes were Z-score normalized and data were subjected to hierarchical clustering (centered correlation distance, centroid linkage) of samples. Genes were ordered from greatest to least magnitude of differential expression between pluripotent and fibroblast cells. Solid circles indicate ES cell samples, empty circles indicate iPS cell samples, and hashed circles represent fibroblast cell samples. Data from individual labs are coded by color as Guenther et al (blue), Maherali et al. (red), Chin et al. (purple), Yu et al. (green). See also Table S5.

Chapter 3

Connecting microRNA Genes to the Core Transcriptional Regulatory Circuitry of Embryonic Stem Cells

Published as: Alexander Marson, Stuart S. Levine, Megan F. Cole, Garrett M. Frampton, Tobias Brambrink, Matthew G. Guenther, Wendy K. Johnston, Marius Wernig, Jamie Newman, Thomas L. Volkert, David P. Bartel, Rudolf Jaenisch, Richard A. Young (2008). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell *134*, 521-33.

Abstract

MicroRNAs (miRNAs) are crucial for normal embryonic stem (ES) cell self-renewal and cellular differentiation, but how miRNA gene expression is controlled by the key transcriptional regulators of ES cells has not been established. We describe here a new map of the transcriptional regulatory circuitry of ES cells that incorporates both protein-coding and miRNA genes, and which is based on high-resolution ChIP-Seq data, systematic identification of miRNA promoters, and quantitative sequencing of short transcripts in multiple cell types. We find that the key ES cell transcription factors are associated with promoters for most miRNAs that are preferentially expressed in ES cells and with promoters for a set of silent miRNA genes. This silent set of miRNA genes is co-occupied by Polycomb Group proteins in ES cells and expressed in a tissue-specific fashion in differentiated cells. These data reveal how key ES cell transcription factors promote the miRNA expression program that contributes to self-renewal and cellular differentiation, and integrate miRNAs and their targets into an expanded model of the regulatory circuitry controlling ES cell identity.

Introduction

Embryonic stem (ES) cells hold significant potential for clinical therapies because of their distinctive capacity to both self-renew and differentiate into a wide range of specialized cell types. Understanding the transcriptional regulatory circuitry of ES cells and early cellular differentiation is fundamental to understanding human development and realizing the therapeutic potential of these cells. Transcription factors that control ES cell pluripotency and self-renewal have been identified (Chambers and Smith, 2004; Niwa, 2007; Silva and Smith, 2008) and a draft of the core regulatory circuitry by which these factors exert their regulatory effects on protein-coding genes has been described (Boyer et al., 2005; Loh et al., 2006; Lee et al., 2006; Boyer et al. 2006; Jiang et al., 2008; Cole et al., 2008; Kim et al., 2008). MicroRNAs (miRNAs) are also likely to play key roles in ES cell gene regulation (Kanellopoulou et al., 2005; Murchison et al., 2005; Wang et al., 2007), but little is known about how miRNAs participate in the core regulatory circuitry controlling self-renewal and pluripotency in ES cells.

Several lines of evidence indicate that miRNAs contribute to the control of early development. miRNAs appear to regulate the expression of a significant percentage of all genes in a wide array of mammalian cell types (Lewis et al., 2005; Lim et al., 2005; Krek et al., 2005; Farh et al., 2005). A subset of miRNAs is preferentially expressed in ES cells or embryonic tissue (Houbaviy et al., 2003; Suh et al., 2004; Houbaviy et al., 2005; Mineno et al., 2006). Dicer-deficient mice fail to develop (Bernstein et al., 2003) and ES cells deficient in miRNA processing enzymes show defects in differentiation, self-renewal and perhaps viability (Kanellopoulou et al., 2005; Murchison et al., 2005; Wang et al., 2007; Calabrese et al., 2008). Specific miRNAs have been shown to participate in mammalian cellular differentiation and embryonic development (Stefani and Slack, 2008). However, how transcription factors and miRNAs function together in the regulatory circuitry that controls early development has not yet been examined.

The major limitation in connecting miRNA genes to the core transcriptional circuitry of ES cells has been sparse annotation of miRNA gene transcriptional start sites and promoter regions. Mature miRNAs, which specify post-transcriptional gene repression, arise from larger transcripts that are then processed (Bartel, 2004). Over 400 mature miRNAs have been confidently identified in the human genome (Landgraf et al., 2007), but only a minority of the

primary transcripts have been identified and annotated. Prior attempts to connect ES cell transcriptional regulators to miRNA genes have searched for transcription factor binding sites only close to the annotated mature miRNA sequences (Boyer et al., 2005; Loh et al., 2006; Lee et al., 2006). Additionally, studies of the core transcriptional circuitry of ES cells have compared transcription factor occupancy to mRNA expression data, but have not systemically examined miRNA expression in ES cells and differentiated cell types, limiting our knowledge of transcriptional regulation of miRNA genes in these cells (Boyer et al., 2005; Loh et al., 2006; Lee et al., 2006; Cole et al. 2008).

To incorporate miRNA gene regulation into the model of transcriptional regulatory circuitry of ES cells, we began by generating new, high-resolution, genome-wide maps of binding sites for key ES cell transcription factors using massive parallel sequencing of chromatin immunoprecipitation (ChIP-Seq). These data reveal highly overlapping occupancy of Oct4, Sox2, Nanog and Tcf3 at the transcriptional start sites of miRNA transcripts, which we systematically mapped based on a method that uses chromatin landmarks and transcript data. We then carried out quantitative sequencing of short transcripts in ES cells, neural precursor cells (NPCs) and mouse embryonic fibroblasts (MEFs), which revealed that Oct4, Sox2, Nanog and Tcf3 occupy the promoters of most miRNAs that are preferentially or uniquely expressed in ES cells. Our data also revealed that a subset of the Oct4/Sox2/Nanog/Tcf3 occupied miRNA genes are silenced in ES cells by Polycomb Group proteins, but are expressed later in development in specific lineages. High-resolution transcription factor location analysis, systematic mapping of the primary miRNA transcriptional start sites in mouse and human, and quantitative sequencing of miRNAs in three different cell types provide a valuable data resource for studies of the gene expression program in ES and other cells and the regulatory mechanisms that control cell fate. The data also produce an expanded model of ES cell core transcriptional regulatory circuitry that now incorporates transcriptional regulation of miRNAs, and post-transcriptional regulation mediated by miRNAs, into the molecular understanding of pluripotency and early cellular differentiation.

Results

High-resolution genome-wide location analysis in ES cells with ChIP-Seq

To connect miRNA genes to the core transcriptional circuitry of ES cells, we first generated high-resolution genome-wide maps of Oct4, Sox2, Nanog, and Tcf3 occupancy (Figure 1). ChIP-Seq allowed us to map transcription factor binding sites and histone modifications across the entire genome at high resolution (Barski et al., 2007; Johnson et al., 2007; Mikkelsen et al., 2007; Robertson et al., 2007), and we optimized the protocol to allow for robust analysis of transcription factor binding in murine ES cells (Supplemental Material). Oct4, Sox2, Nanog and Tcf3 were found to co-occupy 14,230 sites in the genome (Figure 1A, Supplemental Figures S1 and S2, Supplemental Tables S1-S3). Approximately one quarter of these occurred within 8kb of the transcription start site of 3,289 annotated genes, another one quarter occurred within genes but more than 8kb from the start site, and almost half occurred in intergenic regions distal from start sites (Supplemental Text). Binding of the four factors at sites surrounding the *Sox2* gene (Figure 1B) exemplifies two key features of the data: all four transcription factors co-occupied the identified binding sites and the resolution was sufficient to determine the DNA sequence associated with these binding events to a resolution of <25bp. Composite analysis of all bound regions provided higher resolution and suggested how these factors occupy their common DNA-sequence motif (Supplemental Figure S3, Supplemental Table S4). Knowledge of these binding sites provided data necessary to map these key transcription factors to the promoters of miRNA genes.

Identification of miRNA promoters

Imperfect knowledge of the start sites of primary miRNA transcripts has limited our ability to identify the transcription factor binding events that control miRNA gene expression in vertebrates. Previous strategies to identify the 5' ends of primary miRNAs have been hampered because they relied on isolation of transient primary miRNA transcript, required knowledge of the specific cell type in which each given miRNA is transcribed, or focused only on potential start sites proximal to mature miRNAs (Fukao et al., 2007; Mikkelsen et al., 2007; Zhou et al., 2007; Barrera et al., 2008). To systematically identify transcriptional start sites for miRNA genes in the mouse and human genomes, we took advantage of the recent observation that

histone H3 is tri-methylated at its lysine 4 residue (H3K4me3) at the transcriptional start sites of most genes in the genome, even when genes are not productively transcribed, and knowledge that this covalent modification is restricted to sites of transcription initiation (Barski et al., 2007; Guenther et al., 2007). We used the genomic coordinates of the H3K4me3 enriched loci derived from multiple cell types (Supplemental Table S5, Barski et al., 2007; Guenther et al., 2007; Mikkelsen et al., 2007) to create a library of candidate transcription start sites in both human and mouse (Figure 2 and Supplemental Figure S4).

High-confidence promoters were identified for over 80% of miRNAs in both mouse and human (Figure 2, Supplemental Figure S4 and Supplemental Tables S6 and S7). These promoters were associated with 185 murine primary microRNA transcripts (pri-miRNAs) (specifying 336 mature miRNAs), and 294 human pri-miRNAs (specifying 441 mature miRNAs) (Supplemental Table S6 and S7). To identify promoters for miRNA genes, the association of candidate transcriptional start sites with regions encoding mature miRNAs was scored based on proximity to annotated mature miRNA sequences (Landgraf et al., 2007), available EST data, and conservation between species (Figure 2A and Supplemental Figure S5 and Supplemental Text). Four lines of evidence indicate that this approach identified genuine transcriptional start sites for miRNA genes. Existing EST data provided evidence that the predicted transcripts do in fact originate at the identified start sites and continue through the annotated loci of mature miRNAs (Figure 2B and Supplemental Figures S5). In addition to the chromatin signature of promoters, a high fraction of these regions contained CpG islands, a DNA sequence element often associated with promoters (Figure 2B and Supplemental Table S6 and S7). Third, in some instances where evidence of primary miRNA transcripts, which may be present only transiently before processing, were not available in published databases at the identified transcriptional start sites, chromatin marks associated with transcriptional elongation including nucleosomes methylated at H3 lysine 36 (H3K36me3) and H3 lysine 79 (H3K79me2), provided evidence that such transcripts are actively produced (Figure 2C and Mikkelsen et al., 2007). Finally, most miRNA promoters showed evidence of H3K4me3 enrichment in multiple tissues, as observed at the promoters of most protein-coding genes (Barski et al., 2007; Guenther et al., 2007; Heintzman et al., 2007) (Figure 2D).

Occupancy of miRNA promoters by core ES cell transcription factors

The binding sites of the ES cell transcription factors Oct4, Sox2, Nanog and Tcf3 were next mapped to these high-confidence miRNA promoters (Figure 3). In murine ES cells, Oct4, Sox2, Nanog, and Tcf3 co-occupied the promoters for 55 distinct miRNA transcription units, which included three clusters of miRNAs that are expressed as large polycistrons, thus suggesting that these regulators have the potential to directly control the transcription of 81 distinct mature miRNAs (Figure 3A and Supplemental Tables S6). This set of miRNAs occupied by Oct4/Sox2/Nanog/Tcf3 represents roughly 20 percent of annotated mammalian miRNAs, similar to the ~20 percent of protein-coding genes that are bound at their promoters by these key transcription factors (Supplemental Table S2).

To determine if transcription factor occupancy of miRNA promoters is conserved across species, we performed genome-wide location analysis for Oct4 in human ES cells using microarray-based analysis. We found extensive conservation of the set of miRNA genes that were occupied at their promoters by Oct4, as exemplified by the mir-302 cluster (Figure 3A and 3B and Supplemental Tables S7 and S8). Transcription factor occupancy does not necessarily mean that the adjacent gene is regulated by that factor; conserved transcription factor occupancy of a promoter, however, suggests gene regulation by that factor. Thus, our data identify a set of miRNA genes that are bound at their promoters by key ES cell transcription factors in mouse and human cells (Figure 3C), suggesting that core ES cell transcription factor regulation of these particular miRNA transcripts has functional significance.

Regulation of Oct4 bound miRNA transcripts during differentiation

Oct4 and Nanog are rapidly silenced as ES cells begin to differentiate (Chambers and Smith, 2004; Niwa, 2007). If the Oct4/Sox2/Nanog/Tcf3 complex is required for activation or repression of its target miRNAs, the targets should be differentially expressed when ES cells are compared to a differentiated cell-type. To test this hypothesis, Solexa sequencing of 18-30 nucleotide transcripts in ES cells, mouse embryonic fibroblasts (MEFs), and neural precursors (NPCs), was performed to obtain quantitative information on the abundance of miRNAs in pluripotent cells relative to two differentiated cell types (Figure 4).

In each cell type examined, a small subset of mature miRNA transcripts predominated (Figure 4A). Members of the mir-290-295 cluster, which encodes multiple miRNAs with the same seed sequence, constituted approximately two thirds of all mature miRNA transcripts in

murine ES cells. Let-7 family members constituted roughly one quarter and one half of miRNAs in MEFs and NPCs, respectively. The mir-290-295 cluster, which dominated the expression profile of ES cells, but was scarce in both MEFs and NPCs, is occupied at its promoters by Oct4, Sox2, Nanog and Tcf3 (Figure 3A), consistent with the hypothesis that these factors are important for maintaining the expression of the mir-290-295 miRNA cluster in ES cells.

To determine if the behavior of the mir-290-295 cluster is typical of the Oct4/Sox2/Nanog/Tcf3-occupied miRNAs, we further examined the expression of this set of miRNAs in the three cell types. Figure 4B shows how the abundance of this group of miRNAs changed in MEFs and NPCs relative to ES cells. Approximately half of the miRNAs dropped more than an order of magnitude in abundance in MEFs and NPCs relative to ES cells. A small subset of the Oct4/Sox2/Nanog/Tcf3-occupied miRNAs, which will be further discussed below, were expressed only at low levels in ES cells and showed increased abundance in MEFs and NPCs.

Oct4/Sox2/Nanog/Tcf3-occupied miRNAs are, in general, preferentially expressed in embryonic stem cells, as demonstrated by the analysis shown in Figure 4C. Whereas most miRNAs are unchanged in expression in ES cells relative to MEFs or NPCs, a significant portion of Oct4/Sox2/Nanog/Tcf3 occupied miRNAs are 100 fold more abundant in ES cells than in MEFs ($p<5 \times 10^{-15}$), and 1,000 fold more abundant in ES cells than in NPCs ($p<5 \times 10^{-9}$). This group of Oct4/Sox2/Nanog/Tcf3 bound miRNAs that is significantly more abundant in ES cells than in NPCs and MEFs, was also found to be actively expressed in induced pluripotent stem (iPS) cells (generated as described in Wernig et al., 2007), at levels comparable to that in ES cells, consistent with the hypothesis that core ES cell transcription factors maintain the expression of these miRNAs in pluripotent cells (Supplemental Figure S6).

Polycomb Group Proteins co-occupy tissue-specific miRNAs that are silenced in ES cells

We noted that the Oct4/Sox2/Nanog/Tcf3-bound miRNAs include the majority of miRNAs that were preferentially expressed in ES cells, but the data also revealed a second, smaller group of Oct4/Sox2/Nanog/Tcf3-bound miRNA genes that appeared to be transcriptionally inactive in ES cells (Figure 4B). This is reminiscent of previous observations with protein-coding genes in ES cells: Oct4 occupied a set of transcriptionally active genes, but also occupied, with Polycomb Group proteins, a set of transcriptionally repressed genes that are poised for expression upon

cellular differentiation (Lee et al., 2006; Bernstein et al., 2006; Boyer et al., 2006). We reasoned that Polycomb complexes might also co-occupy Oct4 bound promoters for miRNA genes that showed little or no evidence for expression, and thus contribute to their silencing. Indeed, new ChIP-Seq data for the Polycomb Goup protein Suz12 in murine ES cells supported this hypothesis (Figure 5A and Supplemental Tables S6, S7, S10). As expected, these promoters were also enriched for nucleosomes with histone H3K27me3, a chromatin modification catalyzed by Polycomb Group proteins (Figure 5A and Supplemental Table S6 and Mikkelsen et al., 2007). In keeping with the repressive function of the Polycomb Group proteins reported at protein coding genes, miRNAs occupied at their promoters by Suz12 in ES cells were significantly less abundant in ES cells compared to all other miRNAs (Figure 5B). Approximately one quarter of the Oct4/Sox2/Nanog/Tcf3-occupied miRNAs belonged to the repressed set of miRNA genes bound by Suz12 in murine ES cells (Supplemental Tables S6 and S7).

To further examine the behavior of this set of miRNAs during embryonic cell-fate commitment, we returned to our quantitative sequencing data of short transcripts in ES cells, MEFs and NPCs (Figure 5C). Notably, miRNAs that were bound by Polycomb Group proteins in ES cells are among the transcripts that are specifically induced in each of these cell types. For example, transcript levels of miR-9, a miRNA previously identified in neural cells and which promotes neural differentiation (Lagos-Quintana et al., 2002; Krichevsky et al., 2006), are significantly elevated in NPCs relative to ES cells, but this miRNA remains repressed in MEFs. Similarly, miR-218 and miR-34b/34c expression is induced in MEFs, but remains at low levels in NPCs (Figure 5C). Consistent with Polycomb-mediated repression of these lineage-specific miRNAs, the repressive chromatin mark deposited by Polycomb Group proteins, H3K27me3, is selectively lost at the promoters of the miRNAs in the cells in which they are induced (Figure 5C and Mikkelsen et al., 2007).

The tissue-specific expression pattern of miRNAs repressed by Polycomb in ES cells is consistent with these miRNAs serving as determinants of cell-fate decisions in a manner analogous to the developmental regulators whose genes are repressed by Polycomb in ES cells (Lee et al., 2006; Bernstein et al., 2006; Boyer et al., 2006). Such a function in cell-fate determination would require that these miRNAs remain silenced in pluripotent ES cells. Indeed, the miRNAs that are repressed in ES cells by Polycomb Group proteins appear to be induced,

later in development, in a highly restricted subset of differentiated tissues specific to each miRNA (Supplemental Figure S7), unlike the majority of miRNAs identified in mouse (Landgraf et al., 2007). The miRNAs with promoters bound by Polycomb Group proteins in ES cells are significantly enriched ($p<0.005$) among the set of the most tissue-specific mammalian miRNAs (Supplemental Fig. S7 and Landgraf et al., 2007). This suggests a model whereby Polycomb Group proteins repress a set of tissue-specific miRNA genes in ES cells, a subset of which are co-occupied by Oct4, Sox2, Nanog and Tcf3 (Figure 5D).

Discussion

Here we provide new high-resolution, genome-wide maps of core ES cell transcription factors, identify promoter regions for most miRNA genes, and deduce the association of the ES cell transcription factors with these miRNA genes. We also provide quantitative sequence data of short RNAs in ES cells, NPCs and MEFs to examine changes in miRNA transcription. The key transcriptional regulators in ES cells collectively occupied the promoters of many of the miRNAs that were most abundant in ES cells, including those that were down-regulated as ES cells differentiate. In addition, these factors also occupied the promoters of a second, smaller set of miRNAs that were repressed in ES cells and were selectively expressed in specific differentiated cell types. This second group of miRNAs constitutes a subset of the miRNAs that were silenced by the Polycomb group proteins in ES cells, which is also known to silence key lineage-specific, protein-coding developmental regulators. Together these data reveal two key groups of miRNAs that are direct targets of Oct4/Sox2/Nanog/Tcf3, one group of miRNAs that is preferentially expressed in pluripotent cells and a second group that is silenced in ES cells by the Polycomb group proteins, and is poised to contribute to cell fate-decisions during mammalian development.

miRNA contribution to ES cell identity

Several miRNA ploycistrons, which encode the most abundant miRNAs in ES cells and which are silenced during early cellular differentiation (Houbaviy et al., 2003; Suh et al., 2004; Houbaviy et al., 2005), were occupied at their promoters by Oct4, Sox2, Nanog and Tcf3. These

include the mir-290-295 cluster, which contains multiple mature miRNAs that share seed sequences with members of the murine mir-302 cluster, as well as the human mir-371-373 and mir-302 clusters. miRNAs in the 17-92 cluster also share a highly similar seed sequence. miRNAs in this family have been implicated in cell proliferation (O'Donnell et al., 2005; He et al., 2005; Voorhoeve et al., 2006), consistent with the impaired self-renewal phenotype observed in miRNA-deficient ES cells (Kanellopoulou et al., 2005; Murchison et al., 2005; Wang et al., 2007). The zebrafish homologue of this miRNA family, mir-430, contributes to the rapid degradation of maternal transcripts in early zygotic development (Giraldez et al., 2006), and mRNA expression data suggests that this miRNA family also promotes the clearance of transcripts in early mammalian development (Farh et al., 2005).

In addition to promoting the rapid clearance of transcripts as cells transition from one state to another during development, miRNAs also likely contribute to the control of cell identity by fine-tuning the expression of genes. mir-430, the zebrafish homologue of the mammalian mir-302 family, serves to precisely tune the levels of Nodal antagonists Lefty1 and Lefty 2 relative to Nodal, a subtle modulation of protein levels that has pronounced effects on embryonic development (Choi et al., 2007). Recently, a list of ~250 murine ES cell mRNAs that appear to be under the control of miRNAs in the mir-290-295 cluster was reported (Sinkkonen et al., 2008). This study reports that *Lefty1* and *Lefty2* are evolutionarily conserved targets of the mir-290-295 miRNA family. These miRNAs also maintain the expression of *de novo* DNA methyltransferases 3a and 3b (Dnmt3a and Dnmt3b), perhaps by dampening the expression of the transcriptional repressor Rbl2, helping to poise ES cells for efficient methylation of *Oct4* and other pluripotency genes during differentiation.

Knowledge of how the core transcriptional circuitry of ES cells connects to both miRNAs and protein-coding genes, reveals recognizable network motifs downstream of Oct4/Sox2/Nanog/Tcf3, involving both transcriptional and post-transcriptional regulation, that further reveal how this circuitry controls ES cell identity (Figure 6). *Lefty1* and *Lefty2*, both actively expressed in ES cells, are directly occupied at their promoters by Oct4/Sox2/Nanog/Tcf3. Therefore, the core ES cell transcription factors appear to promote the active expression of *Lefty1* and *Lefty2*, but also fine-tune the expression of these important signaling proteins by activating a family miRNAs that target the *Lefty1* and *Lefty2* 3'UTRs. This network motif whereby a regulator exerts both positive and negative effects on its target, termed

"incoherent feed-forward" regulation (Alon, 2007), provides a mechanism to fine-tune the steady-state level or kinetics of a target's activation (Figure 6A). Over a quarter of the proposed targets of the mir-290-295 miRNAs also are likely under the direct transcriptional control of Oct4/Sox2/Nanog/Tcf3 based on our binding maps, suggesting that these miRNAs could participate broadly in tuning the effects of ES cell transcription factors (Figure 6A).

The miRNA expression program directly downstream of Oct4/Sox2/Nanog/Tcf3 could help poise ES cells for rapid and efficient differentiation, consistent with the phenotype of miRNA-deficient cells (Kanellopoulou et al., 2005; Murchison et al., 2005; Wang et al., 2007; Calabrese et al., 2008). Oct4/Sox2/Nanog/Tcf3 also likely contributes to this poising by their occupancy of the Let-7g promoter. Mature Let-7 transcripts are scarce in ES cells, but were among the most abundant miRNAs in both MEFs and NPCs (Figure 3). Primary Let-7g transcript is abundant in ES cells, but its maturation is blocked by Lin28 (Viswanathan et al., 2008 and data not shown). We now report that the promoters of both Let-7g and *Lin28* are occupied by Oct4/Sox2/Nanog/Tcf3, suggesting that the core ES cell transcription factors promote the transcription of both primary Let-7g and *Lin28*, which blocks the maturation of Let-7g. In this way Let-7 and Lin-28 appear to participate in an incoherent feed-forward circuit downstream of Oct4/Sox2/Nanog/Tcf3 to contribute to rapid cellular differentiation (Figure 6B). Notably, ectopic expression of Lin28 in human fibroblasts promotes the induction of pluripotency (Yu et al., 2007), suggesting blocked maturation of pri-Let-7 transcripts plays an important role in the pluripotent state. Additionally, *Dnmt3a and Dnmt3b*, which are indirectly up-regulated by the mir-290-25 miRNAs (Sinkkonen et al., 2008), are also occupied at their promoters by Oct4/Sox2/Nanog/Tcf3, providing examples of "coherent" regulation of important target genes by ES cell transcription factors and the ES cell miRNAs maintained by those transcription factors (Figure 6C).

Multi-layer regulatory circuitry of ES cell identity

The regulatory circuitry we present for miRNAs in ES cells can now be integrated into the model of core regulatory circuitry of pluripotency we have proposed previously (Boyer et al., 2005; Lee et al., 2006; Cole et al., 2008), as illustrated in Figure 7. Our data reveal that Oct4, Sox2, Nanog and Tcf3 occupy the promoters of two key sets of miRNAs, similar to the two sets of protein-coding genes regulated by these factors: one set that is actively expressed in pluripotent ES cells

and another that is silenced in these cells by Polycomb Group proteins and whose later expression might serve to facilitate establishment or maintenance of differentiated cell states.

The expanded circuit diagram presented here integrates transcription factor occupancy of miRNA genes and existing data on miRNA targets into our model of the molecular control of the pluripotent state. These data suggest that miRNAs that are activated in ES cells by Oct4/Sox2/Nanog/Tcf3, serve to modulate the direct effects of these transcription factors, participating in incoherent feed-forward regulation to tune levels of key genes, and modifying the gene expression program to help poise ES cells for efficient differentiation. Core ES cell transcription factors and the miRNAs under their control coordinately contribute transcriptional and post-transcriptional gene regulation to the network that maintains ES cell identity.

Concluding Remarks

The regulatory circuitry controlled by ES cell transcription factors, Oct4, Sox2, Nanog and Tcf3, and the Polycomb Group proteins, which is required for the normal ES cell state, has offered insights into the molecular control of ES cell pluripotency and self-renewal and cellular reprogramming (Jaenisch and Young, 2008). We now provide high-resolution genome-wide location analysis of these factors provided by ChIP-Seq data, and quantitative sequencing of short transcripts in multiple cell types, to connect miRNA genes to the core circuitry of ES cells. This information should prove useful as investigators continue to probe the role of miRNAs in pluripotency, cell-fate decisions, and perhaps regenerative medicine.

Experimental Procedures

A detailed description of all materials and methods used can be found in Supplemental Information.

Cell culture

V6.5 (C57BL/6-129) murine ES cells were grown under typical ES conditions (see Supplemental Information) on irradiated mouse embryonic fibroblasts (MEFs). For location analysis, cells were grown for one passage off of MEFs, on gelatinized tissue-culture plates. To generate neural precursor cells, ES cells were differentiated along the neural lineage using standard protocols (see Supplemental Information). V6.5 ES cells were differentiated into neural progenitor cells (NPCs) through embryoid body formation for 4 days and selection in ITSFn media for 5–7 days, and maintained in FGF2 and EGF2 (R&D Systems) (See Supplemental Information). Mouse embryonic fibroblasts were prepared and cultured from DR-4 strain mice as previously described (See Supplemental Information).

Antibodies and ChIP assays

Detailed descriptions of antibodies, antibody specificity and ChIP methods used in this study are provided in Supplemental Information. Crosslinked cells (~1 x 10^7 per IP) were lysed and sonicated using a Misonix 3000 sonicator to solubilize and shear crosslinked DNA to a 200bp-1000bp fragment size. Batch sonicated whole cell extract was incubated 12-18 hours at 4°C with 100 µl of Dynal Protein G magnetic beads (Dynal) that has been pre-bound to 10 µg of the appropriate antibody. Immunoprecipitates were washed with RIPA buffer and the DNA eluted in 1% SDS at 65°C for 1 hour. Chemical cross-links were reversed for 10 hours to allow isolation of immunoenriched DNA fragments. Immunoprecipitated DNA and control whole cell extract DNA were purified by treatment with RNAse A, proteinase K and two consecutive phenol:chloroform:isoamyl alcohol extractions.

ChIP-Seq

Crosslinked cells (1 x 10^7 per IP) were lysed and sonicated using a Misonix 3000 sonicator to solubilize and shear crosslinked DNA to a 200bp-1000bp fragment size. Batch sonicated whole cell extract was incubated 12-18 hours at 4°C with 100 µl of Dynal Protein G magnetic beads (Dynal) that has been pre-bound to 10 µg of the appropriate antibody. Immunoprecipitates were washed with RIPA buffer and the DNA eluted in 1% SDS at 65°C for 1 hour. Chemical cross-links were reversed for 10 hours to allow isolation of immunoenriched DNA fragments.

Immunoprecipitated DNA and control whole cell extract DNA were purified by treatment with RNAse A, proteinase K and two consecutive phenol:chloroform:isoamyl alcohol extractions.

Purified immunoprecipitated DNA were prepared for sequencing according to a modified version of the Solexa Genomic DNA protocol. Fragmented DNA was end repaired and subjected to 18 cycles of LM-PCR using oligos provided by Illumina. Amplified fragments between 150 and 300bp (representing shear fragments between 50 and 200nt in length and ~100bp of primer sequence) were isolated by agarose gel electrophoresis and purified. High quality samples were confirmed by the appearance of a smooth smear of fragments from 100-1000bp with a peak distribution between 150 and 300bp. 3ng of Linker-ligated DNA was applied to the flow-cell using the Solexa Cluster Station fluidics device. Following bridge amplification the cluster density and morphology were confirmed by microscopic analysis of flow-cells stained with a 1:5000 dilution of SYBR Green I (Invitrogen). Samples were then subjected to 26 bases of sequencing according to Illumina's standard protocols.

Images acquired from the Solexa sequencer were processed through the bundled Solexa image extraction pipeline and aligned to both mouse NCBI build 36 and 37 using ELAND. Only sequences uniquely matching the reference genome without mismatches were used. Mapped reads were extended to 200bp and allocated into 25bp bins. Groups of bins containing statistically significant enrichment for the epigenetic modification were identified by comparison to a Poissonian background model as well as comparison to an empirical distribution of reads obtained from whole cell extract DNA.

Quanititative short RNA sequencing

A method of cloning the 18-30nt transcripts previously described (Lau et al., 2001) was modified to allow for Solexa (Illumina) sequencing (manuscript submitted). Single-stranded cDNA libraries of short transcripts were generated using size selected RNA from mouse embryonic stem cells, mouse neural precursors, and mouse embryonic fibroblasts. RNA extraction was performed using Trizol, followed by RNeasy purification (Qiagen).

5μg of RNA was size selected and gel purified. 3' Adaptor (pTCGTATGCCGTCTTCTGTTG [idT]) was ligated to RNA with T4 RNA ligase and also, separately with RNA Ligase (Rnl2(1-249)k->Q). Ligation products were gel purified and mixed. 5' adaptor (GUUCAGAGUU CUACAGUCCGACGAUC) was ligated with 4 RNA Ligase.

RT-PCR (Superscript II, Invitrogen) was performed with 5' primer (CAAGCAGAAGA CGGCATA). Splicing of overlapping ends PCR (SOEPCR) was performed (Phusion, NEB) with 5' primer and 3' PCR primer (AATGATACGGCGACCACCGACAGGTTCAGAGTTCTAC AGTCCGA), generating cDNA with extended 3' adaptor sequence. PCR product (40 µl) was denatured (85°C, 10 min, formamide loading dye), and the differently sized strands were purified on a 90% formamide, 8% acrylamide gel, yielding single-stranded DNA suitable Solexa sequencing.

The single-stranded DNA samples were resuspended in 10mM Tris (EB buffer)/0.1% Tween and then used as indicated in the standard Solexa sequencing protocol (Illumina). Each library was run on one lane of the Solexa sequencer.

Promoter array design and data extraction

The design of the oligonucleotide-based whole genome array set and data extraction methods are described in Lee et al., 2006. The microarrays used for location analysis in this study were manufactured by Agilent Technologies (http://www.agilent.com).

Acknowledgements

We thank members of the Young, Jaenisch and Bartel laboratories, especially T. Lee, for discussions and critical review of the manuscript. We also thank M. Calabrese and A. Ravi for helpful discussions. We are grateful to S. Gupta and J. Love at The Whitehead Institute Center for Microarray Technology (WICMT) who helped optimize and perform ChIP-Seq, and L.A. Boyer, B. Chevalier, R. Kumar, and T. Lee who were instrumental in performing location analysis in hES cells. We also thank Biology and Research Computing (BaRC), as well as E. Herbolsheimer for computational and technical support and the Whitehead Institute Center for Microarray Technology (WICMT) for assistance with microarray expression analysis. This work was supported in part by NIH grants 5-RO1-HDO45022, 5-R37-CA084198, and 5-RO1-CA087869 to R.J. and by NIH grant HG002668 and a grant from the Whitehead Institute to R.A.Y.

References

Alon, U. (2007). Network motifs: theory and experimental approaches. Nat Rev Genet *8*, 450-461.

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 Res *18*, 46-59.

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

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

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.

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

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.

Chambers, I. (2004). The molecular basis of pluripotency in mouse embryonic stem cells. Cloning Stem Cells *6*, 386-391.

Choi, W. Y., Giraldez, A. J., and Schier, A. F. (2007). Target protectors reveal dampening and balancing of Nodal agonist and antagonist by miR-430. Science *318*, 271-274.

Farh, K. K., Grimson, A., Jan, C., Lewis, B. P., Johnston, W. K., Lim, L. P., Burge, C. B., and Bartel, D. P. (2005). The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. Science *310*, 1817-1821.

Fukao, T., Fukuda, Y., Kiga, K., Sharif, J., Hino, K., Enomoto, Y., Kawamura, A., Nakamura, K., Takeuchi, T., and Tanabe, M. (2007). An evolutionarily conserved mechanism for microRNA-223 expression revealed by microRNA gene profiling. Cell *129*, 617-631.

Giraldez, A. J., Mishima, Y., Rihel, J., Grocock, R. J., Van Dongen, S., Inoue, K., Enright, A. J., and Schier, A. F. (2006). Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. Science *312*, 75-79.

Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R., and Young, R. A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77-88.

He, L., Thomson, J. M., Hemann, M. T., Hernando-Monge, E., Mu, D., Goodson, S., Powers, S., Cordon-Cardo, C., Lowe, S. W., Hannon, G. J., and Hammond, S. M. (2005). A microRNA polycistron as a potential human oncogene. Nature *435*, 828-833.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., *et al.* (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet *39*, 311-318.

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

Houbaviy, H. B., Murray, M. F., and Sharp, P. A. (2003). Embryonic stem cell-specific MicroRNAs. Dev Cell *5*, 351-358.

Jaenisch, R., and Young, R. (2008). Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell *132*, 567-582.

Jiang, J., Chan, Y. S., Loh, Y. H., Cai, J., Tong, G. Q., Lim, C. A., Robson, P., Zhong, S., and Ng, H. H. (2008). A core Klf circuitry regulates self-renewal of embryonic stem cells. Nat Cell Biol *10*, 353-360.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497-1502.

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

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.

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

Krichevsky, A. M., Sonntag, K. C., Isacson, O., and Kosik, K. S. (2006). Specific microRNAs modulate embryonic stem cell-derived neurogenesis. Stem Cells *24*, 857-864.

Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W., and Tuschl, T. (2002). Identification of tissue-specific microRNAs from mouse. Curr Biol *12*, 735-739.

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A. O., Landthaler, M., *et al.* (2007). A mammalian microRNA expression atlas based on small RNA library sequencing. Cell *129*, 1401-1414.

Lee, T. I., Jenner, R. G., Boyer, L. A., Guenther, M. G., Levine, S. S., Kumar, R. M., Chevalier, B., Johnstone, S. E., Cole, M. F., Isono, K., *et al.* (2006). Control of developmental regulators by Polycomb in human embryonic stem cells. Cell *125*, 301-313.

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell *120*, 15-20.

Lim, L. P., Lau, N. C., Garrett-Engele, P., Grimson, A., Schelter, J. M., Castle, J., Bartel, D. P., Linsley, P. S., and Johnson, J. M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature *433*, 769-773.

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.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T. K., Koche, R. P., *et al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-560.

Mineno, J., Okamoto, S., Ando, T., Sato, M., Chono, H., Izu, H., Takayama, M., Asada, K., Mirochnitchenko, O., Inouye, M., and Kato, I. (2006). The expression profile of microRNAs in mouse embryos. Nucleic Acids Res *34*, 1765-1771.

Murchison, E. P., Partridge, J. F., Tam, O. H., Cheloufi, S., and Hannon, G. J. (2005). Characterization of Dicer-deficient murine embryonic stem cells. Proc Natl Acad Sci U S A *102*, 12135-12140.

Niwa, H. (2007). How is pluripotency determined and maintained? Development *134*, 635-646.

O'Donnell, K. A., Wentzel, E. A., Zeller, K. I., Dang, C. V., and Mendell, J. T. (2005). c-Myc-regulated microRNAs modulate E2F1 expression. Nature *435*, 839-843.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., *et al.* (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods *4*, 651-657.

Silva, J., and Smith, A. (2008). Capturing pluripotency. Cell *132*, 532-536.

Sinkkonen, L., Hugenschmidt, T., Berninger, P., Gaidatzis, D., Mohn, F., Artus-Revel, C. G., Zavolan, M., Svoboda, P., and Filipowicz, W. (2008). MicroRNAs control de novo DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells. Nat Struct Mol Biol *15*, 259-267.

Stefani, G., and Slack, F. J. (2008). Small non-coding RNAs in animal development. Nat Rev Mol Cell Biol *9*, 219-230.

Suh, M. R., Lee, Y., Kim, J. Y., Kim, S. K., Moon, S. H., Lee, J. Y., Cha, K. Y., Chung, H. M., Yoon, H. S., Moon, S. Y., *et al.* (2004). Human embryonic stem cells express a unique set of microRNAs. Dev Biol *270*, 488-498.

Viswanathan, S. R., Daley, G. Q., and Gregory, R. I. (2008). Selective Blockade of MicroRNA Processing by Lin-28. Science.

Voorhoeve, P. M., le Sage, C., Schrier, M., Gillis, A. J., Stoop, H., Nagel, R., Liu, Y. P., van Duijse, J., Drost, J., Griekspoor, A., *et al.* (2006). A genetic screen implicates miRNA-372 and miRNA-373 as oncogenes in testicular germ cell tumors. Cell *124*, 1169-1181.

Wang, Y., Medvid, R., Melton, C., Jaenisch, R., and Blelloch, R. (2007). DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. Nat Genet *39*, 380-385.

Wernig, M., Meissner, A., Foreman, R., Brambrink, T., Ku, M., Hochedlinger, K., Bernstein, B. E., and Jaenisch, R. (2007). In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. Nature *448*, 318-324.

Yi, R., O'Carroll, D., Pasolli, H. A., Zhang, Z., Dietrich, F. S., Tarakhovsky, A., and Fuchs, E. (2006). Morphogenesis in skin is governed by discrete sets of differentially expressed microRNAs. Nat Genet *38*, 356-362.

Yi, R., Poy, M. N., Stoffel, M., and Fuchs, E. (2008). A skin microRNA promotes differentiation by repressing 'stemness'. Nature *452*, 225-229.

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.

Zhou, X., Ruan, J., Wang, G., and Zhang, W. (2007). Characterization and identification of microRNA core promoters in four model species. PLoS Comput Biol *3*, e37.

Figure 1

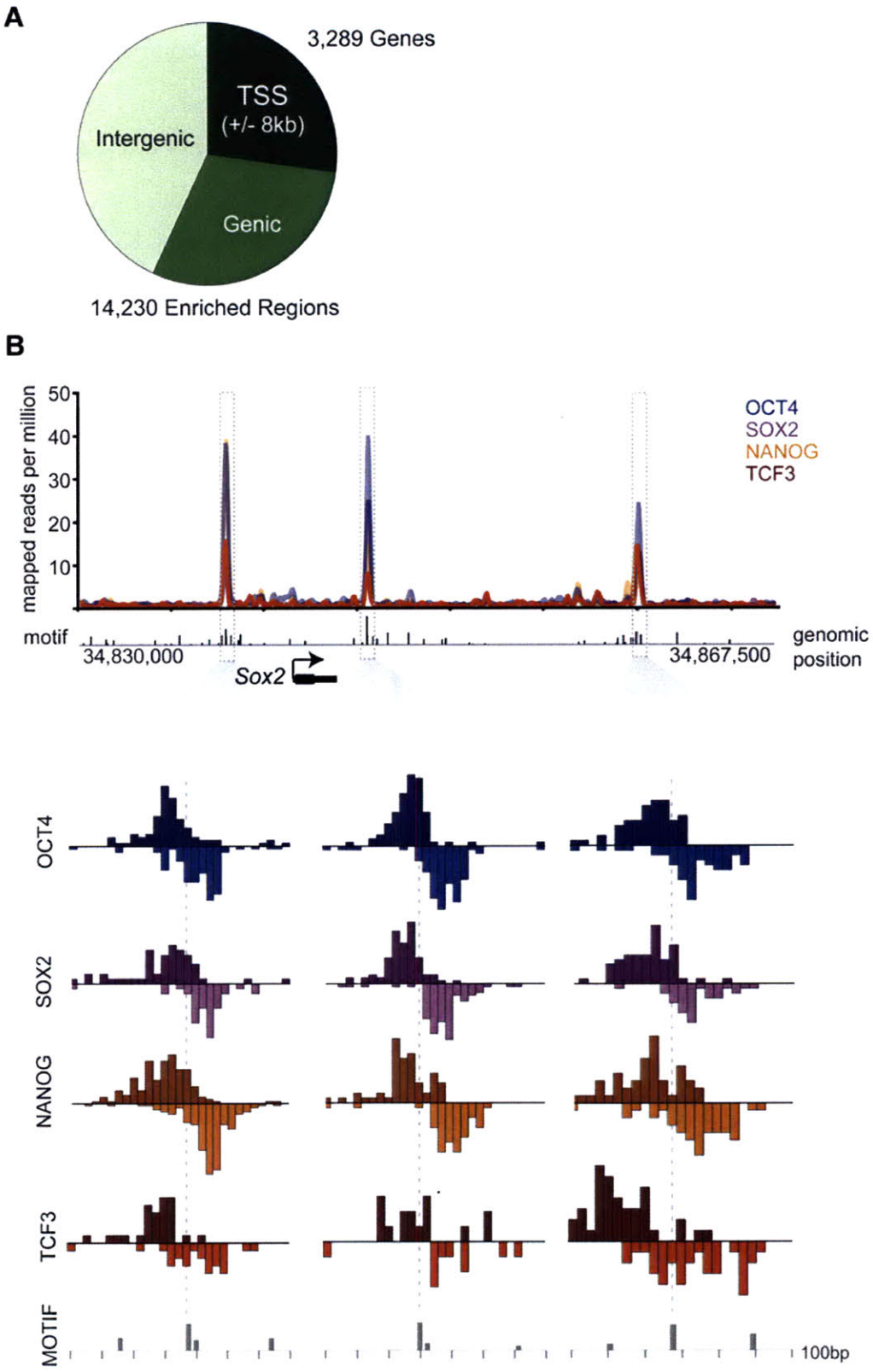

A
3,289 Genes
TSS
(+/- 8kb)
Intergenic
Genic
14,230 Enriched Regions
B
mapped reads per million
50
40
30
20
10
OCT4
SOX2
NANOG
TCF3
motif
34,830,000
34,867,500
genomic
position
Sox2
OCT4
SOX2
NANOG
TCF3
MOTIF
100bp

Figure 1. High-resolution genome-wide mapping of core ES cell transcription factors with ChIP-Seq

A. Summary of binding data for Oct4, Sox2, Nanog and Tcf3. 14,230 sites are co-bound genome wide and mapped to either promoter proximal (TSS +/- 8kb, dark green) (27% of binding sites), genic (>8kb from TSS, middle green) (30% of binding sites), or intergenic (light green) (43% of binding sites). The promoter proximal binding sites are associated with 3,289 genes.

B. (upper) Binding of Oct4 (blue), Sox2 (purple), Nanog (orange) and Tcf3 (red) across 37.5kb of mouse chromosome 3 surrounding the *Sox2* gene (black below the graph, arrow indicates transcription start site). Short sequences uniquely and perfectly mapping to the genome were extended to 200bp (maximum fragment length) and scored in 25bp bins. The score of the bins were then normalized to the total number of reads mapped. Highly enriched regions are highlighted by a dotted box. Oct4/Sox2 DNA binding motifs (Loh et al., 2006) were mapped across the genome and are shown as grey boxes below the graph. Height of the box reflects the quality of the motif. (lower) Detailed analysis of three enriched regions (Chromosome 3: 4,837,600-34,838,300, 34,845,300-34,846,000, and 34,859,900-34,860,500) at the *Sox2* gene indicated with boxes above. The 5' most base from ChIP-Seq were separated by strand and binned into 25bp regions. Sense (darker tone) and anti-sense (light tone) of each of the four factors tested are directed towards the binding site, which in each case occurs at a high-confidence Oct/Sox2 DNA binding motif indicated below.

Figure 2

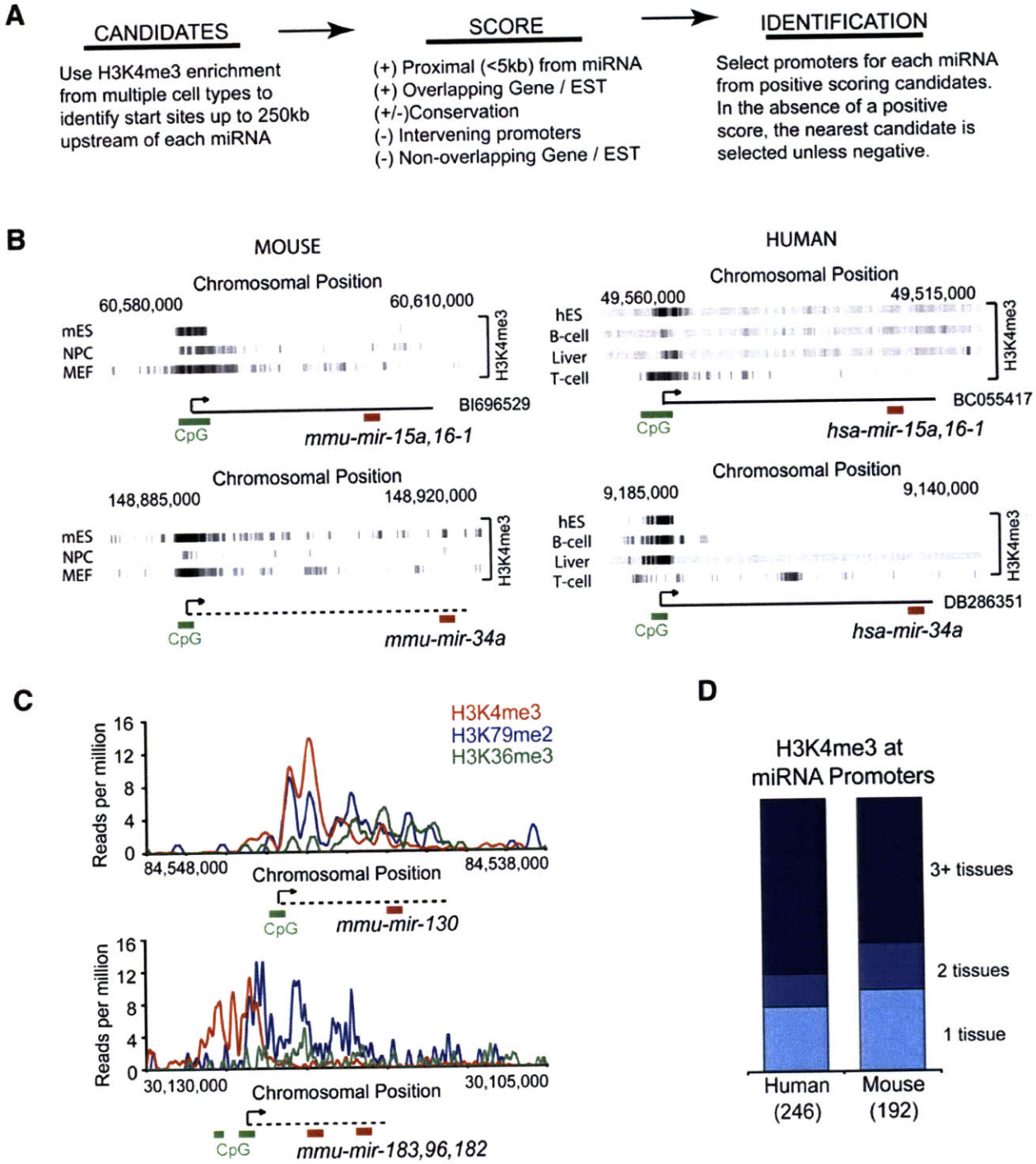

A
CANDIDATES
Use H3K4me3 enrichment from multiple cell types to identify start sites up to 250kb upstream of each miRNA
SCORE
(+) Proximal (<5kb) from miRNA
(+) Overlapping Gene / EST
(+/-)Conservation
(-) Intervening promoters
(-) Non-overlapping Gene / EST
IDENTIFICATION
Select promoters for each miRNA from positive scoring candidates. In the absence of a positive score, the nearest candidate is selected unless negative.
B
MOUSE
Chromosomal Position
60,580,000
60,610,000
mES
NPC
MEF
H3K4me3
BI696529
CpG
mmu-mir-15a,16-1
Chromosomal Position
148,885,000
148,920,000
mES
NPC
MEF
H3K4me3
CpG
mmu-mir-34a
HUMAN
Chromosomal Position
49,560,000
49,515,000
hES
B-cell
Liver
T-cell
H3K4me3
BC055417
CpG
hsa-mir-15a,16-1
Chromosomal Position
9,185,000
9,140,000
hES
B-cell
Liver
T-cell
H3K4me3
DB286351
CpG
hsa-mir-34a
C
H3K4me3
H3K79me2
H3K36me3
Reads per million
16
12
8
4
0
84,548,000
Chromosomal Position
84,538,000
CpG
mmu-mir-130
Reads per million
16
12
8
4
0
30,130,000
Chromosomal Position
30,105,000
CpG
mmu-mir-183,96,182
D
H3K4me3 at miRNA Promoters
3+ tissues
2 tissues
1 tissue
Human (246)
Mouse (192)

Figure 2. Identification of miRNA promoters

A. Description of algorithm for miRNA promoter identification. A library of candidate transcriptional start sites was generated with histone H3 lysine 4 tri-methyl (H3K4me3) location analysis data from multiple tissues (Barski et al., 2007; Guenther et al., 2007; Mikkelsen et al., 2007). Candidates were scored to assess likelihood that they represent true miRNA promoters. Based on scores, a list of mouse and human miRNA promoters was assembled. Additional details can be found in Supplemental Text.

B. Examples of identified miRNA promoter regions are shown. A map of H3K4me3 enrichment is displayed in regions neighbouring selected human and mouse miRNAs for multiple cell types: human ES cells (hES), REH human pro-B cell line (B cell), primary human hepatocytes (Liver), primary human T cells (T cell), mouse ES cells (mES), neural precursor cells (NPCs) and mouse embryonic fibroblasts (MEFs). miRNA promoter coordinates were confirmed by distance to mature miRNA genomic sequence, conservation and EST data (shown as solid line where available). Predicted transcriptional start site and direction of transcription are noted by an arrow, with mature miRNA sequences indicated (red). CpG islands, commonly found at promoters, are indicated (green). Dotted lines denote presumed transcripts.

C. Confirmation of predicted transcription start sites for active miRNAs using chromatin modifications. Normalized ChIP-Seq counts for H3K4me3 (red), H3K79me2 (blue) and H3K36me3 (green) are shown for two miRNA genes where EST data was unavailable. Predicted start site (arrow), CpG islands (green bar), presumed transcript (dotted lines) and miRNA positions (red bar) are shown.

D. Most human and mouse miRNA promoters show evidence of H3K4me3 enrichment in multiple tissues.

Figure 3

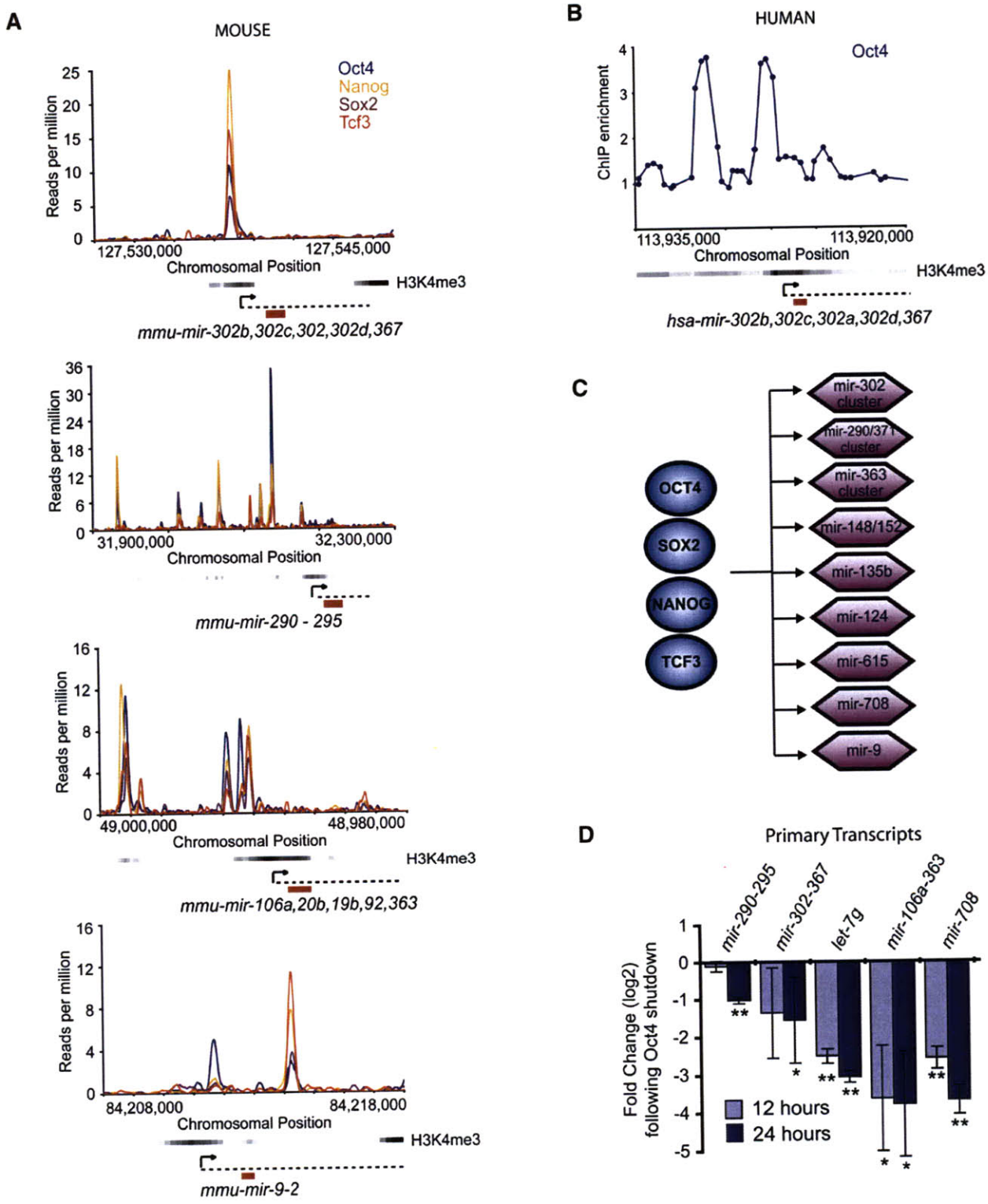

A
MOUSE
Oct4
Nanog
Sox2
Tcf3
Reads per million
Chromosomal Position
127,530,000
127,545,000
H3K4me3
mmu-mir-302b,302c,302,302d,367
31,900,000
32,300,000
mmu-mir-290 - 295
49,000,000
48,980,000
mmu-mir-106a,20b,19b,92,363
84,208,000
84,218,000
mmu-mir-9-2
B
HUMAN
Oct4
ChIP enrichment
113,935,000
113,920,000
hsa-mir-302b,302c,302a,302d,367
C
OCT4
SOX2
NANOG
TCF3
mir-302 cluster
mir-290/371 cluster
mir-363 cluster
mir-148/152
mir-135b
mir-124
mir-615
mir-708
mir-9
D
Primary Transcripts
mir-290-295
mir-302-367
let-7g
mir-106a-363
mir-708
Fold Change (log2) following Oct4 shutdown
12 hours
24 hours

Figure 3. Oct4, Sox2, Nanog and Tcf3 occupancy of miRNA promoters

A. Oct4 (blue), Sox2 (purple), Nanog (orange) and TCF3 (red) binding is shown at four murine miRNA genes as in Figure 1A. H3K4me3 enrichment in ES cells is indicated by shading across genomic region. Presumed transcripts are shown as dotted lines. Coordinates for the mmu-mir-290-295 cluster are derived from NCBI build 37.

B. Oct4 ChIP enrichment ratios (ChIP-enriched versus total genomic DNA) are shown across human miRNA promoter region for the hsa-mir-302 cluster. H3K4me3 enrichment in ES cells is indicated by shading across genomic region.

C. Schematic of miRNAs with conserved binding by the core transcription factors in ES cells. Transcription factors are represented by dark blue circles and miRNAs are represented by purple hexagons. miRNAs from the miR-302 cluster and miR290-295 (mouse)/371-372(human) cluster are selectively expressed in ES cells (Houbaviy et al., 2003).

Figure 4

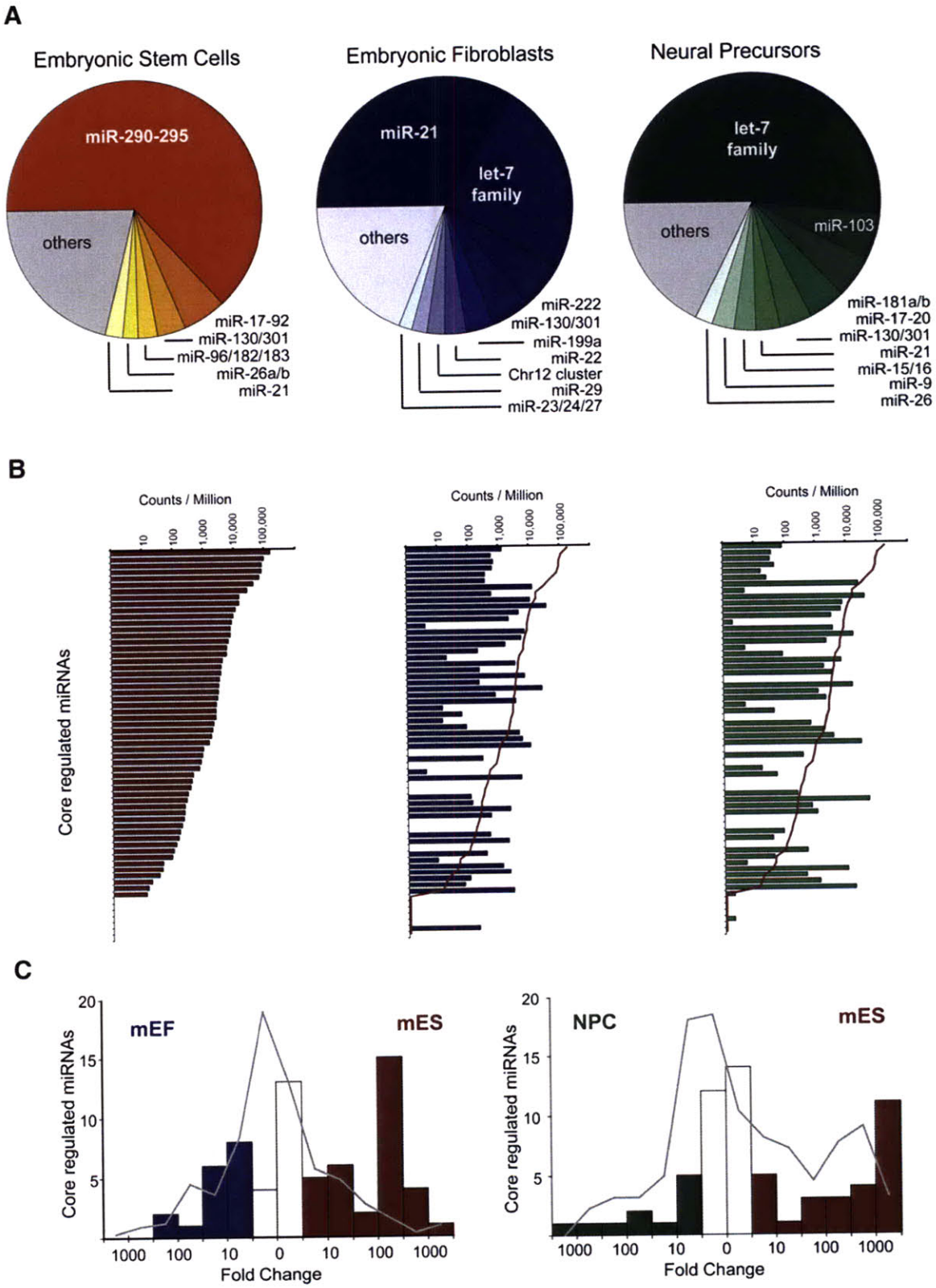

A
Embryonic Stem Cells
miR-290-295
others
miR-17-92
miR-130/301
miR-96/182/183
miR-26a/b
miR-21
Embryonic Fibroblasts
miR-21
let-7 family
others
miR-222
miR-130/301
miR-199a
miR-22
Chr12 cluster
miR-29
miR-23/24/27
Neural Precursors
let-7 family
others
miR-103
miR-181a/b
miR-17-20
miR-130/301
miR-21
miR-15/16
miR-9
miR-26
B
Counts / Million
10
100
1,000
10,000
100,000
Core regulated miRNAs
C
mEF
mES
NPC
mES
Core regulated miRNAs
20
15
10
5
1000
100
10
0
10
100
1000
Fold Change

Figure 4 Regulation of Oct4/Sox2/Nanog/TCF3-bound miRNAs during differentiation

A. Pie charts showing relative contributions of miRNAs to the complete population of miRNAs in mES cells (red) , MEFs (blue) and neural precursors (NPCs, green) based on quantification of miRNAs from by small RNA sequencing. A full list of the miRNAs identified can be found in Supplemental Table S6.

B. Normalized frequency of detection of individual mature miRNAs whose primary transcripts are occupied by Oct4, Sox2, Nanog and Tcf3 in mouse. Red line in center and right panel show the level of detection in ES cells.

C. Histogram of changes in frequency of detection. Changes for miRNAs whose primary transcripts are occupied by Oct4, Sox2, Nanog and Tcf3 in mouse are shown as bars (red for ES enriched, blue for MEF enriched and green for NPC enriched). The background frequency for non-occupied miRNAs is shown as a grey line.

Figure 5

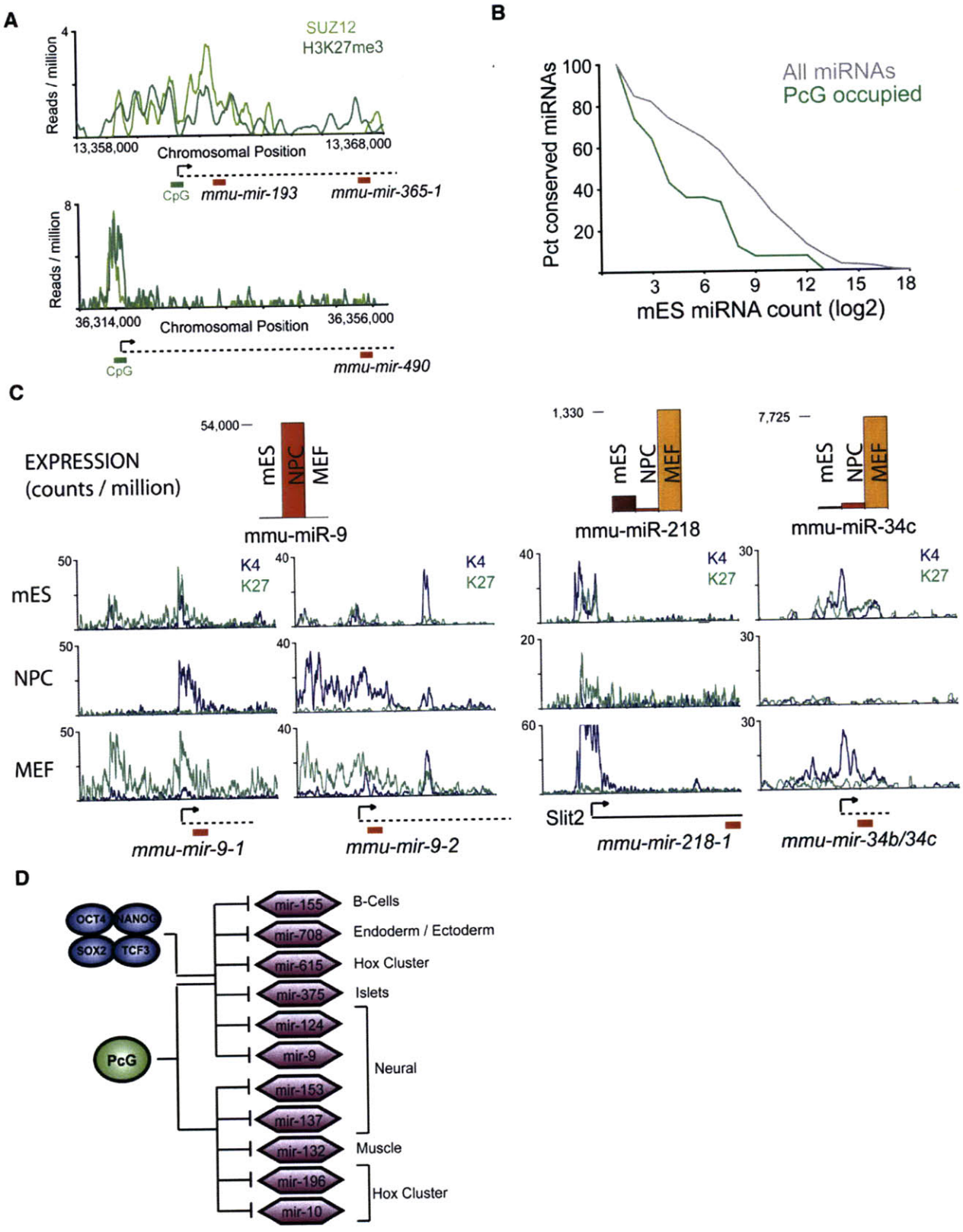

A
SUZ12
H3K27me3
Reads / million
13,358,000
Chromosomal Position
13,368,000
CpG
mmu-mir-193
mmu-mir-365-1
36,314,000
36,356,000
mmu-mir-490
B
All miRNAs
PcG occupied
Pct conserved miRNAs
mES miRNA count (log2)
C
EXPRESSION
(counts / million)
mmu-miR-9
mmu-miR-218
mmu-miR-34c
mES
NPC
MEF
K4
K27
mmu-mir-9-1
mmu-mir-9-2
Slit2
mmu-mir-218-1
mmu-mir-34b/34c
D
OCT4
NANOG
SOX2
TCF3
PcG
mir-155
mir-708
mir-615
mir-375
mir-124
mir-9
mir-153
mir-137
mir-132
mir-196
mir-10
B-Cells
Endoderm / Ectoderm
Hox Cluster
Islets
Neural
Muscle
Hox Cluster

Figure 5. Polycomb represses lineage-specific miRNAs in ES cells

A. Suz12 (light green) and H3K27me3 (dark green, Mikkelsen et al., 2007) binding are shown for two miRNA genes in murine ES cells. Predicted start sites (arrow), CpG islands (green bar), presumed miRNA primary transcript (dotted line) and mature miRNA (red bar) are shown.

B. Expression analysis of miRNAs from mES cells based on quantitative small RNA sequencing. Cumulative distributions for PcG bound miRNAs (green line) and all miRNAs (grey line) are shown.

C. Expression analysis of miRNAs bound by Suz12 in mES cells. Relative counts are shown for mES (red), NPCs (orange) and MEFs (yellow). miR-9 transcript levels were selectively induced in NPCs, while miR-218 and miR-34c were induced in MEFs. H3K27me3 (green line) was lost from the miR-9-1 and the miR-9-2 promoters in NPCs, while the promoters retained H3K4me3 (blue line) (Mikkelsen et al., 2007). H3K27me3 was lost at the miR-218 and miR-34c promoters in MEFs.

D. Schematic of a subset of miRNAs bound by Suz12 in both mES and hES cells. Cells known to selectively express these miRNAs based on computation predictions (Farh et al., 2005) or experimental confirmation (Yi et al., 2006; Yi et al., 2008; Landgraf et al.,2007) are indicated. Transcription factors are represented by dark blue circles, and Suz12 by a green circle. miRNA gene promoters are represented by purple hexagons.

Figure 6

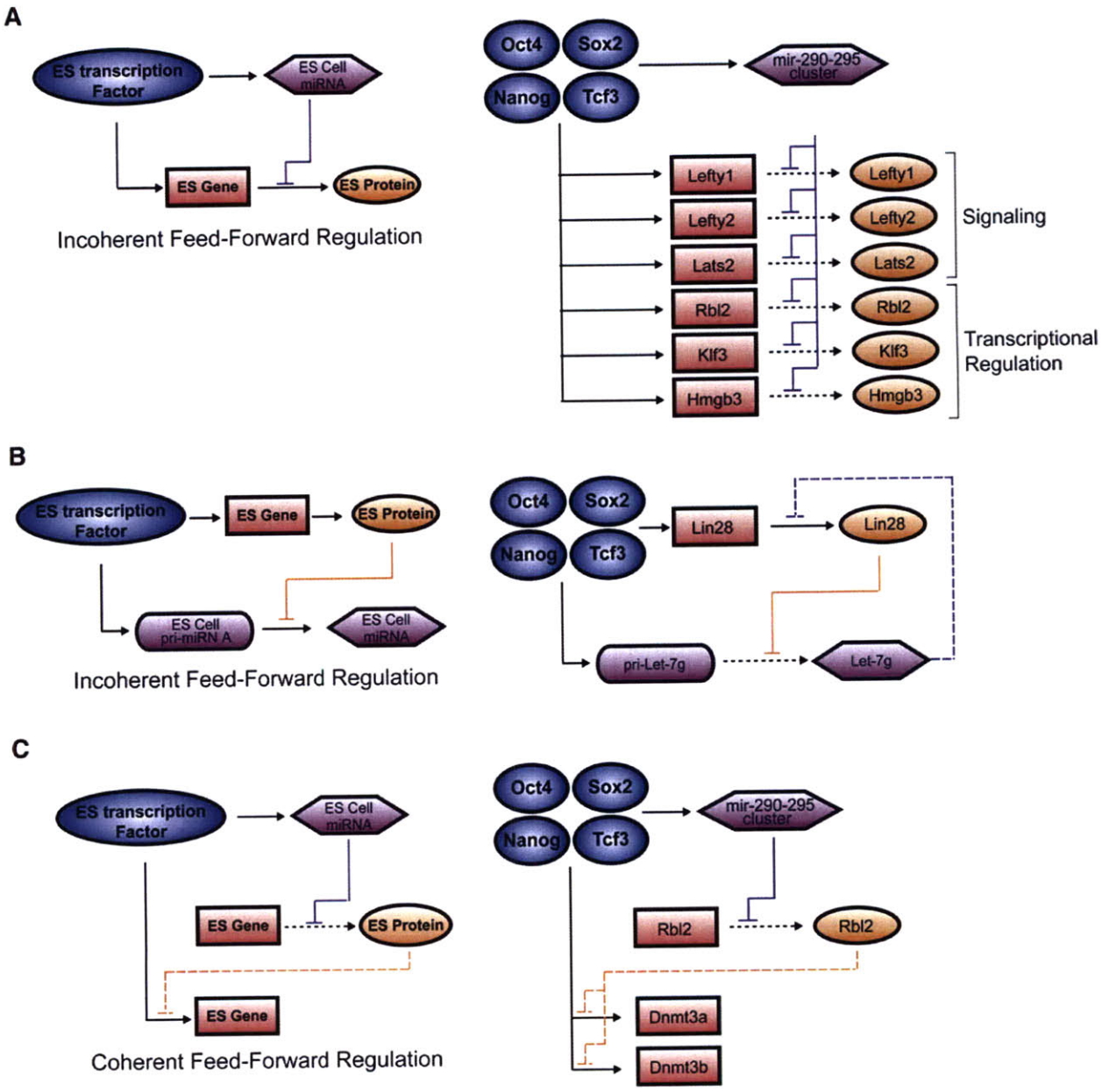

A
ES transcription Factor
ES Cell miRNA
ES Gene
ES Protein
Incoherent Feed-Forward Regulation
Oct4
Sox2
Nanog
Tcf3
mir-290-295 cluster
Lefty1
Lefty2
Lats2
Rbl2
Klf3
Hmgb3
Signaling
Transcriptional Regulation
B
ES transcription Factor
ES Gene
ES Protein
ES Cell pri-miRN A
ES Cell miRNA
Incoherent Feed-Forward Regulation
Oct4
Sox2
Nanog
Tcf3
Lin28
pri-Let-7g
Let-7g
C
ES transcription Factor
ES Cell miRNA
ES Gene
ES Protein
ES Gene
Coherent Feed-Forward Regulation
Oct4
Sox2
Nanog
Tcf3
mir-290-295 cluster
Rbl2
Dnmt3a
Dnmt3b

Figure 6. miRNA modulation of the gene regulatory network in ES cells

A. An incoherent feed-forward motif (Alon 2007) involving a miRNA repression of a transcription factor target gene is illustrated (left). Transcription factors are represented by dark blue circles, miRNAs in purple hexagons, protein-coding gene in pink rectangles and proteins in orange ovals. Particular instances of this netowrks motif identified in ES cells, where signaling molecules or transcriptional regulators directly downstream of Oct4/Sox2/Nanog/Tcf3 are tuned or silenced by miRNAs maintained in ES cells by Oct4/Sox2/Nanog/Tcf3, are illustrated (right).

B. Another incoherent feed-forward motif (Alon 2007) where a protein, encoded by a gene under the control of Oct4/Sox2/Nanog/Tcf3, inhibits the maturation of a primary miRNA transcript maintained in ES cells by Oct4/Sox2/Nanog/Tcf3, is illustrated (left). In ES cells, Lin28 blocks the maturation of primary Let-7g (Visiwanthan et al., 2008). *Lin28* and the Let-7g gene are occupied by Oct4/Sox2/Nanog/Tcf3. Also, noted by the purple dashed line is the Targetscan prediction (Grimson et al., 2007), that mature Let-7g would target Lin28 (right).

C. A coherent feed-forward motif (Alon 2007) where a miRNA represses the expression of transcriptional repressor, which indirectly activates the expression of a gene maintained in ES cells by Oct4/Sox2/Nanog/Tcf3, is illustrated (left). This motif is found in ES cells, where mir-290-295 miRNAs repress Rbl2 indirectly maintaining the expression of *Dnmt3a* and *Dnmt3a*, which are also occupied at their promoters by Oct4/Sox2/Nanog/Tcf3 (right).

Figure 7

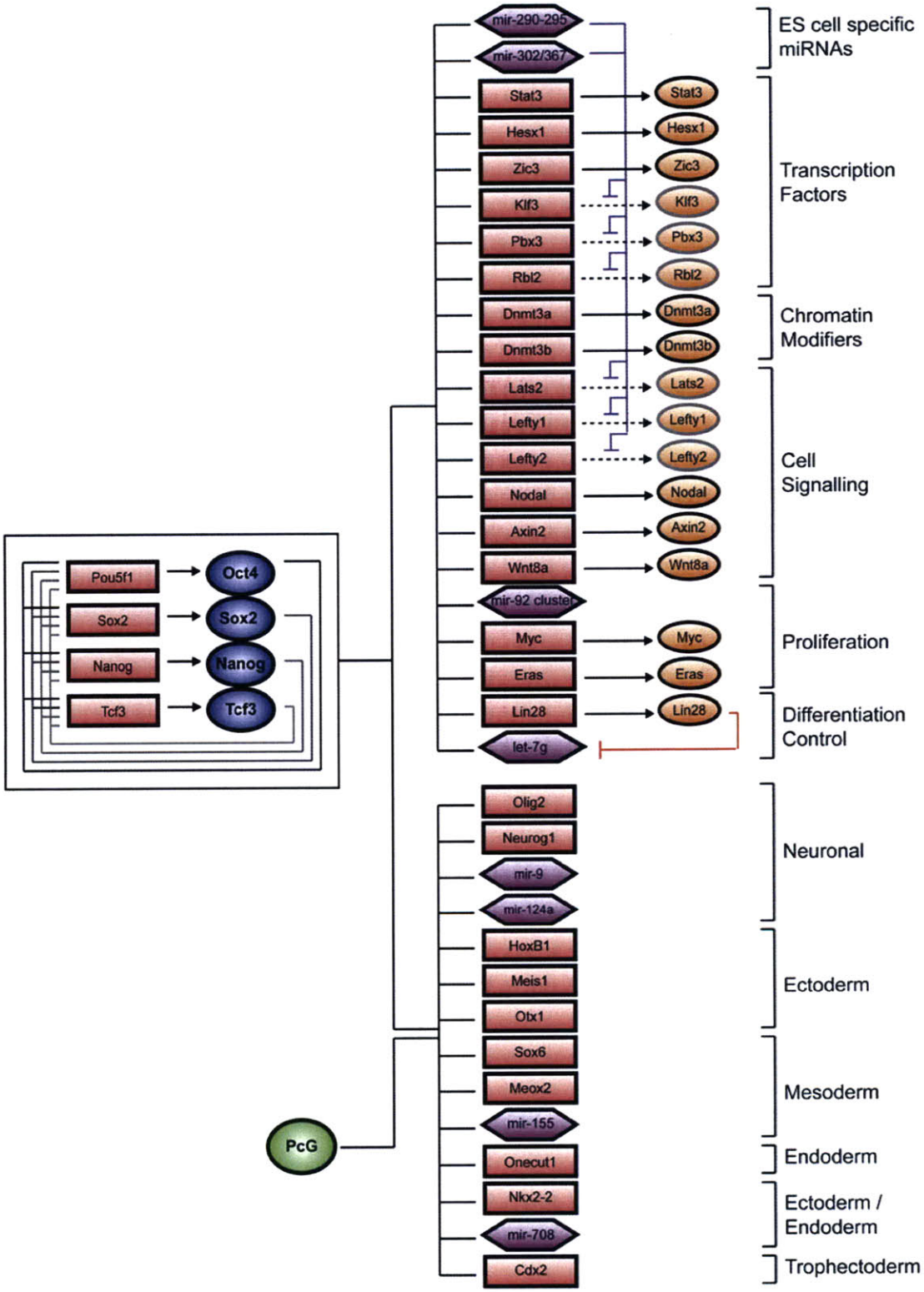

mir-290-295
mir-302/367
Stat3
Hesx1
Zic3
Klf3
Pbx3
Rbl2
Dnmt3a
Dnmt3b
Lats2
Lefty1
Lefty2
Nodal
Axin2
Wnt8a
mir-92 cluster
Myc
Eras
Lin28
let-7g
Pou5f1
Sox2
Nanog
Tcf3
Oct4
Sox2
Nanog
Tcf3
PcG
Olig2
Neurog1
mir-9
mir-124a
HoxB1
Meis1
Otx1
Sox6
Meox2
mir-155
Onecut1
Nkx2-2
mir-708
Cdx2
ES cell specific miRNAs
Transcription Factors
Chromatin Modifiers
Cell Signalling
Proliferation
Differentiation Control
Neuronal
Ectoderm
Mesoderm
Endoderm
Ectoderm / Endoderm
Trophectoderm

Figure 7. Multi-level regulatory network controlling ES cell identity

Updated map of ES cell regulatory circuitry is shown. Interconnected auto-regulatory loop is shown to the left. Active transcripts are shown at the top right, and PcG silenced transcripts are shown at the bottom. Transcription factors are represented by dark blue circles, and Suz12 by a green circle. Gene promoters are represented by red rectangles, gene products by orange circles, and miRNA promoters are represented by purple hexagons.

Chapter 4

Foxp3 Occupancy and Regulation of Key Target Genes During T Cell Stimulation

Published as: Alexander Marson, Karsten Kretschmer, Garrett M. Frampton, Elizabeth S. Jacobsen, Julia K. Polansky, Kenzie D. MacIsaac, Stuart S. Levine, Ernest Fraenkel, Harald von Boehmer, Richard A. Young (2007). Foxp3 occupancy and regulation of key target genes during T cell stimulation. Nature *445,* 931-5.

Abstract

$Foxp3^+CD4^+CD25^+$ regulatory T (T_{reg}) cells are essential for the prevention of autoimmunity (Sakaguchi et al., 1995; Baecher-Allan et al., 2006). T_{reg} cells have an attenuated cytokine response to T cell receptor stimulation, and can suppress the proliferation and effector function of neighbouring T cells (Shevach, 2002; von Boehmer, 2005). The forkhead transcription factor Foxp3 (forkhead box P3) is selectively expressed in T_{reg} cells, is required for T_{reg} development and function, and is sufficient to induce a T_{reg} phenotype in conventional $CD4^+CD25^-$ T cells (Hori et al., 2003; Fontenot et al., 2003; Khattri et al., 2003; Fontenot et al., 2005). Mutations in *Foxp3* cause severe, multi-organ autoimmunity in both human and mouse (Brunkow et al., 2001; Bennett et al., 2001; Wildin et al., 2001). FOXP3 can cooperate in a DNA-binding complex with NFAT (nuclear factor of activated T cells) to regulate the transcription of several known target genes (Wu et al., 2006b). However, the global set of genes regulated directly by Foxp3 is not known and consequently, how this transcription factor controls the gene expression program for T_{reg} function is not understood. Here we identify Foxp3 target genes and report that many of these are key modulators of T cell activation and function. Remarkably, the predominant, although not exclusive, effect of Foxp3 occupancy is to suppress the activation of target genes on T cell stimulation. Foxp3 suppression of its targets appears to be crucial for the normal function of T_{reg} cells, because overactive variants of some target genes are known to be associated with autoimmune disease.

Foxp3 occupancy and regulation of key target genes during T cell stimulation

We developed a strategy to identify genes whose promoters are bound by Foxp3 and whose expression is dependent on that transcription factor (Figure 1). To generate two cell lines that are genetically matched except for Foxp3, we transduced a $Foxp3^{-}CD4^{+}$ murine T cell hybridoma with FLAG-tagged Foxp3. This approach was favoured over comparison of *ex vivo* cells, which are heterogeneous with regard to activation status. The lines provided sufficient numbers of homogeneous cells with appropriate controls to facilitate both location analysis and expression analysis. FACS (fluorescence-activated cell sorting) analysis confirmed that Foxp3 is expressed in the hybridoma at levels comparable to those in *ex vivo* $CD4^{+}CD25^{+}$ T_{reg} cells (Supplemental Figure S1). Previous work has shown that conventional $CD4^{+}$ T cells ectopically expressing Foxp3 do not upregulate interleukin 2 (Il2) secretion following T cell receptor (TCR) dependent stimulation (Schubert et al., 2001). To confirm that the $Foxp3^{+}$ hybridomas contain functional Foxp3, we assayed $Foxp3^{-}$ and $Foxp3^{+}$ cells for Il2 secretion. Indeed, FACS analysis revealed that Il2 secretion is strongly inhibited in phorbol myristate acetate (PMA)/ionomycin stimulated $Foxp3^{+}$ hybridomas compared to stimulated $Foxp3^{-}$ hybridomas (Supplemental Figure S2).

To identify direct targets of Foxp3, DNA sequences occupied by the transcription factor were identified in a replicate set of experiments using chromatin-immunoprecipitation (ChIP) combined with DNA microarrays. For this purpose, DNA microarrays were used that contain 60-mer oligonucleotide probes covering the region from –8 kilobases (kb) to +2 kb relative to the transcript start sites for approximately 16,000 annotated mouse genes (Boyer et al., 2006). The sites occupied by Foxp3 were identified as peaks of ChIP-enriched DNA that span closely neighbouring probes (Figure 2). Foxp3 was found to occupy the promoters of 1,119 genes in PMA/ionomycin stimulated hybridomas (Supplemental Tables S1 and S2). The well-characterized Foxp3 target gene (Wu et al., 2006b; Chen et al., 2006), *Il2*, was among the genes occupied by Foxp3 (Figure 2A). Most of the promoters occupied by Foxp3 in stimulated T cells were also occupied in unstimulated cells (Supplemental Tables S1 and S2, and Supplemental Figure S3). However, at some promoters Foxp3 binding increased considerably in cells stimulated with PMA/ionomycin (Figure 2A and Supplemental Figure S3). Control immunoprecipitation experiments in $Foxp3^{-}$ cells, which produced few positive signals, confirmed the specificity of these results (Supplemental Figure S4). Our confidence in the binding data was further strengthened by the discovery of a DNA sequence motif, which matches

the consensus forkhead motif, at the genomic loci that were bound by Foxp3 (Figure 2B and Supplemental Table S5). This motif distinguishes Foxp3 bound regions from unbound regions tiled on the promoter arrays with a high level of confidence ($p < 10^{-41}$). Instances of this motif are significantly more likely to be conserved in Foxp3 bound regions than in promoter regions that are not bound by Foxp3 ($p < 10^{-23}$), suggesting that these sites serve a functional role (Supplemental Table S5).

To gain insights into the cellular functions that are directly regulated by Foxp3 transcriptional control, we compared the list of genes occupied by Foxp3 in stimulated hybridomas to the biological pathways annotated by the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2000). Among these pathways, Foxp3 target genes are most strongly associated with the TCR signalling pathway ($p = 1.4 \times 10^{-5}$) (Figure 2C). Foxp3 target genes encode proteins that participate at multiple levels of this pathway, including cell surface molecules, signalling components and transcriptional regulators. Many genes with known roles in T cells are missed with existing automated surveys, so we also manually inspected the list of Foxp3 target genes. This revealed that many additional Foxp3 targets are probably important for T cell function, including microRNAs that are differentially expressed between T_{reg} cells and conventional T cells (Figure 2A and Supplemental Table S3 and Cobb et al., 2006). Surprisingly, *Ctla4* was not among the Foxp3 targets, but analysis by RT–PCR (polymerase chain reaction with reverse transcription) revealed no *Ctla4* expression in $Foxp3^-$ and $Foxp3^+$ hybridomas (data not shown). However, the Foxp3 targets included genes previously reported to be upregulated in T_{reg} cells, such as *Il2ra* (*CD25*) (Sakaguchi et al., 1995), *Tnfrsf18* (*GITR*) (McHugh et al, 2002), *Nrp1* (Bruder et al., 2004) and *Ccr4* (Iellem et al., 2001), consistent with predictions that these are directly regulated by Foxp3.

Previous reports have shown that only a portion of transcription factor binding events is associated with transcriptional regulation (Harbison et al., 2004). To identify the set of genes whose expression is dependent on Foxp3, we performed gene expression profiling on $Foxp3^-$ and $Foxp3^+$ T cell hybridoma cells before and after PMA/ionomycin stimulation. Comparison of data from unstimulated $Foxp3^-$ and $Foxp3^+$ cells revealed few differentially expressed genes, suggesting that the transcription factor had little influence on global gene expression in unstimulated hybridomas (Supplemental Table S8). In contrast, PMA/ionomycin-stimulated $Foxp3^-$ and $Foxp3^+$ cells showed significant differences in expression of almost 1% of mouse

genes. Many of these genes were directly occupied by Foxp3, and Foxp3 binding was predominantly, but not exclusively, associated with genes whose expression is downregulated in stimulated Foxp3$^+$ hybridomas. Foxp3 occupied the promoters of approximately half of the genes in this cluster (Figure 3A). This set of downregulated target genes is enriched for genes that are implicated in TCR signalling ($p = 6.1 \times 10^{-3}$). Underscoring the significance of this finding, E2f4, an unrelated control transcription factor that occupies ~800 genes in these hybridomas, was found to occupy only one of these promoters (Supplemental Table S4).

Only a subset of all Foxp3 occupied genes was found to be differentially expressed. One reason for this could be that Foxp3 requires cofactors to modulate transcription. Recently, FOXP3 was shown to cooperate with NFAT in a DNA-binding complex to activate or repress target gene expression (Wu et al., 2006b). Consistent with this report, Foxp3 exerts a more pronounced transcriptional effect in stimulated hybridoma cells than in non-stimulated cells. Importantly, in our experiments, the set of Foxp3 bound genes are enriched ($P < 10^{-19}$) for the presence of an Nfat DNA sequence motif neighbouring the sites of Foxp3 occupancy (Supplemental Table S5).

We next examined whether the genes regulated by Foxp3 in the T cell hybridomas show similar regulatory behaviour in *ex vivo* T cells. Hybridoma cells were used initially because they afforded cells that differ only by Foxp3 status, they provided a homogenous population of non-stimulated T cells, and we could have confidence in genome-wide location data using FLAG-tagged Foxp3. Nonetheless, microarray expression profiling was performed on *ex vivo* T cells from different mice that express a transgenic TCR alone or together with the TCR agonist ligand (Klein et al., 2003), from which relatively pure populations of naïve CD4$^+$CD25$^-$ T helper and Foxp3$^+$CD4$^+$CD25$^+$ T_{reg} cells, respectively, could be isolated. This analysis revealed that many, although not all, of the regulated targets that were identified in the hybridomas show consistent expression patterns in stimulated *ex vivo* cells (Figure 3B and Supplemental Fig. S5). Some differences in gene expression, especially genes that are activated by Foxp3 only in T_{reg} cells, may be due to the fact that *ex vivo* T_{reg} cells are generated through antigenic stimulation (Jordan et al., 2001; Kretschmer et al., 2005) and hence could contain transcriptional cofactors that differ from those in hybridoma cells.

Our findings were further validated for a panel of nine Foxp3 targets. Site-specific primers were used to confirm the binding of Foxp3 to the promoters of these genes (Figure 3C)

and quantitative RT–PCR was used to assay messenger RNA levels in non-stimulated and stimulated hybridomas and *ex vivo* cells, in the presence and absence of 2 μM cyclosporin A (Figure 3D). These experiments confirmed the direct effects of Foxp3 at targets that were identified in the genome scale experiments. Notably, in these experiments (Figure 3D) all genes that were activated following stimulation in $Foxp3^-$ cells and repressed in $Foxp3^+$ cells were activated in a calcineurin dependent manner, consistent with the notion that Nfat is involved in their activation. In addition, just as Foxp3 regulates the protein levels of secreted Il2 (Supplemental Fig. S2), cell surface staining and FACS analysis show that Foxp3 regulates the level of Ly6a protein, demonstrating that Foxp3 transcriptional regulation of its targets modulates protein levels (Supplemental Fig. S6).

Taken together, the results from *ex vivo* T cells and the hybridoma system identify a core set of Foxp3 regulatory targets (Figure 4A), most of which showed suppressed activation in stimulated $Foxp3^+$ cells. A smaller number of Foxp3 target genes was upregulated in stimulated $Foxp3^+$ cells, including some encoding cell surface molecules with known roles in immunoregulation, such as Ly6a (Stanford et al., 1997) and Tnfrsf9 (4-1BB) (Myers et al., 2005). The results from the hybridoma system indicate that Foxp3 occupies regions of most of its target promoters in both unstimulated and stimulated conditions, but increased binding at some promoters and regulation of most targets is observed after stimulation. Furthermore, in hybridomas the major function of Foxp3 at these genes is to suppress the level of gene activation that would occur if this transcription factor were not expressed (Figure 4B). Conceivably, the Foxp3 dependent downregulation of T cell activation and cytokine genes, and upregulation of immunosuppressive cell surface molecules, contribute to both the hyporesponsive and suppressive T_{reg} phenotype.

Mutations in some Foxp3 target genes are already known to be associated with autoimmune disease. The protein tyrosine phosphatase Ptpn22 is a notable example. In our experiments, *Ptpn22* is one of the highest confidence direct targets of Foxp3; it is upregulated on stimulation in $Foxp3^-$ cells, and this upregulation is inhibited in $Foxp3^+$ hybridomas (Figure 3A) and *ex vivo* T_{reg} cells (Figures 3B and 3D). Ptpn22 modulates the signal cascade downstream of the TCR, and mutations in the human *PTPN22* have been associated with type 1 diabetes, rheumatoid arthritis, systemic lupus erythematosus and Graves' disease, as well as other autoimmune diseases (Bottini et al., 2004; Wu et al., 2006a; Bottini et al., 2006). A recent report

suggests that one *PTPN22* single-nucleotide polymorphism associated with autoimmunity is a gain-of-function mutation (Vang et al., 2005). Our findings are compatible with the hypothesis that the gain-of-function mutation might be pathogenic if mutant PTPN22 is overactive in T_{reg} cells (Vang et al., 2005).

In summary, our data indicate that Foxp3 binds to the promoters of well-characterized regulators of T cell activation and function. In the T cell hybridomas studied here, the major role of this transcription factor is to dampen the induction of key genes when T_{reg} cells are stimulated. In *ex vivo* T_{reg} cells, Foxp3 could also activate the expression of a greater number of genes, perhaps owing to the greater abundance of certain transcriptional cofactors in these cells. Some of the identified Foxp3 target genes have been previously implicated in autoimmune diseases, implying that a therapeutic strategy to recapitulate the function of this transcription factor may have clinical utility for these diseases.

Experimental Procedures

A detailed description of all materials and methods used can be found in Supplemental Information.

Growth of murine $CD4^+$ T cell hybridomas and *ex vivo* T cells

$CD4^+$ 5B6-2 hybridoma cells expressing a $PLP_{139-151}$-specific TCR, which was kindly provided by V. Kuchroo, were cultured in Dulbecco's modified Eagle medium (Invitrogen). Primary murine $CD4^+$ T cells were cultured in RPMI-1640 medium (Invitrogen). For gene expression profiling, real-time RT–PCR, and location analysis, cells were cultured in the absence or presence of 50 ng ml^{-1} phorbol 12-myristate 13-acetate (PMA) and 200 ng ml^{-1} ionomycin at 37 °C and harvested after 6 h. Where indicated, cells were preincubated for 1 h with 2μM cyclosporin A. Details of cell generation and isolation are provided in Supplemental Information.

Antibodies and ChIP assays

Detailed descriptions of antibodies, antibody specificity and ChIP methods used in this study are provided in Supplemental Information. All microarray data from this study are available from

ArrayExpress at the EBI (http://www.ebi.ac.uk/arrayexpress) under accession code E-TABM-154.

Promoter array design and data extraction

The design of the oligonucleotide-based promoter array set and data extraction methods are described in Supplemental Information. The microarrays used for location analysis in this study were manufactured by Agilent Technologies (http://www.agilent.com).

Motif analysis

Discovery of the Foxp3 sequence motif from the ChIP-chip binding data was performed using the THEME algorithm. The Foxp3 motif learned by THEME and the Nfat motif from the TRANSFAC database (version 8.3) were used to scan all arrayed sequences to identify matches to the motifs.

Functional classification of bound genes

Comparison of Foxp3 target genes to annotated KEGG biological pathways was performed using the online DAVID tool (http://niaid.abcc.ncifcrf.gov/).

Gene expression profiling

For each hybridoma culture condition, total RNA was prepared from 1×10^7 cells using Trizol (Gibco) followed by additional purification using the RNeasy Mini Kit (Qiagen). Biotinylated antisense cRNA was then prepared according to the Affymetrix standard labelling protocol (one amplification round). For each primary T cell culture condition, total RNA was isolated from 5×10^5 cells with RNeasy. Biotinylated antisense cRNA was prepared by two rounds of *in vitro* amplification using the BioArray RNA Amplification and Labelling System (Enzo Life Sciences) according to the protocol for 10–1,000 ng of input RNA provided by the manufacturer. Biotinylated cRNAs of hybridomas and primary T cells were fragmented and hybridized to Affymetrix GeneChip Mouse Expression Set 430 2.0 arrays at the Microarray Core Facility (Dana-Farber Cancer Institute).

Quantitative RT-PCR

To determine transcript levels in T cell hybridomas and *ex vivo* T cells, RNA was isolated, reverse-transcribed and subjected to real-time PCR performed on an ABI PRISM thermal cycler using SYBR Green PCR core reagents (Applied Biosystems). Detailed information is provided in Supplemental Information.

Acknowledgements

We thank members of the Young, von Boehmer and Fraenkel laboratories, as well as R. Jaenisch and D.K. Gifford, for discussions and critical review of the manuscript, especially T.I. Lee, J. Zeitlinger and D.T. Odom. We also thank Biology and Research Computing (BaRC), especially T. Dicesare for graphic assistance, as well as E. Herbolsheimer for computational and technical support. K.K. was supported in part by the fellowship grant KR2316/1-1 from the German Research Foundation. This work was supported in part by a donation from E. Radutzky, a grant from the Whitaker Foundation to E.F., NIH Grant R37 AI53102 to H.v.B. and NIH grant AI055021 to R.A.Y.

References

Baecher-Allan, C., and Hafler, D. A. (2006). Human regulatory T cells and their role in autoimmune disease. Immunol Rev *212*, 203-216.

Bennett, C. L., Christie, J., Ramsdell, F., Brunkow, M. E., Ferguson, P. J., Whitesell, L., Kelly, T. E., Saulsbury, F. T., Chance, P. F., and Ochs, H. D. (2001). The immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome (IPEX) is caused by mutations of FOXP3. Nat Genet *27*, 20-21.

Bottini, N., Musumeci, L., Alonso, A., Rahmouni, S., Nika, K., Rostamkhani, M., MacMurray, J., Meloni, G. F., Lucarelli, P., Pellecchia, M., *et al.* (2004). A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. Nat Genet *36*, 337-338.

Bottini, N., Vang, T., Cucca, F., and Mustelin, T. (2006). Role of PTPN22 in type 1 diabetes and other autoimmune diseases. Semin Immunol *18*, 207-213.

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.

Bruder, D., Probst-Kepper, M., Westendorf, A. M., Geffers, R., Beissert, S., Loser, K., von Boehmer, H., Buer, J., and Hansen, W. (2004). Neuropilin-1: a surface marker of regulatory T cells. Eur J Immunol *34*, 623-630.

Brunkow, M. E., Jeffery, E. W., Hjerrild, K. A., Paeper, B., Clark, L. B., Yasayko, S. A., Wilkinson, J. E., Galas, D., Ziegler, S. F., and Ramsdell, F. (2001). Disruption of a new forkhead/winged-helix protein, scurfin, results in the fatal lymphoproliferative disorder of the scurfy mouse. Nat Genet *27*, 68-73.

Chen, C., Rowell, E. A., Thomas, R. M., Hancock, W. W., and Wells, A. D. (2006). Transcriptional regulation by Foxp3 is associated with direct promoter occupancy and modulation of histone acetylation. J Biol Chem *281*, 36828-36834.

Cobb, B. S., Hertweck, A., Smith, J., O'Connor, E., Graf, D., Cook, T., Smale, S. T., Sakaguchi, S., Livesey, F. J., Fisher, A. G., and Merkenschlager, M. (2006). A role for Dicer in immune regulation. J Exp Med *203*, 2519-2527.

Fontenot, J. D., Gavin, M. A., and Rudensky, A. Y. (2003). Foxp3 programs the development and function of CD4+CD25+ regulatory T cells. Nat Immunol *4*, 330-336.

Fontenot, J. D., Rasmussen, J. P., Williams, L. M., Dooley, J. L., Farr, A. G., and Rudensky, A. Y. (2005). Regulatory T cell lineage specification by the forkhead transcription factor foxp3. Immunity *22*, 329-341.

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J. B., Reynolds, D. B., Yoo, J., *et al.* (2004). Transcriptional regulatory code of a eukaryotic genome. Nature *431*, 99-104.

Hori, S., Nomura, T., and Sakaguchi, S. (2003). Control of regulatory T cell development by the transcription factor Foxp3. Science *299*, 1057-1061.

Iellem, A., Mariani, M., Lang, R., Recalde, H., Panina-Bordignon, P., Sinigaglia, F., and D'Ambrosio, D. (2001). Unique chemotactic response profile and specific expression of chemokine receptors CCR4 and CCR8 by CD4(+)CD25(+) regulatory T cells. J Exp Med *194*, 847-853.

Jordan, M. S., Boesteanu, A., Reed, A. J., Petrone, A. L., Holenbeck, A. E., Lerman, M. A., Naji, A., and Caton, A. J. (2001). Thymic selection of CD4+CD25+ regulatory T cells induced by an agonist self-peptide. Nat Immunol *2*, 301-306.

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

Khattri, R., Cox, T., Yasayko, S. A., and Ramsdell, F. (2003). An essential role for Scurfin in CD4+CD25+ T regulatory cells. Nat Immunol *4*, 337-342.

Klein, L., Khazaie, K., and von Boehmer, H. (2003). In vivo dynamics of antigen-specific regulatory T cells not predicted from behavior in vitro. Proc Natl Acad Sci U S A *100*, 8886-8891.

Kretschmer, K., Apostolou, I., Hawiger, D., Khazaie, K., Nussenzweig, M. C., and von Boehmer, H. (2005). Inducing and expanding regulatory T cell populations by foreign antigen. Nat Immunol *6*, 1219-1227.

McHugh, R. S., Whitters, M. J., Piccirillo, C. A., Young, D. A., Shevach, E. M., Collins, M., and Byrne, M. C. (2002). CD4(+)CD25(+) immunoregulatory T cells: gene expression analysis reveals a functional role for the glucocorticoid-induced TNF receptor. Immunity *16*, 311-323.

Myers, L. M., and Vella, A. T. (2005). Interfacing T cell effector and regulatory function through CD137 (4-1BB) co-stimulation. Trends Immunol *26*, 440-446.

Sakaguchi, S., Sakaguchi, N., Asano, M., Itoh, M., and Toda, M. (1995). Immunologic self-tolerance maintained by activated T cells expressing IL-2 receptor alpha-chains (CD25). Breakdown of a single mechanism of self-tolerance causes various autoimmune diseases. J Immunol *155*, 1151-1164.

Schubert, L. A., Jeffery, E., Zhang, Y., Ramsdell, F., and Ziegler, S. F. (2001). Scurfin (FOXP3) acts as a repressor of transcription and regulates T cell activation. J Biol Chem *276*, 37672-37679.

Shevach, E. M. (2002). CD4+ CD25+ suppressor T cells: more questions than answers. Nat Rev Immunol *2*, 389-400.

Stanford, W. L., Haque, S., Alexander, R., Liu, X., Latour, A. M., Snodgrass, H. R., Koller, B. H., and Flood, P. M. (1997). Altered proliferative response by T lymphocytes of Ly-6A (Sca-1) null mice. J Exp Med *186*, 705-717.

von Boehmer, H. (2005). Mechanisms of suppression by suppressor T cells. Nat Immunol *6*, 338-344.

Wildin, R. S., Ramsdell, F., Peake, J., Faravelli, F., Casanova, J. L., Buist, N., Levy-Lahad, E., Mazzella, M., Goulet, O., Perroni, L., *et al.* (2001). X-linked neonatal diabetes mellitus, enteropathy and endocrinopathy syndrome is the human equivalent of mouse scurfy. Nat Genet *27*, 18-20.

Wu, J., Katrekar, A., Honigberg, L. A., Smith, A. M., Conn, M. T., Tang, J., Jeffery, D., Mortara, K., Sampang, J., Williams, S. R., *et al.* (2006a). Identification of substrates of human protein-tyrosine phosphatase PTPN22. J Biol Chem *281*, 11002-11010.

Wu, Y., Borde, M., Heissmeyer, V., Feuerer, M., Lapan, A. D., Stroud, J. C., Bates, D. L., Guo, L., Han, A., Ziegler, S. F., *et al.* (2006b). FOXP3 controls regulatory T cell function through cooperation with NFAT. Cell *126*, 375-387.

Figure 1

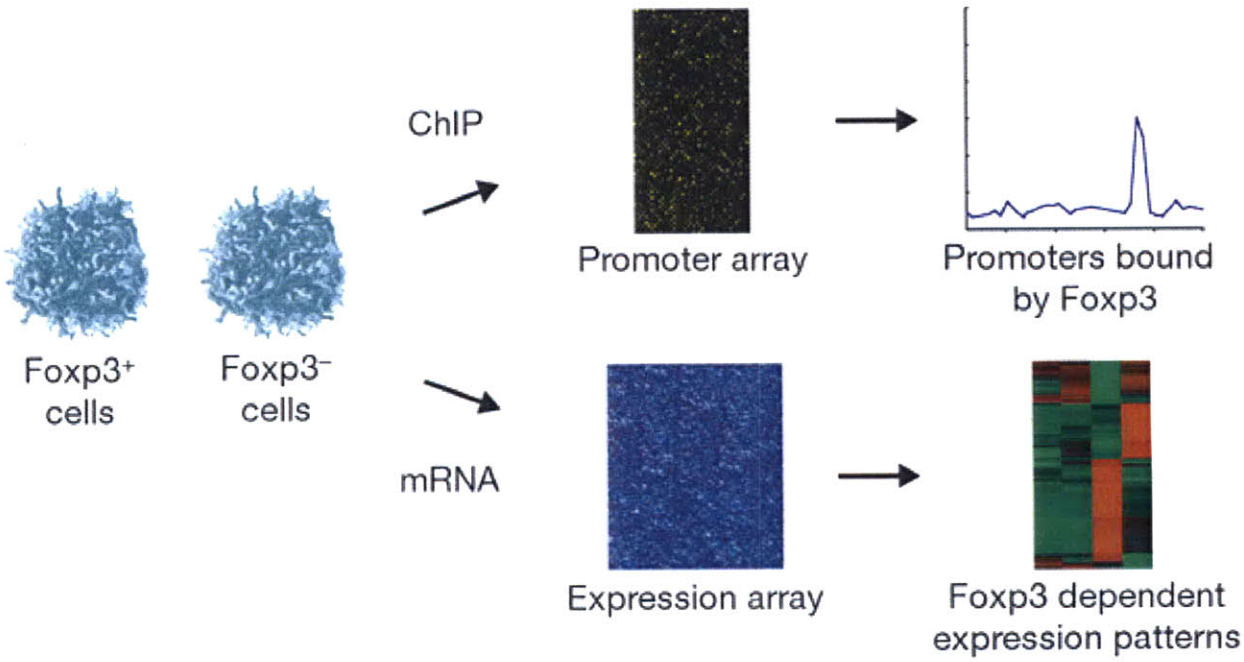


Figure 1. Strategy to identify direct Foxp3 transcriptional effects

Genetically matched $Foxp3^+$ and $Foxp3^-$ cell populations were generated by transduction of FLAG-tagged Foxp3 into a $Foxp3^-$ murine T cell hybridoma. Foxp3 binding sites at promoters across the genome were identified by ChIP experiments with an anti-FLAG antibody. Foxp3 dependent transcriptional regulation was identified by gene expression profiling performed on each of these cell types.

Figure 2

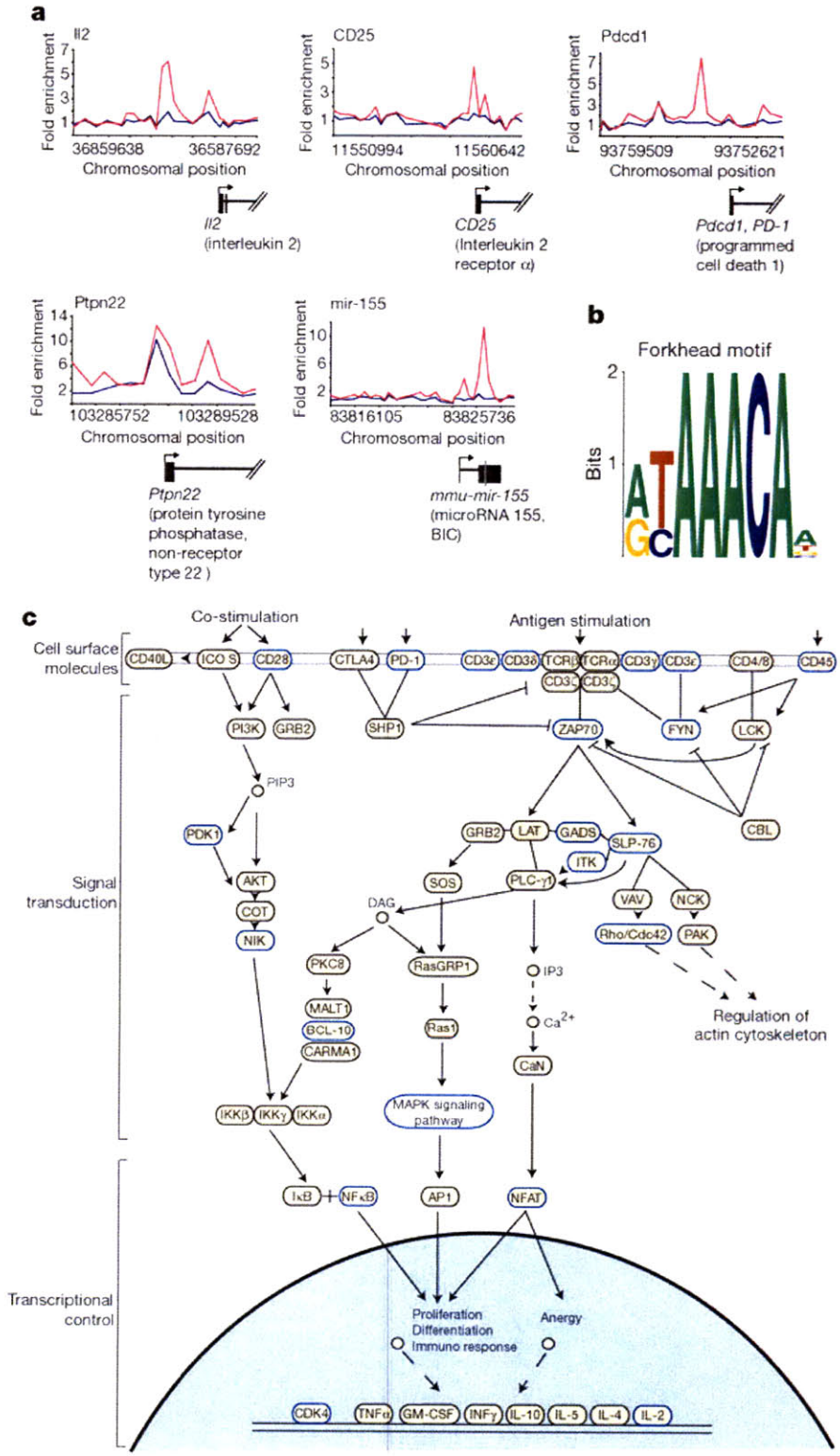

a
Il2
Fold enrichment
Chromosomal position
36859638
36587692
Il2
(interleukin 2)
CD25
11550994
11560642
CD25
(Interleukin 2 receptor α)
Pdcd1
93759509
93752621
Pdcd1, PD-1
(programmed cell death 1)
Ptpn22
103285752
103289528
Ptpn22
(protein tyrosine phosphatase, non-receptor type 22)
mir-155
83816105
83825736
mmu-mir-155
(microRNA 155, BIC)
b
Forkhead motif
Bits
c
Co-stimulation
Antigen stimulation
Cell surface molecules
CD40L
ICOS
CD28
CTLA4
PD-1
CD3ε
CD3δ
TCRβ
TCRα
CD3γ
CD3ζ
CD4/8
CD45
CD3ζ
CD3ζ
PI3K
GRB2
SHP1
ZAP70
FYN
LCK
PIP3
PDK1
GRB2
LAT
GADS
SLP-76
CBL
ITK
AKT
SOS
PLC-γ1
VAV
NCK
Signal transduction
COT
DAG
Rho/Cdc42
PAK
NIK
PKCθ
RasGRP1
IP3
MALT1
BCL-10
CARMA1
Ras1
Ca2+
Regulation of actin cytoskeleton
CaN
IKKβ
IKKγ
IKKα
MAPK signaling pathway
IκB
NFκB
AP1
NFAT
Transcriptional control
Proliferation
Differentiation
Immuno response
Anergy
CDK4
TNFα
GM-CSF
INFγ
IL-10
IL-5
IL-4
IL-2

Figure 2. Direct Foxp3 targets include key modulators of T cell function

A. Foxp3 ChIP enrichment ratios (ChIP-enriched versus total genomic DNA) across indicated promoters are shown for stimulated (pink) and unstimulated (blue) cells. Exons (blocks) and introns (lines) of genes and the mir-155 precursor (grey) are drawn to scale below the plots, with direction of transcription noted by an arrow.

B. Foxp3 bound genomic regions are enriched for the presence of a forkhead DNA motif, represented here in WebLogo (http://weblogo.berkeley.edu).

C. The KEGG (Kanehisa et al, 2000) TCR signalling pathway, enriched ($p = 1.4 \times 10^{-5}$) for proteins encoded by direct targets of Foxp3 (blue outline), is displayed.

Figure 3

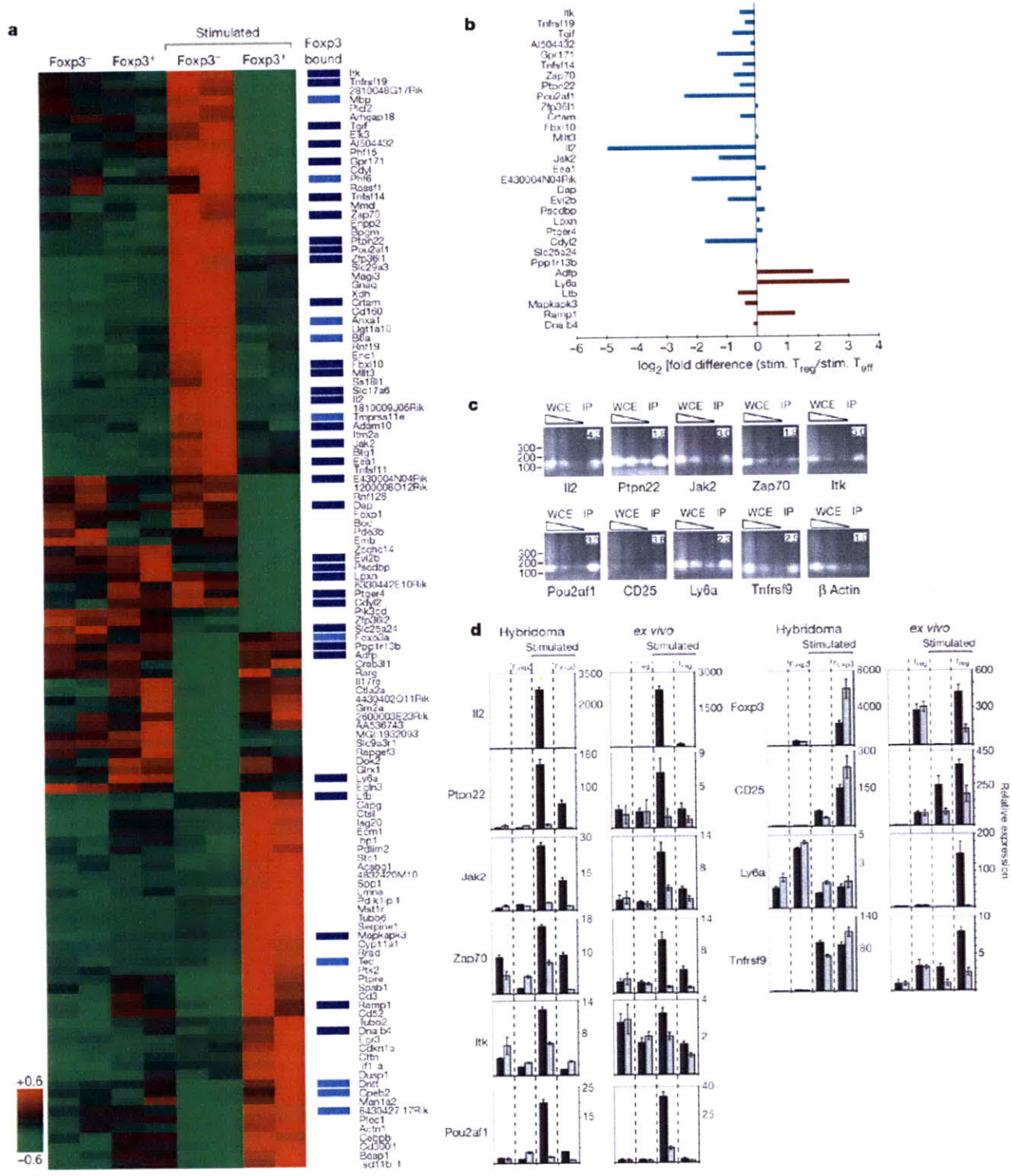

a
Stimulated
Foxp3⁻
Foxp3⁺
Foxp3⁻
Foxp3⁺
Foxp3 bound
+0.6
−0.6
b
log2 [fold difference (stim. Treg/stim. Teff
c
WCE IP
Il2
Ptpn22
Jak2
Zap70
Itk
Pou2af1
CD25
Ly6a
Tnfrsf9
β Actin
d
Hybridoma
ex vivo
Stimulated
Foxp3
Il2
Ptpn22
CD25
Jak2
Ly6a
Zap70
Tnfrsf9
Itk
Pou2af1
Relative expression

Figure 3. Foxp3 directly suppresses the activation of target genes

A. Replicate expression data for the 125 genes with Foxp3 dependent differential expression in stimulated hybridomas (False Disovery Rate [FDR] < 0.05) - were hierarchically clustered and displayed. - The Z-score normalized induction (red) or repression (green) is shown for each gene. Direct targets of Foxp3 in stimulated hybridomas are indicated (dark blue for FDR < 0.05, light blue for FDR < 0.10).

B. For repressed (green) and induced (red) Foxp3-bound targets in **(A)** $\log_2$(fold difference) in expression between stimulated *ex vivo* effector (T_{eff}) T cells and T_{reg} cells is displayed. *Slc17a6* and *Adam10* are not expressed in the *ex vivo* samples.

C. Site-specific PCR on 10 ng of ChIP DNA confirmed selected targets. Immunoprecipitated (IP) DNA was compared to serial dilutions (90, 30 and 10 ng of DNA) of unenriched whole cell extract (WCE) DNA. Enrichment ratios, shown at top right of each sub-panel, are normalized relative to the unenriched beta-actin control. DNA fragment size (bp) is indicated on the left of each row.

D. The transcript levels of the panel of Foxp3 targets presented in **c** and of *Foxp3* were analysed by real-time RT–PCR in stimulated and unstimulated cells, with (grey) and without (black) cyclosporin A. Mean values ± s.d. of relative expression, determined in triplicate, are shown for indicated genes.

Figure 4

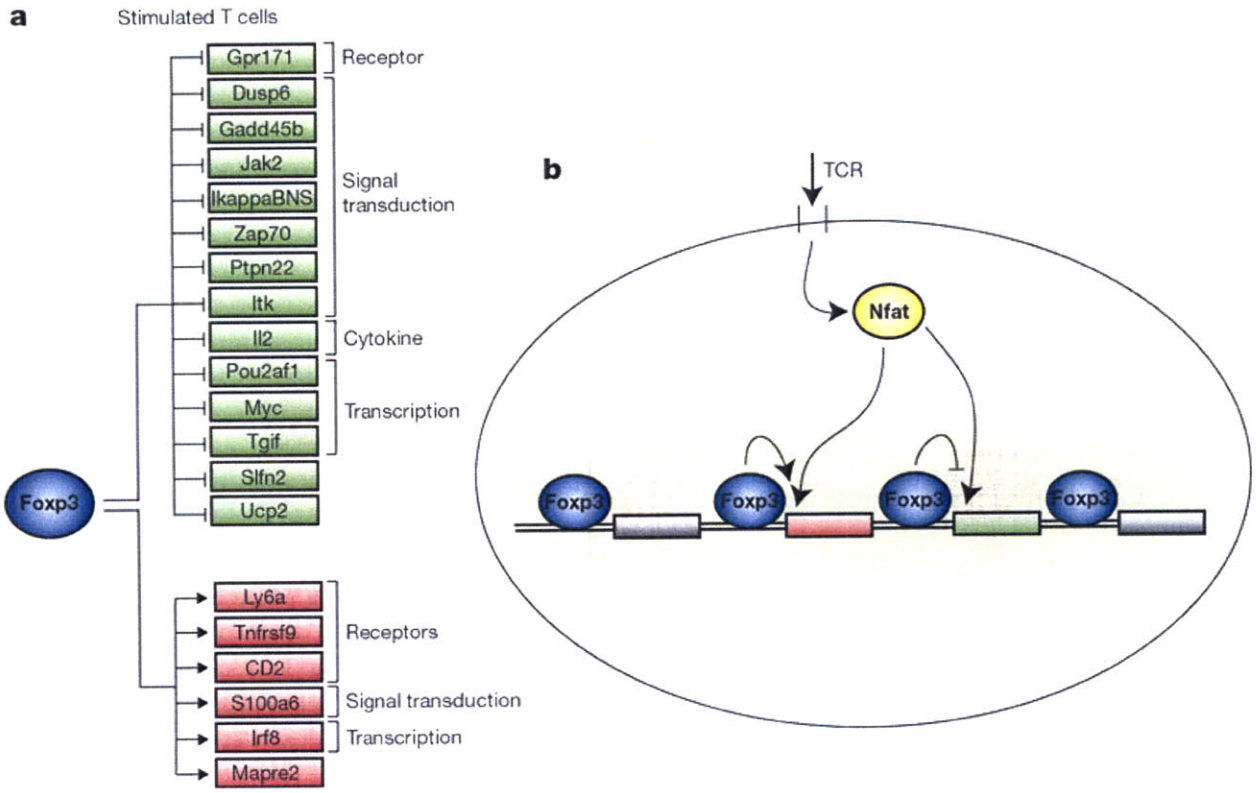


Figure 4. Core direct regulatory effects of Foxp3

A. Shown here are a subset of direct Foxp3 targets that exhibit consistent transcriptional behaviour in hybridomas and in *ex vivo* T cells (Supplemental Fig. S5).

B. Foxp3 binds to a large set of promoters both in unstimulated and stimulated T cells, but Foxp3 transcriptional regulation is more extensive in stimulated T cells. The genomic regions where Foxp3 binds are enriched for an Nfat binding site DNA motif. In the hybridomas, Foxp3 predominantly acts to directly suppress the activation of its target genes.

Chapter 5

Aberrant Chromatin at Genes Encoding Stem Cell Regulators in Human Mixed-Lineage Leukemia

Published as: Matthew G. Guenther, Lee N. Lawton, Tatiana Rozovskaia, Garrett M. Frampton, Stuart S. Levine, Thomas L. Volkert, Carlo M. Croce, Tatsuya Nakamura, Eli Canaani, and Richard A. Young (2008). Aberrant chromatin at genes encoding stem cell regulators in human mixed-lineage leukemia. Genes Dev. *22*, 3403-8.

Abstract

Mixed-lineage leukemia (MLL) fusion proteins are potent inducers of leukemia, but how these proteins generate aberrant gene expression programs is poorly understood. Here we show that the MLL-AF4 fusion protein occupies developmental regulatory genes important for hematopoietic stem cell identity and self-renewal in human leukemia cells. These MLL-AF4-bound regions have grossly altered chromatin structure, with histone modifications catalyzed by trithorax group proteins and DOT1 extending across large domains. Our results define direct targets of the MLL fusion protein, reveal the global role of epigenetic misregulation in leukemia, and identify new targets for therapeutic intervention in cancer.

Introduction

Chromosomal translocations involving the mixed-lineage leukemia gene (*MLL*) are a frequent occurrence in human acute leukemias of both children and adults (Eguchi et al. 2005). In over half of all infant acute leukemias, the MLL protein fuses to one of >50 identified partner genes, resulting in a MLL fusion protein that acts as a potent oncogene (Krivtsov and Armstrong 2007). While extensive gene expression signatures have been determined for primary human leukemia samples (Armstrong et al. 2002; Yeoh et al. 2002; Ferrando et al. 2003; Ross et al. 2003; Rozovskaia et al. 2003; Haferlach et al. 2005), the direct genomic targets of MLL fusion proteins remain unknown. This information is essential to determine how MLL fusion proteins impose oncogenic transcriptional programs and to identify targets for therapeutic intervention in human disease.

Distinct chromatin-modifying complexes and histone modifications are associated with distinct phases of transcription (Li et al. 2007). The trithorax group proteins, including MLL, catalyze histone H3-Lys-4 trimethyl (H3K4me3) modifications at the start sites of transcriptionally engaged genes (Ruthenburg et al. 2007). These H3K4me3-modified regions are largely constrained to the transcription start site regions of genes that are transcriptionally initiated, but not necessarily fully transcribed (Bernstein et al. 2006; Barski et al. 2007; Guenther et al. 2007). As a gene becomes fully transcribed, elongating RNA Polymerase II (Pol II) molecules proceed through gene coding regions along with associated elongation factors including DOT1, which catalyzes dimethylation of histone H3-Lys-79 (H3K79me2) (Li et al. 2007). Physical interactions between the most common MLL partner proteins and transcriptional elongation components suggest that defects in H3K4 and H3K79 methylation might be a key factor in MLL leukemogenesis (Erfurth et al. 2004; Milne et al. 2005b; Okada et al. 2005; Zeisig et al. 2005; Bitoun et al. 2007; Mueller et al. 2007), but the mechanism and extent of H3K79 methylation targeting throughout the genome is poorly understood in human cancer cells.

In order to define the portion of gene regulatory circuitry that is controlled directly by MLL fusion proteins in human leukemia, we determined the binding patterns of an MLL fusion protein and chromatin modifications across the entire human genome. We performed this mapping in leukemic cells harboring the MLL-AF4 fusion gene, because this rearrangement is the most common among MLL fusions and is associated with an extremely poor prognosis in

infants and adults (Eguchi et al. 2005). Our results reveal that the MLL-AF4 oncogene produces gross defects in chromatin structure at a newly defined set of hematopoietic stem cell genes.

Results and Discussion

Identification of MLL-AF4-occupied regions of the genome in human leukemia cells

We used chromatin immunoprecipitation (ChIP) coupled to massively parallel sequencing (ChIP-Seq) to determine how the MLL-AF4 fusion protein was distributed across the entire genome in human leukemia cells. This was performed using two acute lymphoblastic leukemia (ALL) cell lines (Figure 1). The SEM cell line, which was derived from precursor B-cell ALL patient blast cells (Greil et al. 1994) harbors a t(4;11) chromosomal translocation and expresses MLL-AF4 fusion protein, endogenous MLL and endogenous AF4. The REH precursor B-cell ALL patient-derived cell line (Rosenfeld et al. 1977), which expresses only wild-type MLL and wild-type AF4, served as a control to identify regions bound by normal MLL and AF4, but not by the MLL-AF4 fusion protein. Because REH cells were derived from patients with B-cell ALL, these cells also served to control for general effects of B-cell-type leukemia.

We mapped protein–DNA interactions for the MLL-AF4 protein in SEM cells using antibodies to the C terminus of AF4 (anti-$AF4^C$) and the N terminus of MLL (anti-MLL^N) (Figure 1A). Since REH cells do not express MLL-AF4 fusion protein, the sites of MLL and AF4 occupancy that were highly similar in SEM and REH cells are most likely due to binding of normal MLL or AF4 proteins. We expected large numbers of regions in both SEM and REH to be occupied by either AF4 or MLL, since these proteins are widely involved in activated transcription by Pol II (Guenther et al. 2005;Milne et al. 2005a; Bitoun et al. 2007). Indeed, we found that MLL or AF4 localized near the transcription start sites of several thousand genes in both SEM and REH cells (Supplemental Table S1). These bound regions tended to be relatively small, extended 1–3 kb into the 5′ end of the transcribed portion of the gene (Supplemental Figure S1), and were associated with transcriptionally active genes (>80% of transcripts called present by array-based methods in each cell type) (Supplemental Table S2). These data identified the set of genes occupied by MLL and/or AF4 in precursor B ALL cells and confirmed that these genes tend to be transcriptionally active.

We next examined the genome for evidence of binding events specific to MLL fusion protein. To identify MLL-AF4 fusion protein targets, we applied an algorithm that identifies coincident ChIP-Seq signals for MLL^N and $AF4^C$ (Supplemental Material). We identified 226 regions of MLL-AF4 co-occupancy across the genome in SEM cells. This co-occupancy pattern did not occur in REH cells. Many of the MLL-AF4 target regions showed a striking behavior; the signal spanned regions of 5-100 kb (Figure 1B). This binding pattern of MLL-AF4 occurred at genes encoding a variety of important developmental regulators (Figure 2). For example, it occurred at the *PROM1* (prominin-1/CD133) gene (Figure 1B), which encodes a defining antigen of hematopoietic stem cells and is purported to play a role in asymmetric cell divisions in adult stem cells (Wagner et al. 2004; Toren et al. 2005), and *HOXA7, HOXA9, HOXA10* (Figure 2A), whose overexpression are hallmarks of ALLs carrying the t(4;11) chromosome translocation (Rozovskaia et al. 2001, 2003; Armstrong et al. 2002;Yeoh et al. 2002; Ferrando et al. 2003). Overall, MLL-AF4 occupied large domains of 169 known RefSeq genes that were previously unknown as direct targets (Figure 2B,C; complete gene list in Supplemental Table S2).

MLL-AF4 target genes encode hematopoietic stem cell developmental regulators

MLL-AF4 target genes were analyzed by Gene Set Enrichment Analysis (GSEA, htttp://www.broad.mit.edu/gsea) (Subramanian et al. 2005) to determine whether any known phenotypes were associated with the gene set. The most highly enriched subsets were genes overexpressed in leukemia cells and genes that encode transcription factors involved in hematopoiesis (Figure 2D). Genes from the transcription factor group included *HOXA9, RUNX1* and *ETV6,* all of which displayed large areas of MLL-AF4 binding that extended well into the coding regions of these important proleukemia genes. We also discovered other developmental regulators not previously associated with ALL, including the *TWIST1* gene. The *TWIST1* transcription factor is an essential mediator of metastatic growth in human breast cancer cells and also plays a role in nephroblastomas, neuroblastomas, and gastric cancer progression (Pajer et al. 2003; Yang et al. 2004). The TWIST1 protein has also been shown to inhibit apoptosis (Puisieux et al. 2006). Other targets, including the *RUNX2* transcription factor and *JMJD1C* histone demethylase, were previously shown to be up-regulated in MLL-AF9-induced leukemic stem cells (Krivtsov et al. 2006), but were not known to be direct targets of MLL fusion proteins.

Interestingly, *UTX* is a chromatin modifier responsible for activating *HOX* loci (Cloos et al. 2008), suggesting a possible reinforcement of *HOX*A locus overexpression in leukemia. Another category of genes observed in the GSEA analysis was a set of genes differentially expressed in hematopoietic stem cells (Figure 2D; Ramalho-Santos et al. 2002). These included *PROM1/CD133*, which encodes a surface antigen that is a defining marker of hematopoietic stem and progenitor cells and is highly expressed in tumor-initiating cells of the colon and brain (Toren et al. 2005; O'Brien et al. 2007), and the *FLT3* signaling mediator. The enrichment for developmental regulatory factors indicates that the MLL-AF4 oncogene activates specialized transcriptional programs in cancer cells.

MLL-AF4 targets predict leukemia subclass in human patients

We next tested whether MLL-AF4 target genes identified in SEM cells had altered gene expression patterns in human patients with MLL-associated leukemia. If MLL-AF4 target genes specify the MLL-associated leukemia subclass in human cancer, we would expect up-regulation of these genes in patients exhibiting MLL-associated acute leukemia, but not in patients exhibiting non-MLL-linked leukemia. RNA transcript levels from leukemic blasts of 132 pediatric ALL patients of B and T lineages (Ross et al. 2003) were compared for expression of the MLL-AF4 target genes discovered in SEM cells (Figure 3; Supplemental Figure S4). Significantly, about two-thirds of MLL-AF4 targets in SEM cells were at least 50% overexpressed in patients with MLL-associated leukemia, but not in non-MLL-associated leukemia. This overexpression of MLL-AF4 targets was evident not only in MLL-AF4-derived patients' samples, but also in MLL-AF9 and other MLL-derived leukemias (Ross et al. 2003), suggesting a central role of this core gene set in the most common MLL-associated leukemias. The concordance of our MLL-AF4 targets discovered in SEM cells and gene expression signatures in human leukemia patients indicates that the MLL-AF4 target genes discovered in vitro are important for disease progression *in vivo*.

Aberrant chromatin domains occur at regions of MLL-AF4 occupancy

Aberrant modification of chromatin is linked to disease progression in leukemia and other cancers (Jones and Baylin 2007). Among these modifications is the methylation of histone H3 at Lys-79 (H3K79me2), which occurs at the 5′ coding regions of genes that are experiencing

productive transcriptional elongation (Steger et al. 2008), and is a critical checkpoint in transcriptional control (Peterlin and Price 2006; Saunders et al. 2006). Many common MLL partner proteins have been shown to interact with transcriptional elongation components, suggesting that H3K79 methylation might be a key factor in MLL leukemogenesis (Erfurth et al. 2004; Milne et al. 2005b; Okada et al. 2005; Zeisig et al. 2005; Bitoun et al. 2007; Mueller et al. 2007), but the mechanism and extent of H3K79me2 targeting throughout the genome is poorly understood in human cancer cells.

We used ChIP-Seq to determine how the H3K79me2 chromatin modification was distributed across the genome in SEM cells and to ascertain whether this modification was associated with all of the MLL-AF4 target regions (Figure 4). We found ~8,000 regions of H3K79me2 enrichment in MLL-AF4 leukemia cells or control cells, with the vast majority (95%) mapping to known transcripts (Supplemental Tables S14, S19). As expected, most genes (~95%) marked by H3K79me2 in MLL-AF4 leukemia and in control cells produced transcripts that were detectable by microarray-based methods (Supplemental Table S2), with peak enrichment occurring downstream from the transcription start sites (Figure 4D; Supplemental Figure S5). We next extracted the set of H3K79me2-enriched regions in SEM cells and compared them with genomic regions enriched for MLL-AF4 fusion protein. This analysis revealed that ~98% of MLL-AF4 targets were enriched for the H3K79me2 elongation mark. Strikingly, most of these H3K79me2 enrichments formed abnormal domains spanning 5–100 kb extending upstream of and/or downstream from target gene transcriptional start sites in a highly similar pattern to the MLL-AF4 fusion protein (Figure 4A,B). In fact, there was a 92% overlap between the MLL-AF4 target regions and H3K79me2-enriched regions at the DNA base-pair level. Based on the induction of these chromatin modifications at MLL-AF4 target regions, we speculate that the MLL-AF4 fusion protein is directly involved in establishing these aberrant chromatin domains in MLL-linked cancer cells.

Since the elongation-linked H3K79me2 modification forms across large domains with MLL-AF4 (Figure 4A–C), we asked whether regulators of the elongation checkpoint behaved similarly. Indeed, elongation proteins were present across aberrant H3K79me2 domains as determined by ChIP–chip and co-immunoprecipitation experiments. The ENL elongation factor (eleven-nineteen leukemia) bound across aberrant H3K79me2 domains at the *HOXA* and *MEIS1* loci (Supplemental Figure S6; data not shown) and MLL-AF4 associated with ENL and the

pTEFb elongation factor in SEM cells (Supplemental Figure S7; data not shown). Together, these results indicate that the MLL-AF4 protein, the elongation-associated H3K79me2 modification, and additional elongation factors are mistargeted to regions of the genome encoding key developmental regulators.

We next carried out ChIP-Seq experiments in MLL-AF4 leukemia cells using an antibody directed against the histone H3K4me3 modification. This modification is a mark of transcriptional initiation that can be deposited by the MLL complex near the start sites of genes in normal cells (Ruthenburg et al. 2007). Since MLL suffers a monoallelic deletion of its SET-containing the H3K4 methyltransferase domain in MLL-AF4 leukemia, we asked whether MLL-AF4 leukemia cells were able to deposit this histone modification normally across the genome. We found that H3K4me3 modification occurred normally at the start sites of most active genes (~90%) in MLL-AF4 leukemia cells (Supplemental Table S2). Strikingly, at many areas of H3K79me2 mistargeting, H3K4me3 was not only present, but also extended across broad domains of similar size (Figure 4A–C). These chromatin modifications overlapped highly with the MLL-AF4 fusion and occurred at key leukemia and stem cell-associated genes including *HOXA7, HOXA9*, *PROM1*, and *HMGA2* (Figure 4; Supplemental Table S2). This mislocalization of both H3K79me2 and H3K4me3 modifications in MLL-AF4-induced leukemia suggests that the MLL-AF4 fusion protein, a strong transcriptional activator, may be acting at target loci by directly coupling transcriptional initiation and elongation machinery.

A model for MLL-AF4-mediated activation of stem cell-like transcriptional program in leukemia

We describe here the first genome-wide assessment of chromatin modifications and oncogene binding in the most common form of human MLL-associated acute leukemia. The results indicate that MLL-AF4 fusion proteins selectively occupy regions of the genome that contain developmental regulators important for stem cell identity and self-renewal. Our results also show that abnormal patterns of chromatin modifications, including histone H3K79 and histone H3K4 hypermethylation, occur within large domains of MLL-AF4 occupancy. The observation that MLL-AF4 occupies regions of H3K79 hypermethylation and evidence that MLL-AF4 is physically associated with elongation factors including DOT1 strongly suggests that the fusion protein is responsible for generating large domains of H3K79 hypermethylation. The presence of

these aberrant chromatin domains demonstrates that chromatin mistargeting to key regions across the genome is a feature of leukemogenesis.

The presence of aberrant chromatin domains in MLL-linked cancer cells suggests that an abnormal "epigenetic" state exists in these cells. Unlike "genetic lesions" that involve changes in genome sequence in disease cells, "epigenetic lesions" include changes to histone modification states, DNA methylation states, or distribution of chromatin-modifying enzymes (Esteller 2007; Feinberg 2007). By this definition, the aberrant chromatin domains associated with MLL-AF4 binding may thus be considered epigenetic lesions. While it has not been established that these epigenetic lesions cause disease, the H3K79 methylation within these domains likely contributes to oncogenesis since H3K79 methyltransferase activity is required for transformation in MLL-AF10-induced leukemia (Okada et al. 2005). The mechanism of how MLL-AF4 and epigenetic lesions are targeted in the genome remains a central question in leukemia biology.

The binding of the MLL-AF4 fusion protein to a distinct set of developmental genes is of particular interest. Rather than associating with most cellular genes that are engaged in the act of transcription, the fusion protein imposes a more specialized gene expression program. Components of this program include genes not associated previously with MLL-AF4/H3K79me2 mistargeting in leukemia patient-derived cells. These include developmental transcription factors, chromatin regulators, and signaling proteins that are central to leukemia stem cell identity, hematopoietic stem cell identity, and self-renewal (Figure 5). Our findings suggest that MLL-AF4 directly activates a partial hematopoietic stem cell-like transcriptional program found in leukemia stem cells (Krivtsov et al. 2006; Barabe et al. 2007) in concert with an underlying gross defect in chromatin structure.

Materials and Methods

A detailed description of all experimental procedures and data analysis methods can be found in the Supplemental Material.

Cells and cell culture

Human SEM cells with the t(4;11) translocation and REH control cells were purchased from the American Type Culture Collection (ATCC). All cell lines were maintained in RPMI medium 1640 supplemented with 10% FBS. Cells were cross-linked with 1% formaldehyde as described in the Supplemental Material.

ChIP-Seq

ChIP was combined with direct sequencing as described in detail (Supplemental Material). Briefly, DNA from 1×10^8 cells was immunoprecipitated with epitope-specific antibody. Amplified DNA was gel purified and prepared for sequencing using Illumina's Genomic DNA sample kit. Clustering and 26-cycle sequencing were performed using an Illumina Cluster station and 1G analyzer as per the manufacturer's instructions. ChIP-Seq reads were aligned to the human genome and analyzed as described in the Supplemental Material.

Gene expression analysis

Total RNA was isolated from 5×10^6 REH or SEM cells by TRIzol extraction. One microgram of total RNA was labeled according to Affymetrix protocols and hybridized to Affymetrix HG-U133 2.0 plus arrays. The data were analyzed by using Affymetrix Gene Chip Operating Software using default settings. Additional gene expression data from human leukemia patient samples was collated from Ross et al. (2003). Gene expression data and analysis results are provided in the Supplemental Material.

ChIP-Seq, ChIP–chip, and microarray gene expression data and analysis

Complete ChIP-Seq, ChIP–chip, and microarray gene expression data, analysis methods, and results are provided in the Supplemental Material section and in the Gene Expression Omnibus (http://www.ncbi.nlm. nih.gov/geo) database under accession number GSE13313.

Acknowledgements

We thank Y. Zonis, T. Lee, A. Marson, F. Camargo, S. Gupta, and J. Love for help and discussions. We thank the National Cell Culture Center (NCCC) for REH and SEM cells. This

work was supported by grants from the Israel Cancer Research Fund and the Wolfson Foundation (to E.C.), the US-Israel BSF (to E.C. and C.M.C.), NIH grant CA 128609 (to C.M.C. and T.N.), and NIH grant HG002668 (to R.Y.). L.N.L. was supported by individual NRSA fellowship from NHLBI (grant no. F32HL082448).

References

Armstrong S.A., Staunton J.E., Silverman L.B., Pieters R., den Boer M.L., Minden M.D., Sallan S.E., Lander E.S., Golub T.R., Korsmeyer S.J. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat. Genet *30*, 41–47.

Barabe F., Kennedy J.A., Hope K.J., Dick J.E. (2007). Modeling the initiation and progression of human acute leukemia in mice. Science *316*, 600–604.

Barski A., Cuddapah S., Cui K., Roh T.Y., Schones D.E., Wang Z., Wei G., Chepelev I., Zhao K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129,* 823–837.

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.

Bitoun E., Oliver P.L., Davies K.E. (2007). The mixed-lineage leukemia fusion partner AF4 stimulates RNA polymerase II transcriptional elongation and mediates coordinated chromatin remodeling. Hum. Mol. Genet. *16*, 92–106.

Cloos P.A., Christensen J., Agger K., Helin K. (2008). Erasing the methyl mark: Histone demethylases at the center of cellular differentiation and disease. Genes & Dev. *22*, 1115–1140.

Eguchi M., Eguchi-Ishimae M., Greaves M. (2005). Molecular pathogenesis of MLL-associated leukemias.Int. J. Hematol. *82*, 9–20.

Erfurth F., Hemenway C.S., de Erkenez A.C., Domer P.H. (2004). MLL fusion partners AF4 and AF9 interact at subnuclear foci. Leukemia *18*, 92–102.

Esteller M. (2007). Epigenetic gene silencing in cancer: The DNA hypermethylome. Hum. Mol. Genet. *16*, R50–R59.

Feinberg A.P. (2007). Phenotypic plasticity and the epigenetics of human disease. Nature *447*, 433–440.

Ferrando A.A., Armstrong S.A., Neuberg D.S., Sallan S.E., Silverman L.B., Korsmeyer S.J., Look A.T. (2003). Gene expression signatures in MLL-rearranged T-lineage and B-precursor acute leukemias: Dominance of HOX dysregulation. Blood *102*, 262–268.

Greil J., Gramatzki M., Burger R., Marschalek R., Peltner M., Trautmann U., Hansen-Hagge T.E., Bartram C.R., Fey G.H., Stehr K., et al. (1994). The acute lymphoblastic leukemia cell line SEM with t(4;11) chromosomal rearrangement is biphenotypic and responsive to interleukin-7. Br. J. Haematol. *86*, 275–283.

Guenther M.G., Jenner R.G., Chevalier B., Nakamura T., Croce C.M., Canaani E., Young R.A. (2005). Global and Hox-specific roles for the MLL1 methyltransferase. Proc. Natl. Acad. Sci. *102*, 8603–8608.

Guenther M.G., Levine S.S., Boyer L.A., Jaenisch R., Young R.A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77–88.

Haferlach T., Kohlmann A., Schnittger S., Dugas M., Hiddemann W., Kern W., Schoch C. (2005). Global approach to the diagnosis of leukemia using gene expression profiling. Blood *106*, 1189–1198.

Jones P.A., Baylin S.B. (2007). The epigenomics of cancer. Cell *128*, 683–692.

Krivtsov A.V., Armstrong S.A. (2007). MLL translocations, histone modifications and leukaemia stem-cell development. Nat. Rev. Cancer *7*, 823–833.

Krivtsov A.V., Twomey D., Feng Z., Stubbs M.C., Wang Y., Faber J., Levine J.E., Wang J., Hahn W.C., Gilliland D.G., et al. (2006). Transformation from committed progenitor to leukaemia stem cell initiated by MLL-AF9. Nature *442*, 818–822.

Li B., Carey M., Workman J.L. (2007). The role of chromatin during transcription. Cell *128*, 707–719.

Milne T.A., Dou Y., Martin M.E., Brock H.W., Roeder R.G., Hess J.L. (2005a). MLL associates specifically with a subset of transcriptionally active target genes. Proc. Natl. Acad. Sci. *102*, 14765–14770.

Milne T.A., Martin M.E., Brock H.W., Slany R.K., Hess J.L. (2005b). Leukemogenic MLL fusion proteins bind across a broad region of the HoxA9 locus, promoting transcription and multiple histone modifications. Cancer Res. *65*, 11367–11374.

Mueller D., Bach C., Zeisig D., Garcia-Cuellar M.P., Monroe S., Sreekumar A., Zhou R., Nesvizhskii A., Chinnaiyan A., Hess J.L., et al. (2007). A role for the MLL fusion partner ENL in transcriptional elongation and chromatin modification. Blood. *110*, 4445–4454.

O'Brien C.A., Pollett A., Gallinger S., Dick J.E. (2007). A human colon cancer cell capable of initiating tumour growth in immunodeficient mice. Nature *445*, 106–110.

Okada Y., Feng Q., Lin Y., Jiang Q., Li Y., Coffield V.M., Su L., Xu G., Zhang Y. (2005). hDOT1L links histone methylation to leukemogenesis. Cell *121*, 167–178.

Pajer P., Pecenka V., Karafiat V., Kralova J., Horejsi Z., Dvorak M. (2003). The twist gene is a common target of retroviral integration and transcriptional deregulation in experimental nephroblastoma. Oncogene *22*, 665–673.

Peterlin B.M., Price D.H. (2006). Controlling the elongation phase of transcription with P-TEFb. Mol. Cell. *23*, 297–305.

Puisieux A., Valsesia-Wittmann S., Ansieau S. (2006). A twist for survival and cancer progression. Br. J. Cancer. *94*, 13–17.

Ramalho-Santos M., Yoon S., Matsuzaki Y., Mulligan R.C., Melton D.A. (2002). "Stemness": Transcriptional profiling of embryonic and adult stem cells. Science *298*, 597–600.

Rosenfeld C., Goutner A., Choquet C., Venuat A.M., Kayibanda B., Pico J.L., Greaves M.F. (1977). Phenotypic characterisation of a unique non-T, non-B acute lymphoblastic leukaemia cell line. Nature *267*, 841–843.

Ross M.E., Zhou X., Song G., Shurtleff S.A., Girtman K., Williams W.K., Liu H.C., Mahfouz R., Raimondi S.C., Lenny N., et al. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. Blood. 2003;102:2951–2959.

Rozovskaia T., Feinstein E., Mor O., Foa R., Blechman J., Nakamura T., Croce C.M., Cimino G., Canaani E. (2001). Upregulation of Meis1 and HoxA9 in acute lymphocytic leukemias with the t(4: 11) abnormality. Oncogene *20*, 874–878.

Rozovskaia T., Ravid-Amir O., Tillib S., Getz G., Feinstein E., Agrawal H., Nagler A., Rappaport E.F., Issaeva I., Matsuo Y., et al. (2003). Expression profiles of acute lymphoblastic and myeloblastic leukemias with ALL-1 rearrangements. Proc. Natl. Acad. Sci. *100*, 7853–7858.

Ruthenburg A.J., Allis C.D., Wysocka J. (2007). Methylation of lysine 4 on histone H3: Intricacy of writing and reading a single epigenetic mark. Mol. Cell. *25*, 15–30.

Saunders A., Core L.J., Lis J.T. Breaking barriers to transcription elongation. (2006). Nat. Rev. Mol. Cell Biol. *7*, 557–567.

Steger D.J., Lefterova M.I., Ying L., Stonestrom A.J., Schupp M., Zhuo D., Vakoc A.L., Kim J.E., Chen J., Lazar M.A., et al. (2008). DOT1L/KMT4 recruitment and H3K79 methylation are ubiquitously coupled with gene transcription in mammalian cells. Mol. Cell. Biol. *28*, 2825–2839.

Subramanian A., Tamayo P., Mootha V.K., Mukherjee S., Ebert B.L., Gillette M.A., Paulovich A., Pomeroy S.L., Golub T.R., Lander E.S., et al. (2005). Gene set enrichment analysis: A

knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. *102*, 15545–15550.

Toren A., Bielorai B., Jacob-Hirsch J., Fisher T., Kreiser D., Moran O., Zeligson S., Givol D., Yitzhaky A., Itskovitz-Eldor J., et al. (2005). CD133-positive hematopoietic stem cell "stemness" genes contain many genes mutated or abnormally expressed in leukemia. Stem Cells. *23*, 1142–1153.

Wagner W., Ansorge A., Wirkner U., Eckstein V., Schwager C., Blake J., Miesala K., Selig J., Saffrich R., Ansorge W., et al. (2004). Molecular evidence for stem cell function of the slow-dividing fraction among human hematopoietic progenitor cells by genome-wide analysis. Blood *104*, 675–686.

Yang J., Mani S.A., Donaher J.L., Ramaswamy S., Itzykson R.A., Come C., Savagner P., Gitelman I., Richardson A., Weinberg R.A. (2004). Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. Cell *117*, 927–939.

Yeoh E.J., Ross M.E., Shurtleff S.A., Williams W.K., Patel D., Mahfouz R., Behm F.G., Raimondi S.C., Relling M.V., Patel A., et al. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell *1*, 133–143.

Zeisig D.T., Bittner C.B., Zeisig B.B., Garcia-Cuellar M.P., Hess J.L., Slany R.K. The eleven-nineteen-leukemia protein ENL connects nuclear MLL fusion partners with chromatin. (2005). Oncogene *24*, 5525–5532.

Figure 1

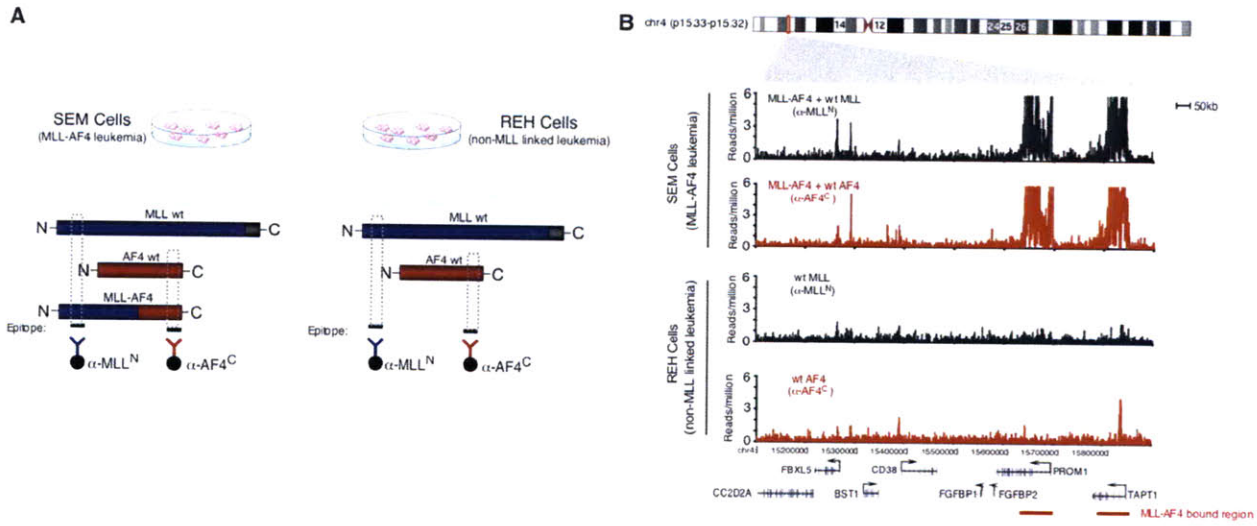


Figure 1. Mapping MLL-AF4 fusion protein-binding sites in human leukemia cells

A. Schematic diagram of strategy for mapping MLL-AF4 fusion protein-binding sites. SEM precursor B acute leukemia cells express the MLL-AF4 fusion protein. REH precursor B acute leukemia cells express only endogenous AF4 and MLL1. The N terminus of MLL (blue) is recognized by ChIP antibody anti-MLL-N (blue) and immunoprecipitates both wild-type MLL and MLL-AF4 fusion protein in SEM cells. The C terminus of AF4 (red) is recognized by ChIP antibody anti-AF4-C and immunoprecipitates both wild-type AF4 and MLL-AF4 fusion protein in SEM cells. Wild-type AF4 and MLL-N are immunoprecipitated by anti-AF4-C and anti-MLL-N, respectively.

B. Binding of AF4 (red) and MLL-N (black) in SEM cells (*top* panels) and REH cells (*bottom* panels) as determined by ChIP-Seq. Binding profiles are shown across an 800-kb portion of the genome surrounding the PROM1 gene (gene models shown in black *below* graph; a black arrow indicates transcription start sites). MLL-AF4 fusion protein binding is indicated by a red bar.

Figure 2

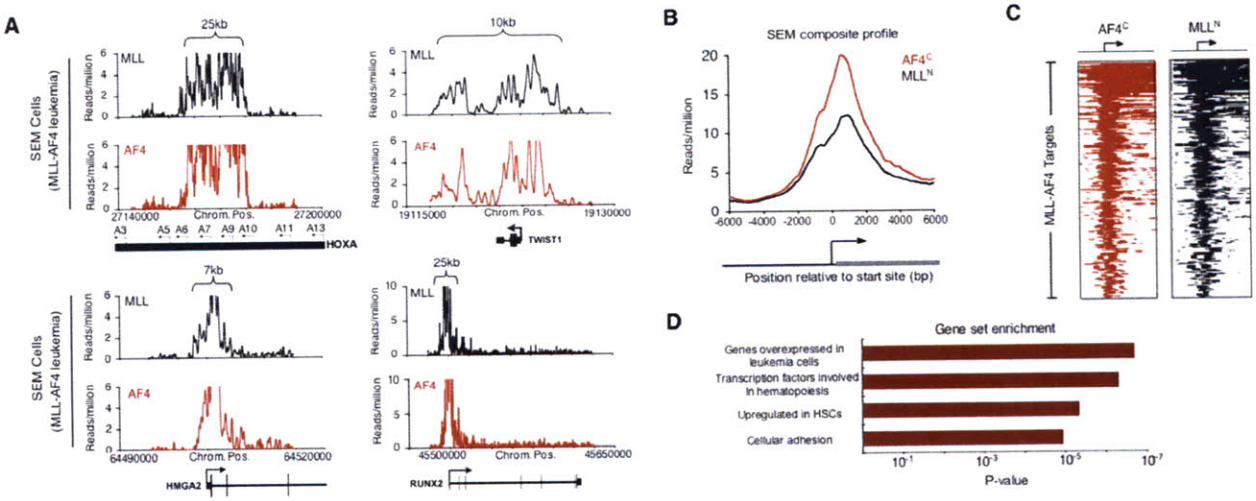


Figure 2. MLL-AF4 target genes are enriched for early developmental regulators

A. Signals for AF4 (red) and MLL-N (black) reflecting binding of the presumptive MLL-AF4 fusion protein in SEM cells as determined by ChIP-Seq. Binding profiles are shown across a 15- to 150-kb portion of the genome surrounding the *HOXA9, TWIST1, HMGA2,* and *RUNX2* genes (gene models shown in black *below* graph; a black arrow indicates transcription start sites). Size of MLL-AF4-enriched region is indicated by *top* brackets. A detailed description of data analysis methods is provided in the Supplemental Material.

B. Composite AF4-C terminus (red) and MLL-N terminus (black) binding profiles for all MLL-AF4 target genes. The start site and direction of transcription of the average gene are indicated by an arrow.

C. ChIP-Seq density heat map of AF4-C terminus (red) and MLL-N terminus (black) for all MLL-AF4 target genes. The genomic region from −5kb to +10kb relative to the transcription start site of each gene is shown. Gene order is determined by highest average MLL/AF4 read density from *top* to *bottom*. The start site and direction of transcription of the genes are indicated by an arrow.

D. Selected results of GSEA (http://www.broad.mit.edu/gsea) of MLL-AF4 target genes.

Figure 3

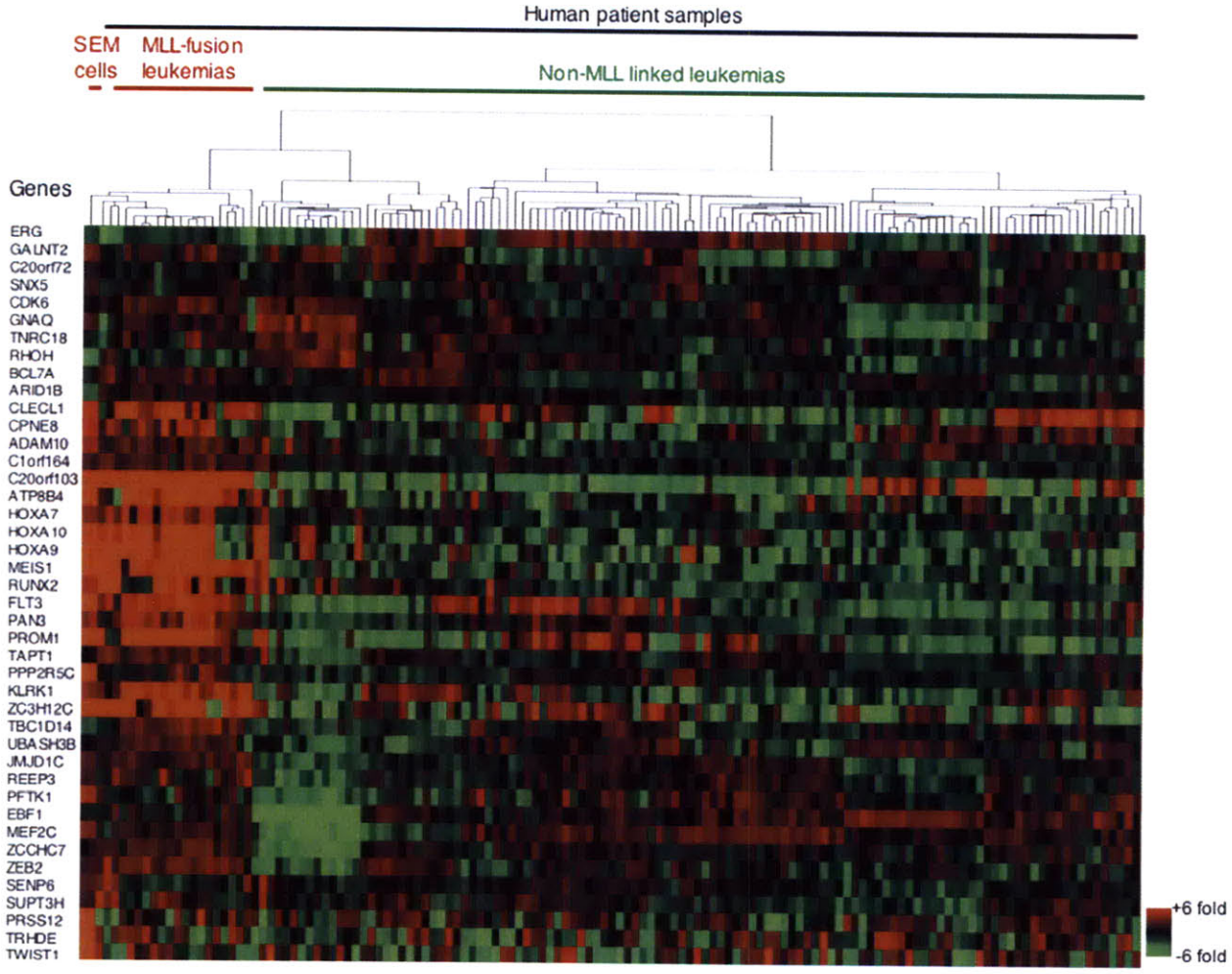


Figure 3. MLL-AF4 target genes define MLL-linked leukemia *in vivo*

Hierarchical clustering of relative expression levels of 42 genes occupied by MLL-AF4 fusion protein target regions >10 kb. Comparisons were made across the SEM and REH cell lines and 132 peripheral blood samples of patients diagnosed with leukemia. Each row corresponds to a gene that is bound by MLL-AF4 for which expression data were available. Each column corresponds to a single gene expression microarray. For each gene, expression is shown relative to the average expression level of that gene across all samples, with shades of red indicating higher than average expression and green lower than average expression. Columns and rows were ordered by unsupervised hierarchical clustering. A detailed description of data analysis methods is provided in the Supplemental Material.

Figure 4

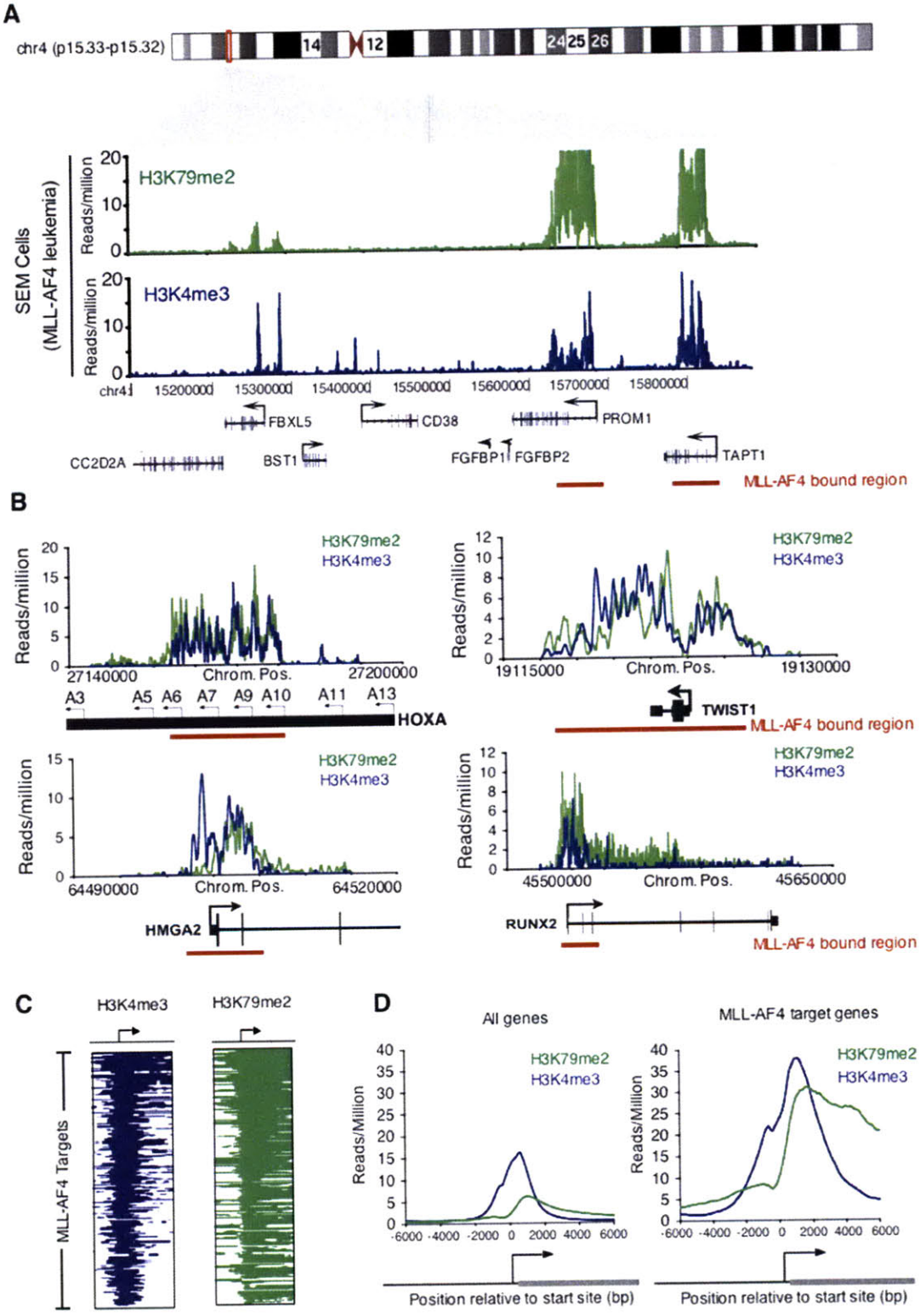

A
chr4 (p15.33-p15.32)
SEM Cells
(MLL-AF4 leukemia)
Reads/million
H3K79me2
H3K4me3
chr4
FBXL5
CD38
PROM1
CC2D2A
BST1
FGFBP1
FGFBP2
TAPT1
MLL-AF4 bound region
B
Chrom. Pos.
HOXA
A3
A5
A6
A7
A9
A10
A11
A13
TWIST1
HMGA2
RUNX2
C
MLL-AF4 Targets
D
All genes
MLL-AF4 target genes
Reads/Million
Position relative to start site (bp)

Figure 4. Mistargeting of chromatin modifications occur as epigenetic lesions at MLL-AF4 target regions

A. Binding of H3K79me2 (green) and H3K4me3 (blue) in SEM cells as determined by ChIP-Seq. Binding profiles are shown across an 800-kb portion of the genome surrounding the *PROM1* gene. Gene models shown in black *below* the graph; a black arrow indicates transcription start sites. MLL-AF4 fusion protein-binding regions are indicated by red bars.

B. Binding of H3K79me2 ChIPs (green) and H3K4me3 ChIPs (blue) in SEM cells as determined by ChIP-Seq. Binding profiles are shown across 15- to 150-kb portions of the genome surrounding the *HOXA9, TWIST1, HMGA2,* and *RUNX2* genes. Gene models are shown in black (*below graph*). A black arrow indicates transcription start sites. MLL-AF4 fusion protein binding is indicated by a red bar.

C. H3K79me2 (green) and H3K4me3 (blue) binding profiles for all MLL-AF4 target genes in SEM cells. Genes are ordered as in Figure 2C.

D. Composite H3K79me2 ChIP enrichments (green) and H3K4me3 ChIP enrichments (blue) for all genes (*left*) and all MLL-AF4 target genes (*right*). The start site and direction of transcription of the average gene are indicated by an arrow.

Figure 5

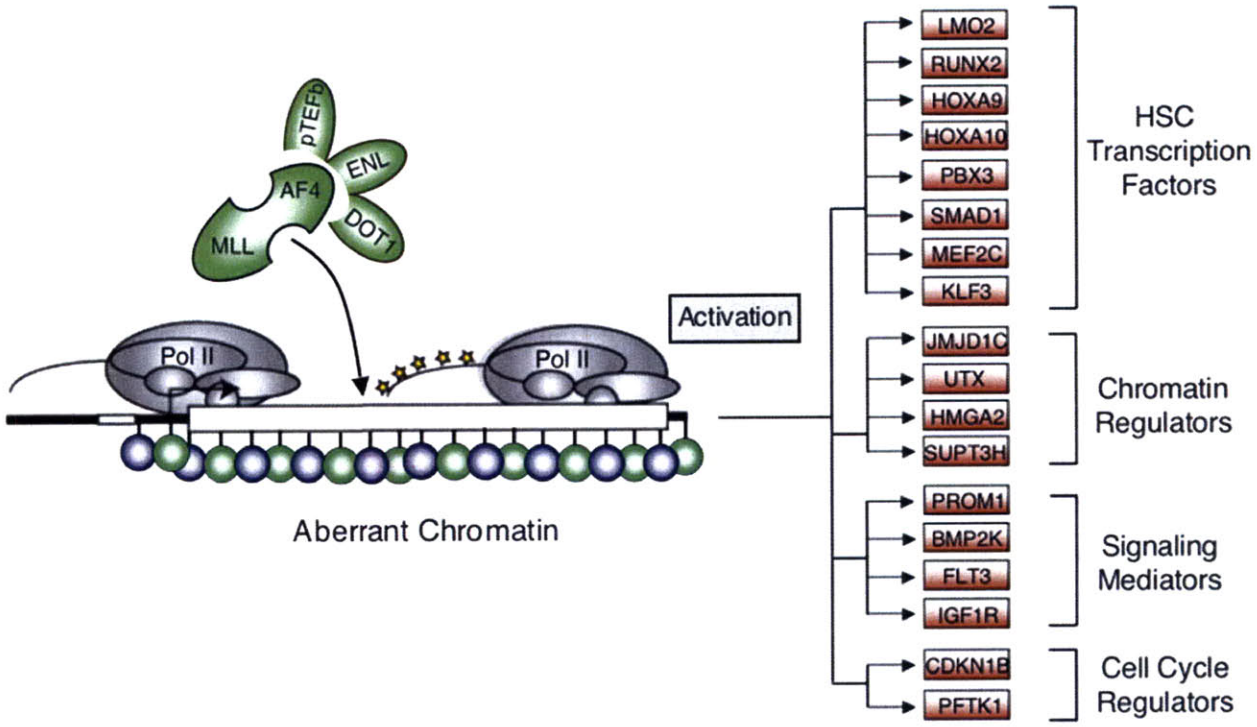


Figure 5. Model of MLL-AF4-mediated gene activation in leukemia

Schematic diagram of transcriptional misregulation in MLL-AF4-induced leukemia. MLL-AF4 associates with elongation proteins (pTEFb, ENL, and DOT1) at sites of epigenetic lesions. Phosphorylated and elongating RNA Pol II shown in yellow. Blue (H3K4me3) and green (H3K79me2) circles represent histone modifications at aberrant chromatin domains. Genes annotated as bound by MLL-AF4 and playing a role in hematopoietic stem cell function are shown to the *right* and subdivided into functional categories.

Chapter 6

CpG Island Structure Defines Polycomb/Trithorax Chromatin Domains in Human ES and iPS Cells

Garrett M. Frampton, David A. Orlando, Matthew G. Guenther, and Richard A. Young.

Abstract

Trithorax Group (TrxG) and Polycomb Group (PcG) protein complexes play key roles in the epigenetic regulation of development. It is not clear however, what directs these chromatin regulators to specific sites throughout the genome to generate H3K4me3 and H3K27me3 modified nucleosomes. We demonstrate here that CpG island structure predicts TrxG/PcG chromatin structure and does this uniquely in human pluripotent stem cells.

CpG island structure defines Polycomb/Trithorax chromatin domains in human ES and iPS cells

Trithorax group (TrxG) and Polycomb group (PcG) genes were discovered in *Drosophila melanogaster* as activators and repressors of Hox transcription factor genes, which specify cell identity along the anteroposterior axis of segmented animals. TrxG proteins catalyze trimethylation of histone H3 lysine 4 (H3K4me3) at the promoters of active genes and facilitate maintenance of active gene states during development, in part by antagonizing the functions of PcG proteins. PcG proteins catalyze trimethylation of histone H3 lysine 27 (H3K27me3) and function to silence genes encoding key regulators of development. TrxG and PcG proteins have been implicated in control of cell identity, proliferation, X inactivation, genomic imprinting and cancer (Schwartz and Pirotta, 2007; Simon and Kingston, 2009; Surface et al., 2010). Further knowledge of how these regulators function is thus fundamentally important for understanding many biological phenomena.

How TrxG and PcG protein complexes are recruited to their sites of action in vertebrate cells is not fully understood. Nucleosomes with H3K4me3 are found immediately downstream of transcription initiation sites (Guenther et al., 2007) consistent with proposals that TrxG complexes are recruited to active promoter regions by transcription factors and/or the transcription initiation apparatus (Li et al., 2007; Thomson et al., 2010). In embryonic stem (ES) cells and induced pluripotent stem (iPS) cells, regions that are occupied by PcG proteins contain a "bivalent" chromatin structure, containing nucleosomes with both H3K4me3 and H3K27me3, and occur at the promoters of genes encoding key developmental regulators (Bernstein et al., 2006; Guenther et al., 2010). There is evidence that PcG protein complexes are recruited to some of these sites by ncRNAs (Rinn et al., 2007; Pandey et al., 2008; Zhao et al., 2008; Khalil et al., 2009; Gupta et al., 2010; Kanhere et al., 2010; Tsai et al., 2010; Yap et al., 2010) and DNA-binding cofactors (Shen et al., 2009; Peng et al., 2009; Li et al., 2010; Pasini et al., 2010). Several models have been proposed to explain the recruitment of PcG proteins genome-wide (Mendenhall and Bernstein, 2008; Margueron and Reinberg, 2010; Guenther and Young, 2010).

CpG islands are small genomic elements, ~1kb in length, which lack the suppression of CG dinucleotides that is found throughout the rest of the genome (Illingworth and Bird, 2009). Two features of CpG islands led us to investigate the potential relationship between them and the chromatin structure catalyzed by TrxG and PcG proteins. First, the majority of CpG islands are

located at the transcription start sites (TSS) of genes, and this is coincident with the genomic locations of nucleosomes with H3K4me3 and H3K27me3. Second, some of the ncRNA species that recruit PcG complexes have characteristic GC-rich stem loop structures that are required for their function (Wutz et al., 2002; Zhao et al., 2008; Kanhere et al, 2010; Yap et al., 2010; Tsai et al., 2010), suggesting that transcription of CpG islands might generally be involved in PcG complex recruitment.

We compared the genome-wide localization of CpG islands to the occupancy of H3K4me3 and H3K27me3 modified nucleosomes for a collection of human ES and iPS cells as well as several differentiated cell types (human primary fibroblasts, primary CD4+ T cells, and IMR90 fetal lung fibroblasts) from published and unpublished ChIP-Seq experiments (Supplemental Information). We observed a striking relationship between H3K4me3 and CpG islands in ES and iPS cells. In human pluripotent stem cells, CpG islands and H3K4me3 nucleosomes occurred in precisely the same regions across the genome (Figure 1A) with almost no H3K4me3 modified nucleosomes occurring outside of a CpG island. This relationship was specific to pluripotent stem cells as compared to several differentiated cell types. Genes occupied by H3K4me3 in pluripotent cells, but not in differentiated cells were nearly always associated (93-94%) with CpG islands. On the other hand, genes occupied by H3K4me3 in differentiated cells but not pluripotent cells were infrequently associated with CpG islands (12-14%; Table S2A).

Comparing H3K27me3 to CpG islands revealed a more complicated relationship. Although most CpG islands (86-88%) in ES cells did not contain H3K27me3 modified nucleosomes, nearly all (86-91%) H3K27me3 occupied regions were associated with CpG islands. Like H3K4me3, the association between H3K27me3 and CpG islands was specific for pluripotent cells. Genes occupied by H3K27me3 in differentiated cells but not in pluripotent cells were frequently not associated with CpG islands (41-52%; Table S2B). Unlike H3K4me3 however, H3K27me3 was not entirely coincident with CpG islands in pluripotent cells and frequently occurred in the regions between closely spaced CpG islands.

To understand how these results relate to genes, we classified the complete set of human genes by the number of CpG islands in their promoter regions (Figure 1B). We found that ~30% of genes do not have a CpG island at their TSS, ~60% of genes have a single CpG island at their TSS, and ~10% of genes have two or more CpG islands at their TSS (Table S2D). As expected,

in ES and iPS cells, the genes with zero CpG islands were occupied by neither H3K4me3 nor H3K27me3 modified nucleosomes. The genes with a single CpG island were occupied by H3K4me3, but very few of these genes were also occupied by H3K27me3. The genes with clusters of multiple CpG islands were nearly all occupied by H3K4me3. Remarkably, the majority of these genes were also occupied by H3K27me3 ($p < 10^{-100}$; Figure 1C) and they were highly enriched for genes encoding key regulators of development ($p < 10^{-100}$). In fact, clusters of CpG islands are found spanning the promoters of nearly every homeobox transcription factor (Table S3A). These results demonstrate that the number of CpG islands that occur at a gene's promoter is highly predictive of the TrxG/PcG chromatin structure at those genes in pluripotent stem cells.

We examined the genes with bivalent chromatin and multiple CpG islands in more detail and observed that the H3K4me3/H3K27me3 modified nucleosomes and CpG island clusters spanned the same genomic regions and that the peaks of H3K4me3 occupancy in these regions were aligned with the individual CpG islands. This phenomenon is most prominent in the four Hox gene clusters, which each contain ~40 CpG islands (Figure S1), but was clearly evident at approximately 1,000 genes encoding developmental regulators and cellular signaling components (Table S3B). These results indicate that the TrxG/PcG chromatin structure at bivalent genes is highly aligned with the local CpG island structure.

Previous studies have identified aspects of the phenomena we describe here, but have not revealed the striking "rules" that apparently govern the genome-wide occupancy of histone H3K4me3 and H3K27me3 in human pluripotent stem cells. Previous work (Mikkelsen et al., 2007) has shown that, in ES cells, H3K4me3 occupancy is correlated with CpG islands. We demonstrate that H3K4me3 nucleosomes and CpG islands are entirely co-incident across the genome in pluripotent cells and that this relationship does not extend to differentiated cell types. Other studies have noted that PcG occupancy can be predicted from the locations, sizes, and motif contents of CpG islands (Ku et al., 2008), by conservation properties (Tanay et al., 2007) or by DNA sequence motifs (Liu et al., 2010), but did not reveal that specifically in pluripotent stem cells, H3K27me3 nucleosomes occur at those promoters that have multiple CpG islands. These rules for histone H3K4me3 and H3K27me3 occupancy in pluripotent cells help to constrain models for the function of PcG and TrxG complexes in pluripotent cells and during differentiation.

Why do nucleosomes with histone H3K4me3 occur at all CpG islands in human pluripotent stem cells? One possibility is that transcription initiation occurs at all these sites (Guenther et al., 2007) and that TrxG proteins are recruited via the transcription apparatus (Essenberg and Shilatifard, 2009). Another possibility is that proteins that bind to CG dinucleotides, such as Cfp1, recruit TrxG proteins to all these sites (Thomson et al., 2010).

Why do nucleosomes with histone H3K27me3 occur at genes with multiple CpG islands in ES and iPS cells? Recent evidence that RNA species containing GC-rich stem loop structures can contribute to PcG complex recruitment and that transcripts from promoter regions frequently contain these structures suggests a general model for establishing PcG domains that involves transcripts from these domains (Guenther and Young, 2010; Wutz et al., 2002; Zhao et al., 2008; Kanhere et al., 2010; Yap et al, 2010; Tsai et al., 2010).

We investigated the possibility that DNA elements with the potential to encode such GC-rich RNA structures were enriched in the promoters of genes with multiple CpG islands. Beginning with the known examples of small ncRNAs that bind to PcG proteins (Figure 2A; Wutz et al., 2002; Zhao et al., 2008), we examined the promoter sequence of every gene looking for sequence elements that would be likely to form the characteristic GC-rich stem loop structure (Figure 2A; Table S2D; Supplemental Information). We found that the promoter regions with more than one CpG island have a much higher probability of forming the GC-rich RNA structures that bind PcG protein complexes than those that contain zero or one CpG island ($p < 10^{-100}$, Figure 2B). Furthermore, we found that among genes with one CpG island and among genes with more than one CpG island, those that were occupied by H3K27me3 had a much higher probability of forming the characteristic stem loops than those that were not occupied by H3K27me3 ($p < 10^{-100}$; Figure 2C).

In summary, we have identified several striking relationships between the histone modifications catalyzed by TrxG and PcG proteins and CpG island structure in human gene promoters. These relationships are specific to pluripotent stem cells, and are not retained in differentiated cells. First, H3K4me3 modified nucleosomes and CpG islands and coincident genome-wide. Second, essentially all H3K27me3 modified nucleosomes occur in close association with CpG islands. Third, genes that do not have a CpG island at their start site are occupied by neither H3K4me3 nor H3K27me3. Fourth, genes that have a single CpG island are occupied by H3K4me3, but not H3K27me3. Fifth, genes with three or more CpG islands are

occupied by H3K4me3 and H3K27me3, and these chromatin domains precisely span the CpG island clusters. We also note that genomic regions containing multiple CpG islands have a higher probability of producing RNA species that could recruit PcG proteins, providing a likely explanation for why genes with multiple CpG islands are occupied by H3K27me3. We conclude that CpG island structure plays a fundamental role in defining TrxG/PcG chromatin structure in human pluripotent stem cells.

Acknowledgements

We thank T. Volkert, J. Love, S. Gupta, J. Kwon and V. Dhanapal for assistance with ChIP-Seq experiments as well as T. Lee, L. Lawton, F. Soldner, and J. Reddy for helpful discussions about the manuscript.

References

Schwartz, Y.B. & Pirrotta, V. (2007). Polycomb silencing mechanisms and the management of genomic programmes. Nat. Rev. Genet. *8*, 9-22.

Simon, J.A. & Kingston, R.E. (2009). Mechanisms of polycomb gene silencing: knowns and unknowns. Nat. Rev. Mol. Cell Biol. *10*, 697-708.

Surface L.E., Thornton S.R., & Boyer L.A. (2010). Polycomb group proteins set the stage for early lineage commitment. Cell Stem Cell *7*, 288-98.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R. & Young, R.A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77-88.

Li, B., Carey, M. & Workman, J.L. (2007). The role of chromatin during transcription. Cell *128*, 707-19.

Thomson , J.S.P., et al. (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature *464*, 1082-1086.

Bernstein, B.E. et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315-26.

Guenther, M.G. et al. (2010). Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells. Cell Stem Cell *7*, 249-57.

Rinn, J.L. et al. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311-23.

Pandey, R.R. et al. (2008). Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol. Cell *32*, 232-46.

Zhao, J., Sun, B.K., Erwin, J.A., Song, J.J. & Lee, J.T. (2008). Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science *322*, 750-6.

Khalil, A.M. et al. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc. Natl. Acad. Sci. *106*, 11667-72.

Gupta, R.A. et al. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature *464*, 1071-6.

Kanhere, A. et al. (2010). Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. Mol. Cell *38*, 675-88.

Tsai, M.C. et al. (2010). Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. Science *329*, 689-93.

Yap, K.L. et al. (2010). Molecular Interplay of the Noncoding RNA ANRIL and Methylated Histone H3 Lysine 27 by Polycomb CBX7 in Transcriptional Silencing of INK4a. Mol. Cell *38*, 662-674.

Shen, X., et al. (2009). Jumonji (JARID2) Modulates Polycomb Activity and Self-Renewal versus Differentiation of Stem Cells. Cell *139*,1303-14.

Peng, J.C., et al. (2009). Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. Cell *139*, 1290-302.

Li, G., et al. (2010). Jarid2 and PRC2, partners in regulating gene expression. Genes Dev. *24*, 368-80.

Pasini, D., et al. (2010). JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. Nature *464*, 306-10.

Mendenhall, E.M. & Bernstein, B.E. (2008). Chromatin state maps: new technologies, new insights. Curr. Opin. Genet. Dev. *18*, 109-15.

Margueron, R. & Reinberg, D. (2010). Chromatin structure and the inheritance of epigenetic information. Nat. Rev. Genet. *11*, 285-96.

Guenther, M.G. & Young, R.A. (2010). Repressive transcription. Science *329*, 150-1.

Illingworth, R.S. & Bird, A.P. (2009). CpG islands--'a rough guide'. FEBS Lett. *583*, 1713-20.

Wutz, A., Rasmussen, T.P. & Jaenisch, R. (2002). Chromosomal silencing and localization are mediated by different domains of Xist RNA. Nat. Genet. *30*, 167-74.

Mikkelsen, T.S. et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-60.

Ku, M. et al. (2008). Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. PLoS Genet. *4*, e1000242.

Tanay, A., O'Donnell, A.H., Damelin, M., & Bestor, T.H. (2007). Hyperconserved CpG domains underlie Polycomb-binding sites. Proc. Natl. Acad. Sci. *104*, 5521-6.

Liu, Y., Shao, Z. & Yuan, G.C. (2010). Prediction of Polycomb target genes in mouse embryonic stem cells. Genomics *96*, 17-26.

Eissenberg, J.C. & Shilatifard, A. (2009). Histone H3 lysine 4 (H3K4) methylation in development and differentiation. Dev. Biol. *339*, 240-9.

Figure 1

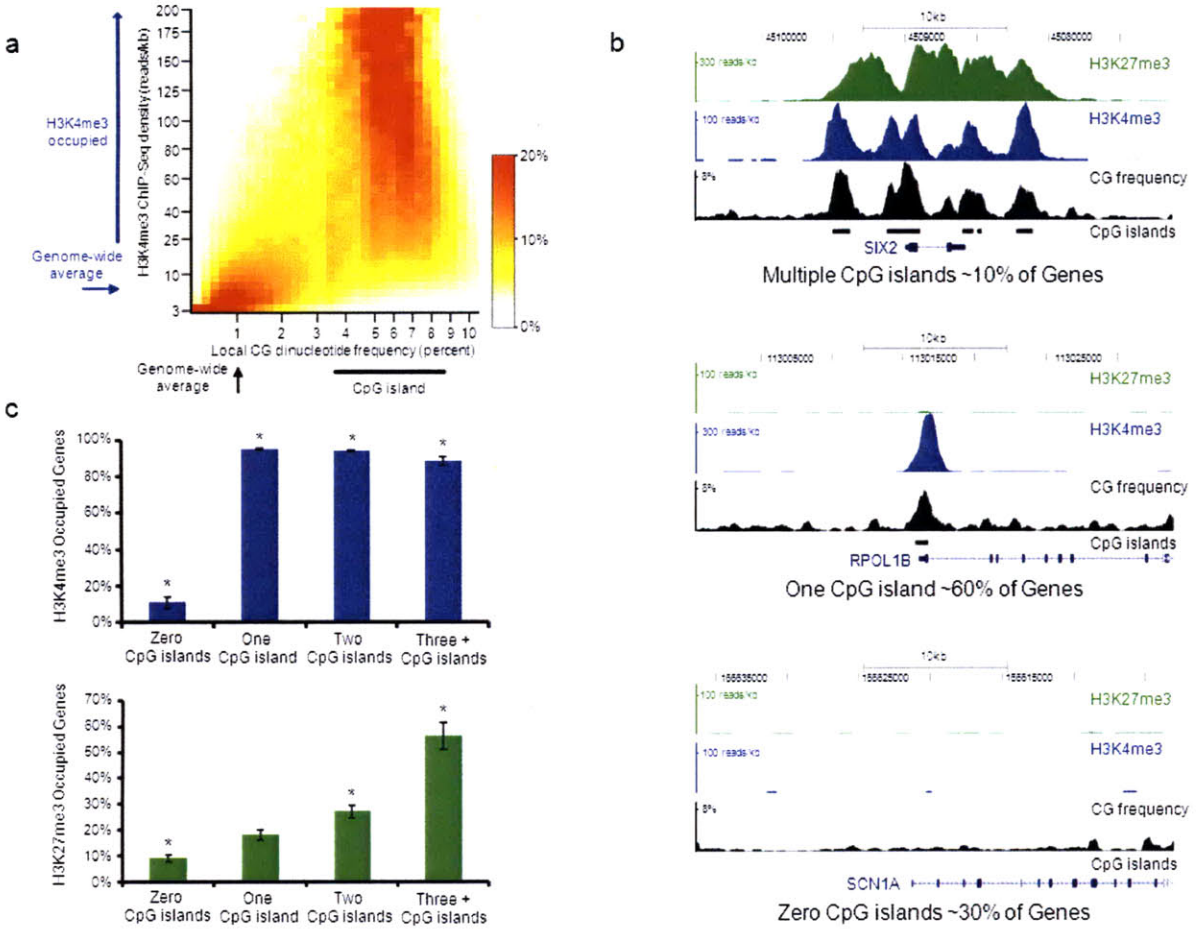

a
H3K4me3 occupied
Genome-wide average
H3K4me3 ChIP-Seq density (reads/kb)
Local CG dinucleotide frequency (percent)
Genome-wide average
CpG island
20%
10%
0%
b
H3K27me3
H3K4me3
CG frequency
CpG islands
SIX2
Multiple CpG islands ~10% of Genes
RPOL1B
One CpG island ~60% of Genes
SCN1A
Zero CpG islands ~30% of Genes
c
H3K4me3 Occupied Genes
H3K27me3 Occupied Genes
Zero CpG islands
One CpG island
Two CpG islands
Three + CpG islands

Figure 1. CpG island structure defines the genomic occupancy of H3K4me3 and H3K27me3 modified nucleosomes in pluripotent stem cells

A. H3K4me3 occupancy is coincident with CpG islands in pluripotent cells. Across the genome, in the human ES cell line WIBR2, local CG dinucleotide density and H3K4me3 ChIP-Seq density were tabulated and are presented in a heatmap.

B. Genes were categorized by the number of CpG islands associated with their transcription start site. The genes encoding the homeobox transcription factor sine oculis homeobox homolog 2 (SIX2), RNA polymerase I polypeptide B (RPOL1B), and sodium channel, voltage-gated, type I alpha (SCN1A), all located on chromosome 2 are shown as examples. H3K27me3 and H3K4me3 ChIP-Seq density in the hES line WIBR2, local CG dinucleotide density, and CpG islands are shown.

C. The portion of genes with zero, one, two, and three or more CpG islands that are occupied by H3K4me3 and H3K27me3 modified nucleosomes in pluripotent stem cells is shown. Error bars indicate the standard deviation of the values for all ES and iPS cell lines. The * symbol indicates that the difference between the values for this class of genes and the genome-wide average is statistically significant ($p < 10^{-10}$) as calculated with a Chi-square test.

Figure 2

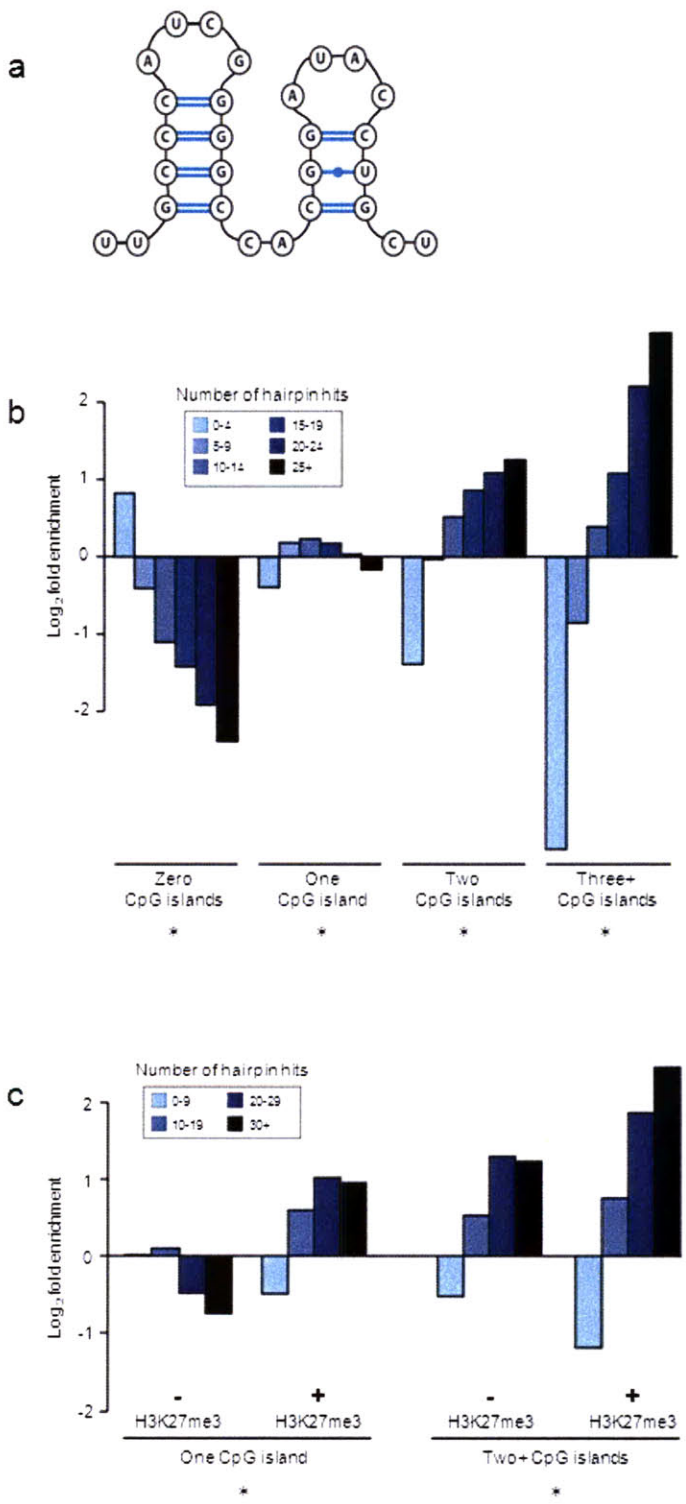

a
b
Number of hairpin hits
0-4
5-9
10-14
15-19
20-24
25+
Log2 fold enrichment
2
1
0
-1
-2
Zero CpG islands
One CpG island
Two CpG islands
Three+ CpG islands
*
*
*
*
c
Number of hairpin hits
0-9
10-19
20-29
30+
Log2 fold enrichment
2
1
0
-1
-2
-
+
H3K27me3
H3K27me3
-
+
H3K27me3
H3K27me3
One CpG island
Two+ CpG islands
*
*

Figure 2. DNA elements encoding RNA species with characteristic GC rich hairpins are enriched at gene promoters with multiple CpG islands

A. The dual stem-loop RNA hairpin that is known to bind PRC2 is shown.

B. The number of RNA hairpin hits around the TSS (+/-5kb) for genes within zero, one, two, and three or more CpG islands shows a bias for more hits in genes with multiple CpG islands (see Supplemental Information for the definition of an RNA hairpin hit). The relative fold enrichment versus all genes is shown for each CpG class. The * symbol indicates that the difference between the values for this class of genes and the genome-wide average is statistically significant ($p < 10^{-100}$) as calculated with a Chi-square test.

C. The number of RNA hairpin hits for genes with one CpG island with or without H3K27me3 occupancy and for genes with more than one CpG island with or without H3K27me3 occupancy is shown. Data are represented as in (b). The * symbol indicates that the difference between the values for genes occupied by H3K27me3 and genes not occupied by H3K27me3 is statistically significant ($p < 10^{-10}$) as calculated with a Mann-Whitney U test.

Chapter 7

Concluding Remarks

Concluding Remarks

Tremendous progress has been made in understanding the molecular mechanisms that cells use to control cell type during development. Despite this strong foundation, there are still immediate and important questions facing biologists. In these concluding remarks, I describe a few of the most important, timely, and interesting areas for future research that I have encountered.

Identification of master regulators of cell type based on mining the GEO database

The epigenetic landscape described by Waddington (Waddington 1957) provides a useful model to describe the development of pluripotent cells from their undifferentiated state to progressively more specialized cell types (Figure 1). Normal development is unidirectional, proceeding from a single pluripotent cell, through multi-potent intermediates, to fully differentiated adult cells. Reprogramming experiments, which allow creation of induced pluripotent stem cells from a large array of somatic cell types, provide a striking example of the power of master regulator transcription factors (TF) to control cell type and even promote cell state transitions that do not occur during normal development (Takahashi and Yamanaka 2006).

Regenerative medicine promises to treat human disease by creating and delivering healthy cells and tissues to substitute for those that are malfunctioning. The ability to manipulate cell type has great potential for regenerative medicine because, perhaps, it will enable the creation of immune-compatible replacements for diseased or damaged cells. Consequently, a system-wide understanding of the sets of TFs that act as master regulators for each cell type will be key to advancing the science of regenerative medicine.

The aim of systems biology is to create predictive models of biological systems in order to understand how they work. Consequently, a detailed account of the complete set of genes encoding master regulator TFs in the human genome, the cell types that they specify, and the regulatory networks that they control is necessary for a systems biology description of human development. Computational methods are likely to be crucial in guiding wet lab research in achieving such an account. Genes encoding master regulator TFs can be identified based on characteristic protein domains, cell type specificity can be determined by mining gene expression data, and regulatory networks architecture and cis-regulatory elements can be predicted using network prediction algorithms (Segal et al. 2003).

Figure 1

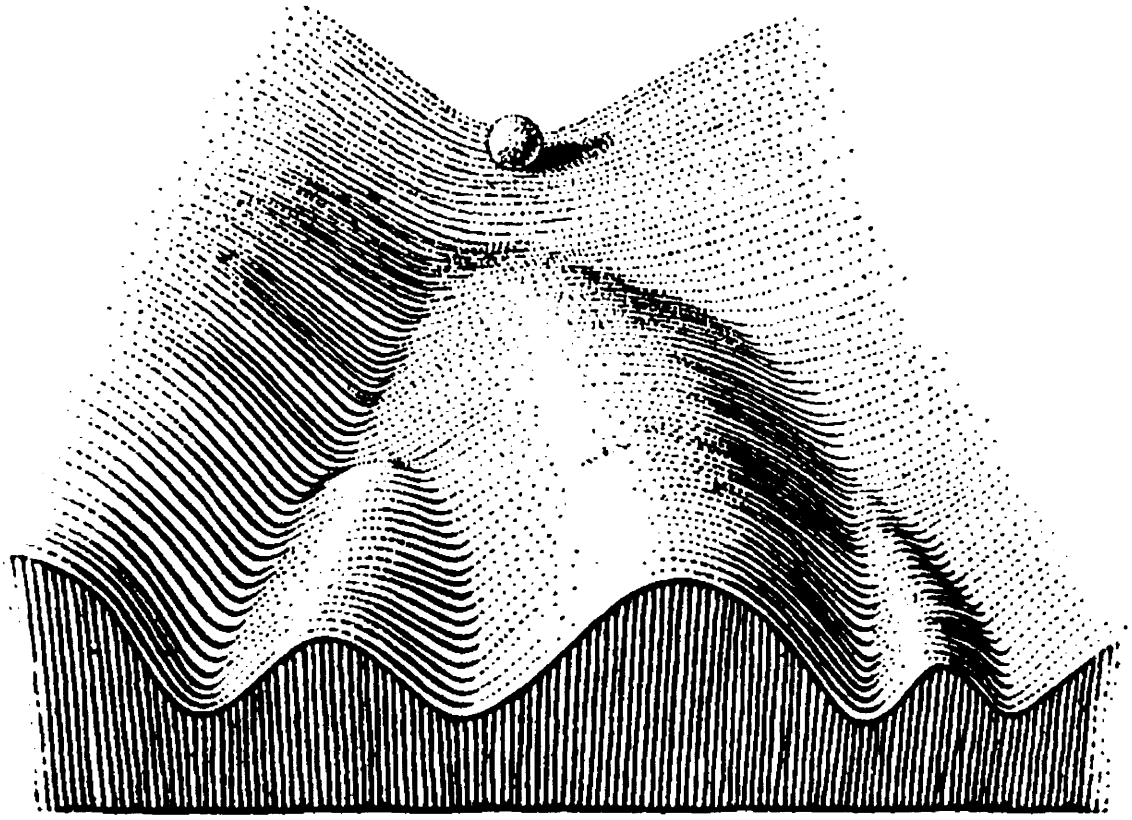


Figure 1. Waddington's epigenetic landscape

A pluripotent stem cell is positioned at the top of the landscape, poised to begin traveling towards a differentiated cell type. As the cell differentiates, it loses those properties of pluripotency and gains the features of a differentiated cell type. (Waddington 1957)

The TFs Oct4 and Nanog were originally identified as master regulators of ES cells because, they are homeobox TFs and because they are expressed at a high level specifically in ES cells and not in other cell types (Scholer et al. 1989a; Scholer et al. 1989b; Nichols et al. 1998). We reasoned that it would be possible to apply a systems biology approach to predict the master regulator TFs for any cell type. We developed a method that will identify candidate master regulator TFs for any cell type based on two properties of master regulator TFs. First, master regulator TFs contain protein domains, such as homeobox and helix-loop-helix domains, which are characteristic of TFs that regulate cell type during development. Second, master regulator TFs are expressed at high levels in the cell types that they specify, and are expressed at low levels or not at all in other cell types. This approach successfully identifies known master regulator TFs and can identify candidate master regulator TFs for any cell type. The algorithm is described in detail below.

First, a list of the protein domains that are characteristic of master regulator TFs was compiled. This was done by identifying all of the genes that are annotated in the Gene Ontology (GO) database (http://www.geneontology.org/) as transcription factors (GO:0003702: RNA polymerase II transcription factor activity) and as participating in development (GO:0032502: developmental process). The PFAM protein domains (http://pfam.sanger.ac.uk/) occurring in these genes were tabulated and then compared to background frequencies of protein domains among all genes in the genome. The protein domains occurring significantly more frequently among the developmental TFs were identified. Subsequently, all genes were examined for the presence of any of these protein domains. This approach identifies a set of 1,088 human and 1,023 mouse genes that contain protein domains characteristic of master regulator TFs.

The second part of the algorithm for identifying candidate master regulator TFs relies on finding genes that are expressed in a highly cell type specific manner. These genes are expressed at high levels in the cells of interest, relative to the full range of expression values observed across all cell types. To determine the range of expression values for each gene we utilized the wealth of publically available gene expression data from the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/) (Edgar et al. 2002). The GEO database contains more than 46,000 human gene expression samples from the Affymetrix U133 Plus 2.0 platform (GPL570) and more than 19,000 mouse gene expression samples from the Affymetrix 430 2.0

platform (GPL1261). These samples provide a rich background that allows accurate determination of the range of possible expression values for each gene across al.l cell types.

In order to accurately compare expression signal values from the large variety of samples in the GEO database it is critical to normalize the samples, so that, for a given gene, an expression signal has the same meaning for each sample. This was accomplished using a quantile/rank normalization approach. For the GPL570 and GPL1261 microarray platforms, the complete set of expression data was downloaded from the GEO database. Samples that were missing data were discarded and control probesets were discarded. Within each sample, the gene with the greatest expression signal was identified. The mean of these values was calculated, and the gene with the greatest signal in each sample was assigned this mean value. This was repeated for all genes from the greatest signal to the least, assigning each the average expression signal for all genes of that rank across all samples. Following normalization of the full GEO datasets, the range of expression values was examined and the 20th and 80^{th} percentile expression signals for each gene were recorded.

In reality, cell type specificity is a continuum, not a binary property of genes. We created a heuristic formula to score the degree of cell type specificity of a given gene in an expression dataset. This score is calculated as follows:

$$\text{Cell Type Specificity Score} = (\text{Expression Signal} - p20) / (p80 - p20)$$

$p80 = 80^{th}$ percentile expression signal

$p20 = 20^{th}$ percentile expression signal

For each gene in an expression dataset, the Cell Type Specificity (CTS) Score is calculated and all genes in the dataset are ranked from greatest to least CTS Score. The genes are then filtered to retain those containing protein domains that are characteristic of TFs that control cell type during development.

In order to test the ability of the CTS algorithm to predict master regulator TFs we required a set of positive controls. To do this we needed to determine precise criteria that would define a set of known master regulator TFs and the cell types that they specify. Though many criteria could be used, the ability to promote trans-differentiation from one cell type to another is a relatively stringent definition of a master regulator TF. In addition to the example of reprogramming to iPS cells using several different set of TFs including Oct4, Sox2, Klf4, c-Myc, Nanog and Esrrb (Takahashi and Yamanaka 2006; Takahashi et al. 2007; Wernig et al. 2007; Yu

et al. 2007; Feng et al. 2009), several other examples of TF mediated trans-differentiation have been described. The TF Myod1 promotes trans-differentiation of many cell types into myoblast cells (Davis et al. 1987; Weintraub 1993; Berkes and Tapscott 2005). The TF PPAR-gamma-2 causes trans-differentiation of fibroblasts into adipocyte cells (Tontonoz et al. 1994; Hu et al. 1995). The TFs PU.1 and CEBP-alpha or CEBP-beta cause trans-differentiation of fibroblast cells into macrophage cells (Feng et al. 2008). Finally, the TF Foxp3 causes conventional helper T cells to assume a suppressive phenotype characteristic of regulatory T cells. The CTS method does an excellent job of predicting most of the known TF that mediate trans-differentiation. For ES cells, Oct4, Sox2, Nanog, and Esrrb are all in the top 10 predicted TFs. Klf4 and c-Myc are not predicted as master regulators of ES cells, though n-Myc is. For myoblast cells, MyoD1 is the 2rd predicted TF. For adipocyte cells, PPAR-gamma-2 is the 3rd predicted TF. For macrophage cells, PU.1 is the 4th predicted TF, but neither CEBP-alpha nor CEBP-beta is predicted. Lastly, Foxp3 in the 3rd predicted TF in regulatory T cells.

One of the weaknesses of microarray gene expression data is that the expression values for different genes are not directly comparable. RNA species have different efficiencies in microarray sample preparation and each probe on a microarray has a different hybridization efficiency and background signal. Furthermore, mRNAs have different translational efficiencies. Consequently, there is a poor correlation between the expression signal of a gene in a microarray experiment and the amount of that protein that is present in a cell. An expression signal of 1,000 units for one gene does not have the same meaning as an expression signal of 1,000 for a different gene. As a result, in expression data analysis, the expression level of a gene can be meaningfully compared across samples, but it is generally not meaningful to compare the expression levels of different genes within one sample. One of the reasons why the CTS approach is so successful is that it transforms all of the expression values for different genes in one expression sample onto the same scale. Within this scale, different genes can be directly compared to each other. Normalization based on the range of observed expression values for each gene across the GEO dataset provides context to interpret absolute expression values.

In the future there are several elements that I plan to add to this project prior to publication. First, I will create a web-tool, hosted on the Whitehead Institute website, which will allow biology researches around the world to apply the Cell Type Specificity algorithm to their datasets of interest. Users would input GEO database accession numbers and the web-tool

would download those datasets from GEO and subject them to CTS analysis with no additional user input necessary. I also plan to allow users to upload and analyze personal datasets that are not in the GEO database.

The Cell Type Specificity web-tool will have significant utility beyond the ability to predict master regulator TFs. The CTS algorithm can also be used to identify other classes of genes, such as signaling components, chromatin regulators, or any other subset of genes, that define a cell type of interest. I am working to define precisely which genes belong to these additional classes beyond transcription factors. Additionally, the web-tool will output CTS scores for all genes in an expression dataset without filtering. Cell type specific genes provide significant of information about a cell type of interest. I plan to find a few strong examples of how identifying cells with highly cell type specific gene expression patterns can give important information to biology researchers outside the field of transcriptional regulation. Finally, the website will be a powerful tool for gene expression data mining, retrieval and normalization, even if users are not interested in conducting CTS analysis.

It has recently been reported that cell type specific genes tend to be targeted by the Mediator and Cohesion protein complexes (Kagey et al. 2010) and by H3K4me2 modified nucleosomes (Pekowska et al. 2010). I plan to test whether the CTS scoring metric is a good predictor of targeting by Mediator, Cohesin, and H3K4me2. In Chapter 1, I hypothesize that targeting cell type specific genes is a general property of master regulator TFs. I plan to test this hypothesis by examining existing location analysis datasets profiling the occupancy of Oct4, Sox, and Nanog in ES cells (Marson et al. 2008), the occupancy of MyoD1 in myoblast cells (Cao et al. 2010), the occupancy of PU.1 in hematopoietic cells (Novershtern et al. 2010), the occupancy of PPAR-gamma in adipocyte cells (Mikkelsen et al. 2010), and the occupancy of Foxp3 in regulatory T cells (A. Marson, personal communication).

Future Directions

It is not well understood why transcription factors occupy some sites in the genome, while other sites with the same DNA sequence are not occupied. For example, in ES cells, the DNA element that is bound by the master regulators Oct4, Sox2, and Nanog (OSN) is well characterized. This 16 basepair DNA motif (Figure 2) occurs at all high confidence ChIP-Seq binding sites for OSN and is significantly more conserved than surrounding sequences (Marson et al. 2008). We do not

currently understand why, however, only a small fraction of the instances of this DNA element genome-wide are occupied by OSN. It is believed that there are additional DNA elements surrounding the core OSN motifs that are occupied by additional TFs, which physically interact with OSN. These additional TF binding sites could theoretically add information specifying which instances of the OSN motif should be occupied, but no other DNA elements that specify the occupancy of the OSN TFs have been described. Whether this phenomenon is truly important for specifying the genome-wide occupancy of OSN has not been investigated. A genome-wide examination of whether there exist additional DNA elements, beyond the core OSN motif, that specify the occupancy of OSN could help to clarify why only some sites in the genome are occupied by master regulator TFs. Alternatively, it is possible that occupancy of OSN in ES cells is specified by an epigenetic mark whose genomic occupancy is not ultimately dependent on the underlying DNA sequence. How this process could occur however, has not been described.

Figure 2

Figure 2. The DNA motif that is bound by OCT4, SOX2, and NANOG in ES cells

One hundred basepairs of genomic sequence, centered at the five hundred largest OCT4 ChIP-Seq peaks (Marson et al. 2008), were submitted to the motif discovery tool MEME (Bailey and Elkan 1995; Bailey et al. 2006) to search for over-represented DNA motifs. A single sixteen basepair motif was discovered by the MEME algorithm. This motif was significantly ($p < 10^{-100}$) over-represented, occurring in 92% of the Oct4 bound sequences. The figure image was generated using the online tool WebLogo (http://weblogo.berkeley.edu/).

Much of the current knowledge of the role of transcriptional regulatory circuitry in control of cell type has been obtained in the context of a static cell state. It is important to remember, however, that during development, the differentiation of cells from one cell type to another is fundamentally important. Consequently, we would like to better understand how transcriptional regulatory circuitry works as a cell transitions from one type to another. In what ways are cells primed to shut down existing circuitry and how are new circuitry elements turned on? Is normal development always unidirectional? What are the logic circuits describing cell fate decisions for all of human development? Though most of these questions are unanswered, great progress in this area is likely to occur over the next few years, and a few general principles have emerged from existing research. Master regulator TFs sometimes occupy sites in progenitor cells that will be important for later cell types (Novershtern et al. 2010). Additionally, master regulator TFs sometimes co-occupy the genome with co-factors that are required for terminating the action of the master regulators upon differentiation (Whyte et al. 2010).

It has become increasingly evident that non-coding RNAs (ncRNA) play an important role in a diverse array of cellular processes including genomic imprinting, heterochromatin formation (Buhler and Moazed 2007), recruitment of histone modifying enzymes (Rinn et al. 2007), translational inhibition, and degradation of mRNAs (Bartel 2009). There are a number of reasons why the importance of ncRNAs may have been underappreciated until recently. First, RNA molecules are generally less chemically stable than DNA or polypeptides, so species of interest are harder to detect. Additionally, the complexity of RNA species that exist within a cell is much greater than the complexity of DNA or protein species due to the huge set of possible transcription start and end sites in combination with the RNA splicing and editing. Finally, the evolutionary pressures acting on many ncRNAs may not require the maintenance of the RNA sequence, so the signal of evolutionary conservation is hard to identify. Bioinformatic analysis in mammals and in *Drosophilae* has revealed that DNA elements can be identified based on sequence conservation (Xie et al. 2005; Stark et al. 2007). A similar approach focused on ncRNAs could reveal functional RNA secondary structures that occur more frequently and are more highly conserved than would be expected by chance.

Acknowledgments

I wish to thank members of the Young lab, especially R.A. Young, T.I. Lee, and J. Reddy for helpful comments during the preparation of this chapter. Figure 2 was obtained from the referenced manuscript.

References

Bailey, T. L. and C. Elkan (1995). Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Machine Learning *21*, 51-80.

Bailey, T. L., N. Williams, C. Misleh and W. W. Li (2006). MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Research *34*, W369-W373.

Bartel, D. P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. Cell *136*, 215-233.

Berkes, C. A. and S. J. Tapscott (2005). MyoD and the transcriptional control of myogenesis. Seminars in Cell & Developmental Biology *16*, 585-595.

Buhler, M. and D. Moazed (2007). Transcription and RNAi in heterochromatic gene silencing. Nature Structural & Molecular Biology *14*, 1041-1048.

Cao, Y., Z. Z. Yao, D. Sarkar, M. Lawrence, G. J. Sanchez, M. H. Parker, K. L. MacQuarrie, J. Davison, M. T. Morgan, W. L. Ruzzo, et al. (2010). Genome-wide MyoD Binding in Skeletal Muscle Cells: A Potential for Broad Cellular Reprogramming. Developmental Cell *18*, 662-674.

Davis, R. L., H. Weintraub and A. B. Lassar (1987). Expression of a single transfected cDNA converts fibroblasts to myoblasts. Cell *51*, 987-1000.

Edgar, R., M. Domrachev and A. E. Lash (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research *30*, 207-210.

Feng, B., J. M. Jiang, P. Kraus, J. H. Ng, J. C. D. Heng, Y. S. Chan, L. P. Yaw, W. W. Zhang, Y. H. Loh, J. Y. Han, et al. (2009). Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb. Nature Cell Biology *11*, 197-U193.

Feng, R., S. C. Desbordes, H. F. Xie, E. S. Tillo, F. Pixley, E. R. Stanley and T. Graf (2008). PUA and C/EBP alpha/beta convert fibroblasts into macrophage-like cells. Proceedings of the National Academy of Sciences of the United States of America *105*, 6057-6062.

Hu, E. D., P. Tontonoz and B. M. Spiegelman (1995). Transdifferentiation of myoblasts by the adipogenic transcription factors PPAR gamma and C/EBP alpha. Proceedings of the National Academy of Sciences of the United States of America *92*, 9856-9860.

Kagey, M. H., J. J. Newman, S. Bilodeau, Y. Zhan, D. A. Orlando, N. L. van Berkum, C. C. Ebmeier, J. Goossens, P. B. Rahl, S. S. Levine, et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430-435.

Marson, A., S. S. Levine, M. F. Cole, G. M. Frampton, T. Brambrink, S. Johnstone, M. G. Guenther, W. K. Johnston, M. Wernig, J. Newman, et al. (2008). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell *134*, 521-533.

Mikkelsen, T. S., Z. Xu, X. L. Zhang, L. Wang, J. M. Gimble, E. S. Lander and E. D. Rosen (2010). Comparative Epigenomic Analysis of Murine and Human Adipogenesis. Cell *143*, 156-169.

Nichols, J., B. Zevnik, K. Anastassiadis, H. Niwa, D. Klewe-Nebenius, I. Chambers, H. Scholer and A. Smith (1998). Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. Cell *95*, 379-391.

Novershtern, N., A. Subramanian, L. N. Lawton, R. H. Mak, W. N. Haining, M. E. McConkey, N. Habib, N. Yosef, C. Y. Chang, T. Shay, et al. (2010). The Transcriptional Architecture of Human Hematopoiesis: Tightly Integrated Circuits and New Transcriptional Regulators Controlling Cell States. Cell (in review).

Pekowska, A., T. Benoukraf, P. Ferrier and S. Spicuglia (2010). A unique H3K4me2 profile marks tissue-specific gene regulation. Genome Res *20*, 1493-502.

Rinn, J. L., M. Kertesz, J. K. Wang, S. L. Squazzo, X. Xu, S. A. Brugmann, L. H. Goodnough, J. A. Helms, P. J. Farnham, E. Segal, et al. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by Noncoding RNAs. Cell *129*, 1311-1323.

Scholer, H. R., R. Balling, A. K. Hatzopoulos, N. Suzuki and P. Gruss (1989a). Octamer binding-proteins confer transcriptional activity in easrly mouse embryogenesis. Embo Journal *8*, 2551-2557.

Scholer, H. R., A. K. Hatzopoulos, R. Balling, N. Suzuki and P. Gruss (1989b). A family of octamer-specific proteins present during mouse embryogenesis - Evidence for germline-specific expression of an OCT factor. Embo Journal *8*, 2543-2550.

Segal, E., M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller and N. Friedman (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature Genetics *34*, 166-176.

Stark, A., M. F. Lin, P. Kheradpour, J. S. Pedersen, L. Parts, J. W. Carlson, M. A. Crosby, M. D. Rasmussen, S. Roy, A. N. Deoras, et al. (2007). Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature *450*, 219-232.

Takahashi, K., K. Tanabe, M. Ohnuki, M. Narita, T. Ichisaka, K. Tomoda and S. Yamanaka (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell *131*, 861-872.

Takahashi, K. and S. Yamanaka (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663-676.

Tontonoz, P., E. D. Hu and B. M. Spiegelman (1994). Stimluation of adipogenesis in fibroblasts by PPAR-gamma-2, a lipid-activated transcription factor. Cell *79*, 1147-1156.

Waddington, C. H. (1957). The Strategy of the Genes. London : George Allen & Unwin.

Weintraub, H. (1993). The MyoD family and myogenesis: redundancy, networks, and thresholds. Cell *75*, 1241-1244.

Wernig, M., A. Meissner, R. Foreman, T. Brambrink, M. C. Ku, K. Hochedlinger, B. E. Bernstein and R. Jaenisch (2007). In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. Nature *448*, 318-U2.

Whyte, W. A., S. Bilodeau, G. M. Frampton, D. A. Orlando and R. A. Young (2010). Enhancer Decommissioning by LSD1 During Embryonic Stem Cell Differentiation. Nature (in review).

Xie, X. H., J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander and M. Kellis (2005). Systematic discovery of regulatory motifs in human promoters and 3 ' UTRs by comparison of several mammals. Nature *434*, 338-345.

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

Appendix A

Supplemental Material for Chapter 2

Chromatin Structure and Gene Expression Programs of Human Embryonic and Induced Pluripotent Stem Cells

Supplemental Experimental Procedures

ES, iPS, and fibroblast cells and cell culture

All primary human fibroblasts cells described in this paper (PDB-AG20442 and GM-M01660) were purchased from the Coriell Cell Repository (Camden, NJ). Fibroblasts were cultured in fibroblast medium (Dulbecco's modified Eagle's medium [DMEM] supplemented with 15% fetal bovine serum [FBS; Hyclone], 1 mM glutamine [Invitrogen], 1% nonessential amino acids [Invitrogen], and penicillin/streptomycin [Invitrogen]).

hiPS cell lines iPS A1, iPS C1, iPS4, iPS A6 (Hockemeyer et al. 2009); hiPS cell lines iPS PDB^{2lox}-17, iPS PDB^{2lox}-21, iPS PDB^{2lox}-5, iPS PDB^{2lox}-22, iPS PDB^{1lox}-17puro-5, iPS PDB^{1lox}-17puro-10, iPS PDB^{1lox}-17puro-33, iPS PDB^{1lox}-21puro-20, iPS PDB^{1lox}-21puro-26, and iPS PDB^{1lox}-21puro-28 (Soldner et al. 2009); hES cell lines BG01 and BG03 (National Institutes of Health code: BG01 and BG03; BresaGen, Inc., Athens, GA); hES cell lines WIBR1, WIBR2, WIBR3, and WIBR7 (Lengner et al., 2010; Whitehead Institute Center for Human Stem Cell Research) and hES cell line H9 (NIH Code:WA09, Wisconsin Alumni Research Foundation, Madison, WI) were maintained on mitomycin C-inactivated mouse embryonic fibroblast (MEF) feeder layers in hESC medium (DMEM/F12 [Invitrogen] supplemented with 15% FBS [Hyclone], 5% KnockOut Serum Replacement [Invitrogen], 1 mM glutamine [Invitrogen], 1% nonessential amino acids [Invitrogen], 0.1 mM β-mercaptoethanol [Sigma], and 4 ng/ml FGF2 [R&D Systems]). Cultures were passaged every 5 to 7 days either manually or enzymatically with collagenase type IV (Invitrogen; 1.5 mg/ml). hiPS cell lines were passaged 15-25 times prior to ChIP-Seq and gene expression analysis.

ChIP-Seq Experiments and Analysis

Chromatin immunoprecipitation

Protocols describing chromatin immunoprecipitation (ChIP) materials and methods can be downloaded from http://web.wi.mit.edu/young/hES_PRC and have previously been described in detail (Lee et al. 2006).

Human ES, iPS or fibroblast cells were grown to a final count of ~$5x10^7$ cells to obtain starting material for six chromatin immunoprecipitations. Cells were chemically cross-linked by the addition of one-tenth volume of fresh 11% formaldehyde solution for 15 minutes at room temperature. Cells were rinsed twice with 1X PBS, harvested by centrifugation, and flash frozen in liquid nitrogen. Cross-linked cells were stored at –80°C prior to use.

Cells were re-suspended, lysed and sonicated to solubilize and shear cross-linked DNA. Sonication was performed using a Misonix Sonicator 3000 at a power of 27W for ten 30 second pulses with a 90 second pause between each pulse. Samples were kept on ice at all times.

The resulting whole cell extract was incubated overnight at 4 degrees C with 10µl of Dynal Protein G magnetic beads that had been pre-incubated with approximately 3 µg of the appropriate antibody. Each individual immunoprecipitation used 1/6 of the 3ml total, or ~$8 x10^6$ cells per IP. The immunoprecipitation was allowed to proceed overnight. Beads were washed three times (3 x 1.5ml) with RIPA buffer and one time (1x 1.5ml) with TE containing 50 mM NaCl. Bound complexes were eluted from the beads by heating at 65 degrees C with occasional vortexing and cross-linking was reversed by overnight incubation at 65 degrees C. Whole cell extract DNA (reserved from the sonication step) was also treated for cross-link reversal. Immunoprecipitated DNA and whole cell extract DNA were then purified by treatment with RNAse A, proteinase K and two phenol:chloroform:isoamyl alcohol extractions.

The ChIP antibodies used were ab8580 (Abcam) for H3K4me3 and ab6002 (Abcam) for H3K27me3.

ChIP-Seq sample preparation

All protocols for Solexa sample preparation and sequencing are provided by Illumina (http://www.illumina.com/). A brief summary of the technique, minor protocol modifications, and data analysis methods are described below.

Purified ChIP DNA was prepared for sequencing according to a modified version of the Illumina/Solexa Genomic DNA protocol. Approximately 50-200ng of IP DNA was prepared for ligation of Solexa linkers by repairing the ends and adding a single adenine nucleotide overhang to allow for directional ligation. A 1:100 dilution of the Adaptor Oligo Mix (Illumina) was used in the ligation step. A subsequent PCR step with 18 amplification cycles added additional linker sequence to the fragments to prepare them for annealing to the Genome Analyzer flow-cell.

Amplified material was purified by Qiaquick MinElute (Qaigen) and a narrow range of fragment sizes was selected by separation on a 2% agarose gel and excision of a band between 150-300 bp, representing IP fragments between 50 and 200nt in length and ~100bp of primer sequence. The DNA was purified from the agarose and diluted to 10 nM for loading on the flow cell.

Solexa sequencing

The DNA library (2-4 pM) was applied to one lane of the flow-cell (eight samples per flow-cell) using a Cluster Station device (Illumina). The concentration of library applied to the flow-cell was calibrated so that polonies generated in the bridge amplification step originate from single strands of DNA. Multiple rounds of amplification reagents were flowed across the cell in the bridge amplification step to generate polonies of approximately 1,000 strands in 1μm diameter spots. Double stranded polonies were visually checked for density and morphology by staining with a 1:5000 dilution of SYBR Green I (Invitrogen) and visualizing with a microscope under fluorescent illumination. Validated flow-cells were stored at 4 degrees C until sequencing.

Flow-cells were removed from storage and subjected to linearization and annealing of sequencing primer on the Cluster Station. Primed flow-cells were loaded into the Genome Analyzer 1G (Illumina). After the first base was incorporated in the sequencing-by-synthesis reaction the process was paused for a key quality control checkpoint. A small section of each lane was imaged and the average intensity value for all four bases was compared to minimum thresholds. Flow-cells with low first base intensities were re-primed and if signal was not recovered the flow-cell was aborted. Flow-cells with signal intensities meeting the minimum thresholds were resumed and sequenced.

Images acquired from the Genome Analyzer were processed through the bundled image extraction pipeline (Illumina), which identified polony positions, performed base-calling and generated QC statistics.

Sequencing of the H3K27me3 ChIP from the hES BG03 cell line failed several quality control metrics and this sample not used for analysis except for gene track comparisons and the profile shown in Figure 1F.

Genomic mapping of ChIP-Seq data

ChIP-Seq reads were aligned using the software Bowtie (Langmead et al., 2009) to NCBI build 36.1 (hg18) of the human genome with default settings. Sequences uniquely mapping to the genome with zero or one mismatch were used in further analysis.

Public availability of ChIP-Seq data

Complete ChIP-Seq data are available from the Gene Expression Omnibus database (http://www.ncbi.nih.gov/geo/) under the accession number GSE22499.

ChIP-Seq density calculation and normalization of ChIP-Seq samples

The analysis methods used were derived from previously published methods (Barski et al., 2007; Johnson et al., 2007; Mikkelsen et al., 2007; Robertson et al., 2007). The genome was divided into bins 100 base pairs in width, beginning at the first base of each chromosome. For identification of genomic regions with statistically significant differential ChIP-Seq occupancy 250 bp bins were used due to computer memory constraints. Each ChIP-Seq read was shifted 100 bp from its mapped genomic position and strand to the approximate middle of the sequenced DNA fragment. The ChIP-Seq density within each genomic bin was then calculated as the number of ChIP-Seq reads mapping within a 1kb window (+/- 500bp) surrounding the middle of that genomic bin.

In order to facilitate comparison of ChIP-Seq samples a quantile normalization method was used. In each ChIP-Seq sample the genomic bin with the greatest ChIP-Seq density was identified. The mean of these values was calculated and the bin with the greatest signal in each sample was assigned this mean value. This was repeated for all genomic bins from the greatest signal to the least, assigning each the average ChIP-Seq signal for all bins of that rank across all samples. H3K4me3 and H3K27me3 samples were subjected to quantile normalization as separate groups.

Identification of ChIP enriched genomic regions and genes

Genomic bins with a normalized ChIP-Seq density greater than a defined threshold were considered enriched. Adjacent enriched bins were combined into enriched regions. For H3K4me3 a threshold of 30 normalized reads per kb and for H3K27me3 a threshold of 25

normalized reads per kb was used. A summary of the H3K4me3 and H3K27me3 occupied regions is provided in Table S1.

The genomic coordinates of the full set of transcripts from the RefSeq database (http://www.ncbi.nlm.nih.gov/RefSeq/) from the March 2006 version of the human genome sequence (NCBI Build 36.1, hg18) was downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTables) on March 1, 2009. Genes were associated with H3K4me3 and H3K27me3 occupied genomic regions if the gene transcription start site (TSS) occurred within the region or if the distance from the TSS to the boundary of the region was less than or equal to 2 kb. If multiple regions were associated with a single gene, all of these gene are reported the region with the greatest peak ChIP-Seq density used. A summary of the genes associated with H3K4me3 and H3K27me3 occupied regions is provided in Table S1.

Pairwise comparisons of H3K4me3 and H3K27me3 occupied genes and peak heights

For each RefSeq gene the peak normalized ChIP-Seq density in the region from -2 kb to +2 kb of the transcription start site was examined. A gene was considered to have different ChIP-Seq occupancy between two cell lines for H3K4me3 if the peak signal at the transcription start site (+/- 2 kb) was greater than or equal to 30 units in one cell line and less than 20 units in the other cell lines. A gene was considered to have different ChIP-Seq occupancy between two cell lines for H3K27me3 if the peak signal at the transcription start site (+/- 2kb) was greater than or equal to 25 units in one cell line and less than 15 units in the other cell line. The percentage of all RefSeq genes with different ChIP-Seq occupancy between the two samples was reported in Table S2.

To compare H3K4me3 or H3K27me3 peak heights between two samples the peak ChIP-Seq density in the region from -2 kb to +2 kb of the transcription start site was examined. For each gene that was occupied by that histone mark in at least one of the two samples the coefficient of variation was recorded. The average coefficient of variation of peak heights between the two samples was reported in Table S2.

ChIP-Seq density heatmaps and composite ChIP-Seq density profiles

For ChIP-Seq density heatmaps, genes were aligned using the position and direction of their transcription start sites. Heatmaps showing the H3K4me3 and H3K27me3 ChIP-Seq density

around gene start sites (-4,500 bp to +4,500 bp) within 500bp bins were generated using Java Treeview (http://jtreeview.sourceforge.net/).

For composite ChIP-Seq density profiles, genes were aligned using the position and direction of their transcription start sites. The average ChIP-Seq density around the transcription start sites of all genes in 500 bp bins was calculated for ES cells and iPS cells.

Statistical method for identifying genomic regions with differential ChIP-Seq occupancy

A test statistic, the 'differential ChIP-Seq score' was created to quantify the degree of differential ChIP-Seq density at a given position in the genome for two groups of ChIP-Seq samples. For each set of four adjacent 250 bp genomic bins the differential ChIP-Seq score was calculated as the absolute value of the mean signal of the samples in group A (A) minus the mean signal in of the samples in the group B (B) divided by an estimate of the noise of these measurements ($NOISE_{AB}$), which is described in more detail below.

$$\text{Differential ChIP-Seq score} = |A - B| \; / \; NOISE_{AB}$$

The value of the $NOISE_{AB}$ was calculated using the following method. First, in each set of four adjacent 250 bp bins across the genome, the mean and standard deviation of the ChIP-Seq signal for all samples, rounded to the nearest integer, was tabulated. Second, for each mean ChIP-Seq signal, the median standard deviation (stdev) was recorded. Third, a power function, predicting the noise in these ChIP-Seq datasets, was fit to this set of mean/stdev pairs. This function was of the form;

$$NOISE(MEAN) = x * MEAN\text{^}y$$

This power function provided a good representation of the dependence between the signal intensity and noise for these ChIP-Seq experiments across the full range of signal intensities, from zero to hundreds of reads per kilobase.

The value of $NOISE_{AB}$ was then calculated as the mean of the maximum of the value of the power function at the mean signal for the samples in group A, NOISE(A), and the actual

standard deviation of these measurements A and the maximum of NOISE(B) and the actual standard deviation of the measurements in group B.

$$NOISE_{AB} = \text{mean}(\max(\sigma_A, NOISE(A)), \max(\sigma_B, NOISE(B))$$

σ_A = STANDARD DEVIATION OF THE SIGNAL IS SAMPLES FROM GROUP A

σ_B = STANDARD DEVIATION OF THE SIGNAL IS SAMPLES FROM GROUP B

To assess the statistical significance of a given differential ChIP-Seq score, a permutation method was used. The distribution of differential ChIP-Seq scores under the null hypothesis was modeled by shuffling the sample to group assignments and re-calculating the test statistic. Based on the permuted sample/group assignments the differential ChIP-Seq scores were re-calculated and tabulated. All possible combinations of sample/group assignments were used to determine the null distribution except for the actual assignments and the inverse of the actual assignments.

Using this null distribution of differential ChIP-Seq scores, a false discovery rate (FDR) associated with any differential ChIP-Seq score could be calculated. This was the fraction of the genomic bins in the null distribution with that score or greater, (P_{NULL}) divided by the fraction in the actual distribution with that score or greater (P_{ACTUAL}).

$$FDR(\text{differential ChIP-Seq score}) = P_{NULL} / P_{ACTUAL}$$

A false discovery rate threshold of 5% was used for the identification of genomic regions with statistically significant differential ChIP-Seq occupancy between male and female cells and between ES and iPS cells and an FDR threshold of 1% was used for the comparison of pluripotent to fibroblast cells. Adjacent sets of genomic bins that were identified as differentially occupied were combined into regions.

Gene Expression Experiments and Analysis

Sample preparation, hybridization, staining, scanning, and image analysis

5 μg total RNA was used to prepare biotinylated cRNA according to the manufacturer's protocol (Affymetrix One Cycle cDNA Synthesis Kit). Briefly, this method involves SuperScript II-

directed reverse transcription using a T7-Oligo-dT promoter primer to create first strand cDNA. RNase H-mediated second strand cDNA synthesis is followed by T7 RNA Polymerase directed *in vitro* transcription, which incorporates a biotinylated nucleotide during cRNA amplification.

Samples were prepared for hybridization using 15 μg biotinylated cRNA in a 1X hybridization cocktail with additional hybridization cocktail components provided in the GeneChip Hybridization, Wash and Stain Kit (Affymetrix). GeneChip arrays (Human U133 Plus 2.0) were hybridized in a GeneChip Hybridization Oven at 45 degrees C for 16 hours at 60 RPM. Washing was performed using a GeneChip Fluidics Station 450 according to the manufacturer's instructions, using the buffers provided in the Affymetrix GeneChip Hybridization, Wash and Stain Kit. Arrays were scanned on a GeneChip Scanner 3000 and images were extracted and analyzed using the default settings of GeneChip Operating Software v1.4.

Public availability of gene expression data

Complete gene expression data are available from the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE22499.

Previously published gene expression datasets

Three previously published datasets comparing gene expression profiles of human ES, iPS and fibroblast cells, using the Affymetrix U133 Plus 2.0 (GPL570) microarray platform were obtained from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) database. For Chin et al. (Chin et al., 2009) data were obtained from the GEO accession numbers GSE9865 and GSE16654 and data from GSE16654 were subjected to an inverse (backwards) logarithmic (base 2) transformation to return them to the same linear scale as the other datasets. For Maherali et al. (Maherali et al., 2008) data were obtained from the GEO database accession GSE12390 and used with no additional processing. For Yu et al. (Yu et al., 2009) data were obtained from the GEO database accession GSE15148 and subjected to an inverse (backwards) logarithmic (base 2) transformation to return them to the same linear scale as the other datasets. Each expression dataset was normalized and analyzed for statistically significant differential expression separately using the methods described below except in Figure 4, where all datasets were normalized and clustered as one group.

Expression data normalization

The data from each gene expression sample were floored at zero and linearly scaled to a mean expression signal of 500 units. Then, all expression signal values were increased by ten units to force all signals to be greater than one unit in logarithmic space. Subsequently, within each dataset, expression signal values were quantile normalized by assigning each probeset the average signal intensity for all probesets of the same rank across all samples. Each expression dataset was normalized separately except in Figure 4, where all datasets were normalized as one group.

Expression data annotation and identification of differentially expressed transcripts

Probeset annotations were downloaded from the NetAffx database (http://www.affymetrix.com/analysis/index.affx) on October 1, 2009. Expression datasets were analyzed for statistically significant differential expression using the online NIA Array Analysis Tool (http://lgsun.grc.nia.nih.gov/ANOVA/). Expression data was transformed into log space by the webtool upon upload. All probesets were tested for differential expression using the following settings.

Threshold z-value to remove outliers: 10000
Error Model: Bayesian
Size of sliding window for averaging error variances: 500
Proportion of highest variance values to be removed: 0
Desirable degrees of freedom for Bayesian error model: 10
Number of permutations: 0

For identification of differentially expressed transcripts between ES and iPS cells an FDR threshold of 0.05 was used. For identification of differentially expressed transcripts between pluripotent cell lines and fibroblast cells an FDR threshold of 0.01 was used. We required that differentially expressed transcripts had at least a 1.5 fold change in signal intensity.

The probesets differentially expressed between human ES and iPS cells and between pluripotent and fibroblast cells in this study, Chin et al., Maherali et al., and Yu et al. are

provided in Table S5. The numbers of overlapping probesets and genes between these four datasets are also provided in Table S5.

Heatmap display and hierarchical clustering of expression data

For heat map display expression data was normalized using a formula similar to a Z-score with the following modifications. Instead of using the mean and standard deviation of all samples, the mean and standard deviation was calculated within each cell type. Then, the mean of the within cell type means, and the mean of the within cell type standard deviations was used for Z-score normalization. This served to create a balanced color range, which was not biased towards groups of samples greater numbers of expression samples.

Centroid linkage, centered correlation distance, hierarchical clustering was performed using the software Cluster 3.0 (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/). Heatmaps were generated using the software Java Treeview (http://jtreeview.sourceforge.net/). Cluster branches were flipped about tree nodes for optimal display.

RT-PCR validation of microarray based expression data

For RNA analysis, hES and hiPS colonies were mechanically isolated and pooled for RNA extraction. Total RNA was isolated from ES, iPS, and fibroblast cells using RNeasy MiniKit (Qiagen). One microgram of total RNA was reverse transcribed using the Invitrogen Superscript III First Strand Synthesis System with oligo-dT primers to produce cDNA. One microliter of cDNA (1/150 of cDNA synthesis reaction) was used for each individual quantitative PCR measurement. cDNA was amplified using TaqMan Pre-developed gene expression assays (20X mixture supplied by Applied Biosystems which included pre-optimized primers and probe;Applied Biosystems). Triplicate reactions were performed in a total of 20μl using Taqman universal PCR master mix in an Applied Biosciences 7500 Real Time PCR Thermocycler. The following probes were used to detect expression in each of four ES cell lines (BG03, WIBR2, WIBR1, WIBR3), four iPS cell lines (iPS 21, iPS C1, iPS 17, iPS A6), and two donor fibroblast cell lines (Fibroblast PDB, Fibroblast GM):

Positive control: POU5F1, Hs00999632_g1

Internal standard: GAPDH, Hs02786624_g1

Test (previously determined by Affymetrix expression array as differential in iPS vs ES cells in Chin et al.) SOX9 - Hs00165814_m1
Test (previously determined by Affymetrix expression array as differential in iPS vs ES cells in Chin et al.) CAT - Hs00156308_m1
Test (previously determined by Affymetrix expression array as differential in iPS vs ES cells in Chin et al.) FN1 - Hs01549980_g1
Test (previously determined by Affymetrix expression array as differential in iPS vs ES cells in Guenther et al.) PUS7L – Hs01094423_m1
Test (previously determined by Affymetrix expression array as differential in iPS vs ES cells in Chin et al.) BMPR2 - Hs00176148_m1
Test (previously determined by Affymetrix expression array as differential in iPS vs ES cells in Maherali et al.) IRX3 – Hs00735523_m1
Test (previously determined by Affymetrix expression array as differential in iPS vs ES cells in Yu et al.) GREM1 – Hs00171951_m1

Detection of abundance was determined by measuring the point during cycling when amplification could first be detected, rather than the endpoint of the 40 cycle reaction. This cycle threshold (Ct) value corresponds to the fractional cycle number where the florescent Taqman probe increases above a fixed threshold (Auto Ct) determined by the ABI Prism 7000 Sequence Detection System software. The measured Ct value was used to calculate the estimated transcripts present in the test sample using relative quantization to the average internal standard GAPDH. Average Ct was calculated for each condition, a "delta Ct" value calculated by subtracting control GAPDH Ct. Expression was calculated as relative to Fibroblast PDB line, with fibroblast PDB expression normalized to 1.

References

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Chin, M.H., Mason, M.J., Xie, W., Volinia, S., Singer, M., Peterson, C., Ambartsumyan, G.,

Aimiuwu, O., Richter, L., Zhang, J., *et al.* (2009). Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. Cell Stem Cell *5*, 111-123.

Ghosh, Z., Wilson, K.D., Wu, Y., Hu, S., Quertermous, T., Wu, J.C. (2010) Persistent donor cell gene expression among human induced pluripotent stem cells contributes to differences with human embryonic stem cells. PLoS One. 5(2):e8975.

Hockemeyer, D., Soldner, F., Cook, E.G., Gao, Q., Mitalipova, M., Jaenisch, R. (2008). A drug-inducible system for direct reprogramming of human somatic cells to pluripotency. Cell Stem Cell *3*, 346-353.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497-1502.

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 Biol *10*, R25.

Lee, T.I., Johnstone, S.E., and Young, R.A. (2006). Chromatin immunoprecipitation and microarray-based analysis of protein location. Nat Protoc *1*, 729-748.

Lengner, C.J., Gimelbrant, A.A., Cheung, W.A., Erwin, J.a., Guenther, M.G., Alagappan, R., Xu, P., Powers, D., Barrett, B.C., Young, R.A., *et al.* (2010). Derivation of pre-X inactivation human embryonic stem cells under physiological oxygen concentrations. Cell *141*, 872-883.

Maherali, N., Ahfeldt, T., Rigamonti, A., Utikal, J., Cowan, C., and Hochedlinger, K. (2008). A high-efficiency system for the generation and study of human induced pluripotent stem cells. Cell Stem Cell *3*, 340-345.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P., *et al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-560.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., *et al.* (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods *4*, 651-657.

Sharov, A.A., Dudekula, D.B., Ko, M.S. (2005) A web-based tool for principal component and significance analysis of microarray data. Bioinformatics. (10):2548-9.

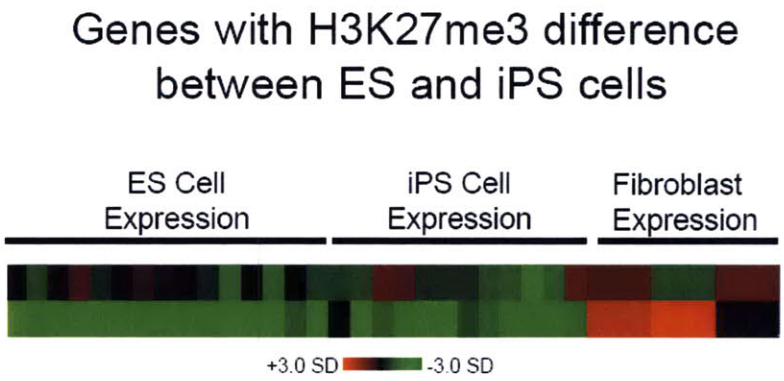
Soldner, F., Hockemeyer, D., Beard, C., Gao, Q., Bell, G.W., Cook, E.G., Hargus, G., Blak, A., Cooper, O., Mitalipova, M., *et al.* (2009). Parkinson's disease patient-derived induced pluripotent stem cells free of viral reprogramming factors. Cell *136*, 964-977.

Yu, J., Hu, K., Smuga-Otto, K., Tian, S., Stewart, R., Slukvin, II, and Thomson, J.A. (2009). Human induced pluripotent stem cells free of vector and transgene sequences. Science *324*, 797-

801.

Figure S1

A.


Genes with H3K27me3 difference between ES and iPS cells
ES Cell Expression
iPS Cell Expression
Fibroblast Expression
+3.0 SD
-3.0 SD

B.

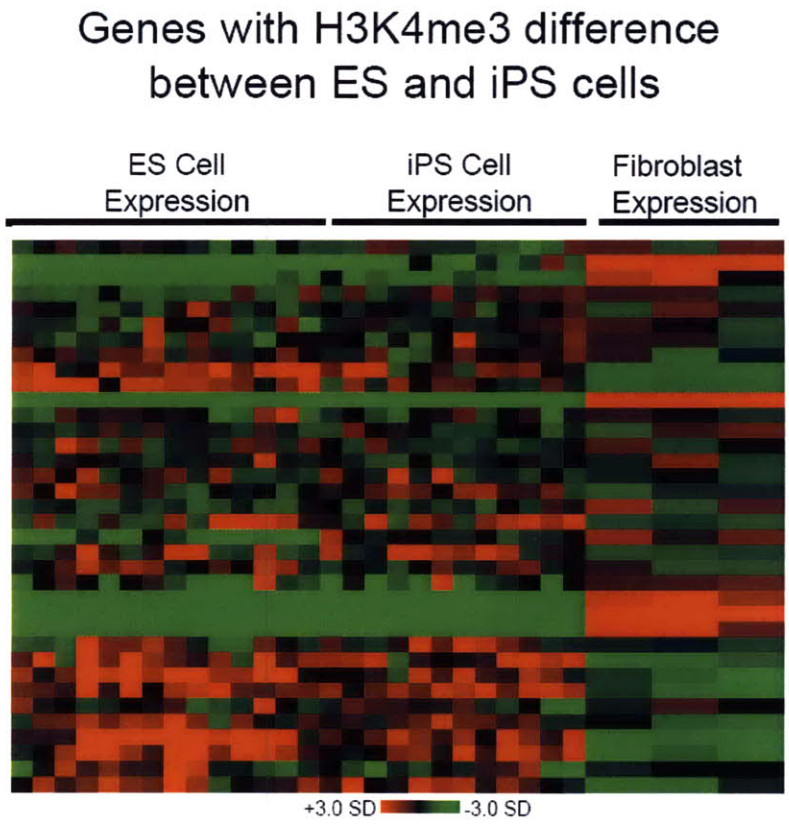

Genes with H3K4me3 difference between ES and iPS cells
ES Cell Expression
iPS Cell Expression
Fibroblast Expression
+3.0 SD
-3.0 SD

Figure S1. Expression of genes associated with genomic regions having differential H3K4me3 and H3K27me3 occupancy between ES and iPS cells

A. Expression data for genes differentially occupied by H3K27me3 between ES and iPS cells. Genes are ordered by the magnitude of differential H3K27me3 occupancy. Samples with higher than average expression are shown in red and samples with lower than average expression are shown in green (scale in standard deviations).

B. Expression data for genes differentially occupied by H3K4me3 between ES and iPS cells. Genes are ordered by the magnitude of differential H3K4me3 occupancy. Samples with higher than average expression are shown in red and samples with lower than average expression are shown in green (scale in standard deviations). See also main text Figure 3.

Figure S2

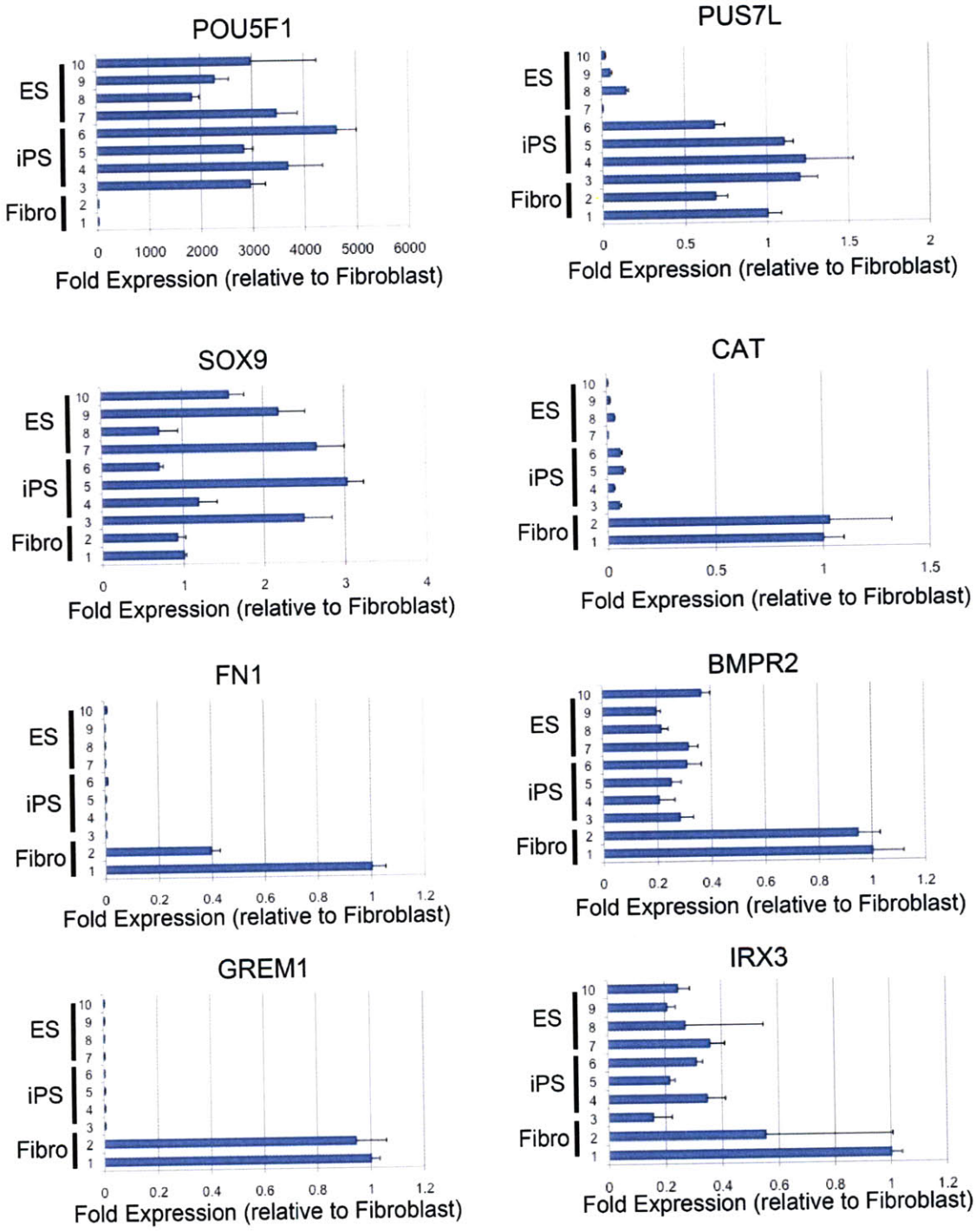

POU5F1
ES
iPS
Fibro
10 9 8 7 6 5 4 3 2 1
0 1000 2000 3000 4000 5000 6000
Fold Expression (relative to Fibroblast)
PUS7L
0 0.5 1 1.5 2
Fold Expression (relative to Fibroblast)
SOX9
0 1 2 3 4
Fold Expression (relative to Fibroblast)
CAT
0 0.5 1 1.5
Fold Expression (relative to Fibroblast)
FN1
0 0.2 0.4 0.6 0.8 1 1.2
Fold Expression (relative to Fibroblast)
BMPR2
0 0.2 0.4 0.6 0.8 1 1.2
Fold Expression (relative to Fibroblast)
GREM1
0 0.2 0.4 0.6 0.8 1 1.2
Fold Expression (relative to Fibroblast)
IRX3
0 0.2 0.4 0.6 0.8 1 1.2
Fold Expression (relative to Fibroblast)

Figure S2. RT-PCR based validation of genes identified as differentially expressed by microarray data

Expression of genes (relative to expression level in fibroblast cells) in ES and iPS cells. Triplicate reactions were performed using Taqman universal expression probes. All samples were normalized to Internal standard GAPDH gene. cDNA from the cell lines 1) Fibroblast PDB, 2) Fibroblast GM, 3) iPS PDB^{2lox}-21, 4) iPS C1, 5) iPS PDB^{2lox}-17, 6) iPS A6, 7) BG03, 8) ES WIBR2, 9) ES WIBR1, 10) ES WIBR3 was used. The Taqman probes used were POU5F1 (Positive control); SOX9 (previously determined by Affymetrix expression array as differential in iPS vs ES cells in Chin et al.); CAT (identified as differentially expressed in iPS vs ES cells in Chin et al.); FN1 (identified as differentially expressed in iPS vs ES cells in Chin et al.); PUS7L (identified as differentially expressed in iPS vs ES cells in Guenther et al.) BMPR2 (identified as differentially expressed in iPS vs ES cells in Chin et al.); IRX3 (identified as differentially expressed in iPS vs ES cells in Maherali et al.); GREM1 (identified as differentially expressed in iPS vs ES cells in Yu et al.). The results show that POU5F1 is highly expressed in ES and iPS cells relative to fibroblasts as expected and that the PUS7L gene is differentially expressed in iPS vs ES cells, consistent with microarray based expression analysis in the same cells. Other genes do not show consistent differences between ES and iPS cells in our cell lines, which is not consistent with previously published results. See also main text Figure 3.

Supplemental Tables

Table S1. H3K4me3 and H3K27me3 occupied genomic regions and genes in ES, iPS, and fibroblast cells

A) H3K4me3 occupied genomic regions

B) H3K27me3 occupied genomic regions

C) H3K4me3 occupied genes

D) H3K27me3 occupied genes

Table S2. Pairwise comparisons of H3K4me3 and H3K27me3 occupied genes and peak heights between ES, iPS, and fibroblast cells and between iPS cells with integrated and excised transgenes

A) Comparisons of H3K4me3 occupied genes

B) Comparisons of H3K27me3 occupied genes

C) Comparisons of H3K4me3 peak heights

D) Comparisons of H3K27me3 peak heights

Table S3. Genomic regions with statistically significant differential H3K4me3 and H3K27me3 occupancy between male and female pluripotent cells, between pluripotent and fibroblast cells, and between ES and iPS cells

A) Regions with different H3K4me3 between male and female pluripotent cells

B) Regions with different H3K27me3 between male and female pluripotent cells

C) Regions with different H3K4me3 between pluripotent and fibroblast cells

D) Regions with different H3K27me3 between pluripotent and fibroblast cells

E) Regions with different H3K4me3 between ES and iPS cells

F) Regions with different H3K27me3 between ES and iPS cells

Table S4. Chromatin differences between ES and iPS cells do not reflect cell of origin

A) H3K4me3 differences do not reflect cell of origin

B) H3K27me3 differences do not reflect cell of origin

Table S5. The probesets differentially expressed between ES and iPS cells and between pluripotent and fibroblast cells in this study, Chin et al., Maherali et al., and Yu et al. and the numbers of differentially expressed genes and probesets overlapping between these datasets

A) Differentially expressed probesets between ES and iPS cells in this study, Guenther et al.

B) Differentially expressed probesets between ES and iPS cells in Chin et al.

C) Differentially expressed probesets between ES and iPS cells in Maherali et al.

D) Differentially expressed probesets between ES and iPS cells in Yu et al.

E) Differentially expressed probesets between pluripotent and fibroblast cells in this study, Guenther et al.

F) Differentially expressed probesets between pluripotent and fibroblast cells in Chin et al.

G) Differentially expressed probesets between pluripotent and fibroblast cells in Maherali et al.

H) Differentially expressed probesets between pluripotent and fibroblast cells in Yu et al.

I) The numbers of differentially expressed genes and probesets overlapping between Guenther et al., Chin et al., Maherali et al., and Yu et al..

Appendix B

Supplemental Material for Chapter 3

Connecting microRNA Genes to the Core Transcriptional Regulatory Circuitry of Embryonic Stem Cells

Supplemental Experimental Procedures

Growth conditions and quality control for human ES cells

Human embryonic stem cells were obtained from WiCell (Madison, WI; NIH Code WA09) and grown as described. Cell culture conditions and harvesting have been described previously (Boyer et al., 2005; Lee et al., 2006; Guenther et al., 2007). Quality control for the H9 cells included immunohistochemical analysis of pluripotency markers, alkaline phosphatase activity, teratoma formation, and formation of embryoid bodies and has been previously published as supplemental material (Boyer et al., 2005; Lee et al., 2006).

Growth conditions for murine ES cells

V6.5 (C57BL/6-129) murine ES cells were grown under typical ES cell culture conditions on irradiated mouse embryonic fibroblasts (MEFs) as previously described (Boyer et al., 2006). Briefly, cells were grown on gelatinized tissue culture plates in DMEM-KO (Gibco/Invitrogen) supplemented with 15% fetal bovine serum (characterized from Hyclone), 1000 U/ml leukemia inhibitory factor (LIF) (Chemicon; ESGRO ESG1106), non-essential amino acids, L-glutamine, Penicillin/Streptomycin andß-mercaptoethanol. Immunostaining was used to confirm expression of pluripotency markers, SSEA 1 (Developmental Studies Hybridoma Bank) and Oct4 (Santa Cruz, SC-5279). For location analysis, cells were grown for one passage off of MEFs, on gelatinized tissue-culture plates.

Antibodies

Oct4-bound genomic DNA was enriched from whole cell lysate using an epitope specific goat polyclonal antibody purchased from Santa Cruz (sc-8628) and compared to a reference whole cell extract (Boyer et al., 2005). A summary of regions occupied with high confidence for this antibody identified by ChIP-Seq in mES cells is provided in Table S3 and by ChIP-chip on genome-wide tiling arrays in hES cells are on Table S8. Oct4 ChIP-Seq data can be visualized on the UCSC browser by uploading supplemental file:
mES_regulator_ChIPseq.mm8.WIG.gz

Sox2-bound genomic DNA was enriched from whole cell lysate using an affinity purified goat polyclonal antibody purchased from R&D Systems (AF2018) and compared to a reference

whole cell extract (Boyer et al., 2005). A summary of regions occupied with high confidence for this antibody identified by ChIP-Seq in mES cells is provided in Table S3. Sox2 ChIP-Seq data can be visualized on the UCSC browser by uploading supplemental file: mES_regulator_ChIPseq.mm8.WIG.gz

Nanog-bound genomic DNA was enriched from whole cell lysate using an affinity purified rabbit polyclonal antibody purchased from Bethyl Labs (bl1662) and compared to a reference whole cell extract (Boyer et al., 2005).. A summary of regions bound with high confidence for this antibody is provided in Table S3. Nanog ChIP-Seq data can be visualized on the UCSC browser by uploading supplemental file: mES_regulator_ChIPseq.mm8.WIG.gz

Tcf3-bound genomic DNA was enriched from whole cell lysate using an epitope specific goat polyclonal antibody purchased from Santa Cruz (sc-8635) and compared to a reference whole cell extract (Cole et al., 2008). A summary of regions occupied with high confidence for this antibody identified by ChIP-Seq in mES cells is provided in Table S3. Tcf3 ChIP-Seq data can be visualized on the UCSC browser by uploading supplemental file: mES_regulator_ChIPseq.mm8.WIG.gz

Suz12-bound genomic DNA was enriched from whole cell lysate using an affinity purified rabbit polyclonal antibody purchased from Abcam (AB12073) and compared to a reference whole cell extract (Lee et al., 2006). A summary of regions bound with high confidence for this antibody is provided in Table S10. Suz12 ChIP-Seq data can be visualized on the UCSC browser by uploading: supplemental file: mES_chomatin_ChIPseq.mm8.WIG.gz

H3K4me3-modified nucleosomes were enriched from whole cell lysate using an epitope-specific rabbit polyclonal antibody purchased from Abcam (AB8580) (Santos-Rosa et al., 2002; Guenther et al., 2007). Samples were analyzed using ChIP-Seq. Comparison of this data with ChIP-Seq published previously (Mikkelsen et al., 2007) showed near identify in profile and bound regions (Table S5). H3K4me3 ChIP-Seq data can be visualized on the UCSC browser by uploading supplemental file: mES_chomatin_ChIPseq.mm8.WIG.gz

H3K79me2-modified nucleosomes were isolated from mES whole cell lysate using Abcam antibody AB3594 (Guenther et al., 2007). Chromatin immunoprecipitations against H3K36me3 were compared to reference WCE DNA obtained from mES cells. Samples were analyzed using ChIP-Seq and were used for visual validation of predicted miRNA promoter association with mature miRNA sequences only (Figure 2). H3K79me2 ChIP-Seq data can be

visualized on the UCSC browser by uploading supplemental file: mES_chomatin_ChIPseq.mm8.WIG.gz

H3K36me3-modified nucleosomes were isolated from mES whole cell lysate using rabbit polyclonal antibody purchased from Abcam (AB9050) (Guenther et al., 2007). Chromatin immunoprecipitations against H3K36me3 were compared to reference WCE DNA obtained from mES cells. Samples were analyzed using ChIP-Seq and were used for visual validation of predicted miRNA promoter association with mature miRNA sequences only (Figure 2). H3K36me3 ChIP-Seq data can be visualized on the UCSC browser by uploading supplemental file: mES_chomatin_ChIPseq.mm8.WIG.gz

Chromatin immunoprecipitation

Protocols describing all materials and methods have been previously described (Lee et al. 2007) and can be downloaded from: http://web.wi.mit.edu/young/hES_PRC/. Briefly, we performed independent immunoprecipitations for each analysis. Embryonic stem cells were grown to a final count of $5x10^7$ – $1x10^8$ cells for each location analysis experiment. Cells were chemically crosslinked by the addition of one-tenth volume of fresh 11% formaldehyde solution for 15 minutes at room temperature. Cells were rinsed twice with 1xPBS and harvested using a silicon scraper and flash frozen in liquid nitrogen. Cells were stored at –80°C prior to use.

Cells were resuspended, lysed in lysis buffers and sonicated to solubilize and shear crosslinked DNA. Sonication conditions vary depending on cells, culture conditions, crosslinking and equipment. We used a Misonix Sonicator 3000 and sonicated at approximately 28 watts for 10 x 30 second pulses (90 second pause between pulses). For ChIP of Oct4, Nanog, Tcf3 and Suz12 in murine ES cells, SDS was added to lysate after sonication to a final concentration of 0.1%. Samples were kept on ice at all times.

The resulting whole cell extract was incubated overnight at 4°C with 100 µl of Dynal Protein G magnetic beads that had been preincubated with approximately 10 µg of the appropriate antibody. Beads were washed 4-5 times with RIPA buffer and 1 time with TE containing 50 mM NaCl. For ChIP of Oct4, Nanog, Tcf3 and Suz12 in murine ES cells, the following 4 washes for 4 minutes each were used instead of RIPA buffer: 1X low salt (20mM Tris pH 8.1, 150mM NaCl, 2mM EDTA, 1% Triton X-100, 0.1% SDS), 1X high salt (20mM Tris pH 8.1, 500mM NaCl, 2mM EDTA, 1% Triton X-100, 0.1% SDS), 1X LiCl (10mM Tris

pH 8.1, 250mM LiCl, 1mM EDTA, 1% deoxycholate, 1% NP-40), and 1X TE+ 50mM NaCl. Bound complexes were eluted from the beads by heating at 65°C with occasional vortexing and crosslinking was reversed by overnight incubation at 65°C. Whole cell extract DNA (reserved from the sonication step) was also treated for crosslink reversal.

ChIP-Seq Sample Preparation and Analysis

All protocols for Illumina/Solexa sequence preparation, sequencing and quality control are provided by Illumina (http://www.illumina.com/pages.ilmn?ID=203). A brief summary of the technique and minor protocol modifications are described below.

Sample preparation

Purified immunoprecipitated (ChIP) DNA were prepared for sequencing according to a modified version of the Illumina/Solexa Genomic DNA protocol. Fragmented DNA was prepared for ligation of Solexa linkers by repairing the ends and adding a single adenine nucleotide overhang to allow for directional ligation. A 1:100 dilution of the Adaptor Oligo Mix (Illumina) was used in the ligation step. A subsequent PCR step with limited (18) amplification cycles added additional linker sequence to the fragments to prepare them for annealing to the Genome Analyzer flow-cell. After amplification, a narrow range of fragment sizes was selected by separation on a 2% agarose gel and excision of a band between 150-300 bp (representing shear fragments between 50 and 200nt in length and ~100bp of primer sequence). The DNA was purified from the agarose and diluted to 10 nM for loading on the flow cell.

Polony generation on Solexa flow-cells

The DNA library (2-4 pM) was applied to the flow-cell (8 samples per flow-cell) using the Cluster Station device from Illumina. The concentration of library applied to the flow-cell was calibrated such that polonies generated in the bridge amplification step originate from single strands of DNA. Multiple rounds of amplification reagents were flowed across the cell in the bridge amplification step to generate polonies of approximately 1,000 strands in 1μm diameter spots. Double stranded polonies were visually checked for density and morphology by staining

with a 1:5000 dilution of SYBR Green I (Invitrogen) and visualizing with a microscope under fluorescent illumination. Validated flow-cells were stored at 4°C until sequencing.

Sequencing

Flow-cells were removed from storage and subjected to linearization and annealing of sequencing primer on the Cluster Station. Primed flow-cells were loaded into the Illumina Genome Analyzer 1G. After the first base was incorporated in the Sequencing-by-Synthesis reaction the process was paused for a key quality control checkpoint. A small section of each lane was imaged and the average intensity value for all four bases was compared to minimum thresholds. Flow-cells with low first base intensities were re-primed and if signal was not recovered the flow-cell was aborted. Flow-cells with signal intensities meeting the minimum thresholds were resumed and sequenced for 26 cycles.

Solexa data analysis

Images acquired from the Illumina/Solexa sequencer were processed through the bundled Solexa image extraction pipeline, which identified polony positions, performed base-calling and generated QC statistics. Sequences were aligned using the bundled ELAND software using murine genome NCBI Build 36 and 37 (UCSC mm8, mm9) as the reference genome. Alignments to build 37 were used for analysis of the mmu-mir-290 cluster only as that cluster is not represented on build 36. Only sequences perfectly and uniquely mapping to the genome were used. A summary of the number of reads used is shown in Table S1.

The analysis methods used were derived from previously published methods (Johnson et al., 2007, Mikkelsen et al., 2007). Sequences from all lanes for each chromatin IP were combined, extended 200bp (maximum fragment length accounting for ~100bp of primer sequence), and allocated into 25 bp bins. Genomic bins containing statistically significant ChIP-Seq enrichment were identified by comparison to a Poissonian background model, using a p-value threshold of 10^{-9}. A list of the numbers of counts in a genomic bins required for each sample to meet this threshold are provided in Table S1. Addtionally, we used and empirical background model obtained from identical Solexa sequencing of DNA from whole cell extract (WCE) from matched cell samples (> 5x normalized enrichment across the entire region, see

below). A summary of the bound regions and their relation to gene targets can be found in Tables S2, S3, S5 and S10.

The p-value threshold was selected to minimize the expected false-positive rate. Assuming background reads are spread randomly throughout the genome, the probability of observing a given number of counts can be modelled as a Poisson process where the expectation can be calculated as the number of mapped reads times the number of bins per read (8) divided by the total number of bins available (we assumed 50% as a very conservative estimate). With the genome divided into ~10^8 bins of 25 bp, a probability of $p<10^{-9}$ represents the likelihood that ~1 experiment in 10 will randomly enrich one bin in the genome.

The Poisson background model assumes a random distribution of binding events, however we have observed significant deviations from this expectation in ChIP-Seq datasets. These non-random events can be detected as sites of enrichment using control IPs and create a significant number of false positive events for actual ChIP-Seq experiments. To remove these regions, we compared genomic bins and regions that meet the statistical threshold for enrichment to an empirical distribution of reads obtained from Solexa sequencing of DNA from whole cell extract (WCE) from matched cell samples. We required that enriched regions have five-fold greater ChIP-Seq density in the specific IP sample as compared with the non-specific WCE sample, normalized for the total number of reads. This served to filter out genomic regions that are biased to having a greater than expected background density of ChIP-Seq reads. We observed that ~200-500 regions in the genome showed non-specific enrichment in these experiments.

Identifications of regions enriched for Oct4/Sox2/Nanog/Tcf3

The identification of enriched regions in ChIP-chip and ChIP-Seq experiments is typically done using threshold for making a binary determination of enriched or not enriched. Unfortunately, there is not actually a clear delineation between truly bound and and unbound regions. Instead, enrichment is a continuum and the threshold is set to minimize false positives (high-confidence sites). This typically requires that thresholds be set at a level that allows a high false-negative rate (~30% for ChIP-chip, Lee et al). When multiple factors are compared, focusing only on the intersection of the different data sets compounds this effect, leading to higher false negative rates and the loss of many critical target genes.

Oct4, Sox2, Nanog and Tcf3 co-occupy promoters throughout the genome (Cole, Figure 1) and cluster analysis of enriched sites reveals apparent co-enrichment for all 4 factors at >90% of sites (Frampton & Young, unpublished data). However, the overlap for any two factors at the cut-off for high-confidence enrichment is only about two thirds (Figure S1, Tables S2 and S3). Therefore many of these sites must have enrichment that is below the high-confidence threshold for at least some of the participating factors. Variability in the enrichment observed for each factor at different binding sites is common in the data (Figures 1b, 3, and S2).

To determine a threshold of binding for multiple factors, we used two complementary methods to examine high-confidence targets of the four regulators. First, the classes of genes enriched by different numbers of factors at high-confidence were compared to the known classes of targets based on gene ontology (Figure S1b, http://gostat.wehi.edu.au/cgi-bin/goStat.pl, Beissbarth and Speed, 2004). The highest confidence targets (those with high levels of immuno-enrichment observed for all for factors) preferentially encoded factors involved in DNA binding, regulation of transcription and development as has been previously shown (Boyer et al., 2005). These gene ontology categories continued to be overrepresented among high-confidence targets of either 3 of the 4 factors or 2 of 4 the factors, albeit at lower levels, but were barely enriched among high confidence targets of only one factor.

As a second test, we examined how different numbers of overlapping high-confidence targets affected the overlap with our previous genome-wide studies using ChIP-chip. Because not all regions of the genome are tiled with equal density on the microarrays used for ChIP-chip, we first determined the minimum probe density required to confirm binding detected by ChIP-Seq (Figure S2). At most genes with high probe density, the ChIP-Seq and ChIP-chip data were very highly correlated. However, regions of the genome with microarray coverage of less than three probes per kilobase were generally unreliable in detecting these enrichment. These regions, which had low probe coverage on the microarrays, represent approximately 1/3 of all sites co-enriched for the four factors by ChIP-Seq. In regions where probe density was greater than three probe per kilobase the fraction of ChIP-Seq sites confirmed by ChIP-chip experiments increased with additional factors co-binding with a large fall off below 2 factors (data not shown). Based on these two analyses, we elected to choose targets occupied at high-confidence by 2 or more of the 4 factors tested for further analysis in this manuscript. (Figures 1a and S1a).

While a majority of the miRNA promoters identified as occupied by

Oct4/Sox2/Nanog/Tcf3 are not occupied by all four factors at high-confidence, it is interesting to note that all of the miRNA genes that share highly similar seeds to miR-302 are occupied at high confidence by all four factors (miR-302 cluster, miR-290 cluster and miR-106a cluster), similar to the promoters of core transcriptional regulators of ES cells. By comparison, promoters also occupied by Suz12 almost never showed high-confidence binding for all four factors (Table S4, see mmu-miR-9-2 in Figure 3). Similar effects were observed for protein-coding genes in mES cells (Lee et al., 2006). Whether this is caused by reduced epitope availability in PcG bound regions or reflects reduced protein binding is unclear.

DNA motif discovery and high-resolution binding-site analysis

DNA motif discovery was performed on the genomic regions that were enriched for Oct4 at high-confidence. In order to obtain maximum resolution, a modified version of the ChIP-Seq read mapping algorithm was used. Genomic bins were reduced in size from 25 bp to 10 bp. Furthermore, a read extension that placed greater weight towards the middle of the 200 bp extension was used. This model placed 1/3 count in the 8 bins from 0-40 and 160-200 bp, 2/3 counts in the 8 bins from 40-80 and 120-160 bp and 1 count in the 4 bins from 80-120 bp. This allowed increased precision for determination of the peak of ChIP-Seq density in each Oct4 bound region. 100bp surrounding the 500 Oct4 bound regions with the greatest peak ChIP-Seq density were submitted to the motif disocvery tool MEME (Bailey and Elkan, 1995; Bailey, 2006) to search for over-represented DNA motifs. A single sixteen basepair motif was discovered by the MEME algorithm (Table S4, Figure S2i). This motif was significantly ($p<10^{-100}$) over-represented in the Oct4 bound input sequences and occurred in 445 of the 500 hundred basepair sequences.

As a default, MEME uses the individual nucleotide frequencies within input sequences to model expected motif frequencies. This simple model might result discovery of motifs which are enriched because of non-random di-, tri-, etc. nucleotide frequencies. Consequently, three different sets of control sequences of identical length were used to ensure the specificity of the motif discovery results. First, the sequences immediately flanking each input sequence were used as control sequences. Second, randomly selected sequences having the same distribution of distances from transcription start sites as the Oct4 input sequences were used as control sequences. Third, sequences from completely random genomic regions were used as control

sequences. Each of these sets of control seqeunces were also examined using MEME. For each of these controls, the motif discovered from actual Oct4 bound sequences was not identified in the control sequences.

The motif discovery process was repeated using different numbers and lengths of sequences, but the same motif was discovered for a wide array of input sequences. Furthermore, when motif discovery was repeated with the top 500 Sox2, Nanog, and Tcf3 occupied regions, the same motif was identified. Overall, the motif occurs within 100 bp of the peak of ChIP-Seq density at more than 90% of the top regions enriched in each experiment, while occuring in the same span at 24-28% of control regions and within 25 bp of the ChIP-Seq peak at more than 80% of regions versus 9-11% of control regions.

We next attempted to determine the precise sites on the genome bound by Oct4, Sox2, Nanog, and Tcf3 at basepair resolution using composite analysis of the bound regions for each factor. In particular, we examined if the different factors tended to associate with specific sequences within the assymetric DNA motif identified at a high fraction of the sites occupied by Oct4, Sox2, Nanog, and Tcf3. A set of ~2,000 of the highest confidence bound regions was determined for each factor based on a count threshold approximately two fold higher than the threshold for high-confidence regions shown in Table S1 (Poisson: $p < 10^{-9}$). Regions without a motif within 50bp of the peak of ChIP-Seq enrichment, typically ~10% of regions, were removed from this analysis. The distance from the first base of the central motif in each bound region to the 5' end of all reads within 250bp was tabulated, keeping reads mapping to the same strand as the motif separate from reads mapping to the oppositie strand. The difference in ChIP-Seq read frequency between reads mapping to the same strand as the motif and the reads mapping to the oppositite strand was calculated at every basepair within the 500 bp window Figure S3. We made the assumption that the precise peak of the ChIP-Seq distribution was the point at which this strand bias was equal to zero.

To determine the precise position where the strand bias was equal to zero, we created a simplified model of the strand bias for each transcription factor. We chose a function with 4 parameters (A,B, C, and M), one of which (M) was the point at which the curve crosses the x-axis.

$$f(x) = A \times \arctan\left(\frac{x-M}{B}\right) \times e^{-\left|\frac{x-M}{B}\right|^{C}}$$

Least squares curve fitting was performed with GNUplot (http://www.gnuplot.info/) using an approximated set of initial conditions (A = -1000, B = 100, C = 2, M = 10). The variablity in M was detemined by monte carlo simulation (n=25) using a random set of half of the ChIP-Seq reads in each dataset and is shown in Figure S3.

Identification of miRNA promoters in human and mouse

To better understand the regulation of miRNAs, we sought to identify the sites of transcription initiation for all miRNAs in both human and mouse, at least to low resolution (~1kb). Most methods used to identify promoters require active transcription of the miRNA and isolation of rare primary miRNA transcripts. We decided to use an approach based on *in vivo* chromatin signature of promoters. This approach has two principle advantages. First, the required data has been published by a variety of laboratories and is readily accessible and second, it does not require the active transcription of the miRNA primary transcript.

Recent results using genome-wide location analysis of H3K4me3 indicate that between 60 and 80% of all protein-coding genes in any cell population have promoters enriched in methylated nucleosomes, even where the gene is not detected by typical transcription profiling (Guenther et al., 2007) Importantly, over 90% of the H3K4me3 enriched regions in these cells map to known or predicted promoters, suggesting that H3K4me3 can be used as a proxy for sites of active initiation. Our strategy to identify miRNA promoters, therefore, uses H3K4me3 enriched sites from as many sources as possible as a collection of promoters. In human, H3K4me3 sites were identified in ES cells (H9), hepatocytes, a pro-B cell line (REH cells) (Guenther et al., 2007) and T cells (Barski et al., 2007). Mouse H3K4me3 sites were identified from ES cells (V6.5), neural precursors, and embryonic fibroblasts (Mikkelsen et al., 2007). In total, we identified 34,793 high-confidence H3K4me3 enriched regions in human and 34,096 high-confidence regions enriched in mouse, collectively present at ~75% of all protein-coding genes.

The list of miRNAs identified in the miRNA atlas (Landgraf et al., 2007) was used as the basis for our identification. The total list consists of 496 miRNAs in human, 382 miRNAs in mouse. ~65% of the murine miRNAs can be found in both species.

For each of these miRNAs, possible start sites were derived from both all H3K4me3 enriched regions within 250kb upstream of the miRNA as well as all known start sites for any miRNAs that were identified as being within known transcripts from RefSeq (Pruitt et al., 2005) Mammalian Gene Collection (MGC) (Gerhard et al., 2004) Ensembl (Hubbard et al., 2005), or University of California Santa Cruz (UCSC) Known Genes (genome.ucsc.edu) (Kent et al., 2002) for which EntrezGene (http://www.ncbi.nlm.nih.gov/entrez/) gene IDs had been generated. Where an annotated start site was found to overlap an H3K4me3 enriched region, the known start was used in place of the enriched region.

A scoring system was derived empirically to select the most likely start sites for each miRNA. Each possible site was given a bonus if it was either the start of a known transcript that spanned the miRNA or of an EST that spanned the miRNA. Scores were reduced if the H3K4me3 enriched region was assignable instead to a transcript or EST that did not overlap the miRNA. Additional positive scores were given to enriched sites within 5kb of the miRNA, while additional negative scores were given based on the number of intervening H3K4me3 sites between the test region and the miRNA. Finally, each enriched region was tested for conservation between human and mouse using the UCSC liftover program (Hinrichs et al., 2006). If two test regions overlapped, they were considered to be conserved (21%). In the cases where human and mouse disagreed on the quality of a site, if the site had an EST or gene overlapping the miRNA, that site was given a high score in both species. Alternatively, if one species had a non-overlapping site, that site was considered to be an unlikely promoter in both species. Finally, for miRNAs where a likely promoter was identified in only one species, we manually checked the homologous region of the other genome to search for regions enriched for H3K4me3-modified nucleosomes that may have fallen below the high-confidence threshold. Start sites were considered to be likely if the total score was ≥ 0 (Figure S4 and S5). In total, we identified likely start sites for ~85% of all miRNAs in both species (Tables S6 and S7). Predicted miRNA genes can be visualised on the UCSC browser by uploading the supplemental files: mouse_miRNA_track.mm8.bed and human_miRNA_track.hg17.bed

Several lines of evidence suggest the high quality of these predictions. First, previous studies have found that miRNAs within 50kb of each other are likely to be co-regulated (Lagos-Quintana et al., 2001; Lau et al., 2001). While the nature of these clusters was not included in our analysis, nearly all miRNAs within a cluster end up identifying the same promoter region (see Figures 2, 3, 5 and S3). The only exceptions to this are found in the large clusters of repeat derived miRNAs found in chromosome 12 of mouse and chromosome 14 in human where a single H3K4me3 enriched region splits the clusters. Second, consistent with the frequent association of CpG islands with the transcriptional start sites for protein-coding genes, ~50% of the miRNA promoters identified here overlap CpG islands (Tables S6 and S7). Finally, for miRNAs that were active in ES cells, histone modifications associated with elongation were able to "connect" the mature miRNAs to the predicted transcription start site (Figure 2).

To further ascertain the accuracy of our promoter predictions, we compared our predicted start sites to those identified in recent studies. Predictions were tested against mmu-mir-34b / mmu-mir-34c (Corney et al., 2007), hsa-mir-34a (Chang et al., 2007) mmu-mir-101a, mmu-mir-202, mmu-mir-22, mmu-mir-124a-1, mmu-mir-433 (Fukao et al., 2007), and hsa-mir-17/18a/19a/20a/19b-1/92a-1 (O'Donnell et al., 2005). Additional miRNA promoters in these manuscripts were not predicted strongly by the above algorithm. For these 14 miRNAs, H3K4me3 sites were identified within 1kb of all but two of the sites. mmu-mir-202 was predicted about 20kb upstream of the annotated start site, but may reflect an H3K4me3 site absent from the tissues sampled. mmu-mir-433 is in the middle of a large cluster of miRNAs on mouse chromosome 12. The annotated TSS lies within the cluster between mir-433 and mir-431 suggesting the promoter may be incorrect. Overall, the accuracy of the promoter predictions is believed to be ~75% (6/8). Additional H3K4me3 data sets and EST data should allow for improved accuracy in predicting and validating these initiation sites.

ChIP-chip sample preparation and analysis

Immunoprecipitated DNA and whole cell extract DNA were purified by treatment with RNAse A, proteinase K and multiple phenol:chloroform:isoamyl alcohol extractions. Purified DNA was blunted and ligated to linker and amplified using a two-stage PCR protocol. Amplified DNA was labeled and purified using Bioprime random primer labeling kits (Invitrogen): immunoenriched DNA was labeled with Cy5 fluorophore, whole cell extract DNA was labeled

with Cy3 fluorophore.

Labeled DNA was mixed (~5 μg each of immunoenriched and whole cell extract DNA) and hybridized to arrays in Agilent hybridization chambers for up to 40 hours at 40°C. Arrays were then washed and scanned.

Slides were scanned using an Agilent DNA microarray scanner BA. PMT settings were set manually to normalize bulk signal in the Cy3 and Cy5 channel. For efficient batch processing of scans, we used Genepix (version 6.0) software. Scans were automatically aligned and then manually examined for abnormal features. Intensity data were then extracted in batch.

44k human whole genome array

The human promoter array was purchased from Agilent Technology (www.agilent.com). The array consists of 115 slides each containing ~44,000 60mer oligos designed to cover the non-repeat portion of the human genome. The design of these arrays are discussed in detail elsewhere (Lee et al., 2006).

Data normalization and analysis

We used GenePix software (Axon) to obtain background-subtracted intensity values for each fluorophore for every feature on the whole genome arrays. Among the Agilent controls is a set of negative control spots that contain 60-mer sequences that do not cross-hybridize to human genomic DNA. We calculated the median intensity of these negative control spots in each channel and then subtracted this number from the intensities of all other features.

To correct for different amounts of each sample of DNA hybridized to the chip, the negative control-subtracted median intensity value of control oligonucleotides from the Cy3-enriched DNA channel was then divided by the median of the control oligonucleotides from the Cy5-enriched DNA channel. This yielded a normalization factor that was applied to each intensity in the Cy5 DNA channel.

Next, we calculated the log of the ratio of intensity in the Cy3-enriched channel to intensity in the Cy5 channel for each probe and used a whole chip error model to calculate confidence values for each spot on each array (single probe p-value). This error model functions by converting the intensity information in both channels to an X score which is dependent on both the absolute value of intensities and background noise in each channel using an f-score

calculated as described for promoter regions or using a score of 0.3 for tiled arrays. When available, replicate data were combined, using the X scores and ratios of individual replicates to weight each replicate's contribution to a combined X score and ratio. The X scores for the combined replicate are assumed to be normally distributed which allows for calculation of a p-value for the enrichment ratio seen at each feature. P-values were also calculated based on a second model assuming that, for any range of signal intensities, IP:control ratios below 1 represent noise (as the immunoprecipitation should only result in enrichment of specific signals) and the distribution of noise among ratios above 1 is the reflection of the distribution of noise among ratios below 1.

High confidence enrichment

To automatically determine bound regions in the datasets, we developed an algorithm to incorporate information from neighboring probes. For each 60-mer, we calculated the average X score of the 60-mer and its two immediate neighbors. If a feature was flagged as abnormal during scanning, we assumed it gave a neutral contribution to the average X score. Similarly, if an adjacent feature was beyond a reasonable distance from the probe (1000 bp), we assumed it gave a neutral contribution to the average X score. The distance threshold of 1000 bp was determined based on the maximum size of labeled DNA fragments put into the hybridization. Since the maximum fragment size was approximately 550 bp, we reasoned that probes separated by 1000 or more bp would not be able to contribute reliable information about a binding event halfway between them.

This set of averaged values gave us a new distribution that was subsequently used to calculate p-values of average X (probe set p-values). If the probe set p-value was less than 0.001, the three probes were marked as potentially bound.

As most probes were spaced within the resolution limit of chromatin immunoprecipitation, we next required that multiple probes in the probe set provide evidence of a binding event. Candidate bound probe sets were required to pass one of two additional filters: two of the three probes in a probe set must each have single probe p-values < 0.005 or the centre probe in the probe set has a single probe p-value < 0.001 and one of the flanking probes has a single point p-value < 0.1. These two filters cover situations where a binding event occurs midway between two probes and each weakly detects the event or where a binding event occurs

very close to one probe and is very weakly detected by a neighboring probe. Individual probe sets that passed these criteria and were spaced closely together were collapsed into bound regions if the centre probes of the probe sets were within 1000 bp of each other.

Comparing enriched regions to known genes and miRNAs

Enriched regions were compared relative to transcript start and stop coordinates of known genes compiled from four different databases: RefSeq (Pruitt et al., 2005), Mammalian Gene Collection (MGC) (Gerhard et al., 2004), Ensembl (Hubbard et al., 2005), and University of California Santa Cruz (UCSC) Known Genes (genome.ucsc.edu) (Kent et al., 2002). All human coordinate information was downloaded in January 2005 from the UCSC Genome Browser (hg17, NCBI build 35). Mouse data was downloaded in June of 2007 (mm8, NCBI build 36).

To convert bound transcription start sites to more useful gene names, we used conversion tables downloaded from UCSC and Ensembl to automatically assign EntrezGene (http://www.ncbi.nlm.nih.gov/entrez/) gene IDs and symbols to the RefSeq, MGC, Ensembl, UCSC Known Gene. Comparisons of Oct4, Sox2, Nanog, Tcf3, H3K4me3 and Suz12 to annotated regions of the genomes can be found in Tables S3, S5, S8 and S10

For miRNAs start sites, two separate windows were used to evaluate overlaps. For chromatin marks and non-sequence specific proteins, miRNA promoters were considered bound if they were within 1kb of an enriched sequence. For sequence specific factors such as Oct4, we used a more relaxed region of 8kb surrounding the promoter, consistent with previous work we have published (Boyer et al., 2005). A full list of the high confidence start sites bound to promoters can be found in Tables S6 and S7.

Growth conditions for neural precursors and mouse embryonic fibroblasts

To generate neural precursor cells, ES cells were differentiated along the neural lineage using standard protocols. V6.5 ES cells were differentiated into neural progenitor cells (NPCs) through embryoid body formation for 4 days and selection in ITSFn media for 5–7 days, and maintained in FGF2 and EGF2 (R&D Systems) (Okabe et al., 1996).

Mouse embryonic fibroblasts were prepared from DR-4 strain mice as previously described (Tucker et al., 1997). Cells were cultured in Dulbecco's modified Eagle medium

supplemented with 10% cosmic calf serum, beta-mercaptoethanol, non-essential amino acids, L-glutamine and penicillin/streptomycin.

Analysis of Mature miRNA Frequency by Solexa Sequencing

Polony generation on Solexa flow-cells

The DNA library (2-4 pM) was applied to the flow-cell (8 samples per flow-cell) using the Cluster Station device from Illumina. The concentration of library applied to the flow-cell was calibrated such that polonies generated in the bridge amplification step originate from single strands of DNA. Multiple rounds of amplification reagents were flowed across the cell in the bridge amplification step to generate polonies of approximately 1,000 strands in 1μm diameter spots. Double stranded polonies were visually checked for density and morphology by staining with a 1:5000 dilution of SYBR Green I (Invitrogen) and visualizing with a microscope under fluorescent illumination. Validated flow-cells were stored at 4°C until sequencing.

Sequencing and analysis

Flow-cells were removed from storage and subjected to linearization and annealing of sequencing primer on the Cluster Station. Primed flow-cells were loaded into the Illumina Genome Analyzer 1G. After the first base was incorporated in the Sequencing-by-Synthesis reaction the process was paused for a key quality control checkpoint. A small section of each lane was imaged and the average intensity value for all four bases was compared to minimum thresholds. Flow-cells with low first base intensities were re-primed and if signal was not recovered the flow-cell was aborted. Flow-cells with signal intensities meeting the minimum thresholds were resumed and sequenced for 36 cycles. Images acquired from the Illumina/Solexa sequencer were processed through the bundled Solexa image extraction pipeline which identified polony positions, performed base-calling and generated QC statistics. Sequences were then assigned to a miRNA if they perfectly matched at least the first 20bp of the mature miRNA sequences downloaded from targetScan (http://www.targetscan.org/). Mature miRNA frequencies were then normalized to each other by determining the expected frequency in mapped reads/million. A full list of the miRNAs detected can be found in Table S9.

miRNA microarray expression analysis

Mouse embryonic fibroblasts and neural precursor cells were cultured as described above. Murine induced pluripotent (iPS) cells, derived as previously described (Wernig et al., 2007), were cultured under the same conditions as murine embryonic stem cells (described above). RNA was extracted with RNeasy (Qiagen) reagents. 5 μg total RNA from treated and control samples were labeled with Hy3™ and Hy5™ fluorescent label, using the miRCURY™ LNA Array labeling kit (Exiqon, Denmark) following the procedure described by the manufacturer. The labeled samples were mixed pair-wise and hybridized to the miRNA arrays printed using miRCURY™ LNA oligoset version 8.1 (Exiqon, Denmark). Each miRNA was printed in duplicate, on codelink slides (GE), using GeneMachines Omnigrid 100. The hybridization was performed at 60C overnight using the Agilent Hybridization system - SurHyb, after which the slides were washed using the miRCURY™ LNA washing buffer kit (Exiqon, Denmark) following the procedure described by the manufacturer. The slides were then scanned using Axon 4000B scanner and the image analysis was performed using Genepix Pro 6.0.

Median minus background signal intensities for all microarray probes were tabulated and quantile normalized. Within each sample, each probe was given a signal value of the average signal of the probe of that rank, across the full dataset. Intensities were then floored at one unit and log normalized. Control probes were removed from further analysis. Statistically significant differential expression was calculated using the online NIA Array Analysis Tool (http://lgsun.grc.nia.nih.gov/ANOVA/). Probes were tested for differential expression using the following settings:

Threshold z-value to remove outliers: 10000
Error Model: Max(Average,Bayesian)
Error variance averaging window: 100
Proportion of highest error variances to be removed: 0.01
Bayesian degrees of freedom: 5
FDR threshold: 0.10

Of 1008 probes, 230 were determined to be differentially expressed between 3 MEF and 2 ES samples. Expression data for the iPS samples were not used for identifying differentially expressed miRNAs.

For clustering and heat map display, expression data were Z-score normalized. Centroid linkage, Spearman rank correlation distance, hierarchical clustering of genes and arrays was performed using Gene Cluster 3.0 (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm#ctv). Heatmaps were generated using Java Treeview (http://jtreeview.sourceforge.net/) with color saturation at 0.6 standard deviations. Complete miRNA microarray expression data, differentially expression results, and clustergram data are provided (Supplemental Tables S99).

Tissue-specificity of miRNAs

To determine the global tissue-specificity for miRNAs we used data from the recent publication of the miRNA atlas[4]. Specificity scores were taken from Table S34 Node 0 from Landgraf et al. (2007). Of the 45 distinct mature miRNAs with specificity scores >1 that are not bound only by Oct4/Sox2/Nanog/Tcf3, 16 were identified as Suz12 targets. These 16 represent over 40% of the distinct mature miRNAs whose promoters are occupied by Suz12 ($p < 5x10^{-4}$ for specificity scores > 1.3)

Identification of Oct4/Sox2/Nanog/Tcf3 occupied feed forward loops

To identify feed forward loops we examined the recent data set identifying functional targets of the miR-290 cluster (Sinkkonen et al., 2008). In their study, Sinkkonen et al. identified miR-290 targets by both looking at mRNAs that increase in level in a Dicer -/- cell line and overlap that data set with mRNAs that decrease in expression when miR-290 is added back to the cells. Because the promoter of the miR-290 gene is occupied by Oct4/Sox2/Nanog/Tcf3, any targets of the miRNA cluster that are also occupied by the 4 factors would represent feed forward targets. Of the 245 miR-290 cluster targets identified in the intersect of the two data sets, promoters for 64 are occupied by Oct4/Sox2/Nanog/Tcf3. This is approximately 50% more interactions then would be expected by random (binomial p-value $< 1x10^{-4}$).

Interestingly, only a small minority of these genes is also occupied by significant quantities of the PRC2 subunit Suz12. Of the 64 targets whose promoters are occupied by

Oct4/Sox2/Nanog/Tcf3, only 5 are occupied by domains of Suz12 binding >500bp (larger region sizes have been correlated with gene silencing, Lee et al., 2006). This may be because PcG bound proteins are not functional targets of mir-290 in mES cells or because these proteins are not expressed following Dicer deletion, they are excluded from the target list, but may be targets at other stages of development. In the later case, the miRNAs may serve as a redundant silencing mechanism for ES cells to help prevent even low levels of expression of the developmental regulators bound by PcG complexes.

References

Bailey, T. L., and Elkan, C. (1995). The value of prior knowledge in discovering motifs with MEME. Proc Int Conf Intell Syst Mol Biol *3*, 21-29.

Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res *34*, W369-373.

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Beissbarth, T., and Speed, T. P. (2004). GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics *20*, 1464-1465.

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.

Chang, T. C., Wentzel, E. A., Kent, O. A., Ramachandran, K., Mullendore, M., Lee, K. H., Feldmann, G., Yamakuchi, M., Ferlito, M., Lowenstein, C. J., *et al.* (2007). Transactivation of miR-34a by p53 broadly influences gene expression and promotes apoptosis. Mol Cell *26*, 745-752.

Cole, M. F., Johnstone, S. E., Newman, J. J., Kagey, M. H., and Young, R. A. (2008). Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells. Genes Dev *22*, 746-755.

Corney, D. C., Flesken-Nikitin, A., Godwin, A. K., Wang, W., and Nikitin, A. Y. (2007). MicroRNA-34b and MicroRNA-34c are targets of p53 and cooperate in control of cell proliferation and adhesion-independent growth. Cancer Res *67*, 8433-8438.

Fukao, T., Fukuda, Y., Kiga, K., Sharif, J., Hino, K., Enomoto, Y., Kawamura, A., Nakamura, K., Takeuchi, T., and Tanabe, M. (2007). An evolutionarily conserved mechanism for microRNA-223 expression revealed by microRNA gene profiling. Cell *129*, 617-631.

Gerhard, D. S., Wagner, L., Feingold, E. A., Shenmen, C. M., Grouse, L. H., Schuler, G., Klein, S. L., Old, S., Rasooly, R., Good, P., *et al.* (2004). The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Res *14*, 2121-2127.

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

Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R., and Young, R. A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77-88.

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., *et al.* (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res *34*, D590-598.

Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., *et al.* (2005). Ensembl 2005. Nucleic Acids Res *33*, D447-453.

Johnson, D., Martazavai, A., Myers, R., Wold, B., (2007). Genome-wide mapping of inn vivo protein-DNA interactions. Science *316*, 1441-2.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res *12*, 996-1006.
Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. Science *294*, 853-858.

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A. O., Landthaler, M., *et al.* (2007). A mammalian microRNA expression atlas based on small RNA library sequencing. Cell *129*, 1401-1414.

Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science *294*, 858-862.

Lee, T. I., Jenner, R. G., Boyer, L. A., Guenther, M. G., Levine, S. S., Kumar, R. M., Chevalier, B., Johnstone, S. E., Cole, M. F., Isono, K., *et al.* (2006a). Control of developmental regulators by Polycomb in human embryonic stem cells. Cell *125*, 301-313.

Lee, T. I., Johnstone, S. E., and Young, R. A. (2006b). Chromatin immunoprecipitation and microarray-based analysis of protein location. Nat Protoc *1*, 729-748.

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell *120*, 15-20.

Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P., and Burge, C. B. (2003). Prediction of mammalian microRNA targets. Cell *115*, 787-798.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T. K., Koche, R. P., *et al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-560.

O'Donnell, K. A., Wentzel, E. A., Zeller, K. I., Dang, C. V., and Mendell, J. T. (2005). c-Myc-regulated microRNAs modulate E2F1 expression. Nature *435*, 839-843.

Okabe, S., Forsberg-Nilsson, K., Spiro, A. C., Segal, M., and McKay, R. D. (1996). Development of neuronal precursor cells and functional postmitotic neurons from embryonic stem cells in vitro. Mech Dev *59*, 89-102.

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res *33*, D501-504.

Santos-Rosa, H., Schneider, R., Bannister, A. J., Sherriff, J., Bernstein, B. E., Emre, N. C., Schreiber, S. L., Mellor, J., and Kouzarides, T. (2002). Active genes are tri-methylated at K4 of histone H3. Nature *419*, 407-411.

Sinkkonen, L., Hugenschmidt, T., Berninger, P., Gaidatzis, D., Mohn, F., Artus-Revel, C. G., Zavolan, M., Svoboda, P., and Filipowicz, W. (2008). MicroRNAs control de novo DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells. Nat Struct Mol Biol *15*, 259-267.

Tucker, K. L., Wang, Y., Dausman, J., and Jaenisch, R. (1997). A transgenic mouse strain expressing four drug-selectable marker genes. Nucleic Acids Res *25*, 3745-3746.

Wernig, M., Meissner, A., Foreman, R., Brambrink, T., Ku, M., Hochedlinger, K., Bernstein, B. E., and Jaenisch, R. (2007). In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. Nature *448*, 318-324.

Figure S1

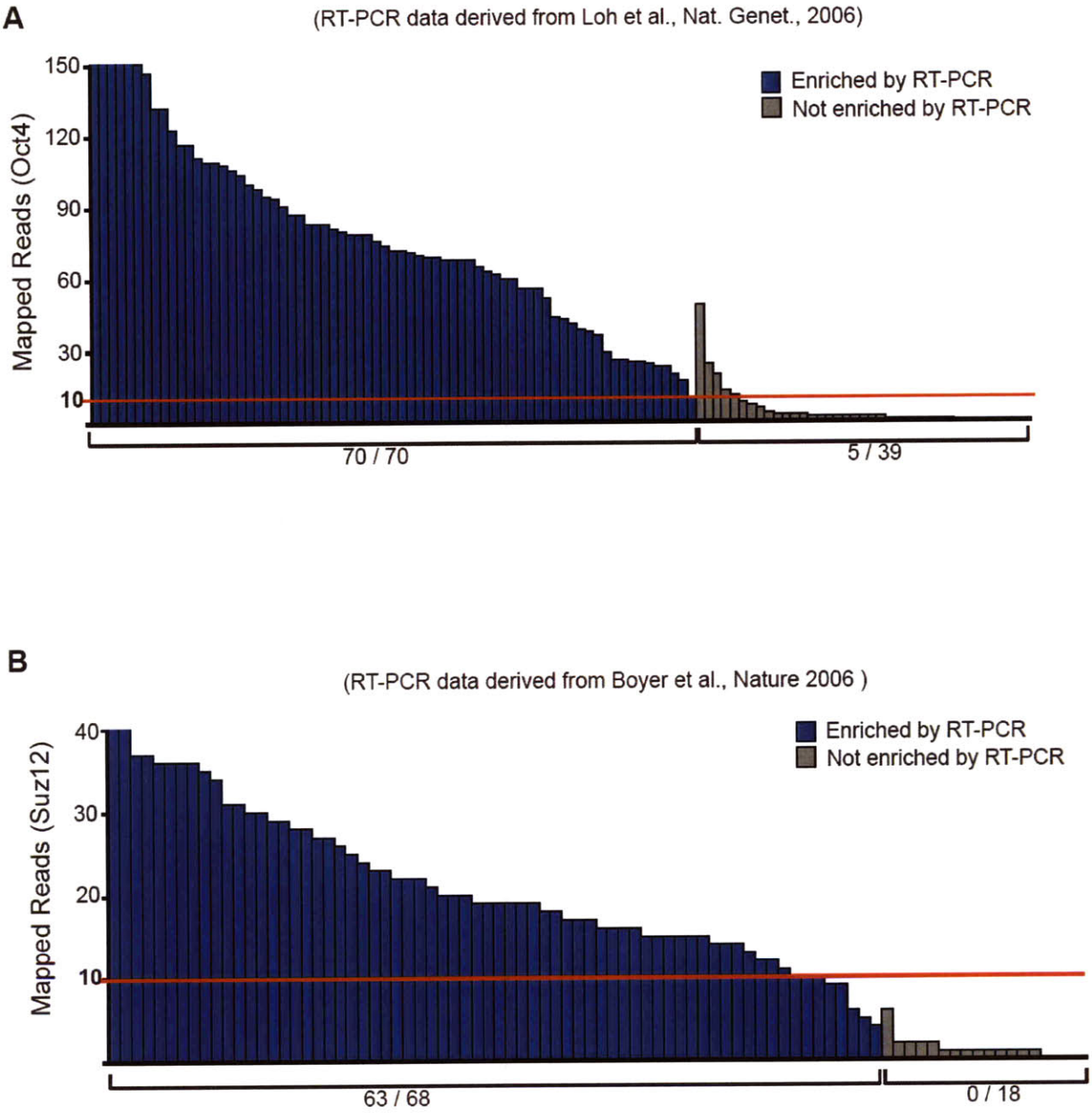

A
(RT-PCR data derived from Loh et al., Nat. Genet., 2006)
Enriched by RT-PCR
Not enriched by RT-PCR
Mapped Reads (Oct4)
150
120
90
60
30
10
70 / 70
5 / 39
B
(RT-PCR data derived from Boyer et al., Nature 2006)
Enriched by RT-PCR
Not enriched by RT-PCR
Mapped Reads (Suz12)
40
30
20
10
63 / 68
0 / 18

Figure S1. Comparison of ChIP-Seq and RT-PCR data for Oct4 and Suz12

A.ChIP-Seq reads for Oct4 enrichment were compared to RT-PCR results previously described in Loh et al. 71 probes that were identified as enriched are shown in blue and 39 regions identified as not-enriched are shown in gray. The maximum number of ChIP-Seq reads assigned within the region is shown on the vertical axis. Red line denoted the threshold of binding with $p < 10^{-9}$. Ambiguous RT-PCR results were excluded.

B. ChIP-Seq reads for Suz12 enrichment were compared to RTPCR results previously described in Boyer et al. as in **A.** ChIP-Seq data within 200bp of 68 probes identified as enriched by RT-PCR and confirmed by ChIP-chip are shown in blue and 18 probes identified as not-enriched are shown in gray.

Figure S2

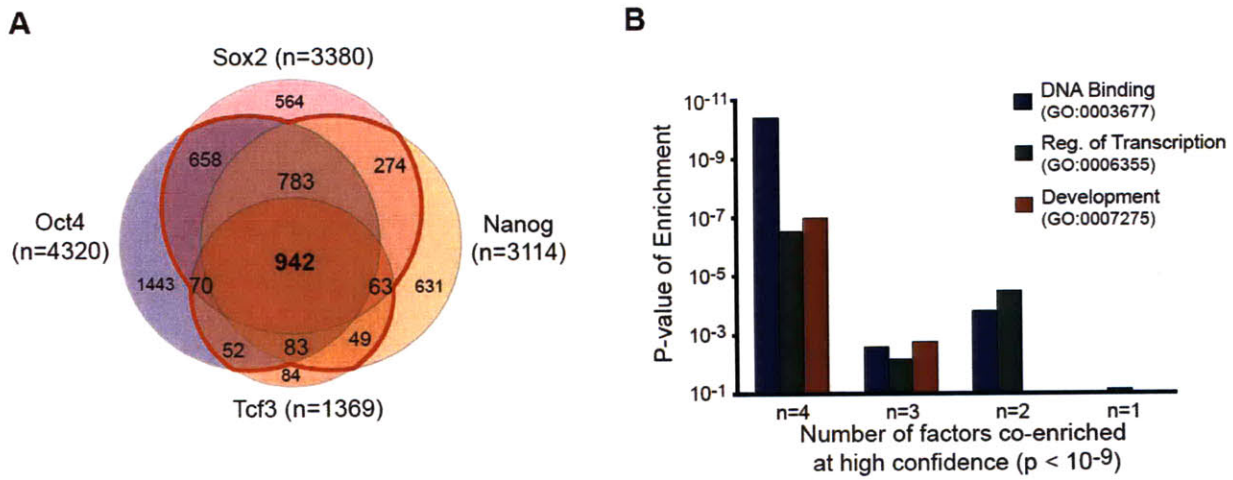


Figure S2. Promoters for known genes occupied by Oct4/Sox2/Nanog/Tcf3 in mES cells

A. Overlap of genes whose promoters are within 8kb of sites enriched for Oct4, Sox2, Nanog, or Tcf3. Not shown are the Nanog:Oct4 overlap (289) and Sox2:Tcf3 overlap (26). Red line deliniates genes considered occupied by Oct4/Sox2/Nanog/Tcf3.

B. Enrichment for selected GO-terms previously reported to be associated with Oct4/Sox2/Nanog binding (Boyer et al., 2005) was tested on the sets of genes occupied at high-confidence for 1 to 4 of the tested DNA binding factors. Hypergeometric p-value is shown for genes annotated for DNA binding (blue), Regulation of Transcription (green) and Development (red).

Figure S3

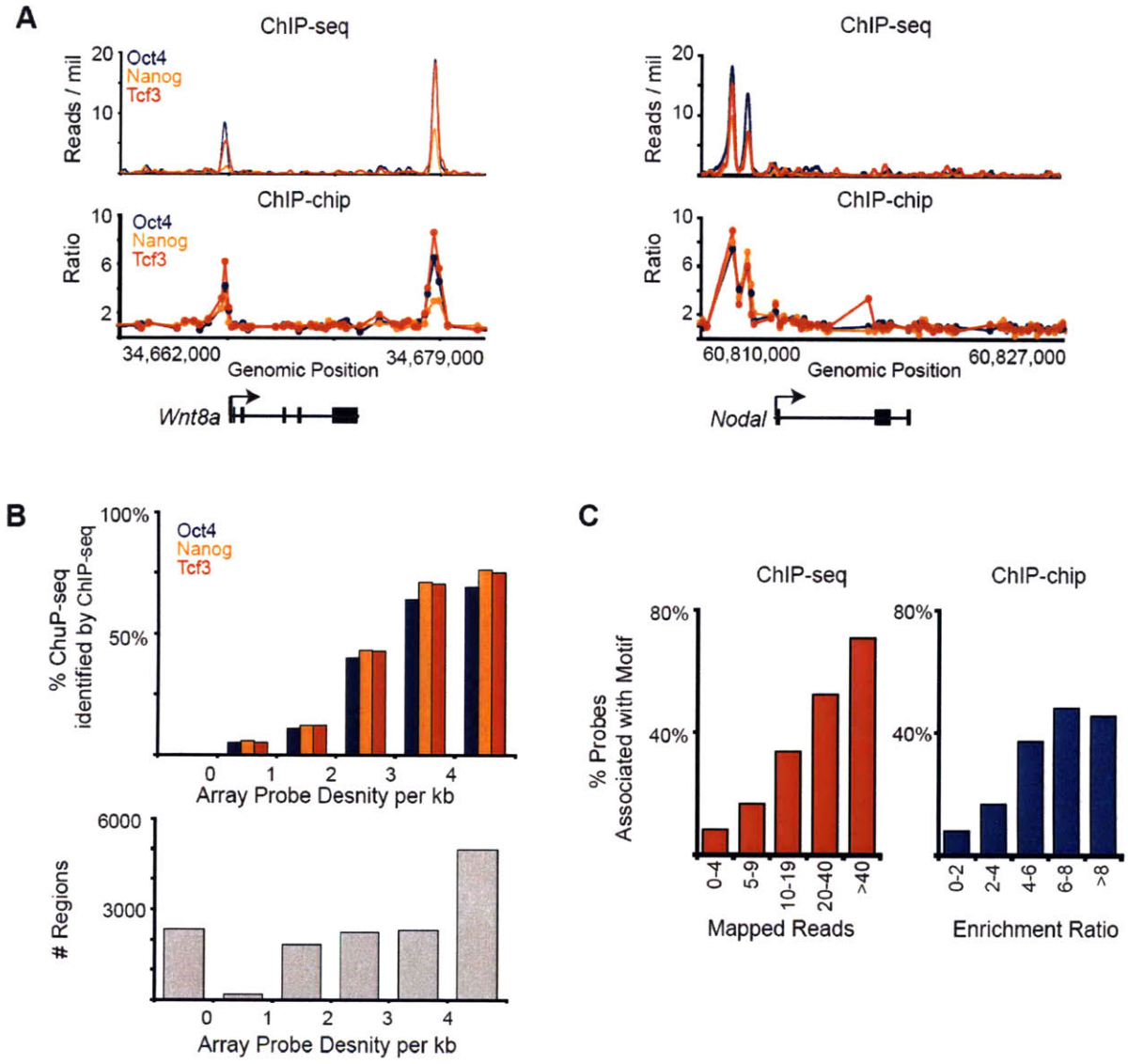

A
ChIP-seq
Reads / mil
Oct4
Nanog
Tcf3
ChIP-chip
Ratio
34,662,000
Genomic Position
34,679,000
Wnt8a
60,810,000
60,827,000
Nodal
B
% ChuP-seq identified by ChIP-seq
100%
50%
Array Probe Desnity per kb
Regions
6000
3000
C
% Probes Associated with Motif
80%
40%
0-4
5-9
10-19
20-40
>40
Mapped Reads
0-2
2-4
4-6
6-8
>8
Enrichment Ratio

Figure S3. Comparison of ChIP-Seq and ChIP-chip genome wide data for Oct4, Nanog and Tcf3

A. Binding of Oct4 (blue), Nanog (orange) and Tcf3 (red) across 17kb surrounding the Wnt8a and Nodal genes (black below the graph, arrow indicates transcription start site) as in Figure 1b. (upper) Binding derived from ChIP-Seq data plotted as reads per million. (lower) Binding derived from ChIP-chip enrichment ratios (Cole et. al., 2008)

B. Poor probe density prevents detection of ~1/3 of ChIP-Seq binding events on Agilent genome-wide tiling arrays. Top panel shows the fraction of regions that are occupied by Oct4/Sox2/Nanog/Tcf3 at high-confidence in mES cells as identified by ChIP-Seq that are enriched for Oct4 (blue), Nanog (orange) and Tcf3 (red) on Agilent genome-wide microarrays (Cole et al., 2008). Numbers on the x-axis define the boundaries used to classify probe densities for the histogram. Bottom panel illustrates a histogram of the microarray probe densities of the enriched regions identified.

Figure S4

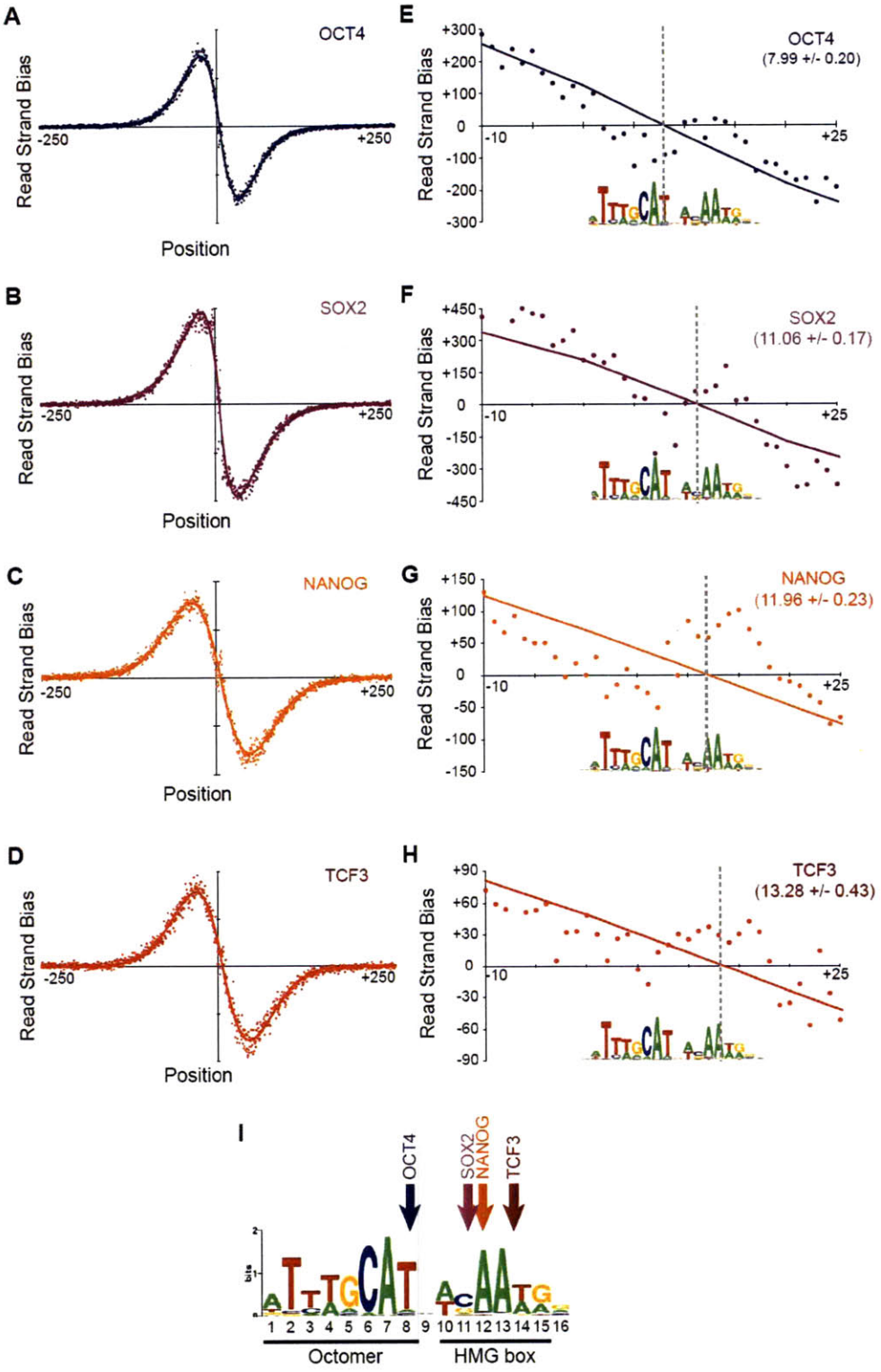

A
OCT4
Read Strand Bias
-250
+250
Position
B
SOX2
Read Strand Bias
-250
+250
Position
C
NANOG
Read Strand Bias
-250
+250
Position
D
TCF3
Read Strand Bias
-250
+250
Position
E
OCT4
(7.99 +/- 0.20)
Read Strand Bias
+300
+200
+100
0
-100
-200
-300
-10
+25
F
SOX2
(11.06 +/- 0.17)
Read Strand Bias
+450
+300
+150
0
-150
-300
-450
-10
+25
G
NANOG
(11.96 +/- 0.23)
Read Strand Bias
+150
+100
+50
0
-50
-100
-150
-10
+25
H
TCF3
(13.28 +/- 0.43)
Read Strand Bias
+90
+60
+30
0
-30
-60
-90
-10
+25
I
OCT4
SOX2
NANOG
TCF3
bits
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Octomer
HMG box

Figure S4. High resolution analysis of Oct4/Sox2/Nanog/Tcf3 binding based on meta-analysis

A-D. Short sequence reads for **A.** Oct4, **B**. Sox2, **C.** Nanog, **D.** Tcf3 mapping within 250bp of 2000 highly enriched regions where the peak of binding was found within 50bp of a high quality Oct4/Sox2 motif were collected. Composite profiles were created at base pair resolution for forward and reverse strand reads centered on the Oct4/Sox2 motif (aligned at +1). The difference between the number of positive and negative strand reads are shown for each base pair (circles). The best fit line is shown for each factor (see Supplemental Text). **E-H.** Zoomed in region of A-D showing 20bp surrounding the Oct4/Sox2 motif. Dashed line indicates the position where the best fit line crosses the X-axis. For reference, the motif is shown below each graph. **I.** Summary of meta-analysis for Oct4, Sox2, Nanog and Tcf3. Arrows indicate the nucleotide where each transcription factor switches from a positive strand bias to a negative strand bias. The octomer and HMG box motifs are indicated.

Figure S5

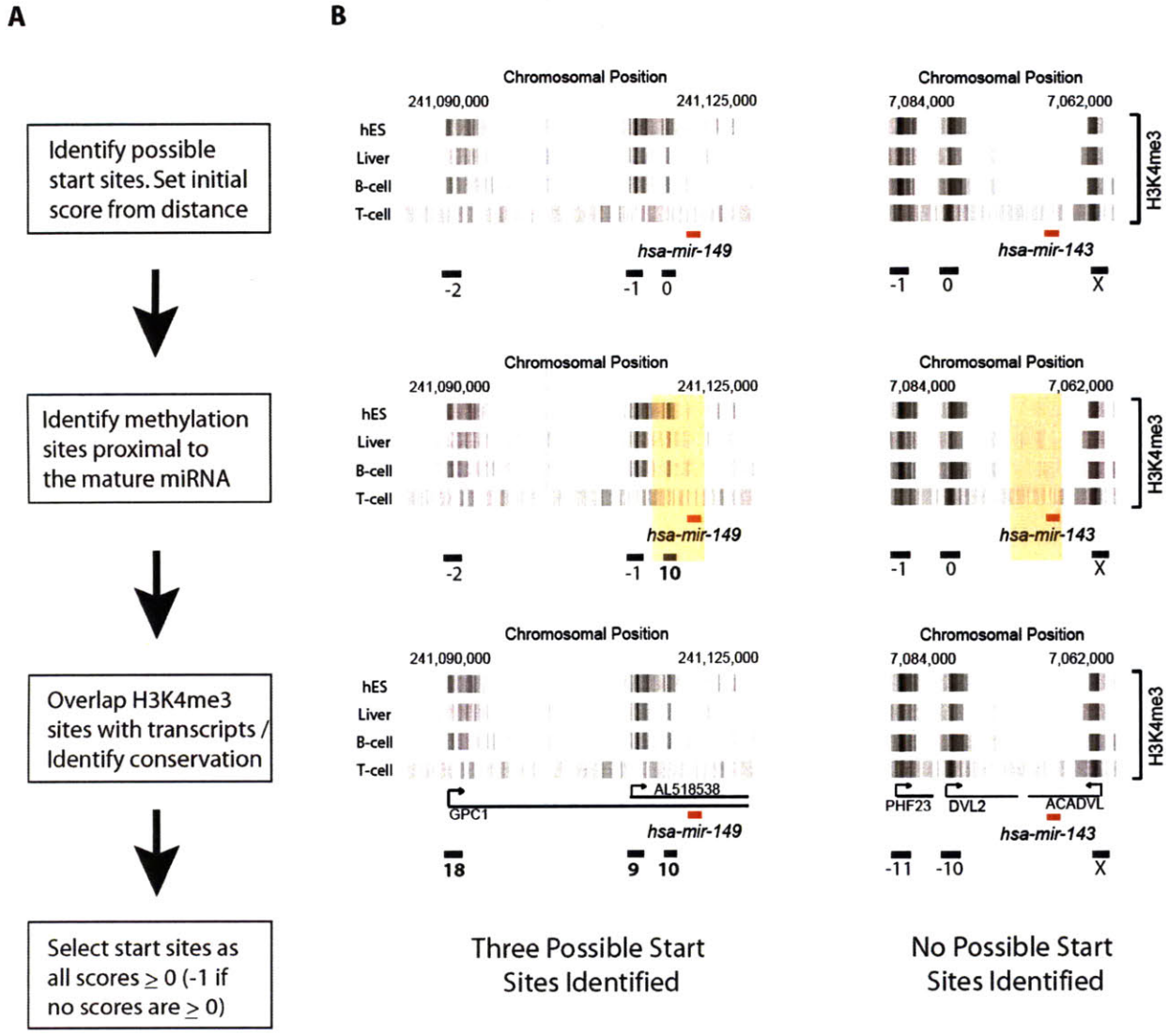

A
Identify possible start sites. Set initial score from distance
Identify methylation sites proximal to the mature miRNA
Overlap H3K4me3 sites with transcripts / Identify conservation
Select start sites as all scores ≥ 0 (-1 if no scores are ≥ 0)
B
Chromosomal Position
241,090,000
241,125,000
hES
Liver
B-cell
T-cell
hsa-mir-149
-2
-1
0
10
18
9
AL518538
GPC1
Three Possible Start Sites Identified
7,084,000
7,062,000
H3K4me3
hsa-mir-143
X
PHF23
DVL2
ACADVL
-11
-10
No Possible Start Sites Identified

Figure S5. Algorithm for identification of miRNA promoters

A. Flowchart describing the method used to identify the promoters for primary miRNA transcripts in human and mouse. For a full description, see supplemental text.

B. Two examples of identification of miRNA promoters. Top, Initial identification of possible start sites based on H3K4me3 enriched regions from four cell types. Enrichment of H3K4me3-modified nucleosomes is shown as shades of gray. Red bar represents the position of the mature miRNA. Black bars below the graph are regions enriched for H3K4me3. Initial scores are shown below the black bars. The region on the far right was excluded from the analysis (score = X) since it is downstream of the mature miRNA. Middle, Identification of candidate start sites <5kb upstream of the mature miRNA (yellow shaded area). Bottom, identification of candidate start sites that either initiate overlapping (left) or non-overlapping (right) transcripts. EST and transcript data is shown. Scores associated with identified genes are shown bold.

Figure S6

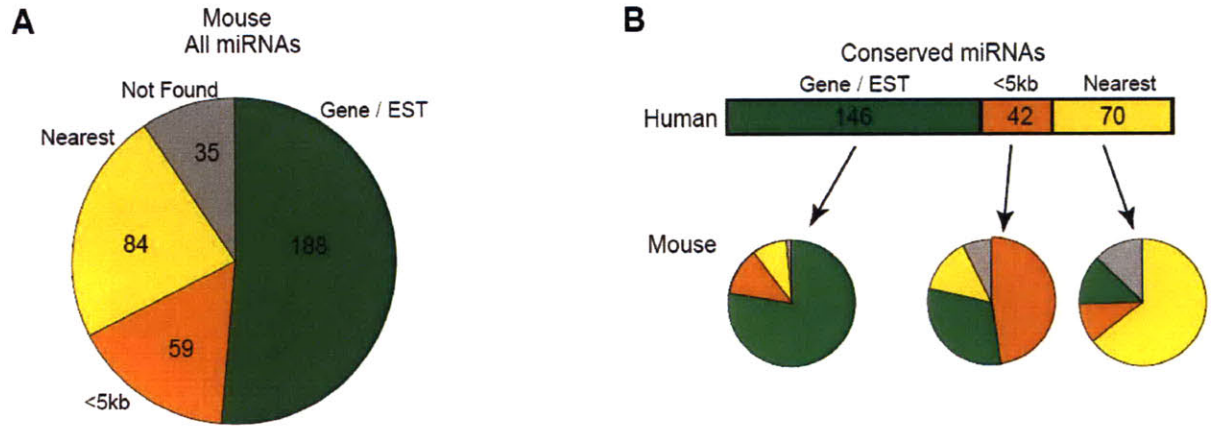


Figure S6. Summary of miRNA promoter classification

A. Promoters assigned to mature miRNAs were classified by the dominant feature of their scoring. Green: miRNAs that were found to have overlapping ESTs or genes confirming their promoters. Orange: miRNAs that were found to have a candidate start site within 5kb of the mature miRNA. Gray: miRNAs with either no candidates within 250kb of the mature miRNA or where all candidates had a score less then zero (see Fig. S4b, right). Yellow: miRNAs for which the closest candidate start site was selected solely on the basis of its proximity.

B. The basis of miRNA promoter identification, including Gene or EST evidence (green), distance of <5 kilobases to mature miRNA (orange), nearest possible promoter to miRNA (yellow), tended to be conserved between human and mouse

Figure S7

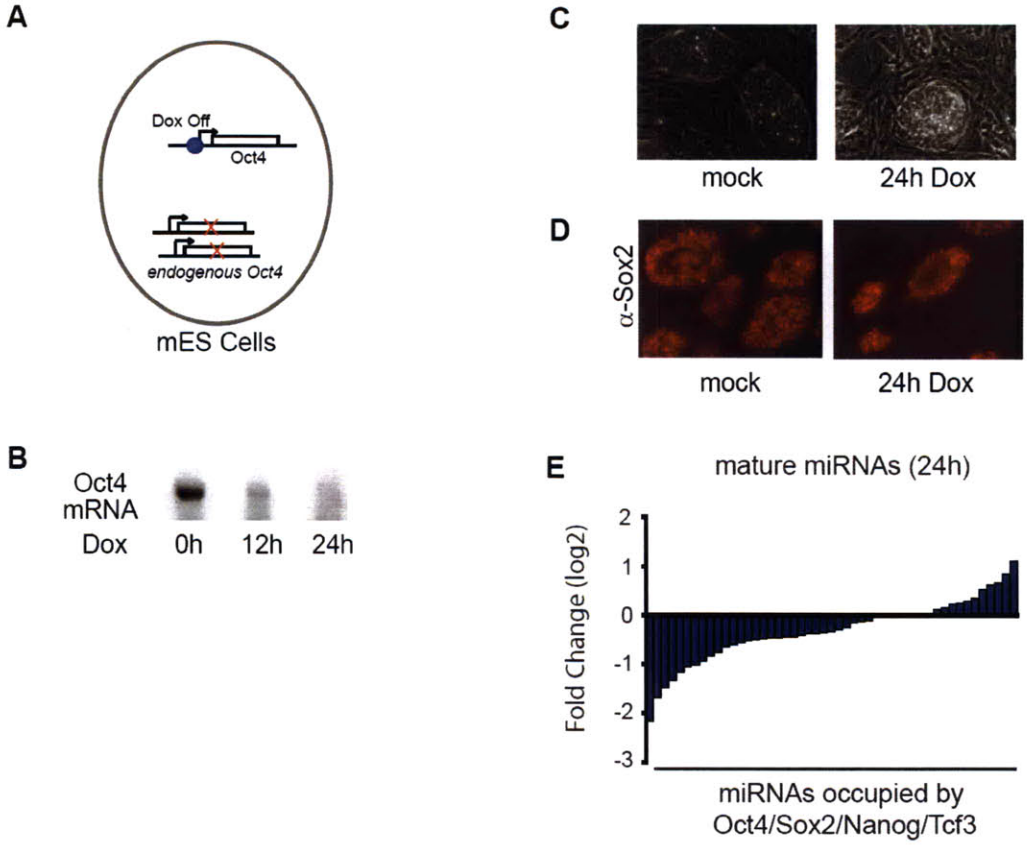


Figure S7. Regulation of miRNAs by Oct4

A. In an engineered murine cell line (Niwa et al., 2000), endogenous *Oct4* is deleted, and Oct4 expression is maintained by a Dox-repressible transgene.

B. By 24 hours of Dox-treatment, Oct4 mRNA levels are reduced as shown by reverse transcription (RT)-PCR.

C. 24 hours following Dox-treatment, cells remain ES-like by morphology.

D. 24 hours following Dox-treatment Sox2 protein can still be detected by immunofluorescence.

E. Changes in levels of Oct4/Sox2/Nanog/Tcf3 occupied mature miRNAs based on Solexa sequencing of small RNAs. Fold change was calculated by comparing normalized read counts from untreated cells and cells 24 hours after Dox treatment. A full list of miRNA reads can be found in Table S9. Details about the normalization procedure are contained in the supplemental text.

Figure S8

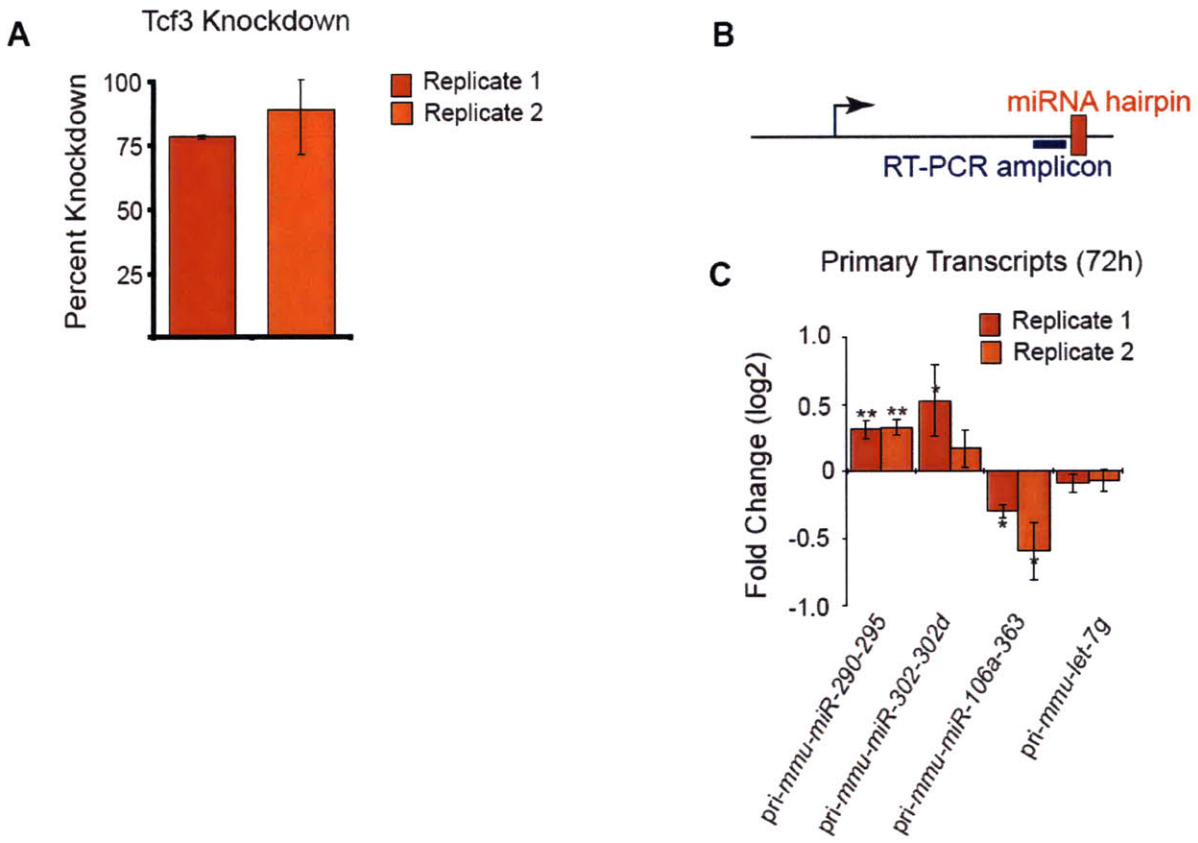


Figure S8. Regulation of miRNAs by Tcf3

Tcf3 was knocked down in V6.5 mES cells using lentiviral vectors containing shRNAs.

A. RT-PCR confirmation of knockdown at 72 hours post-infection using Taqman probes against Tcf3 (relative to levels in cells infected with GFP control lentivirus).

B. Schematic of the position of RT-PCR probes used to measure the levels of pri-miRNA transcripts in Figure 3d and part c

C. Results of quantitative RT-PCR analysis of probes designed to several pri-miRNAs occupied by Oct4/Sox2/Nanog/Tcf3. Change in the level of primary transcript compared to GFP control lentivirus are shown. * = $p < 0.05$, ** = $p < 0.001$ using a two-sampled t-test assuming equal variance. Standard deviation is indicated with error bars.

Figure S9

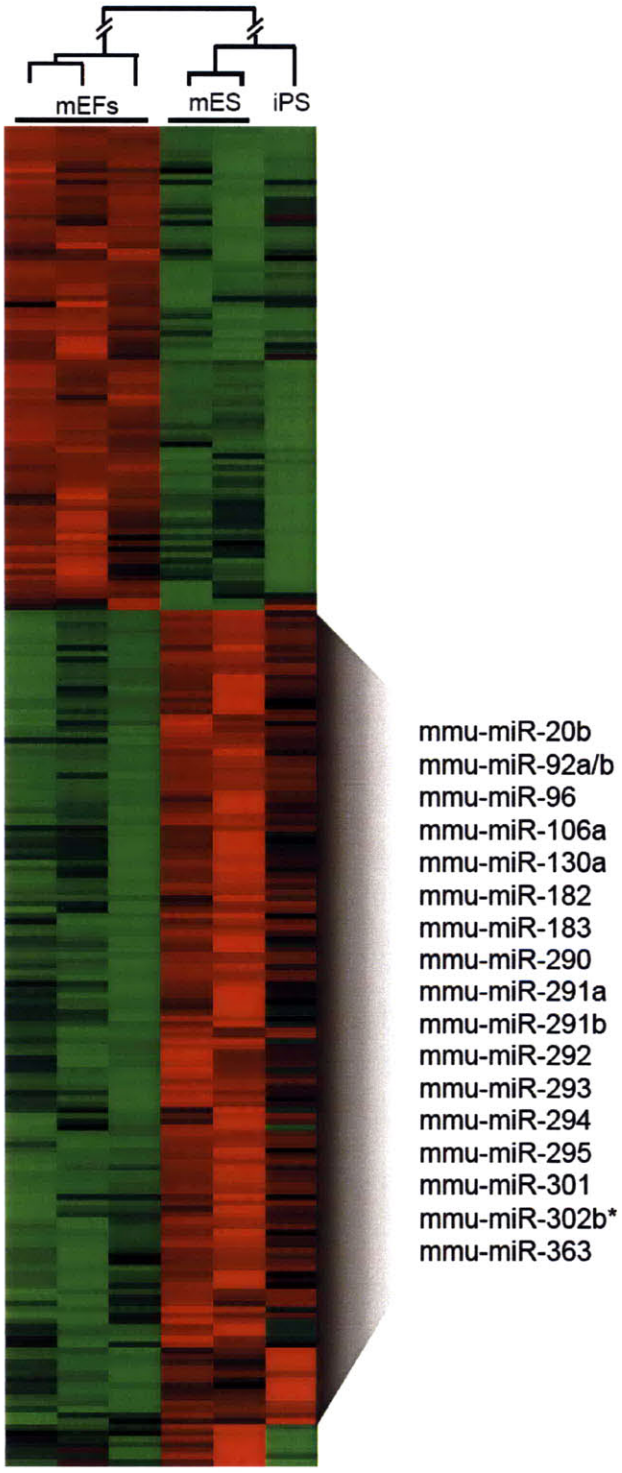

mEFs
mES
iPS
mmu-miR-20b
mmu-miR-92a/b
mmu-miR-96
mmu-miR-106a
mmu-miR-130a
mmu-miR-182
mmu-miR-183
mmu-miR-290
mmu-miR-291a
mmu-miR-291b
mmu-miR-292
mmu-miR-293
mmu-miR-294
mmu-miR-295
mmu-miR-301
mmu-miR-302b*
mmu-miR-363

Figure S9. miRNA genes occupied by the core master regulators in ES cells are expressed in iPS cells

RNA was extracted from MEFs (columns 1-3), mES cells (columns 4,5) and iPS cells (column 6) and hybridized to microarrays with LNA probes targeting all known miRNAs. Differentially expressed miRNAs enriched in either MEFs or mES cells are shown (FDR < 10%, see supplemental text, iPS cells were not used to determine differential expression). Data were Z-score normalized, and cell types were clustered hierarchically (top). Active miRNA promoters associated with Oct4/Sox2/Nanog/Tcf3 are listed to the right.

Figure S10

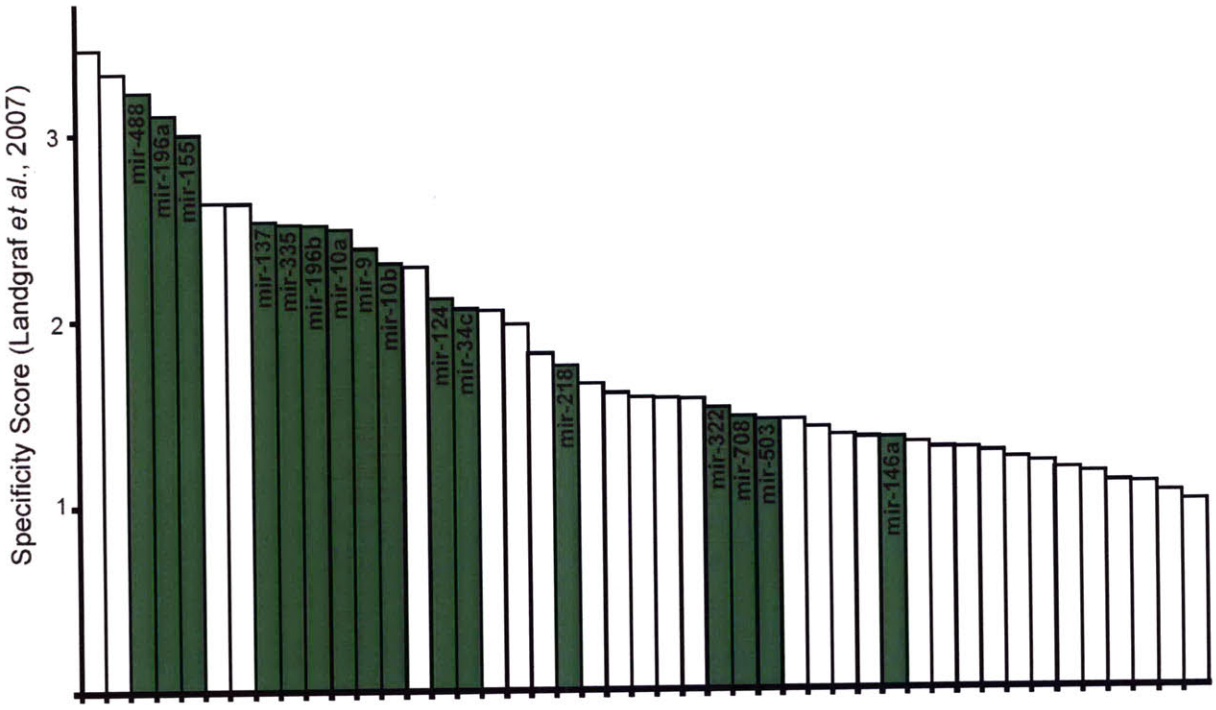


Figure S10. PcG occupied miRNAs are generally expressed in a tissue specific manner

Mature miRNAs derived from genes occupied by Suz12 and H3K27me3-modified nucleosomes were compared to the list of tissue specific miRNAs derived from the miRNA expression atlas (Landgraf et al., 2007). Vertical axis represents tissue-specificity and miRNAs with specificity score ≥1 are shown. miRNAs bound by Oct4/Sox2/Nanog/Tcf3 and expressed in mES cells are not shown (largely ES cell specific miRNAs). Among the tissue-specific miRNAs there is significant enrichment ($p < 0.005$ by hypergeometric distribution) for miRNAs occupied by Suz12 (green)

Index of Supplemental Tables

Appendix C

Supplemental Material for Chapter 4

Foxp3 Occupancy and Regulation of Key Target Genes During T Cell Stimulation

Supplemental Experimental Procedures

Cell Culture, stimulation and analysis of cytokine production

$CD4^+$ 5B6-2 hybridoma cells expressing a $PLP_{139-151}$-specific TCR, which was kindly provided by Vijay Kuchroo, were cultured in Dulbecco's modified Eagle medium (Invitrogen). Primary murine $CD4^+$ T cells were cultured in RPMI-1640 medium (Invitrogen). Media were supplemented with 10% FCS, 2 mM Glutamax, 1 mM HEPES, 1 mM sodium pyruvate, 0.1 mM non-essential amino acids, 0.55 mM 2-mercaptoethanol, 100U/ml penicillin/streptomycin and 0.1 mg/ml gentamycin. For gene expression profiling, real-time RT-PCR, and location analysis, cells were cultured in the absence or presence of 50 ng/ml phorbol 12-myristate 13-acetate (PMA) and 200 ng/ml ionomycin at 37°C and harvested after 6h. Where indicated, cells were preincubated for 1h with 2 μM cyclosporin A. For the analysis of cytokine production of 5B6-2 hybridoma cells, 10 □g/ml brefeldin A was added for the last 4h of 6 and 36h cultures. Cells were harvested at various time points and intracellular cytokine staining was performed using the Cytofix/Cytoperm kit (Becton Dickinson) according to manufacturer's recommendations and phycoerythrin-conjugated anti-mouse IL-2 (JES6–5H4) and TNF-□ (MP6-XT22) antibodies.

Generation of Foxp3-expressing $CD4^+$ T cell hybridoma clones

The murine full-length Foxp3 was amplified from cDNA of purified BALB/c $CD4^+CD25^+$ spleen cells using 5'-ATGCCCAACCCTAGGCCAGCCAA-3' as the sense and 5'-TCAAGGGCAG GGATTGGAGCAC-3' as the antisense primers. A minimal Kozak consensus sequence (double-underlined) just upstream of the initiator codon and a FLAG-tag (single-underlined) and were added using 5'-GAATTCACCATGATGGACTACAAGGACGA CGACGACAAGCCCAACCCTAGGCCAGCCAA-3' as the sense and 5'-GGATCCTCAA GGGCAGGGATTGGAGCAC-3' as the antisense primers. For further cloning sense and antisense primer sequences contained 5' EcoRI and BamHI restriction sites, respectively. Lentiviral vectors encoding the N-terminal FLAG-tagged Foxp3-IRES-GFP or Empty-IRES-GFP control were generated from the pLenti6/V5-D-TOPO vector (Invitrogen). The integrity of cloned cDNA was confirmed by sequencing and sequence comparison to GenBank accession no. NM_054039. Concentrated culture supernatants of 293FT transfected with lentiviral vectors were used to infect 5B6-2 hybridoma cells. Stably transduced 5B6-2-[Foxp3]-IRES-GFP or 5B6-

2-[Empty]-IRES-GFP 5B6-2 hybridoma cell clones were established after flow cytometric single-cell sorting.

Transgenic mice

TCR-hemagglutinin (TCR-HA) BALB/c mice express a transgenic TCR specific for the H2-IE^{d} $HA_{107-119}$ peptide. Double-transgenic TCR-HA x pgk-HA mice additionally express the HA protein under the control of the phosphoglycerate kinase (pgk) promoter and are characterized by high frequencies of TCR-HA-expressing T_{reg} cells (Klein et al., 2003). Mice were bred in the Dana–Farber Cancer Institute animal facility under specific pathogen-free conditions. Animal care and all procedures were in accordance with the guidelines of the Animal Care and Use Committee of the Dana-Farber Cancer Institute.

Purification and FACS analysis of primary $CD4^{+}$ T cells

Single cell suspensions of pooled spleen and lymph node cells were prepared from TCR-HA and TCR-HA x pgk-HA mice for the purification of naïve and T_{reg} cells, respectively. Cells were stained with fluorochrome-conjugated anti-mouse mAbs CD4 (RM4-5), CD25 (PC61) and TCR-HA (6.5). $HA_{107-119}$–specific $CD4^{+}CD25^{-}$ naïve T cells and Foxp3-expressing $CD4^{+}CD25^{high}$ T_{reg} cells were then purified using a FACSAria cell sorter and FACSDiva software (Becton Dickinson). Cells were ≥98% pure upon re-analysis. Intracellular staining with the anti-mouse/rat mAb FJK-16s (eBioscience) revealed that ~95% of purified TCR-$HA^{+}CD4^{+}CD25^{high}$ cells express Foxp3. $CD4^{+}CD25^{-}$ naïve T cells showed negligible staining for Foxp3 (≤0.5%). For the analysis of Ly6a surface expression on freshly isolated or *in vitro* stimulated antigen-specific T_{eff} and T_{reg} cells, primary T cells were purified as described above and additionally stained using a phycoerythrin-conjugated anti-mouse mAb Ly6a/e (E13-161.7).

Foxp3 Location Analysis

Chromatin immunoprecipitation

Protocols describing ChIP methods are available from: http://jura.wi.mit.edu/young_public /hESregulation/ChIP.html Briefly, for each location analysis reaction ~10^{8} N-terminal FLAG-

tagged Foxp3-IRES-GFP or Empty-IRES-GFP 5B6-2 hybridomas were chemically crosslinked by the addition of 11% formaldehyde solution for 20 min at room temperature. Cells were washed twice with 1 x PBS and pellets were stored at -80°C prior to use. Cells were resuspended, lysed, and sonicated to solubilize and shear crosslinked DNA. We used a Misonix Sonicator 3000 and sonicated at power 7 for 10 to 18, 20 second pulses (60 second pause between pulses) at 4°C while samples were immersed in an ice bath. The resulting whole cell extract was incubated overnight at 4°C with 100 ul of Dynal Protein G magnetic beads preincubated with 10 μg of the appropriate antibody for at least 4 hrs. Beads were then washed 4 times with RIPA buffer and 1 time with TBS. Bound complexes were eluted from the beads in elution buffer by heating at 65°C with occasional vortexing, and crosslinking was reversed by ~6 hour incubation at 65°C.Whole cell extract DNA (reserved from the sonication step) was also treated for crosslink reversal. Immunoprecipitated DNA and whole cell extract DNA were then purified by treatment with RNaseA, proteinase K and multiple phenol:chloroform:isoamyl alcohol extractions. Purified DNA was blunted, ligated to a universal linker and amplified using a two-stage PCR protocol. Amplified DNA was labeled using Invitrogen Bioprime random primer labelling kits (immunoenriched DNA was labeled with Cy5 fluorophore, whole cell extract DNA was labeled with Cy3 fluorophore). Labeled and purified DNA was combined (4 – 5 μg each of immunoenriched and whole cell extract DNA) and hybridized to arrays in Agilent hybridization chambers for 40 hours at 40°C. Arrays were then washed and scanned.

Antibodies for ChIP

For ChIP experiments, we used Anti-FLAG (Sigma M2) and anti-E2F4 (Santa Cruz 1082) antibodies. E2F4 antibody has been shown to specifically recognize previously reported E2F4 target genes (Ren et al., 2002; Weinmann et al., 2002). Anti-FLAG antibody has also been demonstrated to work for chromatin immunoprecipitation (Henry et al., 2003).

Ten slide promoter array

This study employed a 10-slide mouse promoter array set that has been used in previously published work (Boyer et al., 2006). These arrays were designed to contain oligonucleotides that cover approximately 10 kb around the transcription start sites of approximately 16,000 of the

best annotated mouse transcription start sites. Arrays were manufactured by Agilent Technologies (www.agilent.com).

Selection of regions and design of subsequences

To define transcription start sites, we first selected transcripts from three of the most commonly used databases for sequence information (Refseq, Ensembl, MGC). Transcription start sites within 500 bp of each other were considered single start sites. To restrict our array to the most likely transcription start sites, we selected only those that were found in at least two of the three databases. We also included microRNAs from the RFAM database.

25 kb of sequence around each transcription start site (20 kb upstream to 5 kb downstream) was initially extracted for analysis from the repeat-masked sequence derived from the May 2004 build of the mouse genome. Because we were balancing feature number, exact number of transcription start sites, tiling density and extent of upstream genomic coverage for our array design, we chose to design oligos across a much larger region than we were likely to fit on the array. Each transcription start site was considered independent, even if the 25 kb region overlapped with the 25 kb region of another transcription start site. While we anticipated not being able to use all of these oligos, this allowed us flexibility in later design steps to add oligos for additional upstream genomic coverage if space became available. The subset of probes from –8 kb to +2 kb was selected for the actual array. We used the program ArrayOligoSelector 10 (AOS; http://arrayoligosel.sourceforge.net/) to score 60-mers for every unmasked subsequence greater than 62 bp across all promoter regions. The scores for each oligo were retained but not put through the built-in AOS selection process. Instead, the collection of scored 60-mers was divided by promoter and sorted by genomic position. Each set of 60-mers was then filtered based on the AOS oligo scoring criteria: GC content, self-binding, complexity and uniqueness. For our most stringent filter, we selected the following ranges for each parameter: GC content between 30 percent and 100 percent, self-binding score less than 100, complexity score less than or equal to 24, uniqueness greater than or equal to –40.

From this subset of 60-mers, we selected oligos designed to cover the promoter region with an estimated density of one probe every 280 bp. At this point, we restricted oligo selection to those oligos found within the region 8 kb upstream to 2 kb downstream of the transcription start site. To achieve more uniform tiling, we instituted a simple method to find probes within a

particular distance from each other. Starting at the upstream end of the region, we selected the first qualified probe and then selected the next qualified probe located 150 bp to 280 bp away. If there were multiple eligible probes, we chose the most distal probe within the 280 bp limit. If no probes were identified within this limit, we continued scanning until we found the next acceptable probe. The process was then repeated with the most recently selected probe until we reached the end of the promoter region.

For regions that were not covered by high quality probes, we returned to the full set of scored 60-mers and filtered using less stringent criteria. This gave us an additional set of 60-mers that we then used to fill gaps in our coverage. After this second pass, we identified gaps in our coverage and added oligos that were properly spaced and best fit our criteria regardless of whether they passed the filter cutoffs. This iterative process gave us a compromise between optimal probe quality and optimal probe spacing.

Compiled probes and controls

There are 407,355 features split over 10 arrays. The probes are arranged such that array 1 begins with the first selected transcript start site on the left arm of chromosome 1, array 2 picks up where array 1 ends, array 3 picks up where array 2 ends, and so on. Over 16,000 genes are represented on the arrays and each promoter region corresponding to a unique start site is covered by approximately 25-27 probes. A true average distance between probes is difficult to calculate due to the presence of large gaps in the probe tiling. Most of these gaps simply represent the distance between the first and last oligos of two different sets of probes designed against two different genes. Other gaps are caused by lack of available sequence information, repeat masking or sequences that are otherwise highly repetitive and not suitable for oligo design.

Several sets of controls were added. A total of 353 oligos representing Arabidopsis thaliana genomic sequence were included. These Arabidopsis oligos were BLASTed against the mouse genome and do not register any significant hits. These oligos were intended to check background signal. We added a total of 186 oligos representing five proximal promoter regions of genes that are known targets of the transcriptional regulator Oct4 (Pipox, Foxh1, Oct4/Pou5f1, Msh2 and Hoxb1). Each of the four promoters is represented by 21 - 32 different oligos that are evenly positioned across the regions. The oligos for the Hoxb1 region are printed an additional two

times. These promoter regions can be used as positive controls. There are 481 gene desert controls. To identify these probes, we identified intergenic regions of 1 Mb or greater and designed probes in the middle of these regions. These are intended to identify genomic regions that are most likely to be unbound by promoter-binding transcriptional regulators (by virtue of their extreme distance from any known gene). We have used these as normalization controls in situations where a factor binds to a large number of promoter regions. There are 224 features printed as intensity controls; 37 oligos are printed twice and 25 of these 37 are printed an additional six times. Based on a limited number of test hybridizations, this set of oligos gives signal intensities that cover the entire dynamic range of the array. Our intention was that this set could serve as a way to normalize intensities across multiple slides throughout the entire signal range. There are 2,256 controls added by Agilent (standard) and the remainder of each array consists of blank spots.

Single slide proximal promoter array

Anti-FLAG Foxp3 IPs were compared to control IPs (E2F4, empty vector) on a single slide array with ~95,000 probes (Supplemental Figure S4). Oligo probes were designed essentially as described above. Probes were designed to tile the entire genome with 1 probe placed every ~250bp. A subset of these probes, covering 800bp upstream and 200bp downstream of annotated transcriptional start sites, were then selected to cover the proximal promoters of approximately 18,000 genes.

Array scanning and data extraction

Slides were scanned using an Agilent DNA microarray scanner BA. PMT settings were set manually to normalize bulk signal in the Cy3 and Cy5 channel. For efficient batch processing of scans, we used GenePix 6.0 software (Molecular Devices). Scans were automatically aligned and then manually examined for abnormal features. Intensity data were then extracted in batch. The complete ChIP-chip datasets have been submitted to the online data repository ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) and are associated with accession code E-TABM-154.

Data normalization and analysis

GenePix was used to obtain background-subtracted intensity values for each fluorophore for every feature on the array. To obtain set-normalized intensities we first calculated, for each slide, the median intensities in each channel for a set of control probes that are included on each array. We then calculated the average of these median intensities for the set of 10 slides. Intensities were then normalized such that the median intensity of each channel for an individual slide equalled the average of the median intensities of that channel across all slides.

Each slide contains a set of negative control spots that contain 60-mer sequences that do not cross-hybridize to murine genomic DNA. We calculated the median intensity of these negative control spots in each channel and then subtracted this number from the set-normalized intensities of all other features.

To correct for different amounts of genomic and immunoprecipitated DNA hybridized to the chip, negative control-subtracted median intensity value of the IP enriched DNA channel for the set of intensity control probes described above was then divided by the median of the genomic DNA channel for the same set of probes. This yielded a normalization factor that was applied to each intensity in the genomic DNA channel.

Because binding events are rare in the genome, DNA fragments can only be enriched using ChIP, and not anti-enriched. Therefore, the distribution of probes that are below the 1:1 axis (X-score = 0) can provide an empirical non-parametric noise distribution for the experiment. For each X score above 0, the probability of enrichment was calculated to be equal to the number of probes with an X score greater then (more enriched) the test score divided by the total number of probes (enriched + noise) with an absolute value of the X score greater then the test score. This calculation removes the assumption that the X scores on a given array are normally distributed.

Identification of bound regions

To automatically determine bound regions in the datasets, we developed an algorithm to incorporate information from neighbouring probes. For each 60-mer, we calculated the average X score of the 60-mer and its two immediate neighbours. If a feature was flagged as abnormal during scanning, we assumed it gave a neutral contribution to the average X score. Similarly, if an adjacent feature was beyond a reasonable distance from the probe (1000 bp), we assumed it gave a neutral contribution to the average X score. The distance threshold of 1000 bp was determined based on the maximum size of labeled DNA fragments put into the hybridization.

Since the maximum fragment size was approximately 550 bp, we reasoned that probes separated by 1000 or more bp would not be able to contribute reliable information about a binding event halfway between them. This set of averaged values gave us a new distribution that was subsequently used to calculate test statistics for each probe set.

As most probes were spaced within the resolution limit of chromatin immunoprecipitation, we next required that multiple probes in the probe set provide evidence of a binding event. Bound probes were required to have a single probe probability of enrichment and a probe set probability of enrichment greater than 0.95 (5% False Discovery Rate) for high stringency binding calls and .90 (10% False Discovery Rate) for low stringency binding calls. Individual probe sets that passed these criteria and were spaced closely together were collapsed into bound regions if the centre probes of the probe sets were within 1000 bp of each other.

Comparing bound regions to known genes

The location of all bound regions was compared to a composite set of transcripts compiled from three databases: RefSeq (Pruitt et al., 2005), Ensembl (Hubbard et al., 2005), and UCSC annotated known genes (http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=knownGene) that were associated with Entrez Gene identifiers and miRNAs downloaded from the the RFAM database (http://www.sanger.ac.uk/Software/Rfam/). By this method approximately 16,000 Entrez Genes have a least one probe within -8 kb to +2 kb of their transcription start site. All coordinate information was downloaded from the UCSC Genome Browser (NCBI build 6; March 2005). For genes, we assigned bound regions within the -8 kb to +2 kb of the transcriptional start site, which is consistent with our array design. MicroRNAs allowed a less stringent binding within 10kb upstream or downstream of the 5' end of the miRNA.

Control location analysis experiments

Control location analysis experiments demonstrate the specificity of our antibodies for the appropriate targets. When Foxp3 location analysis was performed in hybridoma cells that were not transduced with FLAG-tagged Foxp3 almost no enrichment of DNA was detected. Location analysis experiments with anti-E2F4 antibodies identified expected E2F4 targets, which are dissimilar from the Foxp3 targets that were discovered. These control experiments were performed using the single slide proximal promoter arrays.

Site-specific PCR analysis

We used site-specific PCR to confirm binding of Foxp3 to a panel of Foxp3 target genes identified by ChIP-chip. A subset of the bound probe sets was selected and primer pairs were designed to amplify a 150-200 bp region around the genomic location of probes that show peak levels of immunoenrichment. PCR was performed on ligation-mediated PCR amplified IP samples (Figure 3c). 10 ng of immunoprecipitated (IP) DNA was used in PCR reactions. For input whole cell extract (WCE) samples, a range of DNA amounts (90, 30 and 10 ng of DNA) was used. The PCR was performed for 23 cycles and products were visualized on an agarose gel stained with SYBR Gold (Amersham) and quantified using ImageQuant (Amersham). Enrichment ratios shown in Figure 3c were calculated as the ratio of the intensity of the PCR product from 10ng IP DNA to the product from 10ng WCE DNA. Ratios were normalized relative to the ratio observed for the unenriched *ß actin* control. For site-specific PCR in unamplified IP material from cyclosporin A treated cells (Figure S3), ChIP protocol was followed through the proteinase K digestion. For site-specific PCR, 2ul of IP material was used. Again, for whole cell extract (WCE) samples a range of DNA amounts (90, 30 and 10 ng of DNA) was used. 28 cycles of PCR were performed and products were visualized on agarose gel with ethidium bromide. For all site-specific ChIP analysis, a region of the *ß actin* promoter where no Foxp3 binding is expected is used as a negative control.

GENE	5' OLIGO	3' OLIGO	SIZE
Itk	GCTGTTCTTCCAGGAGGATG	AGGCTGGCTGATGCTGATAG	190
Jak2	ACGGCAGGACTAATTGTTGC	GAAAGGGGGAGAAAGAGACG	180
Zap70	TCTAGGACAGGAACACATTGGGTGTCGGGAACACAAGAGGA		158
Ptpn22	TTCTGCCTTTCTTCTGGGAAT	CTAGCGCCTTCCTTTCTCAA	162
Il2	GTCCTCATGGGCTCAACATC	GGGAGGCCAACCTTTGTAAT	156
Pou2af1	TTCATGAGACGGAAACCACA	CACATCTACAGGAGGGAACCA	156
Ly6a	CCCAGCACAGTGGTAAGAGG	GGCAGGGTTTATCACTTGGA	182
Tnfrsf9	TGTGTGTGTGAAGAGGGGTTT	TCCACAGACGTGACAAGGAG	151
CD25	GGGTGAAAAGACAGCTTGGT	GGGTGTGGGATTCACAAATG	151
ß Actin	AGGGTACCACCGGAAAAGTC	CCCCAAAGGCTGAGAAGTTA	150

DNA motif analysis

Discovery of the Foxp3 sequence motif from the ChIP-chip binding data was performed using the THEME algorithm (MacIsaac et al., 2006). THEME tests specific, biologically informed hypotheses about a transcription factor's binding specificity and identifies a motif consistent with both the binding data and prior knowledge regarding the protein's DNA-binding domain structural family.

We extracted genomic sequence corresponding to the regions bound by Foxp3 at high confidence in stimulated cells for use as the "foreground" data set in THEME. We then extracted sequence regions at random from unbound regions on the array for use as the "background" data set. The length of these unbound sequences was matched to the average size of the bound set (700bp) to avoid biasing either set towards motif presence or absence. We then ran THEME, testing hypotheses consistent with the Forkhead DNA-binding domain family, and identified the motif that yielded the lowest bound vs. unbound classification error after 5-fold cross-validation. The statistical significance of the best motif's cross-validation error was assessed by running THEME on randomized data (using the starting hypothesis from which the best motif was derived) and calculating a Z-score for the observed error under the null hypothesis that the sequences were selected at random from the background set. We then tested hypotheses corresponding to 35 other distinct DNA-binding domains in an identical fashion, and found that the best motif identified from the Forkhead family yielded a lower cross-validation error than any motif from all other families tested.

The Foxp3 motif learned by THEME, and the Nfat motif from the TRANSFAC (version 8.3) database (Wingender et al., 1996) were used to scan all arrayed sequences to identify matches to the motif. Each potential site was assessed by summing the position-specific scores from the motif log-odds matrix. Sites were identified as matches if their score was greater than or equal to the threshold, determined by the THEME algorithm, which classified bound and unbound sequences with the lowest error during motif discovery. For the Nfat motif from TRANSFAC a threshold of 60% of the maximum possible score was used. Foxp3 motif conservation in bound regions was determined using mm6/hg17 mouse-human pairwise alignments obtained from the UCSC Genome Browser (Karolchik et al., 2003). If the human sequence directly aligned to the motif match in mouse also met the score threshold we identified that site as a conserved match. We performed the same conservation calculations for randomly selected unbound microarray sequence regions. Statistical significance was determined by fitting

a hypergeometric distribution to the data and testing the null hypothesis that the number of conserved Foxp3 motifs observed in bound regions arose from random selection without replacement from the background population. We then examined the distribution of spacings between Foxp3 and Nfat motifs (when they occurred together in the same Foxp3 bound region). Statistical significance was determined using the hypergeometric test described above. Results are summarized in Supplemental Table S5. The online tool WebLogo (http://weblogo.berkley.edu) was used to generate DNA motif sequence logo in Figure 2b.

Gene Expression Analysis

Gene expression profiling

For each hybridoma culture condition, total RNA was prepared from 1 x 10^7 cells using Trizol (Gibco) followed by additional purification using the RNeasy Mini Kit (Qiagen). Biotinylated antisense cRNA was then prepared according to the Affymetrix standard labelling protocol (one amplification round). For each primary T cell culture condition, total RNA was isolated from 5 x 10^5 cells with RNeasy. Biotinylated antisense cRNA was prepared by two rounds of *in vitro* amplification using the BioArray RNA Amplification and Labeling System (Enzo Life Sciences) according to the protocol for 10-1000 ng of input RNA provided by the manufacturer. Biotinylated cRNAs of hybridomas and primary T cells were fragmented and hybridized to Affymetrix GeneChip Mouse Expression Set 430 2.0 arrays at the Microarray Core Facility (Dana-Farber Cancer Institute). Arrays were stained, scanned, and quantified according to standard Affymetrix protocols. Data were annotated according the NetAffx database (http://www.affymetrix.com/analysis/index.affx) as of March, 2006. The complete expression datasets have been submitted to the online data repository ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) and are associated with accession code E-TABM-154.

Expression data normalization

Quantile normalization was performed separately on the hybridoma and *ex vivo* expression datasets. Expression data were ranked within each sample. Each probeset was given the value of the average signal of the probesets of that rank, across the dataset.

Identification of differentially expressed genes

The hybridoma dataset was analyzed for statistically significant differential expression using the online NIA Array Analysis Tool (Sharov et al., 2005; http://lgsun.grc.nia.nih.gov/ANOVA/). Probesets were tested for differential expression using the following settings:

Threshold z-value to remove outliers: 10000
Error Model: Max(Average,Bayesian)
Error variance averaging window: 200
Proportion of highest error variances to be removed: 0.01
Bayesian degrees of freedom: 5
FDR threshold: .05

Of 45,101 probesets on the Affymetrix Mouse 430 2.0 array, 256 were differentially expressed between Foxp3- and Foxp3+ stimulated hybridomas, while 23 probesets were differentially expressed in the unstimulated cells. The expression data for these probesets are provided (Supplemental Tables S7 and S8). Probesets were excluded that had an average signal intensity that was not in the upper tercile on the arrays (54.4 units). Probesets that did not map to the genes for which we had Foxp3 binding data and probesets that mapped to multiple genes were also excluded. In the cases where multiple probesets mapping to one gene were differentially expressed, only the probeset showing the largest differential expression was displayed.

It is worth noting that a stringent cutoff to call differential expression is used. This identifies expression changes with high confidence but produces an underestimate because there are many genes that show small changes. Our goal is to gain insight into Foxp3 action, and by focusing on only the most pronounced transcriptional effects, we aim to minimize the effects of noise in the expression data. As a result of this approach, some genes that are likely to be regulated directly by Foxp3 exhibit small transcriptional effects that are not called differentially expressed. For example the genes *Bcl10*, *Cd53*, *Rbpsuh*, and *Rgs1* are all direct Foxp3 binding targets that are known to play a role in regulation of T cells, and have the characteristic expression pattern of suppressed activation, but do not meet the FDR < 5% statistical significance cutoff for differential expression.

Genes with consistent Foxp3 dependent differential expression in between stimulated *ex vivo* T helper and T_{reg} cells and between stimulated Foxp3- and Foxp3+ hybridoma cells were determined according to the following method. Probesets were excluded that had an average signal intensity that was smaller than the median signal on the arrays (26.2 units). Probesets that did not map to the genes for which we had Foxp3 binding data and probesets that mapped to multiple genes were also excluded. In the cases where multiple probesets mapping to one gene were differentially expressed, only the probeset showing the largest differential expression was displayed.

A score for Foxp3 dependent differential expression was calculated in the hybridoma and *ex vivo* datasets separately. The product of these scores was used to sort the genes and identify those with the largest Foxp3 dependent differential expression, which was consistent in the two cell types. To calculate each score, the average signal intensity within the two groups being compared was calculated. The difference in signal between the groups being compared was divided by the median signal intensity of all probesets on the array (26.2 units) plus one eighth of the average signal intensity for that probeset.

$$(A - B) / ((A + B) / 16 + \text{median})$$

This generated a differential expression score that is comparable to a signal to noise ratio, where noise is estimated be a linear function of signal intensity. In Figure S5, the 125 genes with the highest overall differential expression score were displayed, to match the number of genes that are shown in Figure 3A.

Hierarchical clustering and heatmap display

For clustering and heat map display, expression data were Z-score normalized separately within the hybridoma and *ex vivo* datasets. For heatmap display in Figure S5 data were Z-score normalized within the full hybridoma and *ex vivo* datasets including the unstimulated samples, though only the stimulated samples are displayed. Average linkage, correlation distance, centered, hierarchical clustering was performed using Gene Cluster (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm#ctv). Heatmaps were generated using Java Treeview (http://jtreeview.sourceforge.net/). Cluster branches were flipped about tree nodes for

optimal display. In Figure 3B, those genes from panel A that are bound by Foxp3 and are expressed in *ex vivo* cells are displayed in panel B. *Slc17a6* and *Adam10* are excluded because they are not expressed in the *ex vivo* samples.

Real-time RT-PCR

Total RNA was prepared from hybridoma cells or FACS purified primary T cell populations using the RNeasy kit (Qiagen) followed by DNase digestion (Qiagen). cDNA was synthesized from total RNA using Superscript II reverse transcriptase and oligo(dT) (Invitrogen Life Technologies) according to the manufacturer's recommendations. Real-time RT-PCR was performed on an ABI PRISM thermal cycler (Applied Biosystems) using SYBR® Green PCR core reagents (Applied Biosystems). Real-time RT-PCR primer sets were either obtained from SuperArray or are available upon request.

Significance of list overlap

Statistical significance of overlap between differentially expressed genes and Foxp3 bound genes was calculated using a standard Chi-Square test.

Functional annotation and statistical significance of gene lists

Functional annotation and statistical significance of gene lists was performed with the on-line tool, DAVID (http://niaid.abcc.ncifcrf.gov/) (Dennis et al., 2003). Genes were imported as EntrezGene IDs and, using the Functional Annotation tool, compared to KEGG pathways (Kanehisa et al., 2000).

The Foxp3 binding targets in PMA/ionomycin-stimulated $CD4^+$ hybridomas were enriched for genes associated with the following KEGG pathways:

KEGG Pathway	**P-Value**
T CELL RECEPTOR SIGNALING PATHWAY	1.4E-5
CELL CYCLE	4.3E-3
FATTY ACID ELONGATION IN MITOCHONDRIA	2.5E-2
CYTOKINE-CYTOKINE RECEPTOR INTERACTION	5.9E-2
PYRIMIDINE METABOLISM	9.2E-2

The Foxp3 target genes that are downregulated in Foxp3+ stimulated hybridomas relative to their levels in Foxp3- stimulated hybridomas were enriched for genes associated with the following KEGG pathways:

KEGG Pathway	**P-Value**
T CELL RECEPTOR SIGNALING PATHWAY	6.1E-3
CYTOKINE-CYTOKINE RECEPTOR INTERACTION	3.6E-2

Supplemental Tables are available online at:

http://www.nature.com/nature/journal/v445/n7130/suppinfo/nature05478.html

References

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.

Chen, C., Rowell, E. A., Thomas, R. M., Hancock, W. W., and Wells, A. D. (2006). Transcriptional regulation by Foxp3 is associated with direct promoter occupancy and modulation of histone acetylation. J Biol Chem *281*, 36828-36834.

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

Henry, K. W., Wyce, A., Lo, W. S., Duggan, L. J., Emre, N. C., Kao, C. F., Pillus, L., Shilatifard, A., Osley, M. A., and Berger, S. L. (2003). Transcriptional activation via sequential histone H2B ubiquitylation and deubiquitylation, mediated by SAGA-associated Ubp8. Genes Dev *17*, 2648-2663.

Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., *et al.* (2005). Ensembl 2005. Nucleic Acids Res *33*, D447-453.

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

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

Klein, L., Khazaie, K., and von Boehmer, H. (2003). In vivo dynamics of antigen-specific regulatory T cells not predicted from behavior in vitro. Proc Natl Acad Sci U S A *100*, 8886-8891.

Macisaac, K. D., Gordon, D. B., Nekludova, L., Odom, D. T., Schreiber, J., Gifford, D. K., Young, R. A., and Fraenkel, E. (2006). A hypothesis-based approach for identifying the binding specificity of regulatory proteins from chromatin immunoprecipitation data. Bioinformatics *22*, 423-429.

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res *33*, D501-504.

Ren, B., Cam, H., Takahashi, Y., Volkert, T., Terragni, J., Young, R. A., and Dynlacht, B. D. (2002). E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. Genes Dev *16*, 245-256.

Sharov, A. A., Dudekula, D. B., and Ko, M. S. (2005). A web-based tool for principal component and significance analysis of microarray data. Bioinformatics *21*, 2548-2549.

Weinmann, A. S., Yan, P. S., Oberley, M. J., Huang, T. H., and Farnham, P. J. (2002). Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. Genes Dev *16*, 235-244.

Wingender, E., Dietze, P., Karas, H., and Knuppel, R. (1996). TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res *24*, 238-241.

Wu, Y., Borde, M., Heissmeyer, V., Feuerer, M., Lapan, A. D., Stroud, J. C., Bates, D. L., Guo, L., Han, A., Ziegler, S. F., *et al.* (2006). FOXP3 controls regulatory T cell function through cooperation with NFAT. Cell *126*, 375-387.

Figure S1

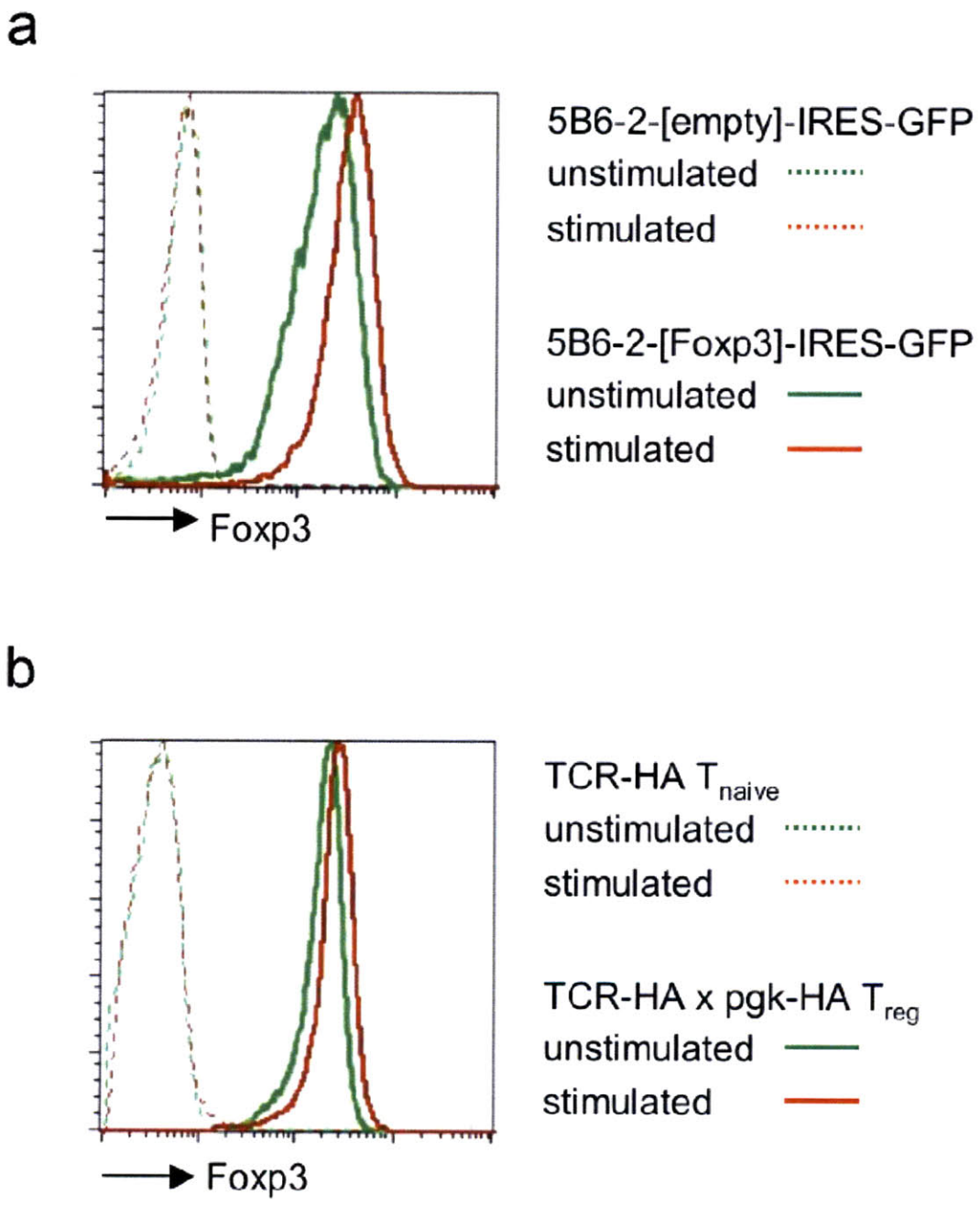

a
5B6-2-[empty]-IRES-GFP
unstimulated
stimulated
5B6-2-[Foxp3]-IRES-GFP
unstimulated
stimulated
Foxp3
b
TCR-HA T_{naive}
unstimulated
stimulated
TCR-HA x pgk-HA T_{reg}
unstimulated
stimulated
Foxp3

Figure S1. Flow cytometric analysis of Foxp3 expression

5B6-2-[Foxp3]-IRES-GFP or 5B6-2-[Empty]-IRES-GFP 5B6-2 hybridoma cells (**A.**) or FACS-purified *ex vivo* TCR-HA transgenic $CD4^+CD25^-$ naïve or $CD4^+CD25^+$ Treg cells (**B.**) were cultured in the absence or presence of PMA/ionomycin. After 6h cells were harvested and intracellular staining was performed using the mAb FJK16s (anti-Foxp3). Histograms show relative levels Foxp3 protein expression, which demonstrates that Foxp3 is expressed at a similar level in the transduced hybridomas as in *ex vivo* T_{reg} cells.

Figure S2

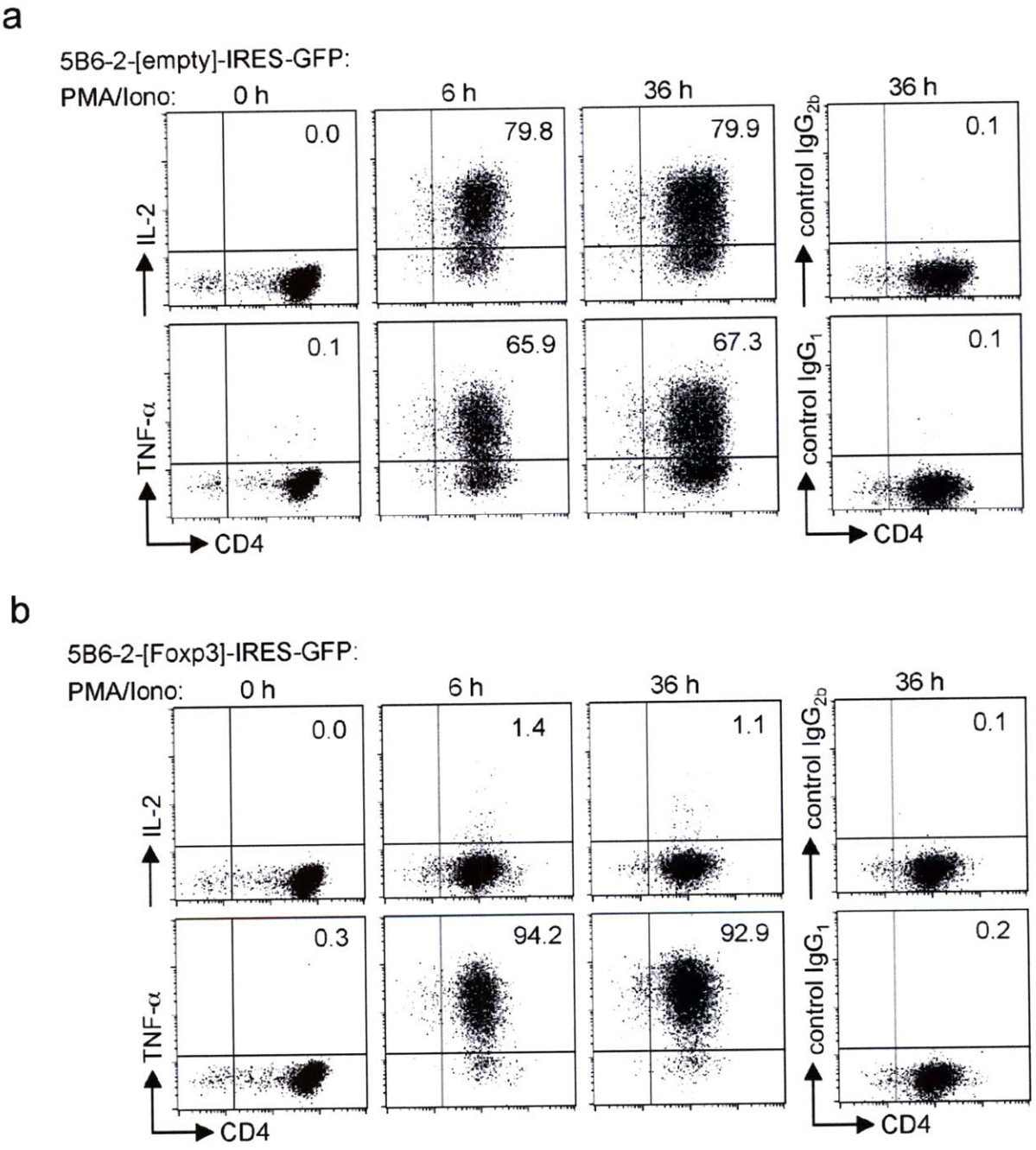

a
5B6-2-[empty]-IRES-GFP:
PMA/Iono:
0 h
6 h
36 h
36 h
0.0
79.8
79.9
0.1
IL-2
control IgG2b
0.1
65.9
67.3
0.1
TNF-α
control IgG1
CD4
CD4
b
5B6-2-[Foxp3]-IRES-GFP:
PMA/Iono:
0 h
6 h
36 h
36 h
0.0
1.4
1.1
0.1
IL-2
control IgG2b
0.3
94.2
92.9
0.2
TNF-α
control IgG1
CD4
CD4

Figure S2. Cytokine production of Foxp3-transduced 5B6-2 hybridomas

5B6-2-[empty]-IRES-GFP (**A.**) or 5B6-2-[Foxp3]-IRES-GFP (**B.**) hybridoma cells were cultured in the presence of PMA/ionomycin for the indicated times. Brefeldin A was added for the last 4h of 6h and 36h cultures. After surface staining for CD4 expression and fixation, intracellular staining with the indicated cytokine antibodies or appropriate isotype controls was performed. Numbers in dot plots indicate the frequencies of cells in the respective quadrant. These data show that Foxp3 transduction suppresses production of Il2, but does not suppress production of TNF-alpha.

Figure S3

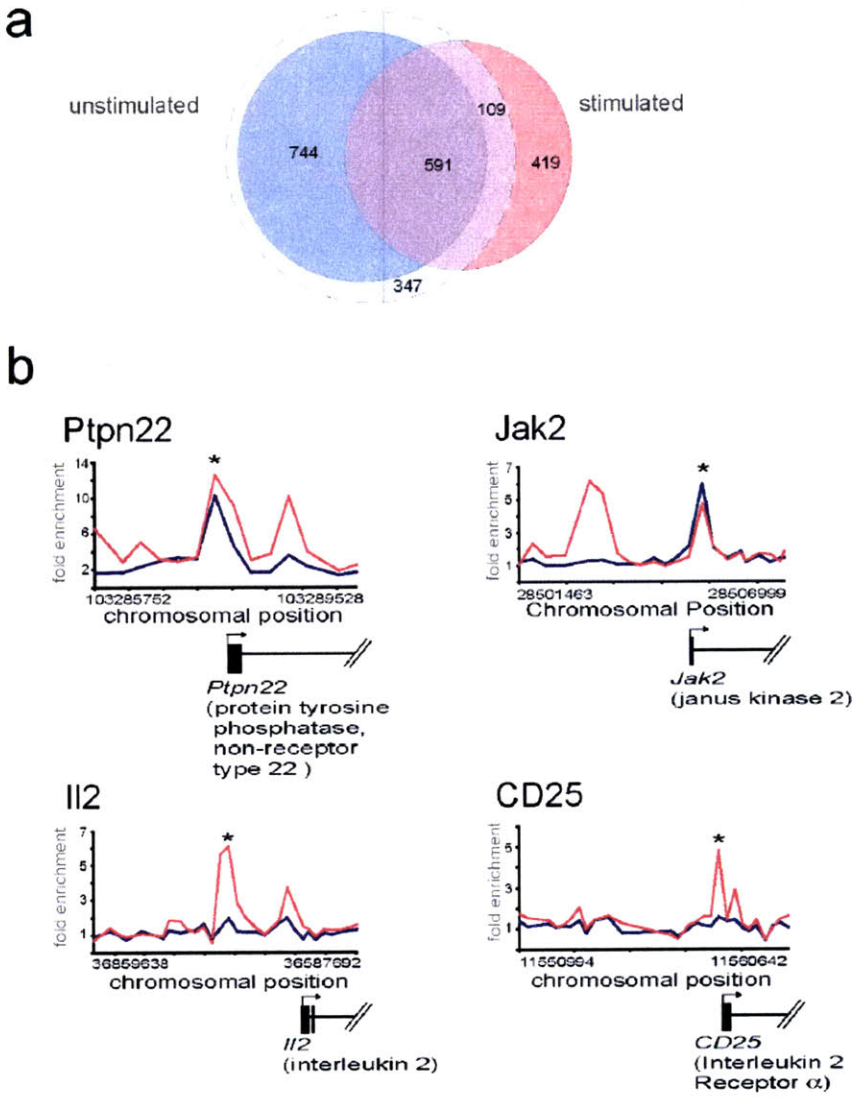

a
unstimulated
stimulated
744
591
109
419
347
b
Ptpn22
fold enrichment
chromosomal position
103285752
103289528
Ptpn22
(protein tyrosine phosphatase, non-receptor type 22)
Jak2
Chromosomal Position
28501463
28506999
Jak2
(janus kinase 2)
Il2
36859638
36587692
Il2
(interleukin 2)
CD25
11560994
11560642
CD25
(Interleukin 2 Receptor α)
c
cyclosporin A treated cells
Foxp3 bound
Foxp3 unbound
WCE IP WCE IP
WCE IP WCE IP WCE IP
200
100
Ptpn22 Jak2
Il2 CD25 B Actin

Figure S3. The promoters of most Foxp3 target genes are bound by Foxp3 before and after T cell stimulation

A. The lists of genes whose promoters are bound by Foxp3 in unstimulated (blue) and stimulated (pink) hybridoma cells are shown in a Venn diagram. The genes occupied by Foxp3 in stimulated T cells (FDR < .05) are represented by the pink circle. The genes occupied by Foxp3 in unstimulated T cells (FDR < .05) are represented by the blue circle. The dotted light blue circle represents the genes occupied by Foxp3 in unstimulated cells if the threshold is relaxed to FDR < .10. Most of the genes bound in stimulated cells are prebound in unstimulated cells.

B. Although most Foxp3 bound genes in stimulated cells were also bound in unstimulated cells, the binding profiles were not identical in the two conditions. *Ptpn22* and *Jak2* were representative of target genes where strong immunoenrichment was observed in both conditions. The binding profile for Foxp3 is shown at these promoters with binding in unstimulated cells displayed with a blue line and binding in stimulated cells displayed with a pink line. The profile across these promoter regions indicates that additional Foxp3 binding events are stabilized in response to PMA/ionomycin stimulation. This phenomenon is observed at several Foxp3 targets. In contrast to the strong immunoenrichment observed in both conditions at *Ptpn22* and *Jak2*, recent reports (Wu et al., 2006; Chen et al., 2006) indicate that Foxp3 is stabilized at the promoters of *Il2* and *CD25* in response to T cell stimulation. This finding is confirmed in our Foxp3 binding data as shown here.

C. Foxp3 binding at the *Ptpn22* and *Jak2* promoters in unstimulated cells was independently confirmed with site-specific ChIP in cyclosporin A treated cells. Primers flanking the binding peaks indicated with asterisks in **B.** were used for ChIP PCR reactions shown here. Immunoenrichment at the *Ptpn22* and *Jak2* promoters was observed. As expected, immunoenrichment was not observed at the *Il2, CD25*, and control *ß actin* promoters.

Figure S4

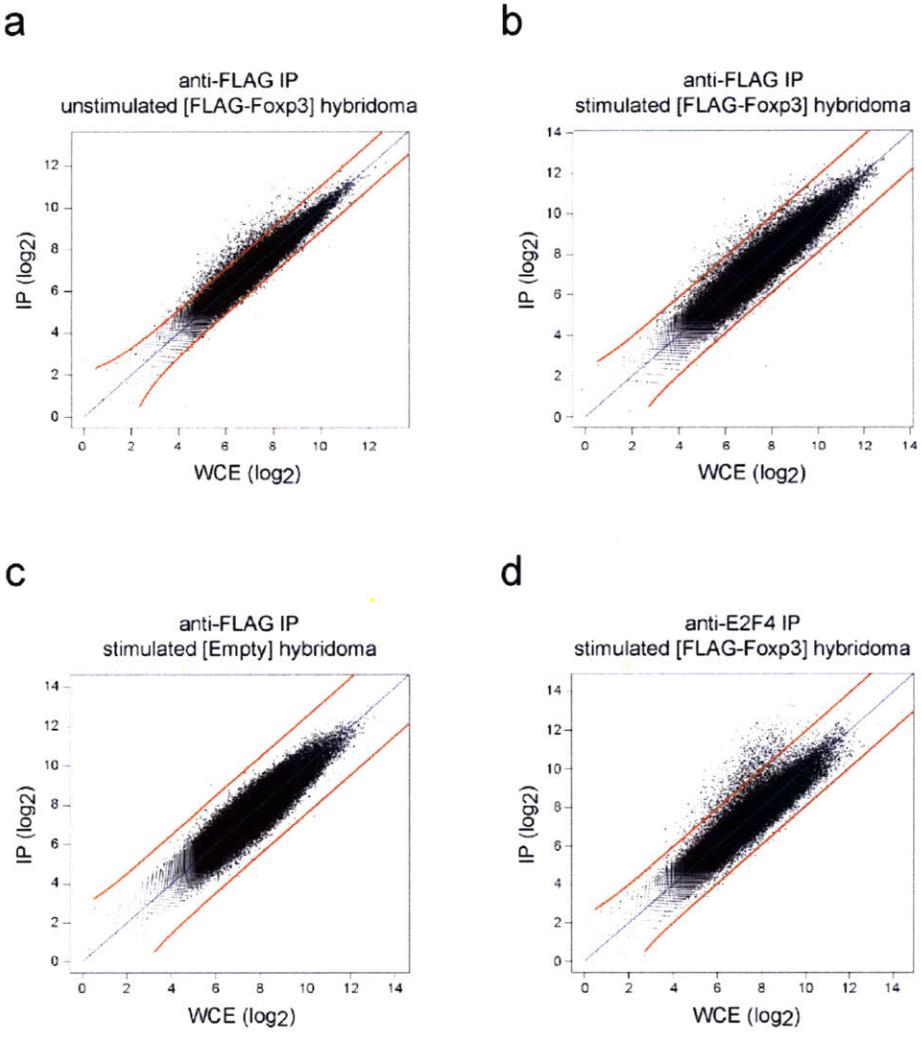


Figure S4. Control experiments confirm the specificity of Foxp3 ChIP-chip

When $\log_2$ intensity values of IP (Cy5 label) material are plotted against $\log_2$ intensity values of whole cell extract (Cy3 label), considerable enrichment of IP material is observed in ChIP experiments in FLAG-tagged Foxp3+ cells both before **(A.)** and after **(B.)** PMA/ionomycin stimulation. In contrast, very little IP enrichment is observed when the same IPs are performed in Foxp3- hybridomas **(C.)**. A positive control IP with an anti-E2F4 antibody identifies expected IP enriched E2F4 targets **(D.)**, which are largely distinct from the identified Foxp3 target genes (see Supplemental Table S4).

Figure S5

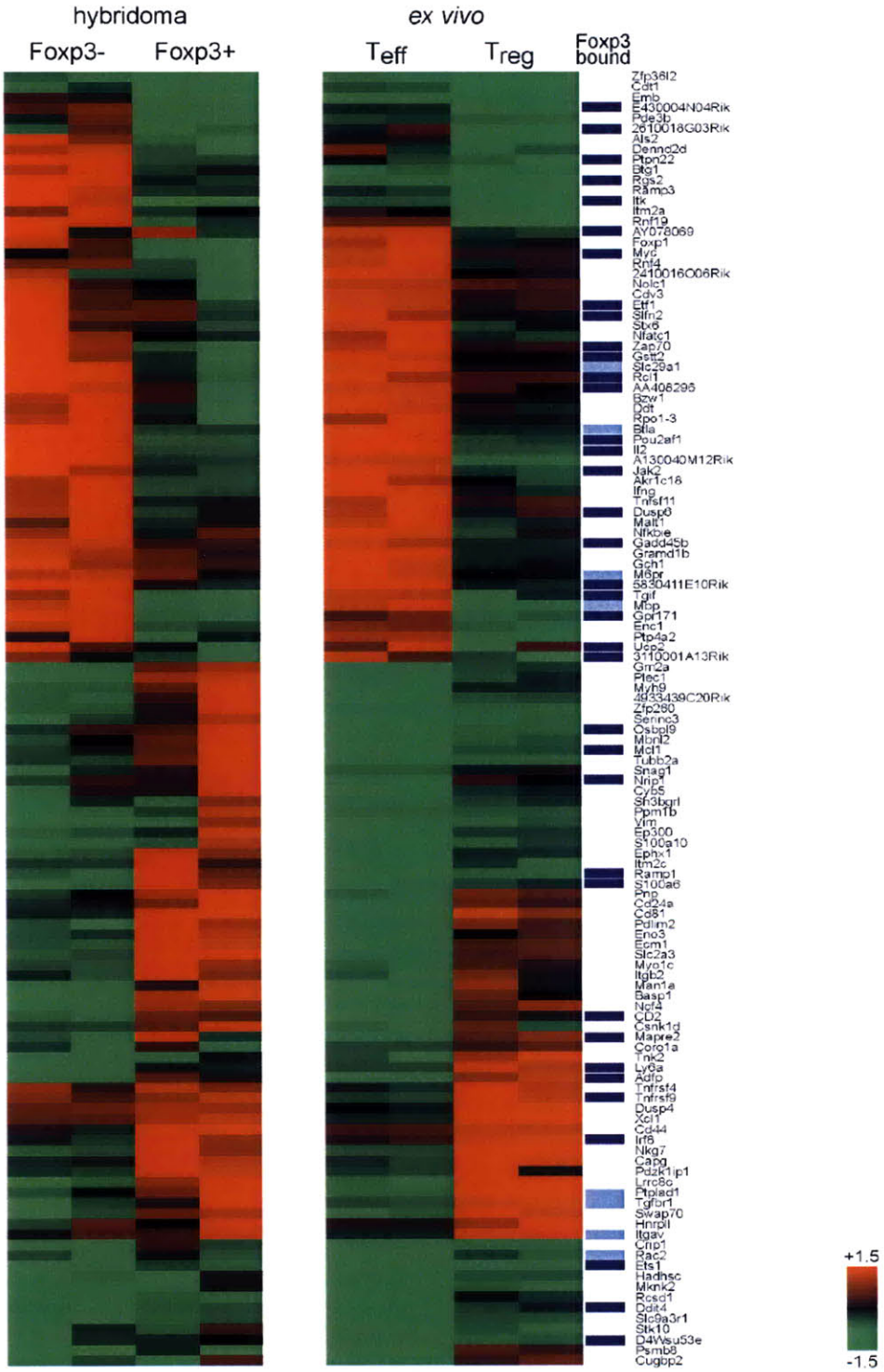


Figure S5. Many Foxp3 targets show consistent Foxp3 dependent differential expression in *ex vivo* and hybridoma cells

Genes were selected that showed consistent Foxp3 dependent differential expression in *ex vivo* and hybridoma cells according to methods described in the Supplemental section, Identification of Differentially Expressed Genes. 125 genes are displayed to match the number of genes in Figure 3a. For clustering and heatmap display data were Z-score normalized within the full hybridoma and *ex vivo* datasets including the unstimulated samples, though only the stimulated samples are displayed. Data were hierarchically clustered and are displayed in a heatmap. The Z-score normalized induction (red) or repression (green) is shown for each gene. Direct targets of Foxp3 are signified with blue bars, with dark blue representing genes called bound with a false discovery rate of 5% and light blue representing a false discovery threshold of 10%. There is a significant enrichment of direct Foxp3 targets among the genes that are downregulated in stimulated Foxp3+ cells ($p < 10^{-19}$).

Figure S6

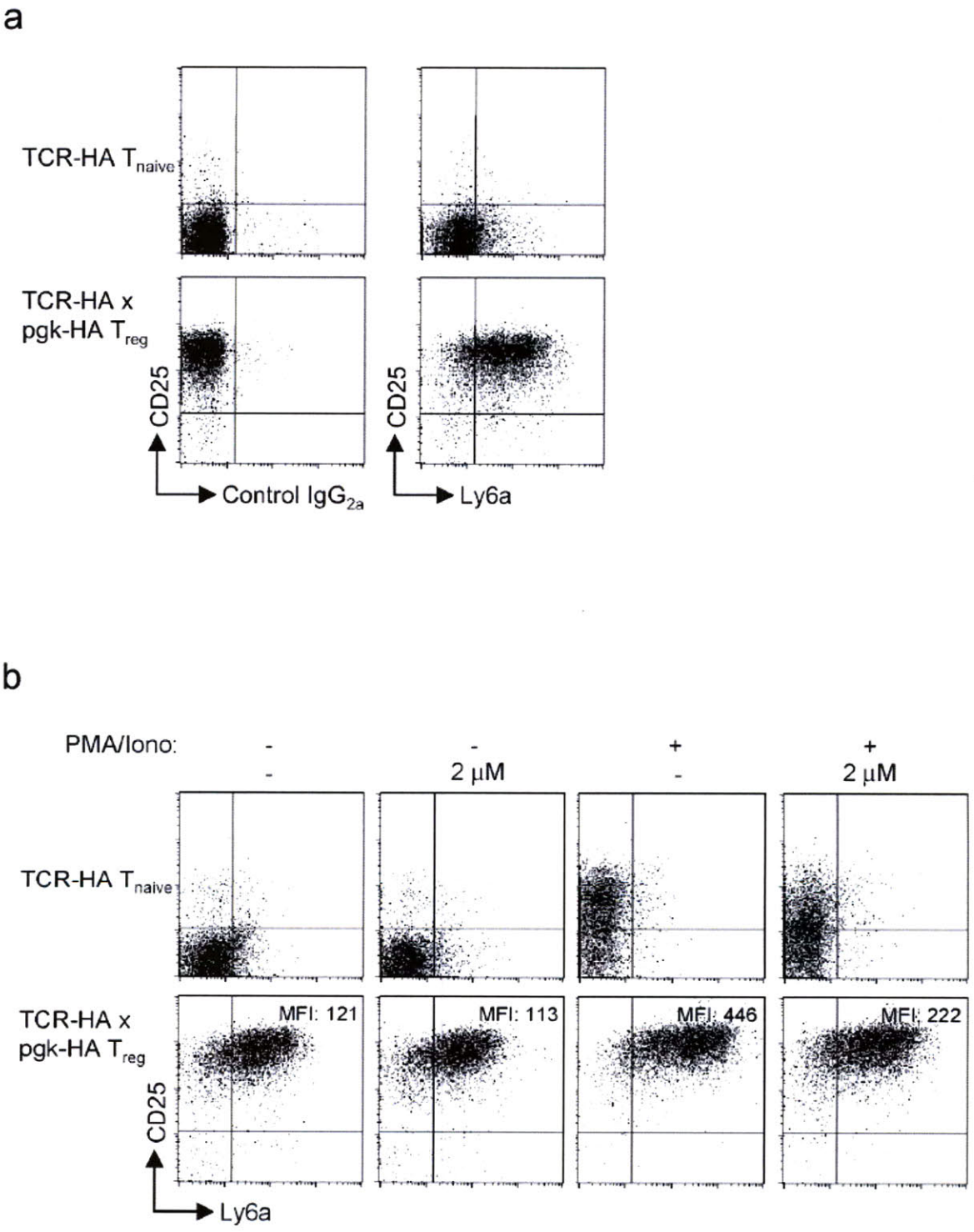

a
TCR-HA T_{naive}
TCR-HA x pgk-HA T_{reg}
CD25
Control IgG_{2a}
CD25
Ly6a
b
PMA/Iono: - - + +
- 2 μM - 2 μM
TCR-HA T_{naive}
TCR-HA x pgk-HA T_{reg}
MFI: 121
MFI: 113
MFI: 446
MFI: 222
CD25
Ly6a

Figure S6. Analysis of Ly6a protein expression on primary T_{reg} cells

Surface expression of Ly6a on $CD4^{+}6.5^{+}$ naïve T cells from TCR-HA mice or Foxp3-expressing $CD4^{+}CD25^{high}6.5^{+}$ T_{reg} cells from double-transgenic TCR-HA x pgk-HA mice was analyzed on (**A.**) freshly FACS-purified cells or after 18h of culture in the absence or (**B.**) presence of 50 ng/ml PMA and 200 ng/ml ionomycin with or without a 1h preincubation with 2 μM cyclosporin A.

Appendix D

Supplemental Material for Chapter 5

Aberrant Chromatin at Genes Encoding Stem Cell Regulators in Human Mixed-lineage Leukemia

Supplemental Experimental Procedures

Antibodies

The antibodies for ChIP were specific for MLL-N terminus (Bethyl BL1289), MLL-C terminus (Nakamura et al. 2002), AF4-C terminus (wis50 Rabbit polyclonal raised against amino acids 767-931), H3K79me2 (Abcam, ab3594). H3K4me3 (Abcam, ab8580) (Santos-Rosa et al. 2002), H3 (Abcam, ab1791), ENL (wis11 provided by Eli Canaani), hypophosphorylated RNA polymerase II (8WG16) (Thompson et al. 1989) and High-affinity HA (Roche clone 3F10).

Chromatin immunoprecipitation

Protocols describing all materials and methods can be downloaded from http://web.wi.mit.edu/young/hES_PRC and have previously been described in detail (Lee et al. 2006).

Human REH or SEM cells were grown to a final count of $5x10^7$ – $1x10^8$ cells for each chromatin immunoprecipitation (ChIP). Cells were chemically crosslinked by the addition of one-tenth volume of fresh 11% formaldehyde solution for 15 minutes at room temperature. Cells were rinsed twice with 1xPBS, harvested by centrifugation and flash frozen in liquid nitrogen. Cells were stored at –80°C prior to use.

Cells were resuspended, lysed and sonicated to solubilize and shear crosslinked DNA. Sonication conditions vary depending on cells, culture conditions, crosslinking and equipment. We used a Misonix Sonicator 3000 and sonicated at a power of 27W for 10 x 30 second pulses (90 second pause between pulses). Samples were kept on ice at all times.

The resulting whole cell extract was incubated overnight at 4°C with 100 µl of Dynal Protein G magnetic beads that had been preincubated with approximately 10 µg of the appropriate antibody. The immunoprecipitation was allowed to proceed overnight. Beads were washed 5 times with RIPA buffer and 1 time with TE containing 50 mM NaCl. Bound complexes were eluted from the beads by heating at 65°C with occasional vortexing and crosslinking was reversed by overnight incubation at 65°C. Whole cell extract DNA (reserved from the sonication step) was also treated for crosslink reversal. Immunoprecipitated DNA and whole cell extract DNA were then purified by treatment with RNAse A, proteinase K and multiple phenol:chloroform:isoamyl alcohol extractions.

HA-ChIPs were performed as described in Lee et al. (2006) with minor modifications made to the sonication/IP buffer (Final: 50 mM Hepes, 140 mM NaCl, 1 mM EDTA, 1% Triton-X, 0.1% DOC). Washes of the immunoprecipitate were performed with IP buffer, with IP buffer containing 500 mM NaCl, RIPA buffer and finally with TE.

ChIP-Seq Experiments and Analysis

All protocols for Illumina/Solexa sequence preparation, sequencing and quality control are provided by Illumina (Illumina, San Diego, CA, http://www.illumina.com/pages.ilmn?ID=252). A brief summary of the technique, minor protocol modifications, and data analysis methods are described below.

Sample preparation

Purified chromatin immunoprecipitated (ChIP) DNA was prepared for sequencing according to a modified version of the Illumina/Solexa Genomic DNA protocol. Approximately 50-200ng of IP DNA was prepared for ligation of Solexa linkers by repairing the ends and adding a single adenine nucleotide overhang to allow for directional ligation. A 1:100 dilution of the Adaptor Oligo Mix (Illumina) was used in the ligation step. A subsequent PCR step with 18 amplification cycles added additional linker sequence to the fragments to prepare them for annealing to the Genome Analyzer flow-cell. Amplified material was purified by Qiaquick MinElute (Qaigen) and a narrow range of fragment sizes was selected by separation on a 2% agarose gel and excision of a band between 150-300 bp, representing IP fragments between 50 and 200nt in length and ~100bp of primer sequence. The DNA was purified from the agarose and diluted to 10 nM for loading on the flow cell.

Polony generation on Solexa flow-cells

The DNA library (2-4 pM) was applied to one lane of the flow-cell (eight samples per flow-cell) using a Cluster Station device (Illumina). The concentration of library applied to the flow-cell was calibrated so that polonies generated in the bridge amplification step originate from single strands of DNA. Multiple rounds of amplification reagents were flowed across the cell in the

bridge amplification step to generate polonies of approximately 1,000 strands in 1μm diameter spots. Double stranded polonies were visually checked for density and morphology by staining with a 1:5000 dilution of SYBR Green I (Invitrogen) and visualizing with a microscope under fluorescent illumination. Validated flow-cells were stored at 4°C until sequencing.

Sequencing

Flow-cells were removed from storage and subjected to linearization and annealing of sequencing primer on the Cluster Station. Primed flow-cells were loaded into the Genome Analyzer 1G (Illumina). After the first base was incorporated in the sequencing-by-synthesis reaction the process was paused for a key quality control checkpoint. A small section of each lane was imaged and the average intensity value for all four bases was compared to minimum thresholds. Flow-cells with low first base intensities were re-primed and if signal was not recovered the flow-cell was aborted. Flow-cells with signal intensities meeting the minimum thresholds were resumed and sequenced for 26 cycles.

Genomic mapping of ChIP-Seq reads

Images acquired from the Genome Analyzer were processed through the bundled image extraction pipeline (Illumina), which identified polony positions, performed base-calling and generated QC statistics. Sequences were aligned to NCBI build 36.1 (hg18) of the human genome using ELAND software (Illumina). Sequences uniquely mapping to the genome with zero or one mismatch were used in further analysis. Twenty-six basepair sequences that did not meet these criteria were truncated by one base and remapped. Truncated sequences, which uniquely mapped to the genome with zero or one mismatches were added to the set to be used in further analysis. Truncation and remapping were repeated down to a sequence length of 15 bases. Sequences from two flow cells for each chromatin IP target were combined. A summary of the number of reads used in each ChIP-Seq experiment is provided in Table S1. Complete ChIP-Seq data is available in the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE13313.

Identification of enriched genomic regions

The analysis methods used were derived from previously published methods (Barski et al. 2007; Johnson et al. 2007; Mikkelsen et al. 2007; Robertson et al. 2007). Each ChIP-Seq read was extended 100 bp from its mapped genomic position and strand to approximate the middle of the sequenced DNA fragment. The genome was divided into 25 bp wide bins. The ChIP-Seq density within each genomic bin was then calculated as the number of ChIP-Seq reads mapping within a 1kb window (+/- 500bp) surrounding the middle of that the genomic bin.

Genomic bins containing statistically significant ChIP-Seq enrichment were identified using a Poissonian background model using a p-value threshold of 10^{-8}. Assuming the background ChIP-Seq density was spread randomly throughout the genome, the probability of observing a given density can be modeled as a Poisson process with an expectation value equal to the total number of ChIP-Seq reads times the number of genomic bins into which each read maps (1,000bp / 25bp per bin = 40 bins) divided by the total number of genomic bins with unique sequence (estimated to be 70% of total bins) available for mapping. A list of the minimum ChIP-Seq density required in each dataset to meet this threshold is provided in Table S1. ChIP enriched genomic bins within 2kb were compressed into enriched regions. Regions less than 100 bp in length were discarded.

The Poisson background model assumes a random distribution of background ChIP-Seq density. This model has numerous shortcomings, both statistical and biological, and we have observed significant deviations from its expectations. These non-random events create a large number of false positive events for actual ChIP-Seq experiments analyzed using a Poisson model. To remove these regions, we compared genomic bins and regions that meet the statistical threshold for enrichment to an empirical distribution of reads obtained from Solexa sequencing of DNA from whole cell extract. We required that enriched genomic bins and regions have five-fold greater ChIP-Seq density in the specific IP sample as compared with the non-specific background, normalized for the total number mapped ChIP-Seq reads in each sample. We observed that ~100-500 regions in the genome showed non-specific enrichment in these experiments.

A summary of the ChIP-Seq enriched regions in each experiment is provided in Tables S11 through S20.

Comparing enriched regions to gene annotation

The genomic coordinates of the full set of ~ 26,000 transcripts from the RefSeq database (http://www.ncbi.nlm.nih.gov/RefSeq/) was downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTables?) on May 12, 2008. Transcripts were assigned to ChIP enriched genomic regions for AF4, MLL-N, MLL-C, and H3K79me2 based on the following algorithm. First, for each ChIP enriched region, all transcripts with transcription start sites (TSS) occurring within the region were assigned to that region. Second, all transcripts for which the region from 2kb upstream of the TSS to the transcription stop site overlapped the ChIP enriched region were recorded. Of these overlapping transcripts, the single transcript whose TSS was closest to the enriched region's peak of ChIP density was assigned to the region. A different algorithm focused on transcription starts sites was used to assign transcripts to H3K4me3 ChIP enriched regions. All transcripts for which the region from 2kb upstream to 2kb downstream of the TSS overlapped a H3K4me3 ChIP enriched region were assigned to that region. This resulted in the assignment of zero, one, or many transcripts to each ChIP enriched region. In some cases more than one ChIP enriched region mapped to the same transcript.

A summary of the genes called enriched in each ChIP-Seq experiment is provided in Table S3. Additionally, supplemental data files are provided that contain genome browser tracks showing genome-wide Chip-seq density and enriched regions for all experiments.

Identification of MLL-AF4 fusion protein target regions

Locating MLL-AF4 fusions in cancer cells is complicated by the fact that both wild type and fusion proteins are present in the cells, making it difficult to distinguish between a 'normal' binding site, and the site of the fusion protein. In order to locate MLL-AF4 binding sites in the genome of SEM cells, we therefore sought to find patterns of binding that were found uniquely in SEM cells and not in the comparable REH cells that lack the fusion. Because these patterns were unique to SEM cells, we concluded that they must be derived, at least in part, from the fusion protein.

It is important to note that, while the loci we identify a very likely to be bound by AF4-MLL, we cannot definitively evaluate binding of MLL-AF4 at many other places in the genome. To specifically find the fusion protein, we created a strict set of criteria that would be certain to eliminate any regions that might be falsely identified as fusion protein target regions. Regions that do not satisfy the rules of our algorithm but are enriched for both AF4-C and MLL-N

epitopes may well be bound by the fusion protein, but we cannot rule out the possibility that they are bound only by the normal proteins. Thus, to minimize false positives, we have excluded these regions, even when there is a high likelihood that they are derived from the fusion protein,

In order to identify the MLL-AF4 enriched regions, we used the following algorithm. For each AF4 ChIP enriched region larger than 3kb in SEM and REH cells, the pattern of AF4 ChIP-Seq density was compared to the pattern of MLL-N ChIP-Seq density using a correlation based similarity metric, which is described in detail below. Large regions were selected because they were more prevalent in cells containing the fusion protein than in control cancer cells. Regions were called candidate fusion targets if the AF4 to MLL-N similarity was greater than 0.8, indicating high similarity. Of 5,062 AF4 ChIP enriched regions in SEM cells, 308 regions met these criteria and of 1,057 AF4 ChIP enriched regions in REH cells, only one region met these criteria.

We then removed candidate MLL-AF4 fusion protein target regions if they appeared to have similar AF4 and MLL-N occupancy in SEM and REH cells, since REH cells do not contain the MLL-AF4 fusion protein. This likely results in the removal of some real MLL-AF4 targets and as well as regions that are targets of endogenous AF4 and MLL. This was accomplished by eliminating regions if the SEM to REH AF4 similarity or the SEM to REH MLL-N similarity was greater than 0.8 or if the SEM to REH AF4 similarity and the SEM to REH MLL-N similarity were greater than 0.4. This resulted in the elimination of 82 candidate fusion protein target regions in SEM cells and the single region in REH cells. This left 226 MLL-AF4 fusion protein target regions in SEM cells none in REH cells. A summary of the identification of MLL-AF4 target regions in SEM cells is provided in Table S4. A summary of the identification of MLL-AF4 target regions in REH cells is provided in Table S5. A list of the MLL-AF4 fusion protein target regions in SEM cells is provided in Table S6. A list of the RefSeq transcripts and gene symbols that are MLL-AF4 targets is provided in Table S2.

We considered the possibility that the enrichment of signal for anti-MLL-N and anti-AF4-C at "MLL-AF4" sites in SEM cells reflected the recruitment of endogenous (full length, wild type) MLL and endogenous (full length, wild type) AF4 and not the true MLL-AF4 fusion protein. Three lines of evidence argue against this possibility. First, ChIP-Seq using an antibody (anti-MLL-C) that recognized only non-fusion (native) MLL antibody resulted in promoter proximal binding that did not overlap completely with AF4 binding (Figure S2). In contrast,

there was very high concordance between the AF4-C and MLL-N signals consistent with the occupancy of a single protein (MLL-AF4) rather than two independent proteins. (Figure 1B; Figure 2A,B; Figure S2). In fact, the binding patterns of AF4-C and MLL-N were nearly indistinguishable at MLL-AF4 target regions (Figure 1B;Figure 2A,B). A composite binding profile for all 169 MLL-AF4 target genes confirmed this by showing a striking positional overlap of the MLL-N and AF4-C epitopes, but no such overlap in control cells (Figure 2B, Tables S4 and S5). Second, if detection of MLL-AF4 target genes were actually an artifact of highly expressed genes recruiting endogenous MLL-N and AF4-C, we would expect the binding behavior of the MLL-AF4 target set to be the same as a random set of highly expressed genes. This was not found to be the case when a set of the 500 most highly expressed genes in SEM cells was compared to the MLL-AF4 target set. Here, the MLL-AF4 gene set was dramatically different in binding profile and extended >6kb into the coding regions of target genes (Figure S2A). In fact, we found many active genes in both SEM and REH cells that are bound by AF4 (Table S11,S16), but these genes show little co-recruitment of normal (non-fusion) MLL at the same genes (Table S2), indicating that co-occurrence of MLL and AF4 is not a general characteristic of active genes in ALL cells. Third, we tested a subset of MLL-AF4 target regions for the ability to recruit exogenous MLL-AF4 protein. HA-MLL-AF4 plasmids were introduced into SEM cells and anti-HA ChIP analysis was performed. Quantitative PCR detection of a randomly selected group MLL-AF4 targets revealed that 85% of binding events could be verified using this assay (Figure S3). Taken together, our results identify more than 200 genomic regions that are targeted by the MLL-AF4 fusion protein in leukemia cells.

ChIP-Seq density similarity metric

In order to compare the ChIP-Seq density profiles between different ChIP targets (AF4, MLL-N, MLL-C) and different cell types (SEM and REH) we developed an algorithm, based on a correlation coefficient, to score the similarity between two ChIP-Seq density profiles in a given genomic region. First, the genomic region was extended 2kb on either side. Since the similarity algorithm is based on a Pearson correlation it was important to include un-enriched regions surrounding regions of ChIP enrichment. Second, each ChIP-Seq density profile was transformed by subtracting the density threshold for calling ChIP-Seq enrichment in that experiment (14-17) and then imposing a minimum of zero counts. This was done to ensure that

similarities between two profiles within background signal range have a small influence on the similarity score. Third, the Pearson correlation coefficient between the two extended, transformed ChIP-Seq profiles was calculated. This similarity score was used to make comparison of the following ChIP-Seq profiles; AF4:MLL-N in SEM cells, AF4:MLL-N in REH cells, AF4 in SEM:REH cells, and MLL-N in SEM:REH cells. A score of 1 indicates perfect similarity between two profiles while a score of zero or below indicates no similarity between to profiles. We used a threshold of greater than 0.8, which indicates high similarity, and a threshold of less than 0.4, which indicates low similarity between two profiles.

Composite ChIP-Seq density profiles

Composite profiles of enrichment were generated as described previously (Pokholok et al. 2005; Guenther et al. 2007). Selected genes were aligned with each other according to the position and direction of their transcription start sites. The average ChIP-Seq density in 25 bp bins was calculated. For each experiment, the composite profile was normalized to the density per million total reads.

ChIP-Seq density heatmaps

Selected genes were aligned with each other according to the position and direction of their transcription start site. For each experiment, the ChIP-Seq density profiles were normalized to the density per million total reads. Additionally, ChIP-Seq density profiles were background subtracted using a background signal value of two reads per million. Genes were sorted by the total AF4 and MLL-N ChIP-Seq density in the displayed region. Heatmaps were generated using Java Treeview (http://jtreeview.sourceforge.net/) with color saturation at six reads per million.

Gene Set enrichment analysis

The gene symbols of transcripts targeted by MLL-AF4 were submitted to the online tool Gene Set Enrichment Analysis, Molecular Signatures Database (http://www.broad.mit.edu/gsea/msigdb/index.jsp) (Subramanian et al. 2005) on May 15, 2008. The MLL-AF4 target genes were compared to curated (n=1,892), computational (n=883), and Gene Ontology (n=1,454) gene sets.

Selected statistically significant results are presented in Figure 2. Complete Gene Set Enrichment Analysis results are provided in Table S10.

ChIP-Chip Experiments and Analysis

Microarray design

The 95K feature human promoter array (H23a) used in Figure S5 was purchased from Agilent Technology (Santa Clara, CA). This microarray (ID# 13950) contains ~95,000 60mer oligonucleotide probes tiled at a density of approximately 1 probe per 250 bp. The array included all four HOX clusters (HOXA-D with 100kb flanking each clusters 3' and 5' ends) as well as the promoter regions of ~17,000 genes from ~750 bp upstream to ~250 bp downstream of the transcription start sites.

Sample preparation, hybridization, and imaging

Purified DNA was blunted and ligated to linker and amplified using a two-stage PCR protocol. Amplified DNA was labeled and purified using Bioprime random primer labeling kits (Invitrogen), RNA Polymerase II, MLL-N, AF4, ENL, H3K79me2, and H3K4me3 immunoenriched DNA was labeled with Cy5 fluorophore, whole cell extract and H3 (control ChIP for histone modifications) immunoenriched DNA was labeled with Cy3 fluorophore. Labeled DNA (5-6 μg) was mixed and hybridized to arrays in Agilent hybridization chambers for 36-42 hours at 40°C. Arrays were then washed and scanned using an Agilent DNA microarray scanner BA. PMT settings were set manually to normalize bulk signal in the Cy3 and Cy5 channel.

Image quantification, data normalization and analysis

We used GenePix Pro software (Molecular Devices, Sunnyvale, CA) to obtain background-subtracted intensity values for each fluorophore for every feature on the arrays respectively. Among the Agilent controls is a set of negative control spots that contain 60-mer sequences that do not cross-hybridize to human genomic DNA. We calculated the median intensity of these negative control spots in each channel and then subtracted this number from the intensities of all

other features. To correct for different amounts of each sample of DNA hybridized to the chip, the negative control-subtracted median intensity value of control oligonucleotides from the Cy3-enriched DNA channel was then divided by the median of the control oligonucleotides from the Cy5-enriched DNA channel. This yielded a normalization factor that was applied to each intensity in the Cy5 DNA channel. For graphical display, the enrichment ratio of each probe was plotted as the average of the raw enrichment ratio of that probe and the nearest 5' and 3' probes (sliding average of three probes).

Complete ChIP-chip data is available in the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE13313.

Gene-specific realtime-PCR

For relative quantification by real-time PCR, dilutions of the whole cell extract DNA were used to construct standard curves for the amplifications of the SEM immunoprecipitates for each primer pair. PCRs were performed in duplicate reactions on the 7700 ABI Detection System using the POWER SYBR Green PCR master mix (ABI) and standard deviations were calculated. Enrichment was calculated from the difference between the threshold cycle of the target region and ACTB control primers for the ChIP DNA minus the difference between the threshold cycle of the target region and ACTB control primers for the WCE DNA. Primers were selected that are within 1kb of the peak of the MLLAF4 fusion domains from ChIP-Seq data. Genes were selected that were expected to be bound by the fusion gene as well as several control genes that were negative by ChIP-Seq and expressed (ACTB, RPS3, SUZ12) or not expressed (IL1F10, HOXD9, NANOG, POU5F1) in SEM cells. Results are displayed in Figure S3. A table of the primer sequences used is provided in Table S21.

GeneChip Expression Experiments and Analysis

Sample preparation

5 μg total RNA was used to prepare biotinylated cRNA according to the manufacturer's protocol (Affymetrix One Cycle cDNA Synthesis Kit). Briefly, this method involves SuperScript II-directed reverse transcription using a T7-Oligo-dT promoter primer to create first strand cDNA.

RNase H-mediated second strand cDNA synthesis is followed by T7 RNA Polymerase directed *in vitro* transcription, which incorporates a biotinylated nucleotide during cRNA amplification.

Hybridization, staining, scanning, and image analysis

Samples were prepared for hybridization using 15 μg biotinylated cRNA in a 1X hybridization cocktail with additional hybridization cocktail components provided in the GeneChip Hybridization, Wash and Stain Kit (Affymetrix). GeneChip arrays (Human U133 2.0) were hybridized in a GeneChip Hybridization Oven at 45°C for 16 hours at 60 RPM. Washing was performed using a GeneChip Fluidics Station 450 according to the manufacturer's instructions, using the buffers provided in the Affymetrix GeneChip Hybridization, Wash and Stain Kit. Arrays were scanned on a GeneChip Scanner 3000 and images were extracted and analyzed using the default settings of GeneChip Operating Software v1.4 (GCOS).

Expression signal values were quantile normalized by assigning each probeset the average signal intensity for all probesets of the same rank across the four experiments. Control probes were removed from further analysis. Probeset annotations were downloaded from the NetAffx (http://www.affymetrix.com/analysis/index.affx) database on May 14, 2008.

Complete expression data are provided in Table S3 and available in the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE13313.

Identification of top 500 expressed genes

The average expression signal of each probeset in each cell type was calculated and the probesets were ranked. Transcripts corresponding to the 500 genes with the highest signal in each cell type were used. Transcripts that were identified as MLL-AF4 fusion protein targets were excluded. This identified ~400 of the most highly expressed Refseq transcripts in each cell type.

Expression based present/absent calls

Present (P), marginal (M), absent (A) transcript detection calls from GCOS for each probeset were tabulated. Transcripts were called present if at least one corresponding probeset was called present in at least one of the two experiments for each cell type.

MLL-AF4 fusion target expression hierarchical clustering

Expression data from SEM and REH cells was combined with gene expression data profiling 132 leukemia samples (Ross et al. 2003). The full dataset was quantile normalized by assigning each probeset the average signal intensity for all probesets of the same rank across all experiments. Expression data was extracted for all transcripts assigned to MLL-AF4 fusion target regions larger than 10kb for Figure 3. This cutoff was used to enable sufficient space for each target gene to be listed. Expression data was extracted for all transcripts assigned to all MLL-AF4 fusion target regions for Figure S4. Genes with no expression data were excluded. Expression signals from multiple probesets mapping to the same gene were averaged to produce a single expression profile for each gene. Expression values were log (base 10) transformed and then mean centered for each gene. Centroid linkage, centered Pearson correlation distance, hierarchical clustering of arrays and centroid linkage, centered Pearson correlation distance, hierarchical clustering of genes was performed using Gene Cluster 3.0 (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm#ctv). Heatmaps were generated using Java Treeview (http://jtreeview.sourceforge.net/). Array tree branches were flipped for optimal display. Complete input data and clustering results are provided in Table S8 and Table S9.

Identification of binding partners by IP-Western blot

Nuclear extracts were prepared as described in Nakamura et al. 2002 (Nakamura et al. 2002). Aliquots of 0.5 - 0.7 mg of nuclear extract protein were immunoprecipitated. Polyclonal antibodies were raised in rabbits against AF4 polypeptides spanning residues 3-148 (N-AF4), 767-931 [C-AF4 (1)], and 727-885 [C-AF4 (2)]. Polyclonal antibodies to ENL were directed against polypeptides spanning residues 145-298 (ENL1) or 215-366 ENL2). Mouse monoclonal antibodies for MLL immunoprecipitation were purchased from Upstate and were directed against N-terminal and C-terminal sequences 05-764 and 05-765, respectively). For western analysis we used antibody 169 raised against MLL segment encompassing residues 79-290. Santa Cruz antibody T-18 sc-8127 was used for immunoprecipitation and detection of cyclin T1. Note: antibodies against MLL-N, which precipitates the MLL-AF4 fusion and normal ENL in SEM cells, does not precipitate normal AF4 (Figure S5). This, in conjunction with the failure to precipitate MLL-AF4 by AF4-N antibodies, indicate that normal AF4 is not complexed to MLL-AF4. Thus, MLL-AF4 does not recruit the entire normal AF4 complex , but rather the necessary components (e.g. ENL, cyclin T1). These results provide the first demonstration that an MLL

fusion protein is associated with a normal partner protein (ENL) and with pTEFb.

References

Barski, A., S. Cuddapah, K. Cui, T.Y. Roh, D.E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. 2007. High-resolution profiling of histone methylations in the human genome. Cell *129,* 823-37.

Guenther, M.G., S.S. Levine, L.A. Boyer, R. Jaenisch, and R.A. Young. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130,* 77-88.

Johnson, D.S., A. Mortazavi, R.M. Myers, and B. Wold. 2007. Genome-wide mapping of in vivo protein-DNA interactions. Science *316,* 1497-502.

Lee, T.I., S.E. Johnstone, and R.A. Young. 2006. Chromatin immunoprecipitation and microarray-based analysis of protein location. Nat Protoc *1,* 729-48.

Mikkelsen, T.S., M. Ku, D.B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.K. Kim, R.P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E.S. Lander, and B.E. Bernstein. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448,* 553-60.

Nakamura, T., T. Mori, S. Tada, W. Krajewski, T. Rozovskaia, R. Wassell, G. Dubois, A. Mazo, C.M. Croce, and E. Canaani. 2002. ALL-1 is a histone methyltransferase that assembles a supercomplex of proteins involved in transcriptional regulation. Molecular Cell *10,* 1119-1128.

Pokholok, D.K., C.T. Harbison, S. Levine, M. Cole, N.M. Hannett, T.I. Lee, G.W. Bell, K. Walker, P.A. Rolfe, E. Herbolsheimer, J. Zeitlinger, F. Lewitter, D.K. Gifford, and R.A. Young. 2005. Genome-wide map of nucleosome acetylation and methylation in yeast. Cell *122,* 517-27.

Robertson, G., M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O.L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods *4,* 651-7.

Ross, M.E., X. Zhou, G. Song, S.A. Shurtleff, K. Girtman, W.K. Williams, H.C. Liu, R. Mahfouz, S.C. Raimondi, N. Lenny, A. Patel, and J.R. Downing. 2003. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. Blood *102,* 2951-9.

Santos-Rosa, H., R. Schneider, A.J. Bannister, J. Sherriff, B.E. Bernstein, N.C. Emre, S.L. Schreiber, J. Mellor, and T. Kouzarides. 2002. Active genes are tri-methylated at K4 of histone H3. Nature *419,* 407-11.

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

Thompson, N.E., T.H. Steinberg, D.B. Aronson, and R.R. Burgess. 1989. Inhibition of in vivo and in vitro transcription by monoclonal antibodies prepared against wheat germ RNA polymerase II that react with the heptapeptide repeat of eukaryotic RNA polymerase II. J Biol Chem *264,* 11511-20.

Supplemental Tables

Table S1: Summary of ChIP-Seq experiments

Table S2: Summary of ChIP-Seq enrichment calls and GeneChip expression present/absent calls for all Refseq transcripts

Table S3: Complete gene expression dataset

Table S4: Summary of identification of MLL-AF4 target regions in SEM cells

Table S5: Summary of identification of MLL-AF4 target regions in REH cells

Table S6: MLL-AF4 fusion protein target regions in SEM cells

Table S7: ChIP-chip enrichment ratios for H3K4me3, H3K79me2, ENL, AF4, and MLL-N within the HOXA locus in REH and SEM cells

Table S8: Hierarchical clustering of MLL-AF4 fusion target gene expression in leukemia patient samples for selected target genes, input data and results

Table S9: Hierarchical clustering of MLL-AF4 fusion target gene expression in leukemia patient samples for all target genes, input data and results

Table S10: MLL-AF4 fusion target Gene Set Enrichment Analysis results

Table S11: Genomic regions enriched in AF4 ChIP in SEM cells

Table S12: Genomic regions enriched in MLL-N ChIP in SEM cells

Table S13: Genomic regions enriched in MLL-C ChIP in SEM cells

Table S14: Genomic regions enriched in H3K79me2 ChIP in SEM cells

Table S15: Genomic regions enriched in H3K4me3 ChIP in SEM cells

Table S16: Genomic regions enriched in AF4 ChIP in REH cells

Table S17: Genomic regions enriched in MLL-N ChIP in REH cells

Table S18: Genomic regions enriched in MLL-C ChIP in REH cells

Table S19: Genomic regions enriched in H3K79me2 ChIP in REH cells

Table S20: Genomic regions enriched in H3K4me3 ChIP in REH cells

Table S21. Gene-specific PCR primer pairs used in real-time PCR for HA-ChIP experiment.

Figure S1

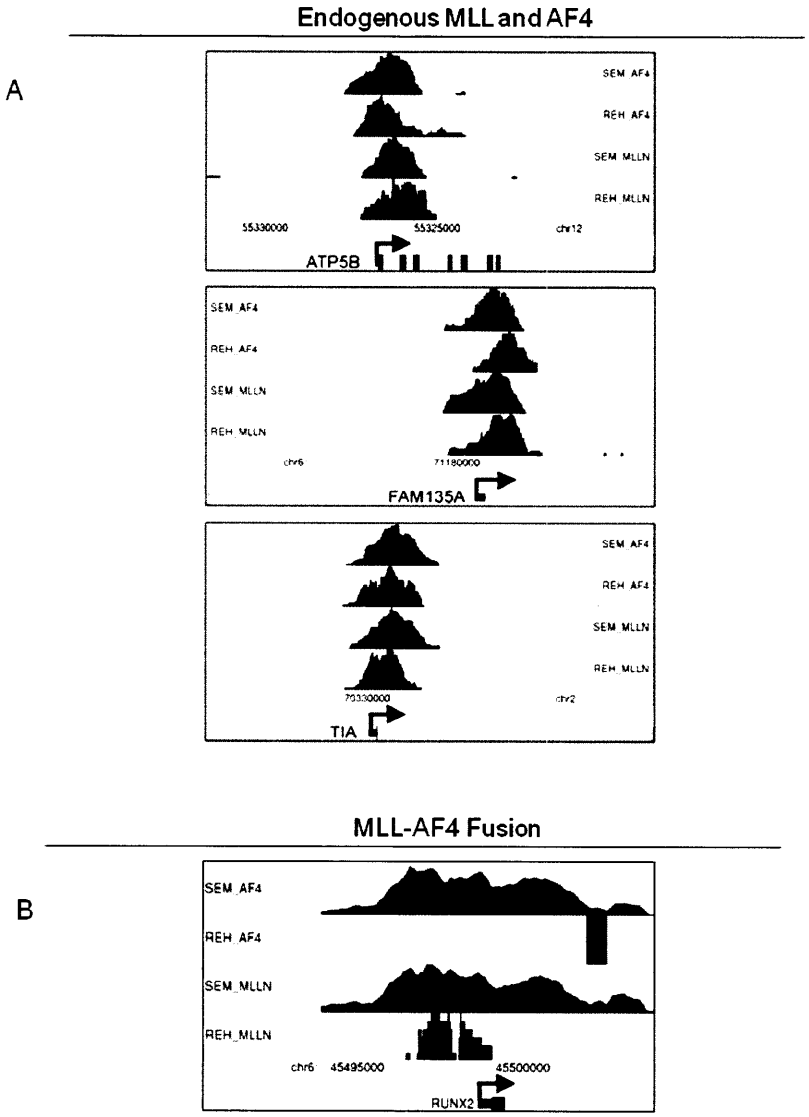


Figure S1. Endogenous MLL and AF4 bind near the 5' end of active genes

Mapped ChIP-Seq reads within 10 kb window surrounding the transcriptional start site of representative genes bound by MLL-N terminus and AF4-C in both SEM and REH cells

A. Endogenous MLL and AF4 fusion protein binding in REH cells.

B. MLL-AF4 fusion binding in SEM cells. ChIP-Seq tracks normalized to reads per million for each factor. Gene models are below each example. The start site and direction of transcription is indicated by an arrow.

Figure S2

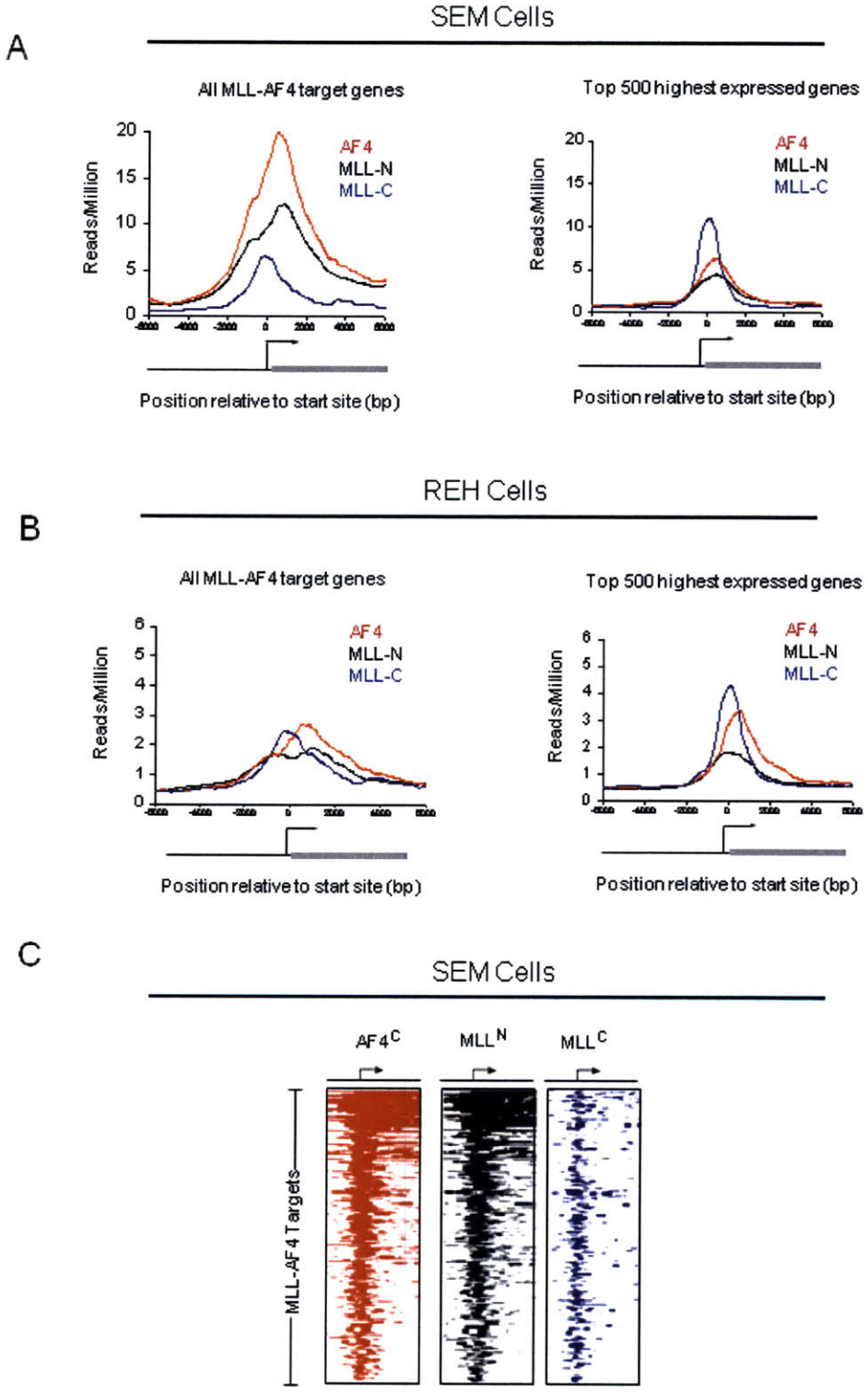

A
SEM Cells
All MLL-AF4 target genes
Top 500 highest expressed genes
Reads/Million
AF4
MLL-N
MLL-C
Position relative to start site (bp)
B
REH Cells
All MLL-AF4 target genes
Top 500 highest expressed genes
Reads/Million
AF4
MLL-N
MLL-C
Position relative to start site (bp)
C
SEM Cells
AF4C
MLLN
MLLC
MLL-AF4 Targets

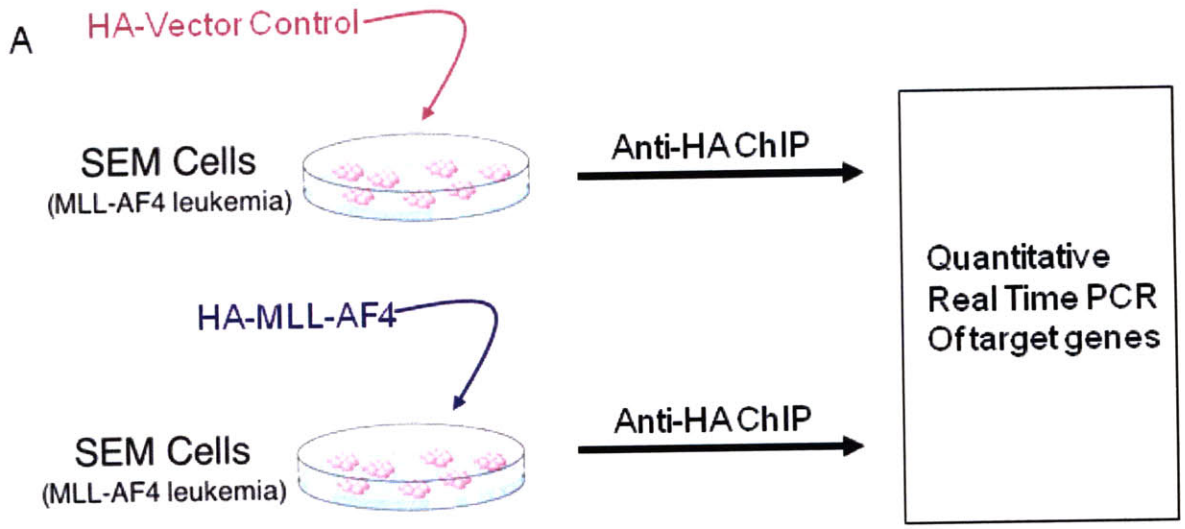
Figure S2. MLL-N and AF4 signals co-occupy at MLL-AF4 target genes in SEM cells

A. AF4 (red), MLL-N (black), and MLL-C (blue) composite ChIP-Seq density profiles for all MLL-AF4 targets (left) and the top 500 expressed genes (right) in SEM cells.

B. AF4 (red), MLL-N (black), and MLL-C (blue) composite ChIP-Seq density profiles for all MLL-AF4 targets (left) and the top 500 expressed genes (right) in REH cells

C. ChIP-Seq density heatmap of AF4 (red), MLL-N terminus (black) and MLL-C terminus (blue) for all MLL-AF4 target genes in SEM cells. The genomic region from -5kb to +10kb relative to the transcription start site of each gene is shown. Gene order is determined by highest average MLL/AF4 read density from top to bottom. The start site and direction of transcription of the genes are indicated by an arrow. Methods for generating composite profiles and for identifying top 500 expressed and fusion target genes are described in detail in the Supplemental Experimental Procedures section.

Figure S3


A
HA-Vector Control
SEM Cells
(MLL-AF4 leukemia)
Anti-HA ChIP
HA-MLL-AF4
SEM Cells
(MLL-AF4 leukemia)
Anti-HA ChIP
Quantitative
Real Time PCR
Of target genes

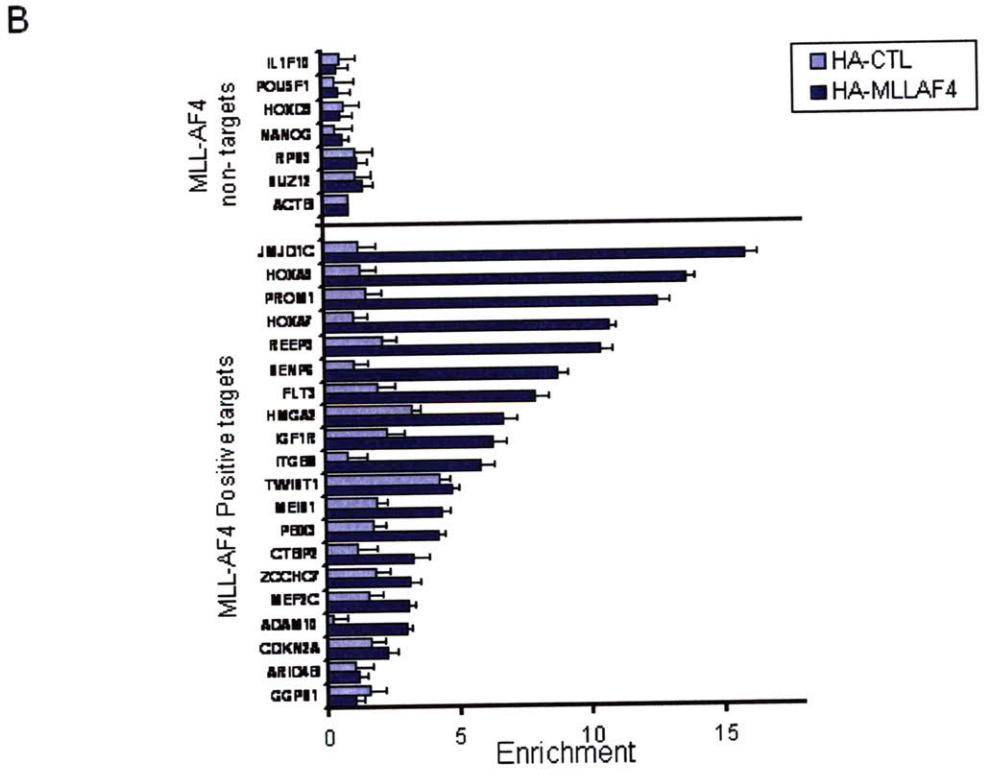

B
HA-CTL
HA-MLLAF4
MLL-AF4 non- targets
MLL-AF4 Positive targets
IL1F10
POU5F1
HOXC8
NANOG
RPS3
SUZ12
ACTB
JMJD1C
HOXA9
PROM1
HOXA7
REEP5
SENP6
FLT3
HMGA2
IGF1R
ITGB8
TWIST1
MEIS1
PBX3
CTBP2
ZCCHC7
MEF2C
ADAM19
CDKN2A
ARID4B
GGPS1
0
5
10
15
Enrichment

Figure S3. Verification of MLL-AF4 fusion targets by HA-MLL-AF4 ChIP

A. Schematic diagram of experimental verification of MLL-AF4 targets .Chromatin immunoprecipitation and gene-specific PCR were performed in SEM cells electroporated with constructs expressing either HA-tagged MLL-AF4 or HA-control vector. Cells were collected 30 hrs after electroporation, crosslinked, and ChIP was performed using an antibody against HA (Roche 3F10).

B. Site-specific quantitative PCR analysis of HA-MLL-AF4 chromatin immunoprecipitations or HA-Control (HA-CNL) immunoprecipitations. This analysis confirmed that genomic regions incorporated the tagged MLL-AF4 fusion construct. Immunoprecipitated (IP) DNA was compared to serial dilutions (100, 10 and 1 ng) of whole cell extract (WCE) DNA. Enrichment ratios of target genes are normalized relative to beta-actin (ACTB) control.

Figure S4

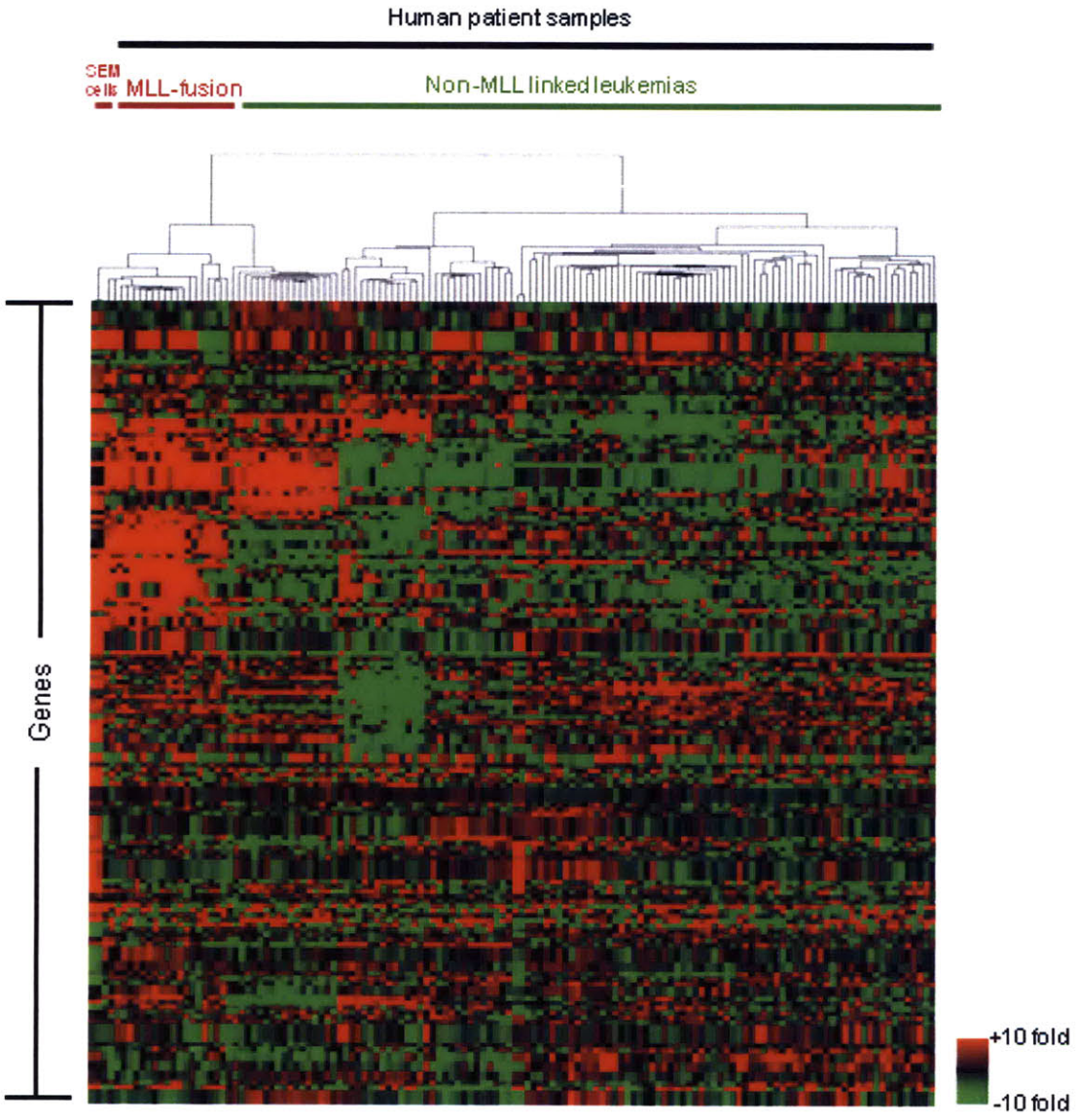

Human patient samples
SEM cells
MLL-fusion
Non-MLL linked leukemias
Genes
+10 fold
-10 fold

Figure S4. MLL-AF4 target genes define MLL-linked leukemia *in vivo*, hierarchical clustering of all fusion protein target genes

Hierarchical clustering of relative expression levels of 169 transcripts occupied by MLL-AF4 fusion protein target regions. Comparisons were made across the SEM and REH cell lines and 132 peripheral blood samples of patients diagnosed with leukemia. Each row corresponds to a gene that is bound by MLL-AF4 for which expression data was available. Each column corresponds to a single gene expression microarray. For each gene, expression is shown relative to the average expression level of that gene across all samples, with shades of red indicating higher than average expression and green lower than average expression. Columns and rows were ordered by unsupervised hierarchical clustering. A detailed description of data analysis methods is provided in the Supplemental Experimental Procedures section.

Figure S5

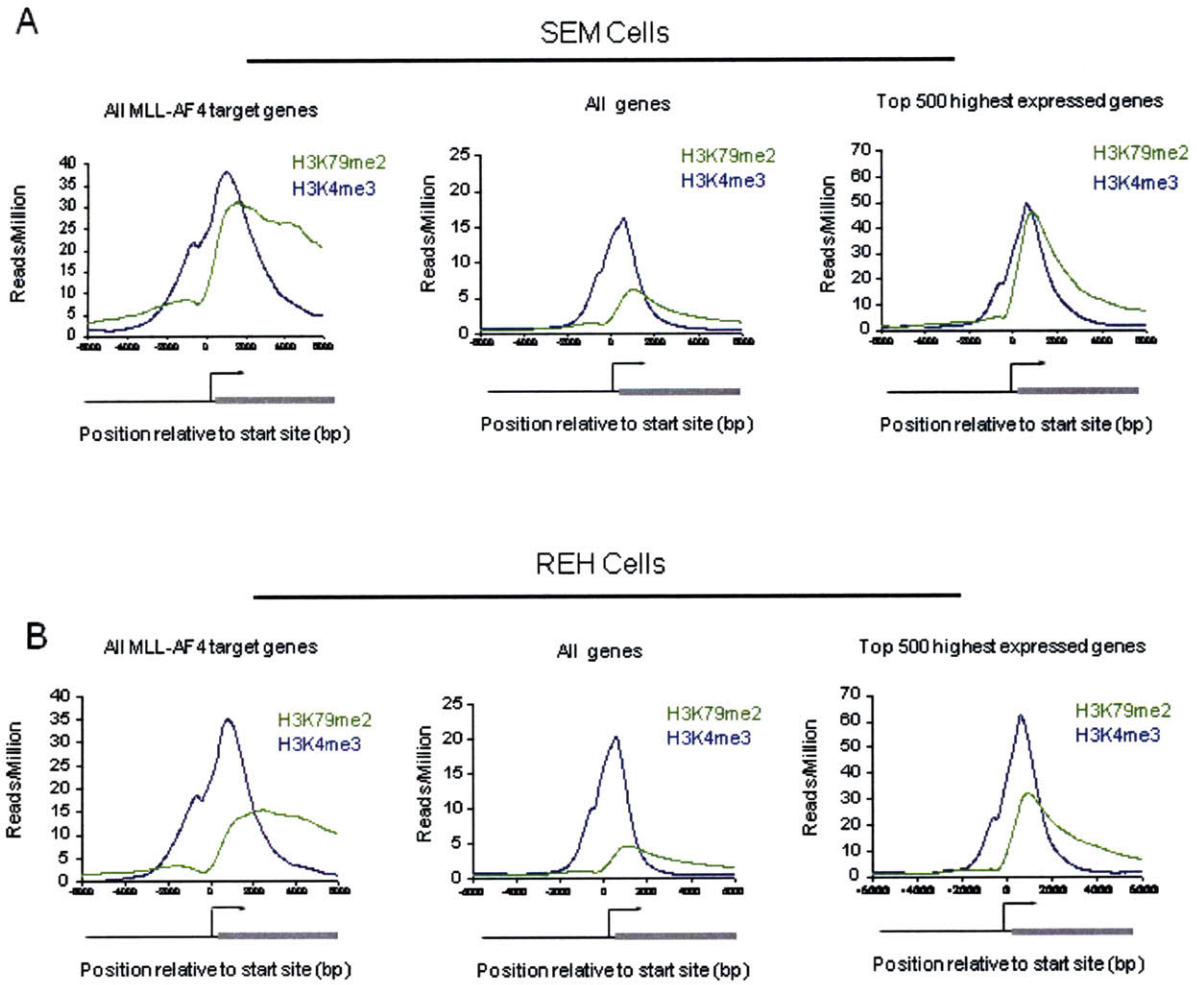

A
SEM Cells
All MLL-AF4 target genes
All genes
Top 500 highest expressed genes
H3K79me2
H3K4me3
Reads/Million
Position relative to start site (bp)
B
REH Cells
All MLL-AF4 target genes
All genes
Top 500 highest expressed genes
H3K79me2
H3K4me3
Reads/Million
Position relative to start site (bp)

Figure S5. Wide domains of H3K4me3 and H3K79me2 in SEM cells are not due only to high levels of expression

Composite ChIP-Seq density profiles for SEM (a) and REH (b) cells from –6kb to +6kb relative to transcription start site. The start site and direction of transcription of the average gene are indicated by an arrow. (Left) H3K4me3 (blue) and H3K79me2 (green) composite ChIP-Seq density profiles for MLL-AF4 targets in each cell type. (Middle) H3K4me3 (blue) and H3K79me2 (green) composite ChIP-Seq density profiles for all genes in each cell type. (Right) H3K4me3 (blue) and H3K79me2 (green) composite ChIP-Seq density profiles for the top 500 expressed probesets in SEM (top) and REH (bottom) cells. Methods for generating composite profiles and for identifying top 500 expressed and fusion target genes are described in detail in the Supplemental Experimental Procedures section.

Figure S6

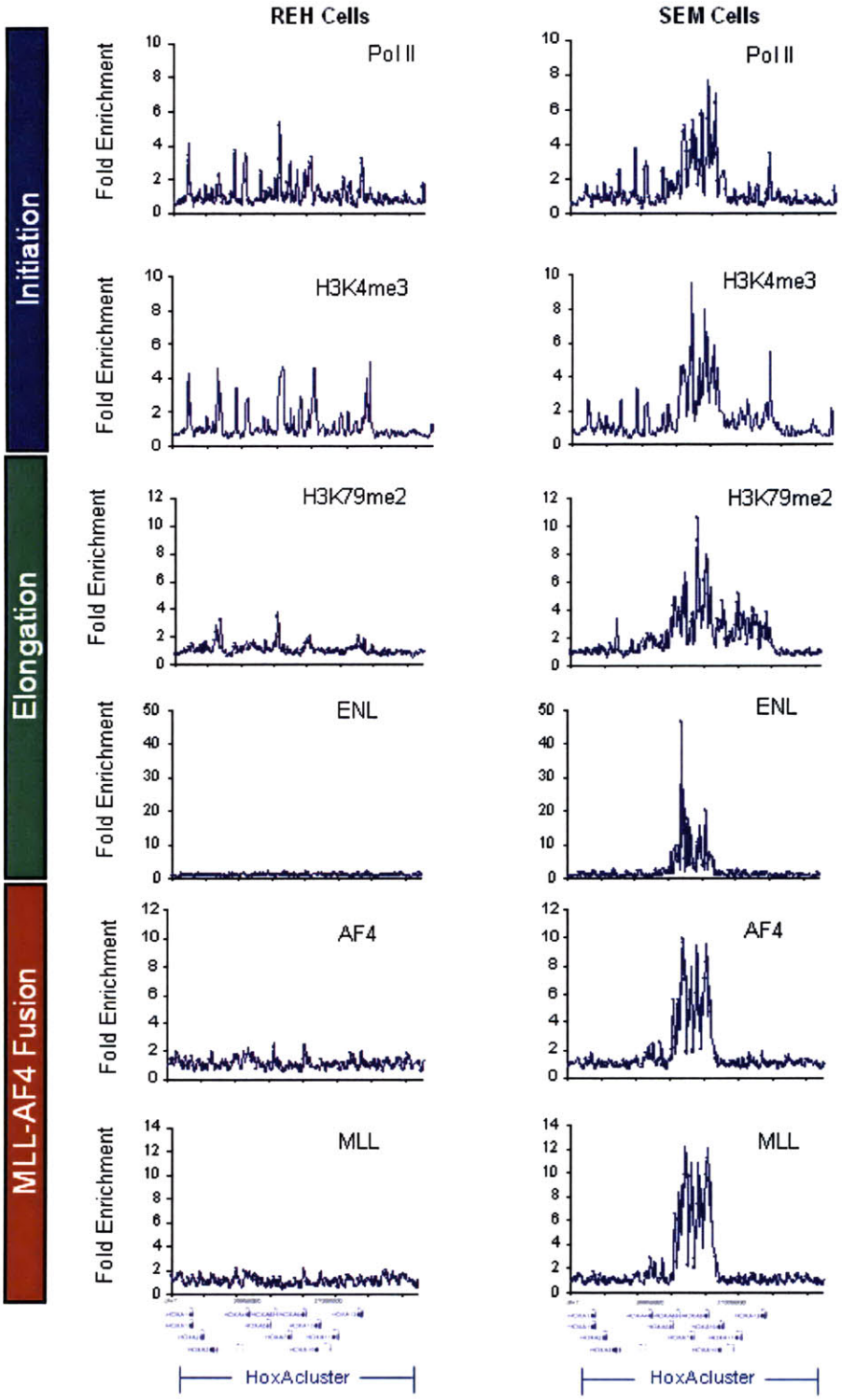

REH Cells
SEM Cells
Initiation
Elongation
MLL-AF4 Fusion
Pol II
H3K4me3
H3K79me2
ENL
AF4
MLL
Fold Enrichment
HoxAcluster

Figure S6. The elongation factor ENL associates with Pol II and H3K79me2at MLL-AF4 targets

ChIP-chip enrichment for RNAP2, H3K4me3, H3K79me2, ENL, AF4, and MLL-N antibodies across a 100 Kb region of the HOXA cluster in REH cells (left) and SEM cells (right). Individual gene models within HOXA cluster are shown at bottom. H3K4me3 and H3K79me2 ChIPs were normalized to total H3 ChIPs in each cell line.

Figure S7

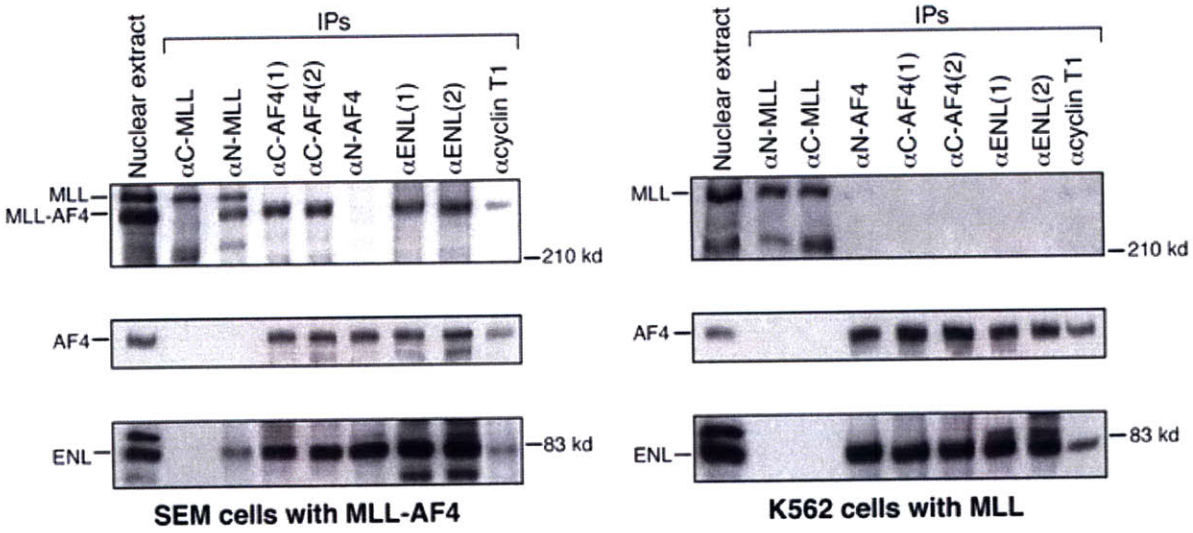


Figure S7. MLL-AF4 associates with ENL and pTEFb

Coimmunoprecipitation of MLL-AF4,ENL and the pTEFb component Cyclin T1 in SEM cells containing the MLL-AF4 fusion protein (left). Antibodies used for immunoprecipitation are shown at top and all eluates were detected by immunoblot using anti-MLL, anti-AF4, or anti-ENL antibodies. AF4,ENL and Cyclin T1 do not coimmunoprecipitate with normal MLL, as evident in analysis of the SEM cells and of control K562 cells (right). Note that normal AF4 does not coprecipitate with MLL-AF4 and it's associated proteins, indicating that the fusion protein does not recruit the entire normal AF4 complex (BItoun et al. 2007), but rather the necessary components (e.g. ENL and Cyclin T1).

Appendix E

Supplemental Material for Chapter 6

CpG Island Structure Defines Polycomb/Trithorax Chromatin Domains in Human ES and iPS Cells

Supplemental Experimental Procedures

Cells and cell culture

The human induced pluripotent stem (iPS) cell line $M2^{3F}$ was derived as described previously (Soldner et al. 2009) and was maintained on mitomycin C-inactivated mouse embryonic fibroblast (MEF) feeder layers in hESC medium (DMEM/F12 [Invitrogen] supplemented with 15% FBS [Hyclone], 5% KnockOut Serum Replacement [Invitrogen], 1 mM glutamine [Invitrogen], 1% nonessential amino acids [Invitrogen], 0.1 mM β-mercaptoethanol [Sigma], and 4 ng/ml FGF2 [R&D Systems]). Cultures were passaged every 5 to 7 days either manually or enzymatically with collagenase type IV (Invitrogen; 1.5 mg/ml). The hiPS cell line $M2^{3F}$ was passaged 15-25 times prior to ChIP-Seq analysis.

Information about cells and cell culture for human ES cell lines, WIBR1, WIBR2, WIBR3, and WIBR7, human iPS cell lines, iPS PDB^{1lox}-17puro-5, and iPS PDB^{1lox}-21puro-26, primary human fibroblast cells, GM-M01660 (Guenther et al., 2010), human fetal lung fibroblast cells, IMR90(GEO accession number GSM469970), and human primary CD4+ T cells (Barski et al., 2007) was described previously.

Chromatin immunoprecipitation

Protocols describing chromatin immunoprecipitation (ChIP) materials are freely available on the internet (http://web.wi.mit.edu/young/hES_PRC) and have been described previously in detail (Lee et al. 2006).

Cells were grown to a final count of ~$5x10^7$ to obtain starting material for six chromatin immunoprecipitations. Cells were chemically cross-linked by the addition of one-tenth volume of fresh 11% formaldehyde solution for 15 minutes at room temperature. Cells were rinsed twice with 1X PBS, harvested by centrifugation, and flash frozen in liquid nitrogen. Cross-linked cells were stored at –80°C prior to use.

Cells were re-suspended, lysed and sonicated to solubilize and shear cross-linked DNA. Sonication was performed using a Misonix Sonicator 3000 at a power of 27 watts for ten 30 second pulses with a 90 second pause between each pulse. Samples were kept on ice at all times.

The resulting whole cell extract was incubated overnight at 4 degrees C with 10μl of Dynal Protein G magnetic beads that had been pre-incubated with approximately 3 μg of the

appropriate antibody. Each individual immunoprecipitation used 1/6 of the 3ml total, or ~8 x10^6 cells per IP. The immunoprecipitation was allowed to proceed overnight. Beads were washed three times (3 x 1.5ml) with RIPA buffer and one time (1x 1.5ml) with TE containing 50 mM NaCl. Bound complexes were eluted from the beads by heating at 65 degrees C with occasional vortexing and cross-linking was reversed by overnight incubation at 65 degrees C. Whole cell extract DNA reserved from the sonication step was also treated for cross-link reversal. Immunoprecipitated DNA and whole cell extract DNA were then purified by treatment with RNAse A, proteinase K and two phenol:chloroform:isoamyl alcohol extractions.

The ChIP antibodies used were ab8580 (Abcam) for H3K4me3 and ab6002 (Abcam) for H3K27me3.

ChIP-Seq sample preparation and Solexa sequencing

All protocols for Illumina library preparation and sequencing are provided by Illumina (http://www.illumina.com/). A brief summary of the technique, minor protocol modifications, and data analysis methods are described below.

Purified ChIP DNA was prepared for sequencing according to a modified version of the Illumina Genomic DNA protocol. Approximately 50-200ng of IP DNA was ligated to a 1:100 dilution of the Illumina Adaptor Oligo Mix. After 18 cycles of PCR amplification library fragments between 150-300bp were purified on a 2% agarose gel. Between 4-6 pmoles of DNA library was applied to each lane of the flow cell and sequenced on Illumina GAII sequencers according to standard Illumina protocols. Images acquired from the Genome Analyzer were processed through the bundled image extraction pipeline.

Public availability of ChIP-Seq data

Complete H3K4me3 and H3K27me3 ChIP-Seq data for the $M2^{3F}$ iPS cell line has been submitted to the Gene Expression Omnibus database (http://www.ncbi.nih.gov/geo/).

Previously published ChIP-Seq datasets

ChIP-Seq data from human ES cell lines WIBR1, WIBR2, WIBR3, and WIBR7, human iPS cell lines iPS PDB^{1lox}-17puro-5, and iPS PDB^{1lox}-21puro-26, and from primary human fibroblast cells GM-M01660 were obtained from the GEO database accession number GSE22499

(Guenther et al., 2010). ChIP-Seq data from human lung fibroblast IMR90 cells were obtained from the GEO database accession number GSM469970. ChIP-Seq data from human primary CD4+ T cells were obtained from were obtained from the authors directly (Barski et al., 2007).

Genomic mapping of ChIP-Seq data

ChIP-Seq reads were aligned using the software Bowtie (Langmead et al., 2009) to NCBI build 36.1 (hg18) of the human genome with default settings. Sequences uniquely mapping to the genome with zero or one mismatch were used in further analysis. In the cases where multiple reads mapped to the same position and strand all beyond the first two were discarded.

ChIP-Seq density calculation

The genome was divided into bins 25 base pairs in width, beginning at the first base of each chromosome. Each mapped ChIP-Seq read was shifted 100 bp from its mapped genomic position and strand to the approximate middle of the sequenced DNA fragment. Subsequently, the ChIP-Seq density within each genomic bin was calculated as the number of ChIP-Seq reads mapping within a 1kb window (+/- 500bp) surrounding the middle of that genomic bin.

Identification of genes occupied by H3K4me3 and H3K27me3

The genomic coordinates of the full set of transcripts from the RefSeq database (http://www.ncbi.nlm.nih.gov/RefSeq/) from the March 2006 version of the human genome sequence (NCBI Build 36.1, hg18) was downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTables) on September 1, 2010.

For each RefSeq gene the peak ChIP-Seq density in the region +/-1 kb around the transcription start site (TSS) for H3K4me3 and +/- 25kb around the TSS for H3K27me3 was examined. A gene was considered to be occupied by H3K4me3 or H3K27me3 if the peak ChIP-Seq density in this region was greater than a defined threshold (defined below) and the ratio of specific IP to background signal at this position was greater than 3. Two different thresholds were used in each experiment to call H3K4me3 and H3K27me3 occupied genes at both high and low confidence. Since the ChIP and sequencing efficiencies were different for each experiment, appropriate thresholds were different for each dataset. The total number of ChIP-Seq reads used from each experiment as well as the high and low confidence thresholds are reported (Table S1).

A gene was considered differentially occupied by H3K4me3 or H3K27me3 between two experiments if it was called occupied at high confidence in one experiment and was not called occupied at low confidence in the other experiment. We examined only autosomal genes, since iPS cell lines exhibit varying degrees of X-inactivation and H3K27me3 occupancy of the X-chromosome. A summary of the genes occupied by H3K4me3 and H3K27me3 in each cell line is provided (Tables S2A and S2B).

Identification of CpG islands and assignment to genes

The base composition criteria that were originally used to define CpG inlands (Gardiner-Garden and Frommer, 1987) were created long before the complete sequence of the genome was known. Since then, new definitions have been developed, which offer greatly enhanced sensitivity and specificity in the genome-wide identification of CpG islands (Takai and Jones, 2002; Ponger et al., 2002; Hackenberg et al, 2006; Glass et al., 2007; Wu et al., 2010).

We used a modified version of these methods optimized to provide the greatest sensitivity and specificity in comparing CpG island positions to ChIP-Seq datasets. We tabulated the local CG dinucleotide frequency in a 1 kb. window (+/- 500bp) at every position in the genome. CG dinucleotides in protein coding regions of the genome were excluded. Scanning the genome in 25 bp. bins, using the CG frequency at the middle position of each bin, we identified bins in which the CG frequency was greater than or equal to 4.6%. Adjoining bins were collapsed into regions and regions that were less than 300 bp. in length were excluded. This method identifies 42,371 CpG islands in the NCBI build 36.1 (hg18) of the human genome (Table S2C).

In order to assign CpG islands to genes, the following method was used. The genome was scanned and CpG islands within 4 kb. of one another were merged into CpG island clusters. Most of these CpG island clusters consisted of only one CpG island, but there were several thousand clusters of multiple CpG islands. If a CpG island cluster overlapped with the transcription start site (+/-1 kb.) of a gene all CpG islands in that cluster were assigned to that gene. A summary of the number of CpG islands for every gene is provided (Table S2D).

Gene ontology analysis

For gene ontology analysis ChIP-Seq results from the human ES cell line WIBR2 were used. Gene ontology analysis was performed using the online tool DAVID (http://david.abcc.ncifcrf.gov/; Huang et al., 2009). A summary of the genes annotated as encoding regulators of development and homeobox transcription factors is provided (Table S3).

Analysis of RNA secondary structure

For each gene, the sequence +/- 5kb around the TSS was analyzed for sequences which may form the characteristic CG rich, PcG recruiting, hairpin structure (Wutz et al., 2002; Zhao et al., 2008). A 28bp window was slid across each sequence in 1bp increments and for each window the minimum free energy of that sequence folding into the structure shown in Figure 2A was calculated using rnaEval (Hofacker et al., 1994). In order to account for the possibility of the hairpin forming transcript being generated from transcription of either strand and/or in either direction, the free energy was calculated for four different sequence/structure pairs: the 28bp sequence and its reverse complement were compared to the structure in Figure 2A as well as its mirror image (with the smaller loop on the left). The minimum free energy of those 4 combinations was used as the free energy for that window. Windows with a GC mononucleotide content of at least 50% and a minimum free energy at or below -5 kcal/mol were counted as a potential PcG recruiting hairpin sequence. A summary of the number of hairpin hits for each gene is provided (Table S2D).

The fold enrichment shown in Figure 2B and 2C, is a measure of the enrichment of genes with a particular range of potential hairpin hits versus the expected number of genes. As an example of how this is calculated, consider the calculation for the fold enrichment of genes with 15-20 hairpin hits in the multiple CpG class. First we calculate the percentage of multiple CpG genes with between 15 and 20 possible hairpins, ~15%. Then, we calculate the percentage of all genes with this range of hairpin hits, ~8%. The fold enrichment for 15-20 hairpin hits in the multiple CpG class is then simply the observed percentage divided by the expected percentage, or 15/8, for a fold enrichment of ~1.9. This calculation was done for each RNA hairpin range in each gene class.

References

Soldner, F., Hockemeyer, D., Beard, C., Gao, Q., Bell, G.W., Cook, E.G., Hargus, G., Blak, A., Cooper, O., Mitalipova, M., Isacson, O., Jaenisch, R. Parkinson's disease patient-derived induced pluripotent stem cells free of viral reprogramming factors. *Cell* 136, 964-77 (2009).

Guenther, M.G., Frampton, G.M., Soldner, F., Hockemeyer, D., Mitalipova, M., Jaenisch, R., Young, R.A. Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells. *Cell Stem Cell* 7, 249-57 (2010).

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837 (2007).

Lee, T.I., Johnstone, S.E., and Young, R.A. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nat. Protoc.* 1, 729-748 (2006).

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

Gardiner-Garden, M., Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* 196, 261-82 (1987).

Takai, D., Jones, P.A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci.* 99, 3740–3745 (2002).

Ponger, L., Mouchiroud, D., CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics.* 18, 631–633 (2002).

Hackenberg, M., Previti, C., Luque-Escamilla, P.L., Carpena, P., Martinez-Aroza, J., Oliver, J.L. CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics* 7, 446 (2006).

Glass, J.L., Thompson, R.F., Khulan, B., Figueroa, M.E., Olivier, E.N., Oakley, E.J., Van Zant, G., Bouhassira, E.E., Melnick, A., Golden, A., Fazzari, M.J., Greally, J.M. CG dinucleotide clustering is a species-specific property of the genome. *Nucleic Acids Res.* 35, 6798–6807 (2007).

Wu, H., Caffo, B., Jaffee, H.A., Irizarry, R.A., Feinberg, A.P. Redefining CpG islands using hidden Markov models. *Biostatistics* 11, 499-514 (2010).

Huang, D.W., Sherman, B.T., Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nat. Protoc.* 4, 44-57 (2009).

Wutz, A., Rasmussen, T.P. & Jaenisch, R. Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nat. Genet.* 30, 167-74 (2002).

Zhao, J., Sun, B.K., Erwin, J.A., Song, J.J. & Lee, J.T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* 322, 750-6 (2008).

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

Figure S1

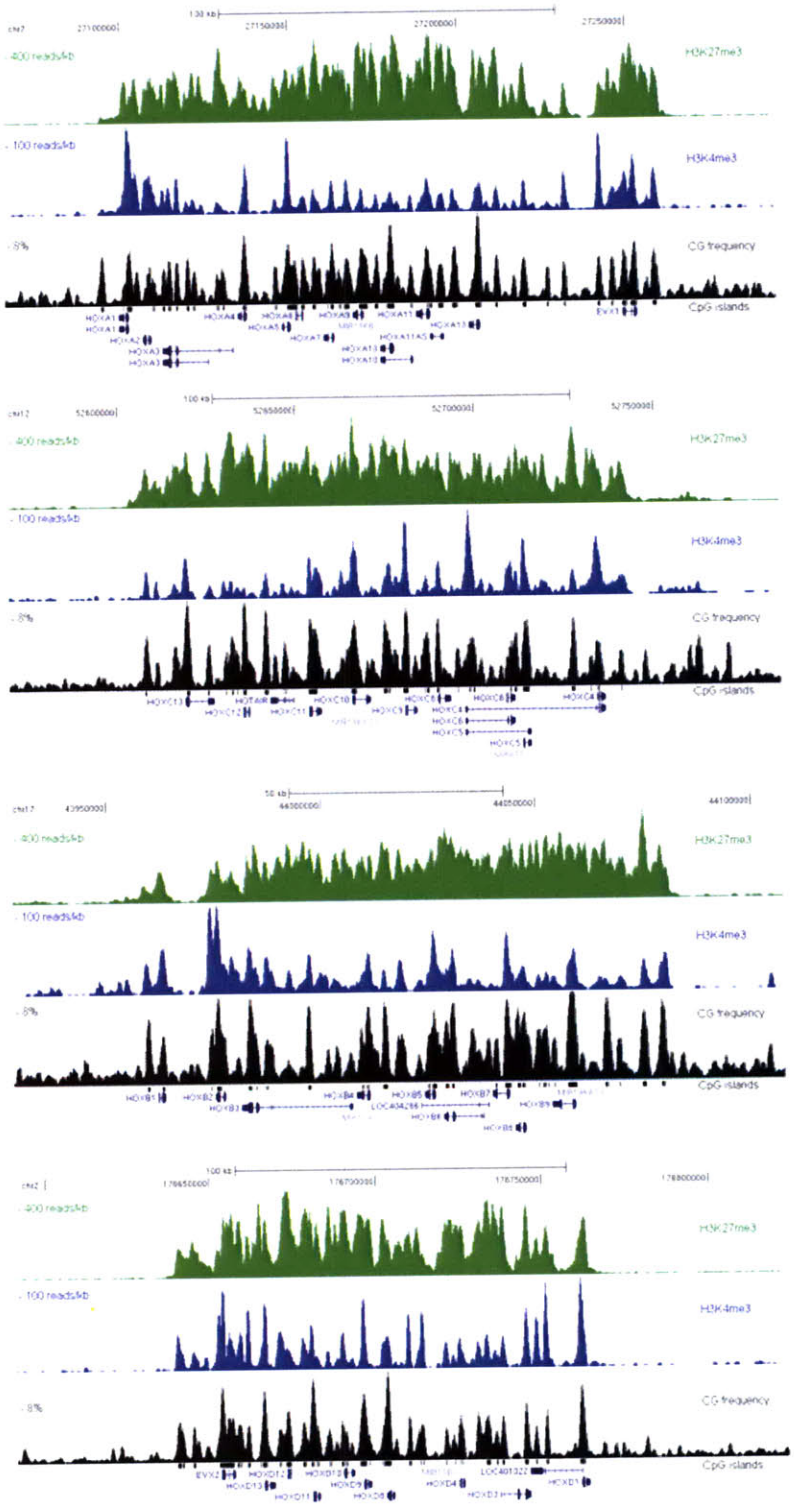


Figure S1. CpG island structure and ES cell TrxG/PcG chromatin structure in the four human Hox gene clusters.

For each of the four human Hox gene clusters H3K27me3 and H3K4me3 ChIP-Seq density in the hES cell line WIBR2, local CG dinucleotide density, and CpG islands are shown.

Supplemental Tables

Table S1: Summary of ChIP-Seq experiments and datasets. The number of ChIP-Seq reads is for each ChIP-Seq experiment and the high and low confidence thresholds for H3K4me3 and H3K27me3 occupancy for each experiment are provided.

Table S2: H3K4me3 and H3K27me3 occupancy in all cells, the number of CpG islands, and the number of RNA hairpin hits for all genes are provided. The genomic coordinates of all CpG islands are also provided.

E) H3K4me3 occupancy for every gene

F) H3K27me3 occupancy for every gene

G) Genomic coordinates of all CpG islands

H) The number of CpG islands and RNA hairpin hits for every gene

Table S3: The number of CpG islands for genes annotated as homeobox transcription factors and as development regulators.

A) The number of CpG islands for genes annotated as homeobox transcription factors

B) The number of CpG islands for genes annotated as development regulators