Dissecting the Transcriptional Regulatory Network of Embryonic Stem Cells

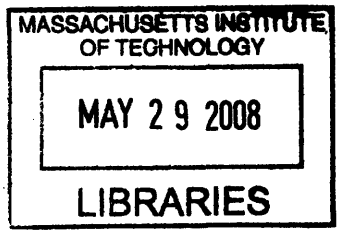
By

Megan F. Cole
B.A., Biology; B.A., Computer Science
Amherst College, 2003

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

DOCTOR OF PHILOSOPHY
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2008



Signature of Author _____
Department of Biology
May 14, 2008

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

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

Dissecting the Transcriptional Regulatory Network of Embryonic Stem Cells

by

Megan F. Cole

Submitted to the Department of Biology on May 14th, 2008
in partial fulfillment of the requirements for the Degree of
Doctor of Philosophy in Biology

Abstract

The process by which a single fertilized egg develops into a human being with over 200 cell types, each with a distinct gene expression pattern controlling its cellular state, is poorly understood. An understanding of the transcriptional regulatory networks that establish and maintain gene expression programs in mammalian cells is fundamental to understand development and should provide the foundation for improved diagnosis and treatment of disease. Although it is not yet feasible to map the entirety of these networks in vertebrate cells, recent work in embryonic stem (ES) cells has demonstrated that core features of the network can be discovered by focusing on key transcriptional regulators and their target genes. Here, I describe important insights that have emerged from such studies and highlight how similar approaches can be used to discover the core networks of other vertebrate cell types. Knowledge of the regulatory networks controlling gene expression programs and cell states can guide efforts to reprogram cell states and holds great promise for both disease therapeutics and regenerative medicine.

Thesis Supervisor: Dr. Richard A. Young
Title: Professor of Biology

Dedication

To my parents Steve and Mary Lou Cole,
for their eternal love, support, and encouragement.

Acknowledgements

No one achieves a PhD without the support of others and I have many people to thank who have helped me along this journey.

I must first thank my advisor, Rick Young, for encouraging my growth as a scientist, teaching me the skills I will need to run my own lab, and for providing me with honest and constructive feedback. Rick has also shown great compassion in supporting me through personally difficult times, for which my family and I are extremely grateful.

Thanks to several other faculty members who have helped shape my career; Mike Laub for encouraging me to pursue a PhD, Laurie Boyer for showing me what can be achieved, Phil Sharp for his insightful advice and caring nature, Hazel Sive for her passion for teaching, and Nancy Hopkins and Gerry Fink for their perceptive comments.

Big thanks to the entire Young Lab for being my colleagues, mentors, friends and Boston family. Special thanks to Stuart Levine for his computational relief and selfless and caring nature; Jamie Newman for her hard work, strong spirit and true friendship; and Tony Lee for being an all around terrific bay-mate, mentor, and friend.

Thanks to my friends, Brett and Susan for keeping the grad class together and to my Girls Night ladies, Amy, Morgan and Denise, for maintaining my sanity.

Thanks to my grandmother Betty Waghelstein, for showing me how to be a strong woman in the professional world. Thanks also to my brother Ryan Cole for his inspiring excitement for life and the new.

Thanks to Mark Styczynski, my fiancé, who more than anyone has seen me through my struggles, and kept me laughing during my MIT years.

Most of all I would like to thank my parents, Steve and Mary Lou Cole. Without their sacrifice, confidence, and love I would not be obtaining this PhD. All my achievements are also a testament to them.

Table of Contents

Preamble

In the first chapter of this thesis I discuss relevant background information, key themes uncovered, and future challenges pertaining to my thesis work on the transcriptional regulatory network of embryonic stem cells. I first introduce transcriptional regulatory networks; discussing mechanisms of eukaryotic transcription, their study in vertebrate cells, and computational insights stemming from studies of networks in model organisms. I then highlight the importance of discovering the transcriptional regulatory network of embryonic stem cells. Next I discuss key themes uncovered from my, and others' work, underscoring their potential implications. Throughout the first chapter I also offer future challenges, goals and possible approaches that logically follow my thesis work. In the following chapters and appendices of this thesis I then present more detailed accounts of the body of work constituting my thesis and discussed in Chapter 1.

Chapter 1

Dissecting Transcriptional Regulatory Networks of Vertebrate Cells

Introduction

With a few exceptions, the diverse array of cell types that constitute the human body have identical genetic material and it is the regulated expression of this genetic material that confers unique properties onto different cell types. During development the genetic material from the initial fertilized egg is methodically passed on to each daughter cell. However, differences in expression of the genetic material between daughter cells are induced in order to initiate different cell lineages. In this manner, different cell types are specified during development by the precise establishment of their unique gene expression programs.

Complex organisms such as humans, that have to specify over 200 diverse cell types, must therefore have robust mechanisms to tightly control the unique expression pattern of each cell. These expression patterns are controlled by transcriptional regulators, which act in a combinatorial manner to control gene expression. Throughout development the set of transcriptional regulators expressed is adjusted in order to establish the proper combination of regulators to specify a given cell type and expression program.

While proper gene expression directs cell fate during development, abnormal gene expression can affect cell state to induce disease. Inappropriate gene expression patterns are the underlying cause of many devastating human diseases including some cancers, autoimmune diseases, diabetes and cardiovascular disease, and many regulators that control gene expression patterns within a cell have been linked to these diseases (Arce et al., 2006; Bennett et al., 2001; Chahrour and Zoghbi, 2007; Couzin, 2008; Kloosterman and Plasterk, 2006; Ryffel, 2001; Villard, 2004; Vogelstein et al., 2000; Wang et al., 2007). Either genetic mutation or aberrant expression of these regulators can cause disease development. Although some disease related expression regulators have been extensively studied, our understanding of how they act within the cell to induce a disease state is limited because we cannot place their contribution to gene expression within the larger context of cellular gene expression control.

The control of expression patterns in vertebrate cells is almost entirely unmapped due to experimental and conceptual limitations that attend its study. Although initial maps of transcriptional regulatory networks have been created for several model organisms (Davidson et al., 2002; Lee et al., 2002; Salgado et al., 2006), it is not yet possible to create equivalent maps in vertebrate cells due to experimental limitations. Recently, however, several groups have begun to tackle the challenge of mapping the control of gene expression patterns in vertebrate cells, particularly embryonic stem (ES) cells, and have demonstrated that significant insights can be obtained using current technologies. A paradigm for studying vertebrate regulatory networks, established by studies in ES cells, together with key themes uncovered from these studies, are described here and provide a possible guide for future studies.

Elucidation of vertebrate control of gene expression patterns must be aggressively pursued as it will lead to advances in our understanding of development, control of cellular state, dysregulation of cell state in disease, as well as our ability to manipulate cell state for disease treatment and regenerative medicine.

Molecular Mechanisms Controlling Eukaryotic Transcription

The level and types of RNA species produced within a cell determines cellular state and is a tightly regulated process. In eukaryotes the control of RNA expression includes inputs from transcription factors, chromatin regulators, signaling pathways and non-coding RNAs (Figure 1). These components form a complex network, the transcriptional regulatory network (TRN), which regulates the expression of RNA within a cell. The mechanisms by which these components affect gene expression are largely understood and can be used to guide the process of mapping the TRNs of vertebrate cells. Here we will only briefly discuss the role of various inputs in the control of gene expression. For more detailed descriptions of eukaryotic transcription and gene expression the reader is referred to several excellent reviews (Kornberg, 2007; Lee and Young, 2000; Lemon and Tijan, 2000; Li et al., 2007; Orphanides and Reinberg, 2002).

In a simplified model, eukaryotic transcription can be viewed as a sequence of events, each step being a regulated process (Figure 2). The initial steps include binding of transcription factors (also called sequence-specific DNA binding factors) to cis regulatory DNA elements, acetylation of histones, and recruitment of the transcription apparatus (Gill, 2001; Santos-Rosa et al., 2002; Struhl, 2005). Serine 5 of the heptapeptide repeat on the C-Terminal Domain (CTD) of RNA Pol II then becomes phosphorylated, allowing Pol II to initiate transcription, and leading to methylation of lysine 4 of histone 3 by a histone methylase of the trithorax family (Hirose and Ohkuma, 2007; Pokholok et al., 2005; Santos-Rosa et al., 2002). The Nelf and DSIF complexes cause a promoter-proximal pause of transcription that is relieved upon phosphorylation of PolII's CTD at serine 2 (Sims et al., 2004). This hyperphosphorylated form of Pol II can then transcribe through the gene, and during this process a histone methyltransferase methylates lysine 36 of histone 3 (Bannister et al., 2005; Pokholok et al., 2005). Methylated H3K36 residues are recognized by histone deacetylases, which deacetylate histones in the coding region, essentially resetting the chromatin for the next round of transcription (Lee and Shilatifard, 2007). Various component classes are involved in the regulation of different steps in the process of gene expression and the manner in which these components regulate this process affects their role in the transcriptional regulatory network controlling cellular gene expression.

Transcription factors lie at the heart of transcriptional regulation and anchor transcriptional regulatory networks. Due to their ability to recognize specific DNA sequences transcription factors serve as the mechanism whereby the genetic information encoding instructions for proper gene expression is interpreted. Pol II cannot itself recognize promoter sequences and instead must be recruited by transcription factors. Transcription factors are also the single largest protein family encoded in the human genome, where they account for approximately 10% of protein-coding genes (Babu et al., 2004; Lander et al., 2001; Levine and Tijan, 2003). They bind to both promoter proximal and distal (even as far away as 100Kb) regulatory DNA sequences and can both aid or inhibit recruitment of the transcription apparatus at target genes (Blackwood and Kadonga, 1998; West and Fraser, 2005).

Chromatin regulators are typically recruited to specific portions of the genome by DNA-binding transcription factors or the transcription apparatus where they can affect

Figure 1

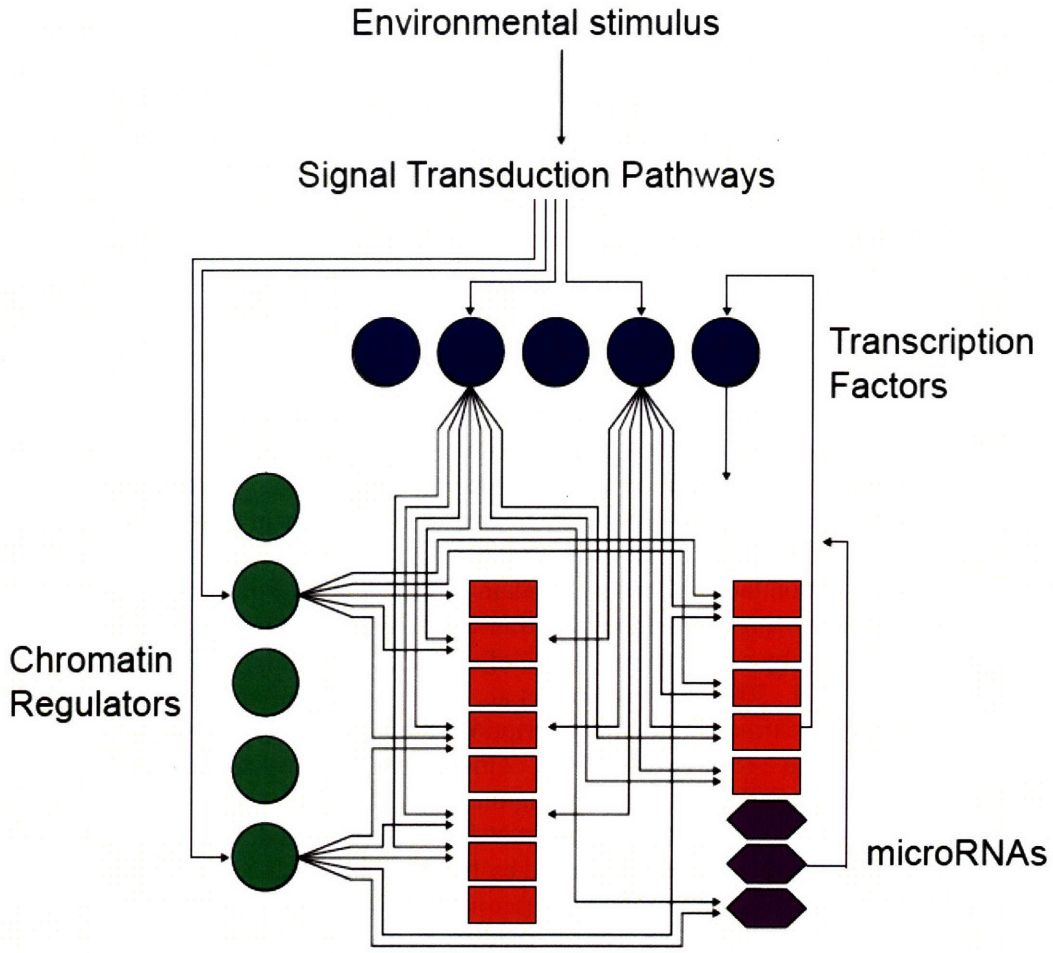

Environmental stimulus
Signal Transduction Pathways
Transcription Factors
Chromatin Regulators
microRNAs

Figure 1 Abstract Map of a Transcriptional Regulatory Network.
Diagram depicting the inputs from transcription factors (blue circles), chromatin modifiers (green circles), signaling pathways, and miRNAs (purple diamonds) to form a network regulating the expression of target genes (target protein coding genes are represented by orange rectangles and target non-coding RNAs are represented by purple diamonds).

Figure 2

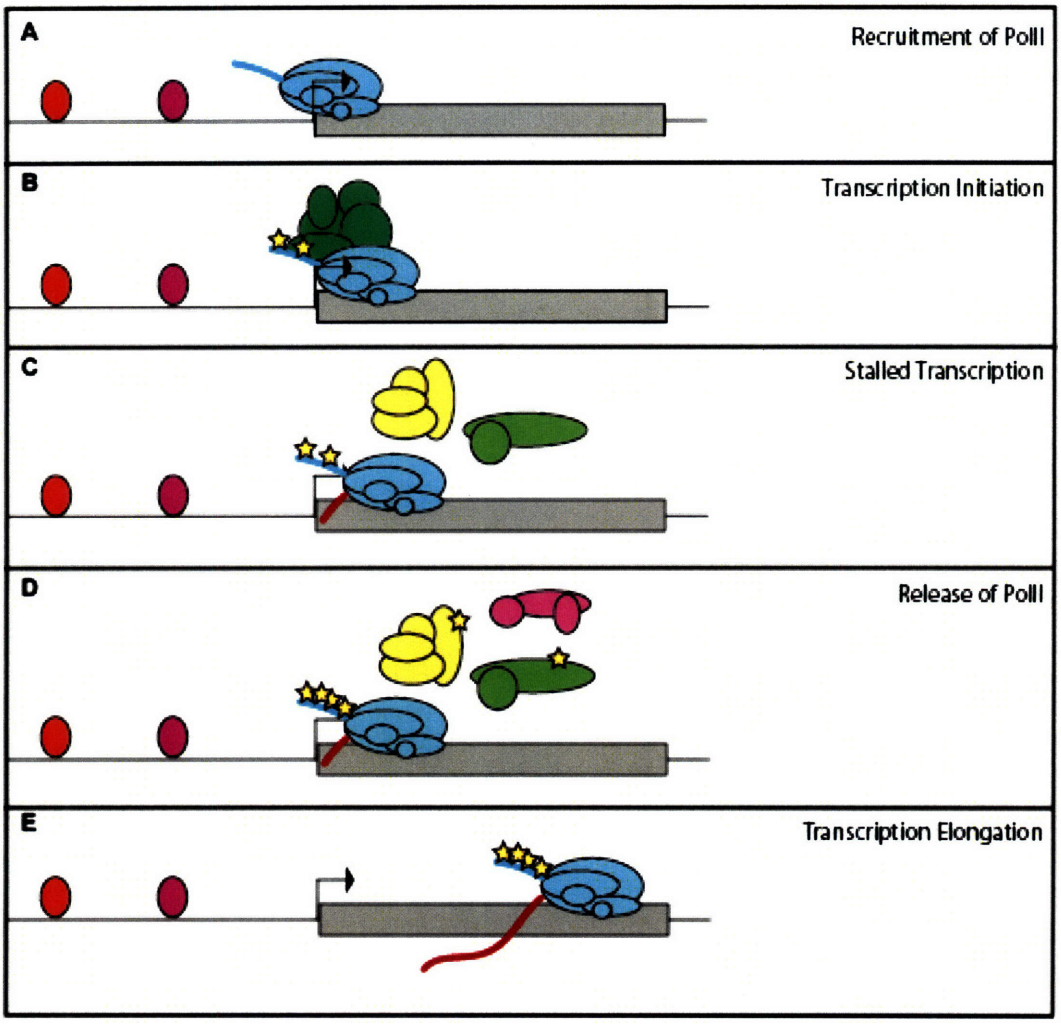

A
Recruitment of PolII
B
Transcription Initiation
C
Stalled Transcription
D
Release of PolII
E
Transcription Elongation

Figure 2 Model of eukaryotic transcription.
(A) Transcription factors (red and purple ovals) bind to DNA motifs and recruit transcription apparatus (blue complex) and histones H3 and H4 become acetylated.
(B) Transcription is initiated by the phosphorylation (yellow stars) of serine 5 on the heptapeptide repeat on the C-terminal domain of PolII by TFIIH (green complex), histone H3 is tri-methylated at lysine 4 by the Trithorax complex and PolII starts transcribing.
(C) Transcription is stalled by NELF (yellow complex) and DSIF (light green complex).
(D) Stall is released by phosphorylation of serine 2 on the PolII C-terminal domain, NELF and DSIF by PTEFb (pink complex).
(E) PolII now transcribes through the entire entire gene creating full-length transcript and leading to methylation of lysine 36 on histone H3.

gene expression and hence act to augment the transcriptional regulatory network. Unlike many transcription factors, most chromatin regulators function in the same fashion at all target genes, either repressing or activating their expression. Chromatin regulators often function by depositing methylation or acetylation marks on histones making it possible to deduce the role of many chromatin regulators at target genes in the transcriptional regulatory network by examining a small number of histone modifications. Cells are able to maintain histone marks and so the affect of chromatin modifiers can be long lasting and has been proposed as a possible mechanism of cellular memory (Bantignies and Cavalli, 2005; Hirose, 2007; Turner, 2002).

Signaling pathways connect environmental signals to the transcriptional regulatory network to effect changes in gene expression and potentially the network itself. They can transduce extracellular signals through a series of intracellular components eventually ending with a terminal component capable of affecting cellular state through a variety of mechanisms. Terminal components are often protein kinases that activate transcriptional regulators, or are transcriptional regulators themselves such as transcription factors and chromatin modifiers. In this manner, signaling pathways can initiate new gene expression programs and can be viewed as the most upstream components of transcriptional regulatory networks.

Although the mechanisms by which non-coding RNAs (ncRNAs) affect gene expression are less understood, the need to incorporating their influence into the transcriptional regulatory network is clear. Non-coding RNAs can affect chromatin state and transcriptional regulators (Amaral et al., 2008; Barrandon et al., 2008; Goodrich and Kugel, 2006; Hawkins and Morris, 2008). A large class of ncRNAs, termed micro RNAs (miRNAs), modify gene expression by inducing degradation of mRNA transcripts (Ambros and Chen, 2007; McManus and Sharp, 2002; Wu and Belasco, 2008). In this way ncRNAs can act at multiple stages of transcription regulation, with miRNAs acting downstream of other component classes in the transcriptional regulatory network.

Through an understanding of the mechanistic control of gene expression by these component classes, scientists can gain insight into how each type of component controls the transcriptional regulatory network genome-wide and use this knowledge to guide studies in this area. Transcription factors, by recognizing genetic information encoded in DNA, act as the anchor of the network. Chromatin regulators, by modifying chromatin structure are able to augment the landscape of the network and can also serve to implement more long-term or stable gene expression regulation. Signaling pathways act to maintain or initiate changes in the regulatory network in response to environmental or developmental cues. miRNAs and other ncRNAs may act to fine-tune expression levels and serve as a rapid means of control or change in the network as they are faster to produce then protein regulators and can immediately affect protein production by eliminating existing mRNAs rather than simply eliminating the production of mRNAs. Given these unique roles it is clear that studies of transcriptional regulatory networks must begin by examining key transcription factors but that regulators from each component class should be examined to gain a more thorough understanding of the multi-level and combinatorial control of vertebrate gene expression.

Core Transcriptional Regulatory Networks

Levine and Tijan proposed that "organismal complexity arises from progressively more elaborate regulation of gene expression" (Levine and Tijan, 2003). Organismal complexity does not correlate with genome size but rather with the ratio of transcriptional regulators to genes (Babu et al., 2004; Levine and Tijan, 2003; Nimwegen, 2003; Tupler et al., 2001). Given the combinatorial nature of the control of gene expression, a small increase in the number of transcription factors could lead to a large increase in regulatory complexity. With the large number of regulators in vertebrates this implies that their transcriptional regulatory networks will be enormously complex.

Transcription factors alone account for approximately 10% of protein-coding genes in humans (Babu et al., 2004; Lander et al., 2001; Levine and Tijan, 2003). If one assumes that 1/3 of protein-coding genes are expressed in each cell type (Brandenberger et al., 2004; Guenther et al., 2007; Sato et al., 2003; Su et al., 2004), and that a similar fraction of transcription factor genes are expressed, then roughly 700 transcription factors are expressed in each cell. Even attempting to map this network, without accounting for the other types of network components such as chromatin regulators or ncRNAs, is currently not feasible experimentally.

While each regulator contributes to the proper expression of genes within a cell, only a small subset of these regulators play key roles in establishment or maintenance of cell state. Many regulators can be removed without dire consequences for the cell (Giaever et al., 2002; Kemphues, 2005; Winzeler et al., 1999). This may be a biological mechanism to protect against mutations and to ensure robust gene expression patterns, and may be due to the scale-free architecture of biological networks (discussed below). The small set of regulators that play a vital role in a particular transcriptional regulatory network are termed here 'key regulators' and are defined by their important role in maintenance or establishment of the cell state governed by the network.

A simplified version of a transcriptional regulatory network that still captures many important themes can be deduced by discovering the population of genes that are controlled by the key regulators for that cell type. We call this simplified network the 'core transcriptional regulatory network'. Given current experimental limitations in studying complete vertebrate networks, discussed below, more focused studies of core networks, guided by identification of key regulators, should predominate scientific efforts. These studies of core networks, though missing many inputs from other regulators, can nonetheless still elucidate key network themes, as has been shown to be the case in ES cells.

Transcriptional Regulatory Networks in Vertebrate Cells

Studies of transcriptional regulatory networks in vertebrate cells are of vital importance to studies of development, disease treatment and regenerative medicine but are hindered by substantial experimental and conceptual challenges. Recent technological advances have allowed for high-throughput studies in model organisms such as *E. coli* and *S. cerevisiae,* but due to the increased complexity, the same brute-force approaches cannot efficiently map the TRNs of vertebrate cells. In order to achieve success in vertebrate studies it is therefore important to identify a clear set of goals, to understand the capabilities and limitations of currents technologies and to identify pivotal challenges faced by vertebrate studies.

Goals

Knowledge of vertebrate TRNs throughout development is essential to understanding this process as well as being able to manipulate cell fates for medicinal purposes. The overall goal of mapping the complete transcriptional regulatory network of every cell type and understanding how these networks are dynamically established during development is a large challenge that cannot, with current technologies and resources, be solved within a short timeframe. Given the extreme complexity in deciphering these networks it is important to establish goals that can guide scientists to choose which portions of the overall goal to tackle first. Although studies in ES cells did not explicitly lay out a series of objectives, they have nonetheless established a clear set of goals that can guide future studies.

As scientists cannot simultaneously tackle all possible cell-types, an initial goal must be to identify cell-types that will yield the highest rewards. These cell-types should be selected based on medical relevance and ease of study. Cell-types that hold the most promise for regenerative medicine or for disease treatment should be prioritized. However, practical concerns regarding the ease of obtaining high numbers of homogenous cells should also be taken into account. Although primary tissues would be ideal, comparable cell lines will often be the best choice due to the experimental benefits they offer.

As discussed above, it is also of primary importance to assemble an initial map of the core transcriptional regulatory network using a subset of components that are most likely to play a key role in the control of gene expression for the particular cell type. Given the enormous number of transcriptional regulators that are involved in the control of gene expression, it is vital to identify a manageable number of these components that could quickly be mapped and that are likely to reveal key themes in the transcriptional regulatory network. The first components to be studied are likely to be cell type specific transcription factors, but it will also be important to add inputs from other component classes.

Another goal should be to identify how cells dynamically manipulate their transcriptional regulatory networks in response to extracellular cues. These studies will be quite challenging as capturing dynamic changes in a complex network is experimentally much more difficult than mapping more static cell states. However, this understanding is of vital importance because it can guide efforts to manipulate cell state for medical purposes.

The ultimate goal is for scientists to be able to manipulate cell state, ideally without the use of recombinant DNA. Knowledge of how cells in vivo direct changes in the transcriptional regulatory network to affect changes in cell state should be able to guide efforts by scientists to do the same. This goal is of vital importance as the controlled manipulation of cell state is a necessary step to being able to harness the therapeutic potential that modern biology promises.

By prioritizing the order in which scientists tackle the monumental challenge of mapping the complete vertebrate TRN throughout development it is likely that the conceptual and medical benefits that stem from this knowledge will be realized more rapidly.

Current Technologies

Generally two types of experiments are used to study transcriptional regulatory networks in vertebrates: location analysis and expression studies. Location analysis identifies the location of proteins along the genome and so can identify potential direct target genes of transcriptional regulators. Expression studies, such as the measurement of expressional changes upon loss of a transcriptional regulator, are also used to identify potential direct or indirect target genes or to examine the functional consequence of regulator binding at target gene promoters.

The location analysis assay identifies the location of regulatory proteins along the genome and so can be used to identify the direct target genes of a transcriptional regulator. The technology behind this assay is quickly evolving to improve both data quality and high-throughput capabilities. Detailed descriptions of location analysis assays can be found in several manuscripts and are only briefly described below (Acevedo et al., 2007; Bulyk, 2006; Gordon et al., 2007; Lee et al., 2006).

The location analysis protocol builds off of the classic pull-down experiment. Protein-DNA interactions are first immobilized by chemical crosslinking in vivo. Crosslinked chromatin is then sheared into smaller fragments, typically 100-500bp in length. Sequences bound by the protein of interest are then enriched by immunoprecipitation, either of native protein or a tagged protein construct. Enriched sequences are then identified by DNA microarrays (using non-immunoprecipitated chromatin as a control) or by direct sequencing.

The experimental ease and cost associated with the detection of bound DNA sequences by either array or sequencing have been reduced recently, making the use of these techniques by many labs now possible. The increased resolution offered by sequencing based technologies makes it likely that this method will predominate in future studies. In fact, this technology can pinpoint protein binding sites to within 25bp or less (Marson et al., submitted).

Expression analyses also play a vital role in mapping transcriptional regulatory networks (Bar-Joseph et al., 2003; Gao et al., 2004; Goutsias and Lee, 2007; Li et al., 2008). Advances in measurement of transcript levels as well as manipulation of transcriptional regulators to induce changes in expression have been made recently and improve the ability of scientists to study transcriptional regulatory networks in vertebrate cells. Sequencing based technologies now allow for the efficient measurement of small RNA transcripts in addition to the measurement of mRNA levels (Marson et al., submitted). This allows for small non-coding RNAs, such as miRNAs, to be

incorporated into maps of transcriptional regulatory networks. Advances in the manipulation of transcriptional regulators through RNAi allow for more efficient studies examining the functional consequences of transcriptional regulators on target gene expression (Cockrell, 2007; Dann, 2007; Ding and Buchholz, 2006).

Although these genomic technologies continue to improve, there remain fundamental challenges that have not yet been solved that impede more high-throughput studies.

Challenges

While maps of the transcriptional regulatory networks of model organisms such as *E. coli*, *S. cerevisiae* and even the multicellular organism sea urchin, are rapidly being discovered, little is known about the TRNs of vertebrate cells due to enormous experimental and conceptual challenges associated with studying these networks. A key aspect of mapping the TRNs of vertebrate cells will likely be the development of technologies and analyses that will make these studies more feasible. In order to push forward the development of these technologies it is vital to first clearly establish what the specific challenges in creating maps of vertebrate TRNs are.

An obvious experimental challenge faced in vertebrates is the shear number of networks to map. It is estimated that humans have over 200 cell types, each with a unique regulatory network. Given the difficulty of creating a complete TRN for even a single vertebrate cell type, the added challenge of performing this task in many cell types only multiplies the experimental time and expense.

A related experimental challenge is the difficulty in obtaining the necessary purity and number of cells to perform the requisite experiments. Most chromatin immunoprecipitation experiments require at least one million relatively pure population of cells. Obtaining this number and purity of cells can be very difficult, if not nearly impossible, for certain cell types. It would be particularly challenging to attempt to obtain relatively homogenous cell populations for cells transitioning from one cell type to another. Currently, scientists can approximate this process only by attempting to string together static TRN maps made from cells frozen in different stages of development.

Another immense experimental challenge faced is simply the incredible number of components to map in vertebrate cells. Even if researchers were to map only the role of transcription factors, this would require mapping roughly 700 factors. With current technologies, mapping the regulatory relations for this number of factors would be infeasible due to both experimental time and expense.

A related experimental challenge is the difficulty in mapping regulatory relations for many factors. Current assays only map one factor at a time. A major experimental challenge will be to develop high-throughput techniques that allow scientists to map many factors in a single experiment.

Another limitation with current location analysis assays is the difficulty in immunoprecipitating certain factors. These difficulties can be caused by many reasons such as lack of a strong antibody, obstruction of antigen by other components and indirect or temporal interaction of the factor with the chromatin. Some labs have avoided issues associated with antibody performance by instead tagging proteins. However, this technique is hampered by the difficulty of tagging endogenous proteins and by potential biological effects due to the tag. Limitations caused by the immunoprecipitation step will

likely be solved in the near future by a combination of better antibody production, more high-throughput and quality checked tagging strategies, and better techniques for capturing indirect and transient protein-DNA interactions.

There are also analytical challenges associated with network studies in vertebrate cells; a major one being the assignment of target genes. Data gathered from location analysis experiments merely reveals where along the chromosome a regulator binds, which does not automatically translate into a clear list of target genes. In higher organisms, regulatory binding sites can be quite far from the transcription start site, making it difficult to assign binding sites to target genes without further information (Blackwood and Kadonga, 1998; West and Fraser, 2005). Current studies generally use simple algorithms to identify target genes based solely on promoter proximal binding events and so likely miss assignment of many target genes. New technologies, such as Chromatin Conformation Capture (3C) hold great promise for use in assigning binding sites to target genes (Dekker et al., 2002).

Another conceptual challenge is to decipher the logic controlling target gene expression. Current technologies can identify the regulators of a target gene but it is difficult to determine how they act in combination to determine expression level. The solution to this question will likely come from a combination of more collected data to analyze, new analysis techniques and also new types of experiments aimed at deciphering this logic.

A final, and perhaps most important, conceptual challenge is the useful interpretation and representation of network maps. Even in bacteria these maps are too complex to be represented in a figure that can be comprehended. Just as a series of tools had to be developed in order to allow researchers to easily make use of genome sequence data, it will also be necessary to develop tools to allow researchers to easily make use of TRN data. It seems likely that the production of vertebrate network data and the development of user-friendly tools, will fundamentally change the process of biological research just as genome sequence data has.

Box 1. Goals and Challenges of Elucidating Vertebrate Networks

Goals:

- Identify medically relevent cell types to map
- Identify key regulators in each cell type to map
- Determine cellular mechanisms for network manipulations
- Manipulate networks to reprogram cell states

Challenges:

Experimental

- Number of vertebrate cell types to map
- Difficulty in obtaining large populations of homogenous cells
- Inability to capture network transition states
- Number of vertebrate transcriptional regulators to map
- Lack of high-throughput mapping techniques for multiple factors
- Immunoprecipitation of factors

Conceptual

- Target gene assignment
- Deciphering logic of gene control
- Development of useful representations and tools

Insights from Model Organisms

The transcriptional regulatory networks of several model organisms have largely been mapped and analyses of these networks offer important insights that will likely translate to vertebrate cells. Researchers often represent and analyze these networks as graphs where nodes represent transcriptional regulators (regulatory proteins or RNAs) and target genes/transcripts (Figure 3A). Edges connect regulators to target genes and imply regulation of expression of the target gene by the regulator. In some representations, edges can also be used to depict the regulation of a factor by another factor, or to depict the fact that a gene is the transcript encoding a regulator. A few key insights from studies in model organisms are discussed below and, for more thorough reviews, the reader is referred to several excellent papers (Alon, 2007; Balaji et al., 2006; Davidson et al., 2002; Lee et al., 2002; Levine and Davidson, 2005; Salgado et al., 2006; Thieffry et al. 1998).

Scale-free/hub network architecture

Canonical network or graph-theory terms and measures are also used to describe biological transcriptional regulatory networks. An important measure associated with every network node is its degree, k, also referred to as its connectivity; the number of edges connecting the node to other network nodes. The average node degree is equal to 2L/N, where L is the total number of network edges and N is the total number of network nodes. The degree distribution of a network describes the probability, P(k), that a node will have k edges. When P(k) follows a Poisson distribution (i.e. edges are distributed randomly) then the network is termed random whereas when P(k) follows a Power Law distribution the network is termed scale-free (Figure 3B).

Transcriptional regulatory networks, along with many other types of biological and real-world networks tend to be scale-free or hub networks (Barabasi and Oltvai, 2004). Unlike random networks, where most nodes have close to the average degree, scale-free networks have a subset of nodes that have many connections. These nodes act as network hubs. Scale-free networks also describe protein interaction networks, social networks and the network of internet web pages (Goh et al., 2002).

A hub network architecture creates network stability in that the removal or disruption of a large number of nodes at random will not greatly affect the overall network (Barabasi and Oltvai, 2004). However, these networks are vulnerable to the directed disruption of key hubs. This type of network architecture would therefore allow cells to maintain proper gene expression even with mutations in several genes. However, when a set of key regulators are disrupted, as is often the case in cancer and disease, then the network becomes unstable and could lead to improper gene expression and a change in cell state.

Network Motifs

The determination of a large amount of the TRNs for several model organisms has allowed computational analyses of these networks to identify general network motifs. A network motif is a pattern in the connection between nodes that occurs in the network at a higher than random frequency. Several important network motifs have been identified in

Figure 3

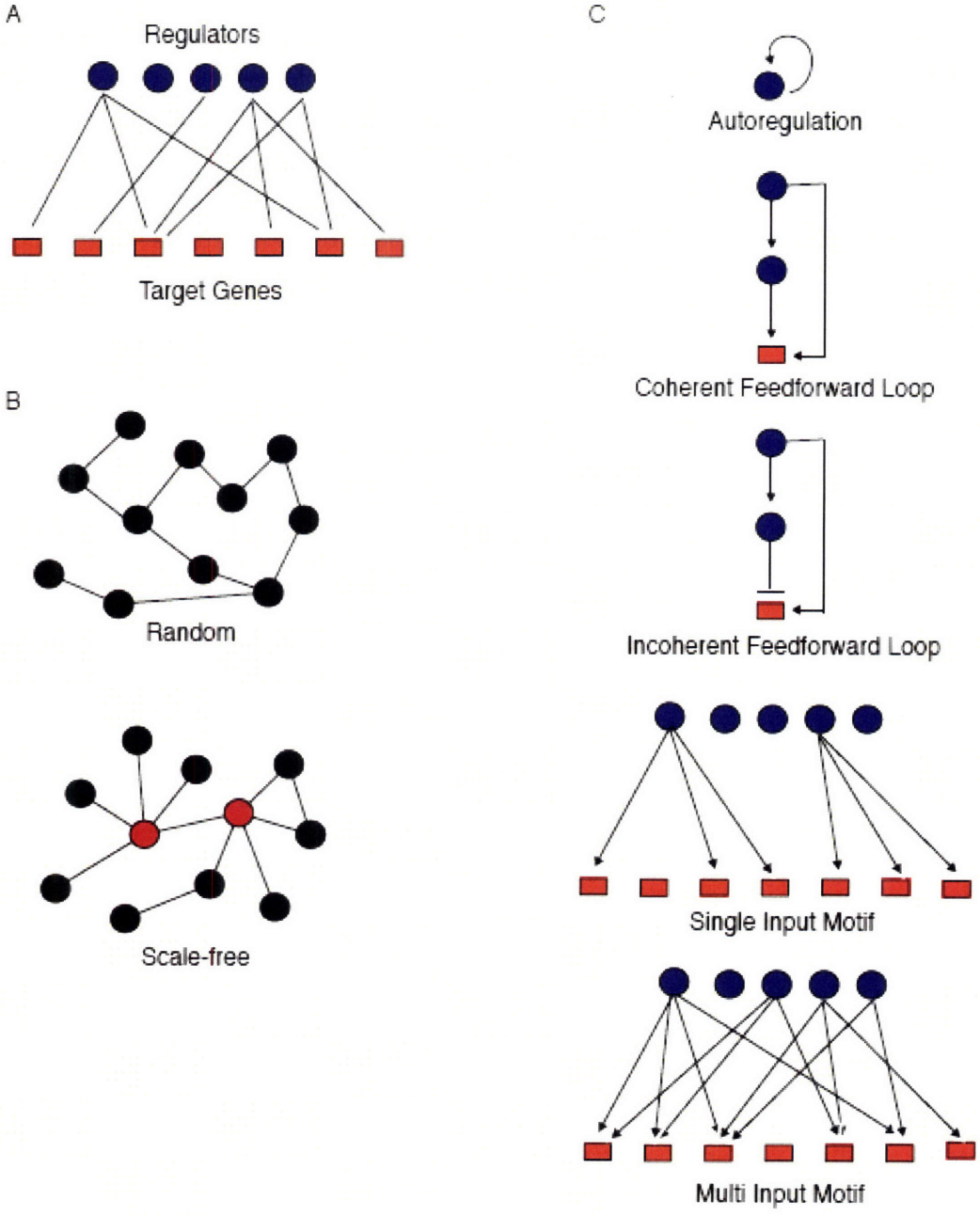

A
Regulators
Target Genes
B
Random
Scale-free
C
Autoregulation
Coherent Feedforward Loop
Incoherent Feedforward Loop
Single Input Motif
Multi Input Motif

Figure 3 Network diagrams and motifs.
(A) Graph representation of a transcriptional regulatory network where nodes represent transcriptional regulators (blue circles) or target genes (orange rectangles). Regulation of a target gene by a transcriptional regulator is depicted by a black line.
(B) Depictions of random and scale-free/hub networks, each with 11 regulators (circles) and 11 edges (black lines). Hub nodes are colored red.
(C) Depictions of network motifs. Regulators are represented as blue circles, target genes as orange rectangles, positive regulation is represented by arrows and negative regulation by t-bars.

model organism TRNs and are depicted in Figure 3C (Alon et al., 2007; Balaji et al., 2007; Lee et al., 2002).

Autoregulation is the simplest possible network motif and involves a regulator regulating its own transcript, either positively or negatively. Autoregulation can be used to affect response time and variation (Alon et al., 2007). Negative autoregulation decreases the amount of time it takes for a regulator to reach its stable expression level upon induction and leads to tight control of expression level. Alternatively, positive autoregulation increases the amount of time it takes for a regulator to reach its stable expression level and causes higher variation in expression level and potentially even a bimodal distribution of expression in a population of cells. Autoregulation can therefore be used by cells to control the stability and dynamics of key regulators' expression levels.

Feedforward loops are another common network motif found in the TRNs of model organisms. There are 8 possible feedforward motifs involving 3 nodes, but 2 of these motifs occur at a high frequency (Alon et al., 2007). A common motif is a coherent feedforward motif where regulator A positively regulates the transcripts of B and C and regulator B additionally positively regulates the transcript of C. Another common motif is an incoherent feedforward motif where regulator A positively regulates the transcripts of B and C but regulator B negatively regulates the transcript of C. Coherent feedforward motifs can be used by cells to control the response time to both addition and removal of signal and incoherent feedforward motifs can be used to create a pulse of expression.

Other important network motifs are the single-input and multi-input motifs (Balaji et al., 2007; Lee et al., 2002). Single-input motifs involve a single transcriptional regulator node connecting to many target gene nodes that are not controlled by other regulators. This motif allows a cell to turn on a set of related target genes simply by turning on a single regulator. This motif is useful for the control of genes with simple expression patterns where their expression is reliant on a single condition/cell type. Conversely, multi-input motifs involve regulation of target genes by a combination of regulators and allow the cell to integrate multiple signals to control complex expression profiles.

Combinatorial Control of Gene Expression

In order for organisms to establish complex patterns of gene expression using a limited number of transcriptional regulators these regulators must act in various combinations to produce differing expression patterns. Complex patterns of gene expression are of obvious importance for multicellular organisms that must establish many different cell fates. However, even unicellular organisms rely on the combinatorial control of gene expression to allow them to adjust their expression pattern in response to varying environmental conditions.

Analysis of the *S. cerevisiae* transcriptional regulatory network revealed the large extent of combinatorial control of gene expression in eukaryotic cells (Balaji et al. 2006). The network analyzed included 157 transcription factors, 4410 target genes and a total of 12873 interactions. The average gene was controlled by 2.9 regulators, with 45 genes being regulated by 15 or more regulators. Further analysis uncovered the extent to which pairs of regulators controlled significantly overlapping target gene sets. They found that a surprising number of regulators controlled overlapping gene sets, forming multi-input motifs. These analyses demonstrate the large extent to which transcriptional regulators

act in combination at target genes, even in relatively simple organisms, to create complex patterns of gene expression and it is postulated that higher organisms will display even more complex and combinatorial regulation of gene expression.

ES Cells Provide an Initial Transcriptional Regulatory Network

An important first goal stated above is to map the transcriptional regulatory networks of developmentally and medically important cells types. Embryonic stem cells are perhaps the most developmentally and medically relevant cell type as they are capable of differentiating into all other cell types within the vertebrate body. They nearly mark the starting point for the dynamic map of regulatory networks throughout development and are currently the earliest experimentally feasible cell type that can be studied, as relatively pure populations can be grown in culture. For these reasons, embryonic stem cells are a natural choice for initial efforts dissecting vertebrate TRNs. In fact, several labs have already made great advances in mapping this network and their work sets a precedent for how future studies of vertebrate networks could be successfully conducted.

ES cells are defined by two properties; the ability to self-renew and the ability to differentiate into many cell types (plutipotency). ES cells are derived from the inner cell mass of the developing blastocyst and can be propagated in culture under appropriate conditions. In culture these cells can be induced to undergo differentiation towards various cell states and, to some extent, can be directed down specific lineages, although our current understanding and ability to manipulate this process is limited. Better understanding the transcriptional regulatory network of ES cells and how it is dynamically altered during development will not only aid our understanding of this process, but will likely also aid our ability to manipulate this process in order to create therapeutically relevant cell types.

With the recent advance in the ability to induce pluripotent stem cells from somatic cells, the use of ES cells in human regenerative medicine is now even more of a possibility that must be aggressively explored. The ability to derive ES cells from a patient and to then manipulate these cells into medically relevant cell types holds great promise for many diseases. However, realization of this potential requires an understanding of the network controlling ES cells and how it is manipulated by external signals to differentiate down various pathways.

Studies in both human and mouse ES cells have begun to assemble a map of their transcriptional regulatory network. Although more work must to be done to further fill in this initial network diagram, already many important insights into the control of the pluripotent state as well as general insights into vertebrate TRNs have been discovered. This review will outline studies of the TRN of ES cells in order to highlight important insights into vertebrate networks and to offer a paradigm for how further studies in vertebrate cells can be efficiently and effectively conducted.

Selection of Key Regulators

A crucial aspect of the approach used to study networks in ES cells is careful identification of key transcriptional regulators that must be mapped. This has allowed for the elucidation of fundamental themes found within this network even though only a small fraction of the regulatory relations have been mapped. Key transcriptional regulators play a vital role in establishing or maintaining a cell state, acting to define the specific network needed for a particular cell type. Given the importance of identifying key regulators to map, it is essential to define criteria that can be used to guide this process. Studies in ES cells suggest that several lines of evidence prove useful for identification of key regulators, including genetic phenotypes, expression profiles and knowledge of molecular characteristics.

Genetic Phenotypes

Genetic perturbation of key network components often results in a phenotype that highlights their vital role in a particular cell type. There are generally three types of genetic evidence that one can use to identify potential key regulators; genetic perturbation in the cell type of interest, genetic perturbation in a whole animal model, and ectopic expression of the component in a different cell type. One can draw upon the array of knowledge gained from these three types of studies in model organisms to identify potential key regulators of a particular network.

Improper levels of key component expression within the cell type(s) they control can lead to loss of cellular identity or function. In ES cells either loss or overexpression of the key regulator Oct4 leads to inappropriate differentiation (Niwa et al. 2000).

Animals lacking a key regulator are often unable to properly produce or maintain that particular cell type. For example, mice with mutations in the Pax-6 gene, a master regulator of sensory organs display the small eye phenotype and mutations in Pax-6 are also associated with the human eye disease aniridia (Callaerts et al. 1997). Similarly, mice missing Oct4 are unable to develop a normal inner cell mass (Nichols et al., 1998).

Certain components are so key that their improper expression can transform one cell type to another. A classic example of this phenomenon is the ability of MyoD to transform fibroblasts into muscle cells (Weintraub et al. 1989). Recent studies have also demonstrated the ability of several key ES cell regulators to transform somatic cells into ES cells (Jaenisch and Young, 2008; Okita et al., 2007; Park et al., 2008; Takahashi and Yamanaka, 2006; Takahashi et al., 2007; Wernig et al., 2007; Yu et al., 2007).

Expression Profiles

Key components that act to establish or maintain a cells' transcriptional regulatory network are often uniquely expressed in that particular cell type. Expression of these components in other cell types could transform these networks and so their expression is highly specific to a single or small subset of cell types. The large number of expression datasets publicly available can therefore be used to identify potential key components of transcriptional regulatory networks.

Simple clustering of expression data can often identify regulators that are uniquely expressed in a cell type. For example, the transcriptional regulator NeuroD1 is uniquely expressed in neural cells and pancreatic islet cells and has been shown to be

capable of reprogramming hepatocytes to pancreatic like cells (Kojima et al., 2003). Work from the Yamanaka lab leveraged this type of analysis to identify a small number of potential key regulators of ES cells for reprogramming purposes (Takahashi and Yamanaka, 2006).

Molecular Knowledge

Knowledge of molecule relations and interactions can also aid the identification of key components. There are many published experimental and computational studies that offer a wealth of information to scientists regarding the relationships and interactions between proteins. This information can easily be exploited to identify new key components through their relation to other known key components

Regulators within certain gene families often share important roles in the regulation of transcription and development. For example, proteins in the homeodomain family have been shown to be key regulators of cell fate specification (Hombria and Lovegrove, 2003). The fact that the ES cell regulators Oct4 and Nanog are homeodomain proteins further implicated a key role for these factors in maintaining the ES cell state.

Regulators often act in complexes to exert their influence on gene expression and the knowledge of a key role of one coregulator can imply an equally important role for the other. For example, the fact that Sox2 forms a heterodimer with the key ES cell regulator Oct4 highly suggested that Sox2 is also a key regulator of ES cells (Ambrosetti et al., 1997). A recent study in ES cells has further exploited this concept by identifying proteins that interact with the ES cell key regulator Nanog in an attempt to identify other key ES cell regulators (Wang et al., 2006).

Transcription Factors

Initial studies of transcription factors in the ES cell transcriptional regulatory network focused on the key regulators Oct4, Sox2 and Nanog. Knowledge of genetic phenotypes, expression profiles and molecular relations were all leveraged to identify these factors as potential key components of the ES cell network. Genetic studies demonstrated functional consequences in ES cells of inappropriate expression of these factors, the expression of Oct4 and Nanog was found to be specific to pluripotent cells, and Sox2 was known to form a heterodimer with Oct4 (Ambrosetti et al., 1997; Avilion et al., 2003; Chambers et al., 2003; Hart et al., 2003; Lee et al., 2004; Mitsui et al., 2003; Nichols et al., 1998; Scholer et al., 1990). Due to the overwhelming evidence suggesting a key role for these regulators in the ES cell network, two groups mapped their target genes in ES cells and uncovered several important network themes (Boyer et al., 2005; Loh et al., 2006).

The genomic binding sites for these three factors were identified in ES cells using location analysis and revealed a striking co-regulation of target genes. Although Oct4 and Sox2 were expected to bind the same target genes, as they form a heterodimer, researchers were surprised to find that Nanog also occupied a large percentage of Oct4-Sox2 bound genes. More recent studies that map additional transcription factors in ES cells have found that these factors also follow the theme of target gene co-occupancy (Cole et al., 2008; Kim et al., 2008).

These studies suggest that the multi-input motif may be a powerful network motif utilized in vertebrate cells. While the average target gene in *E. coli* is regulated by roughly 2 transcription factors, and the average gene in *S. cerevisiae* is regulated by roughly 3 transcription factors, these early studies in ES cells suggest that this number will be significantly higher for vertebrate cells. This finding supports the hypothesis put forth by Levine and Tijan that increased organism complexity is due in large part to increased complexity of gene expression patterns (Levine and Tijan, 2003).

The large extent of co-regulation of vertebrate genes also implies complex logic controlling genes' expression patterns that may allow cells to delicately adjust their expression programs. Although a correlation between the number of transcription factors regulating a gene and its expression level has been noted, it seems clear that future work, both experimental and computational, will need to be performed in order to tease apart the clearly complex logic controlling vertebrate gene expression (Kim et al., 2008). A potential advantage offered cells by a high level of co-regulation of target genes is that gene expression may be more robust to changes in single transcription factors or sequence motifs. This co-regulation may also help explain why genetic perturbation of regulators or perturbation of motifs in promoter sequences does not always result in large expression changes for many target genes.

Studies using advanced sequencing technologies further reveal that Oct4, Sox2 and Nanog often bind to DNA regions in extremely close proximity, within a region of roughly 25 base pairs (Marson et al., submitted). This proximity provokes the hypotheses that these, and perhaps other factors may be forming complexes on DNA to coordinately affect transcription, or that they may be competing for binding to DNA sequences with the data representing a population of cells' binding events rather than co-binding within

individual cells. More studies into the biochemical nature of these binding events are needed in order to test these possibilities.

Initial studies also revealed that the key regulators Oct4, Sox2 and Nanog form a fully connected (complete) interconnected autoregulatory loop (Figure 4A). A motif such as this could act to stabilize the core ES cell transcriptional regulatory network while also allowing for its rapid change upon the correct set of differentiation signals. For example, Chickarmane et al. demonstrate how the circuit formed by Oct4, Sox2 and Nanog could act as a bistable switch controlling ES cell maintenance versus differentiation (Chickarmane et al., 2006). The connection of key regulators to each other could be a general network mechanism permitting the integration of multiple signals to produce a coordinated response in gene expression.

Another important theme uncovered through these studies was the control of other transcriptional regulators by key components. The target genes of Oct4, Sox2 and Nanog were significantly enriched for transcription factors and developmental regulators (Boyer et al., 2005). The control of these secondary regulators allows key transcription factors to indirectly affect a much larger set of genes. This hierarchical network structure has been described in model organisms and may likely prove to be a common vertebrate network architecture (Martinez-Antonio and Collado-Vidas, 2003). This network structure may allow for greater genome sequence variability as genes could be regulated by transcription factors other than the key regulators, with a diversity of binding motifs. This network structure would also allow for rapid large-scale changes in transcription program in response to signals that may only directly target a handful of key regulators.

Closer analysis of the transcriptional regulators bound by Oct4, Sox2 and Nanog revealed that they bound to both active transcription factors and repressed developmental regulators. This observation demonstrated the importance of regulation of silent developmental regulators in ES cells as well as other cell types. Inappropriate expression of these regulators could initiate transcriptional regulatory networks for other cell types, so it is vital to maintain these key regulators of other cell types in the repressed state. This repression by key regulators in ES cells, would therefore act to maintain the ES cell state and ensure that improper differentiation did not occur. This network structure stabilizes cell state and is likely an architecture employed by many vertebrate cell types.

Chromatin Regulators

Several groups have studied important chromatin regulators and marks in ES cells and their combined work reveals important insights into developmental networks (Lee et al., 2006; Guenther et al., 2007; Boyer et al., 2006; Bernstein et al., 2006). These studies focused on general chromatin marks associated with transcription as well as Polycomb and Trithorax Group (PcG and trxG) proteins. These regulators were identified as key regulators of the ES cell network due to an abundance of genetic evidence demonstrating an essential role for these regulators in early development (Breiling et al., 2007; Faust et al., 1998; O'Carroll et al., 2001; Pasini et al., 2004). Studies of their role in the ES cell network have revealed important insights that translate to general themes of vertebrate transcriptional regulatory networks.

Studies of repressive Polycomb and activating Trithorax complexes in ES cells revealed a crucial role for these chromatin modifiers in the ES cell transcriptional regulatory network. The set of silent developmental regulators bound by key transcription factors in ES cells were found to also be targets of both Polycomb and Trithorax complexes. These developmental regulators were therefore described as being bivalently marked by both activating and repressive marks.

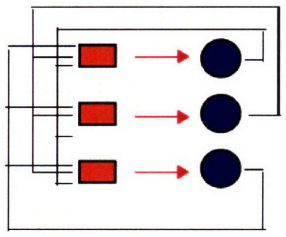
Further chromatin studies revealed that the transcription apparatus was recruited to the promoters of these bivalently marked developmental regulators but that full-length transcript was not produced. This implies that while Pol II is recruited and transcription initiated, that transcription elongation is stalled and therefore no transcript is produced. This stalled transcription suggests that developmental regulators are perhaps poised to be expressed and that their expression is much more dynamic than previously believed.

This feature of developmental regulators may offer several benefits to ES cells. First, it seems to maintain the silence of these regulators so as not to allow induction of transcriptional networks of other cell types governed by them. Second, by maintaining poised Pol II at the promoters of these genes, it allows for the rapid expression and establishment of a new network upon induction of differentiation and removal of repressive Polycomb at a subset of these developmental regulators (Figure 4B). Finally, it may even function to mark these regulators within the cell, serving to especially protect them against mutations or improper expression without permanently shutting down their transcriptional capability. This last hypothesis is supported by the observation that Polycomb binding at these developmental regulators differs from its binding at other target genes by binding to the entire gene region, rather than a punctate binding site within a gene's promoter.

As Polycomb repression of developmental regulators has an important role in maintaining cell state the key question of what directs Polycomb to these regulators emerges. There seem to be two hypotheses that could explain the targeting of Polycomb Group Complexes. The first is that sequence specific transcription factors bind motifs within these genes and recruit PcG. The second is that non-coding RNA transcripts direct the binding of PcG. A role for non-coding RNAs in directing PcG mediated repression is supported by the work of the Chang lab (Rinn et al., 2007). It seems likely that PcG recruitment to developmental regulators may be mediated by a distinct mechanism from its recruitment to other target genes given its unique binding profile at these regulators. An equally pressing question that must be addressed is what selectively

Figure 4

A



B

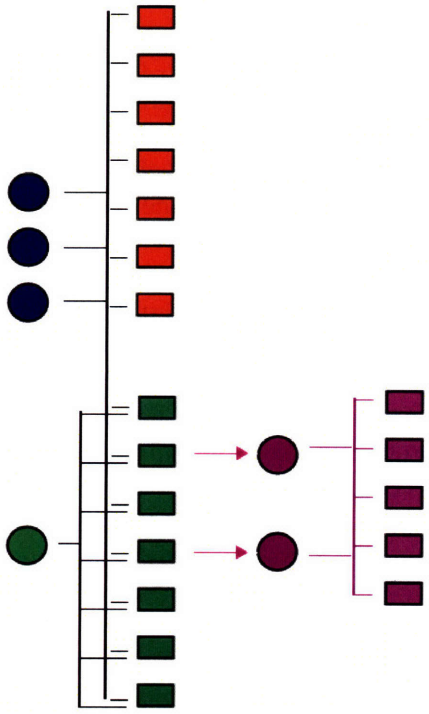


Figure 4 Diagrams of network themes/concepts.
(A) Interconnected autoregulatory loop. Genes (orange rectangles) encoding transcriptional regulators (blue circles) are themselves controlled by the transcriptional regulators, forming an interconnected autoregulatory loop.
(B) Diagram of the transcriptional regulatory network in ES cells and induction of a new network upon differentiation. Key transcription factors (blue circles) regulate active genes (orange rectangles) and along with Polycomb chromatin modifier (green circle) regulate silent developmental regulators (green rectangles) in ES cells. Upon induction of differentiation, a subset of developmental regulators become activated and produce transcriptional regulators (purple circles), which control target genes (purple rectangles) of the transcriptional regulatory network for the new cell type.

removes polycomb from a subset of developmental regulators upon differentiation. A better understanding of this process will likely aid our ability to direct differentiation for therapeutic purposes.

Along with the discovery of paused Pol II at developmental regulators scientists also found that regulation of transcript elongation is a pervasive regulatory mechanism in ES and other cells at a wide set of genes. Data from Guenther et al. 2007 estimates that approximately 30% of genes in ES and other cells have transcription initiation without elongation. Regulation of transcription elongation rather than initiation would allow cells to respond more quickly to environmental stimuli to produce full-length transcripts. The pervasiveness of this level of regulation suggests that the decrease in response time is of great benefit to cells. These discoveries in ES cells also call for further examination of the detailed mechanism of this type of regulation.

Signaling Pathways

A recent pioneering study of the role of a signaling pathway in the ES cell regulatory network has revealed the importance of pathway components in the network and the necessity for their further study (Cole et al., 2008). Genetic evidence demonstrates a clear role for the Wnt signaling pathway throughout development and in ES cells (Logan and Nusse, 2004; Reya and Clevers, 2005). A terminal component of this pathway, Tcf3, was identified as a likely key regulator in ES cells due to its genetic and expression phenotypes (Korinek et al., 1998; Merrill et al., 2004; Pereira et al., 2006). The incorporation of Tcf3 in the ES cell network map reveals a potential means for environmental signals to affect changes in the network for differentiation in vivo or for reprogramming in culture.

Tcf3 was discovered to co-occupy the genome with the key transcriptional regulators Oct4, Sox2, and Nanog. Genomic analyses suggested that these four regulators largely bind to the same promoter regions and even that they bind in extremely close proximity within promoter regions. This co-occupancy of target genes suggests that the Wnt signaling pathway, through Tcf3, intimately connects to the core regulatory network of ES cells and is, in fact, part of this core network. The direct connection of signaling pathways to key target genes throughout the genome offers the cellular benefit of decreased response time to developmental cues. It is likely that this finding will apply to many other key signaling pathways and other cell types but further studies are needed to explore this possibility.

In addition to joining the key transcription factors Oct4, Sox2 and Nanog at target genes, Tcf3 also joins these regulators in the interconnected autoregulatory loop formed by them. Tcf3 both regulates and is regulated by these key transcription factors. In this manner, cells can respond to Wnt signaling through a feedforward loop where the key ES cell regulators as well as their targets are immediately targeted by Tcf3. This network structure would allow for both rapid and stable response to environmental stimuli.

Expression analyses of both Tcf3 and Wnt pathway perturbation revealed that Tcf3 mainly functions to repress target genes under standard conditions in ES cells but that it can activate these genes upon Wnt stimulation. This dynamic positive and negative regulation by Tcf3 allows Wnt signals to quickly adjust the gene expression program of and influence the delicate balance between pluripotency and differentiation in ES cells. The direct connection of factors whose influence on gene expression can be dynamically regulated by environmental signals to the transcriptional regulatory network suggests mechanisms whereby cells undergo changes in gene expression programs and cell state as well as mechanisms whereby scientists may be able to manipulate networks to reprogram cell states.

Non-coding RNAs

Several recent studies have revealed an important role for the miRNA class of non-coding RNAs in ES cells. Genetic studies of the miRNA machinery revealed an important role for this pathway in ES cells and expression studies have identified several miRNAs with ES cell specific expression (Bernstein et al., 2003; Houbaviy et al., 2003; Houbaviy et al., 2005; Kanellopoulou et al., 2005; Mineno et al., 2006; Murchison et al., 2005; Suh et al., 2004; Wang et al., 2007). This evidence argues that miRNAs also play a key role in the regulatory network of ES cells. Studies examining their role are beginning to highlight some important insights.

The miRNA machinery has been shown to be essential for the proper down-regulation upon differentiation of the key ES cell transcription factor Oct4 (Stadler and Ruohola-Baker, 2008). This regulation has been linked to chromatin modifications (Sinkkonen et al., 2008). These studies therefore suggest that miRNAs directly or indirectly regulate both key transcription factors and chromatin modifiers. An important implication of these results is that miRNAs may play a key role in down-regulating key regulators upon induction of a new cell fate in order to allow establishment of a new transcriptional regulatory network. The possibility of miRNAs shutting down key network components upon signals to differentiate suggests that they may also be able to aid reprogramming of cell state in culture.

Work examining the regulation of miRNA genes by other network components has revealed that they are largely regulated in a manner similar to that of protein-coding genes (Marson et al., submitted). A major aspect of this work was the systematic identification of miRNA promoters, which had previously been unmapped. This demonstrated that miRNA promoters can largely be identified by markers of transcription initiation such as the presence of Pol II and transcription-associated chromatin marks. miRNA genes, like protein-coding genes, were also found to be targets of key transcription factors and chromatin modifiers. A subset of silent miRNA genes were also found to be bivalently marked in ES cells, presumably due to their key role in other cell types.

These findings suggest that miRNA genes can and should be incorporated into maps of transcriptional regulatory networks. A key aspect to mapping their role in this network, however, will be the accurate mapping of miRNA target genes either computationally or experimentally. The relatively recent addition of miRNA genes to the ES cell transcriptional regulatory network map highlights both their key role in the network and also the great need for more studies in this area in order to uncover further themes.

Box 2. Insights from Studies in ES Cells

Transcription Factors:
- Have overlapping target gene sets
- Create complex logic to control gene expression
- Bind in close proximity to each other on DNA
- Form interconnected autoregulatory loop
- Control secondary regulators to indirectly regulate many genes
- Regulate silent developmental regulators

Chromatin Modifiers
- Bivalently mark silent developmental regulators
- Developmental regulators have paused transcription
- Regulation of elongation is important mechanism in vertebrate cells

Signaling Pathways
- Control target genes of key transcription factors
- Bind to same DNA region as key transcription factors
- Join key transcription factors in interconnected autoregulatory loop
- Bring dynamic developmental signals directly to core network

miRNAs
- Downregulate key transcription factors upon differentiation
- Are regulated in manner similar to protein-coding genes

Using ES Cell Studies to Guide Future Efforts

The body of work performed in ES cells can be used to guide future efforts to uncover transcriptional regulatory networks. Studies in ES cells have led to great advances in our understanding of these cells and their control. The importance of the themes uncovered highlights the necessity to continue further studies of transcriptional regulatory networks.

Network themes discovered in one cell type often apply to others and so can guide studies in other cell types. Several themes uncovered in the ES cell transcriptional regulatory network also apply to other cell types. For example, the themes of regulatory loops of key regulators, co-occupancy of target genes and silencing of developmental regulators are found in multiple cell-types (Bracken et al., 2006; Kim et al., 2008; Lynn et al., 2007; Odom et al., 2004; Olson, 2006; Rajasekhar and Begemann, 2007). Therefore, a theme uncovered in one network could be examined to determine whether it is found in other networks as well. This could greatly increase the efficiency of mapping key network themes in various cell types.

While only a handful of network components have been mapped in ES cells many important themes have been uncovered and our understanding of the network and its structure has drastically improved with each new component mapped. Initial network maps in ES cells involving only 3 transcription factors identified several fundamental network themes including autoregulation, co-regulation of target genes and regulation of developmental regulators. The universality of these themes suggests that initial studies in other vertebrate cell types that map only a few key transcriptional regulators will also identify important network themes.

The success of studies in ES cells is largely due to the careful selection of network components to map. In order to uncover as much of the core circuitry as possible with each new factor mapped it is critical to select components that are most likely to play key roles in the network. Studies in ES cells have demonstrated that key components can be identified using information from genetic phenotypes, expression profiles and molecular knowledge. Given current experimental limitations, the directed selection of components to map initially in various cell types will be essential.

Studies in ES cells have also demonstrated the importance of incorporating different types of inputs into network diagrams. The addition to the ES cell network map of inputs from chromatin modifiers, signaling pathways, and non-coding RNAs added additional layers of regulation to the map that both deepened and expanded our understanding of it. The addition of chromatin modifiers for example elucidated the transcriptional state of target genes and suggested a mechanism whereby networks could quickly initiate changes to induce differentiation down specific lineages. This work highlights the importance of incorporating a diversity of components into maps of regulatory networks.

Studies in ES cells have also demonstrated the importance of utilizing standardized experimental techniques in order to facilitate the incorporation of different components into a single network map. All experimental techniques have their own biases and noise associated with them. In genomic studies such as the mapping of transcriptional regulatory networks where new data must be layered on top of existing data, it is especially critical to minimize additional noise created by experimental variation in order to fully interpret the results. The maintenance of uniform data

collection when switching to new technologies should be upheld for future network studies.

Concluding Remarks

Although the genome-wide study of transcriptional regulatory networks in vertebrate cells is a young field its importance and application to developmental biology and disease treatment is already evident. As scientists have begun to manipulate regulatory networks to induce certain cell states, a better understanding of these networks and their manipulation is critical. Although the experimental and conceptual tools to allow scientists to efficiently map transcriptional regulatory networks throughout human development and disease do not yet exist, important advances within this field can nonetheless be made with directed studies. It will be important to both continue dissecting piece-by-piece vertebrate transcriptional regulatory networks and to advance technologies to allow for more high-throughput studies. The ultimate goal being to improve our manipulation and modeling of these networks as well as to fit them into the larger context of systems biology which encompasses all cellular components and interactions.

References

Acevedo, L., Iniguez, A., Holster, H., Zhang, X., Green, R., and Farnham, P. (2007). Genome-scale ChIP-chip analysis using 10,000 human cells. Biotechniques *43*: 791-797.

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

Amaral, P., Dinger, M., Mercer, T., and Mattick, J. (2008). The eukaryotic genome as an RNA machine. Science *319*: 1787-1789.

Ambros, V., and Chen, X. (2007). The regulation of genes and genomes by small RNAs. Development *134*: 1635-1641.

Ambrosetti, D., Basilico, C., and Dailey, L. (1997). Synergistic activation of the fibroblast growth factor 4 enhancer by Sox2 and Oct3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites. Mol. Cell Biol. *17*: 6321-6329.

Arce, L., Yokoyama, N., and Waterman, M. (2006). Diversity of LEF/TCF action in development and disease. Oncogene *25*: 7492-7504.

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

Babu, M., Luscombe, N., Aravind, L., Gerstein, M., and Teichmann, S. (2004). Structure and evolution of transcriptional regulatory networks. Curr Op Structural Biol *14*: 283-291.

Balaji, S., Babu, M., and Aravind, L. (2007). Interplay between network structures, regulatory modes and sensing mechanisms of transcription factors in transcriptional regulatory network of E. coli. J. Mol. Biol. *372*: 1108-1122.

Balaji, S., Babu, M., Iyer, L., Luscombe, N., and Aravind, L. (2006). Comprehenseive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. J. Mol. Bio. *360*: 213-227.

Bannister, A., Schneider, R., Myers, F., Thorne, A., Crane-Robinson, C., and Kouzarides, T. (2005). Spatial distribution of di- and tri- methyl lysine 36 of histone H3 at active genes. J. Biol. Chem. *280*: 17732-17736.

Bantignies, F., and Cavalli, G. (2006). Cellular memory and dynamic regulation of polycomb group proteins. Curr. Opin. Cell Biol. *18*: 275-283.

Barabasi, A., and Oltvai, Z. (2004). Network biology: understanding the cell's functional organization. Nat. Rev. Genet. *5*: 101-113.

Bar-Joseph, Z., Gerber, G., Lee, T., Rinaldi, N., Yoo, J., Robert, F., Gordon, D., Fraenkel, E., Jaakola, T., Young, R., and Gifford, D. (2003). Computational discovery of gene modules and regulatory networks. Nat. Biotechnol. *21*: 1337-1342.

Barrandon, C., Spiluttini, B., and Bensaude, O. (2008). Non-coding RNAs regulating the transcriptional machinery. Biol. Cell *100*: 83-95.

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

Bernstein, B., Mikkelsen, T., Xie, X., Kamal, M., Huebert, D., 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.

Blackwood, E., and Kadonga, J. (1998). Going the distance: a current view of enhancer action. Science *281*: 60-63.

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

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

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

Brandenberger, R., Khrebtukova, I., Thies, R., Miura, T., Jingli, C., Puri, R., Vasicek, T., Lebkowski, J., and Rao, M. (2004). MPSS profiling of human embryonic stem cells. BMC Dev. Biol. *4*: 10.

Breiling, S., Sessa, L., and Orlando, V. (2007) Biology of polycomb and trithorax group proteins. Int. Rev. Cytol. *258*: 83-136.

Bulyk, B. (2006). DNA microarray technologies for measuring protein-DNA interactions. Curr. Op. Biotechnol. *17*: 422-430.

Callaerts, P., Halder, G., and Gehring, W. (1997). PAX-6 in development and evolution. Annu. Rev. Neurosci. *20*: 483-532.

Chahrour, M., and Zoghbi, H. (2007). The story of Rett syndrome: from clinic to neurobiology. Neuron *56*: 422-437.

Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., and Smith, A. (2003). Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell *113*, 643–655.

Chickarmane, V., Troein, C., Nuber, U., Sauro, H., and Peterson, C. (2006). Transcriptional dynamics of the embryonic stem cell switch. PloS Comput. Biol. *2*: 1080-1092.

Cockrell, A., and Kafri, T. (2007). Gene delivery by lentivirus vectors. Mol. Biotechnol. *36*: 184-204.

Cole, M., Johnstone, S., Newman, J., Kagey, M., and Young, R. (2008). Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells. Genes & Dev *22*: 746-755.

Couzin, J. (2008). MicroRNAs make big impression in disease after disease. Science *319*: 1782-1784.

Dann, C. (2007) New technology for an old favorite: lentiviral transgenesis and RNAi in rats. Transgenic Res. *16*: 571-580.

Davidson, E., Rast, J., Oliveri, P., Ransick, A., Calestani, C., Yuh, C., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., et al. (2002). A genomic regulatory map for development. Science *295*: 1669-1678.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*: 1306-1311.

Ding, L., and Buchholz, F. (2006). RNAi in embryonic stem cells. Stem Cell Rev. *2*: 11-18.

Faust, C., Lawson, K.A., Schork, N.J., Thiel, B., and Magnuson, T. (1998). The Polycomb-group gene eed is required for normal morphogenetic movements during gastrulation in the mouse embryo. Development 125, 4495–4506.

Gao, F., Foat, B., and Bussemaker, H. (2004). Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. BMC Bioinfor. *5*: 31.

Giaever, G., Chu, A., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, L., Anderson, K., Andre, B., et al. (2002). Functional profiling of the Saccharomyces cerevisiae genome. Nature *418*: 387-391.

Gill, G. (2001). Regulation of the initiation of eukaryotic transcription. Essays Biochem. *37*: 33-43.

Goh, K., Oh, E., Jeong, H., Kahng, B., and Kim, D. (2002). Classification of scale-free networks. PNAS *99*: 12583-12588.

Goodrich, J., and Kugel, J. (2006). Non-coding-RNA regulators and RNA polymerase II transcription. Nat Rev Mol Cell Biol *7*: 612-616.

Gordon, R., 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.

Goutsias, J., and Lee, N. (2007). Computational and experimental approaches for modeling gene regulatory networks. Curr. Pharm. Des. *13*: 1415-1436.

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

Hart, A.H., Hartley, L., Ibrahim, M., and Robb, L. (2004). Identification, cloning and expression analysis of the pluripotency promoting Nanog genes in mouse and human. Dev. Dyn. *230*, 187–198.

Hawkins, P., and Morris, K. (2008). RNA and transcriptional modulation of gene expression. Cell Cycle *7*: 602-607.

Hirose, S. (2007). Crucial roles for chromatin dynamics in cellular memory. J Biochem *141*: 615-619.

Hirose, Y., and Ohkuma, Y. (2007). Phosphorylation of the C-terminal domain of RNA polymerase II plays central roles in the integrated events of eukaryotic gene expression. J. Biochem. *141*: 601-608.

Hobert, O. (2008). Gene regulation by transcription factors and microRNAs. Science *319*: 1785-1786.

Hombria, J., and Lovegrove, B. (2003). Beyond homeosis-HOX function in morphogenesis and organogenesis. Differentiation *71*: 461-476.

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

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.

Kemphues, K. (2005). Essential genes. WormBook 1-7.

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

Kloosterman, W., and Plasterk, R. (2006). The diverse functions of microRNAs in animal development and disease. Dev. Cell *11*: 441-450.

Kojima, H., Fujimiya, M., Matsumura, K., Younan, P., Imaeda, H., Maeda, M., and Chan, L. (2003). NeuroD-betacellulin gene therapy induces islet neogenesis in the liver and reverses diabetes in mice. Nat. Med. *9*: 596-603.

Korinek, V., Barker, N., Willert, K., Molenaar, M., Roose, J., Wagenaar, G., Markman, M., Lamers, W., Destree, O., and Clevers, H. (1998). Two members of the Tcf family implicated in Wnt/B-catenin signaling during embryogenesis in the mouse. Mol. Cell. Biol. *18*: 1248-1256.

Kornberg, R. (2007). The molecular basis of eukaryotic transcription. PNAS *104*: 12955-12961.

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

Lee, J.H., Hart, S.R., and Skalnik, D.G. (2004). Histone deacetylase activity is required for embryonic stem cell differentiation. Genesis *38*, 32–38.

Lee, T., Rinaldi, N., Robert, F., Odom, D., Bar-Joseph, Z., Gerber, G., Hannett, N., Harbison, C., Thompson, C., Simon, I., et al. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science *298*: 799-804.

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

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

Lee, J., and Shilatifard, A. (2007). A site to remember: H3K36 methylation a mark for histone deacetylation. Mutat. Res. *618:* 130-134.

Lee, T., and Young, R. (2000). Transcription of eukaryotic protein-coding genes. Annu Rev Genet *34*: 77-137.

Lemon, B., and Tijan, R. (2000). Orchestrated response: a symphony of transcription factors for gene control. Genes & Dev *14*: 2551-2569.

Levine, M., and Davidson, E. (2005). Proc. Natl. Acad. Sci. *102*: 4936-4942.

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

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

Li, H., Xuan, J., Wang, Y., and Zhan, M. (2008). Inferring regulatory networks. Front. Biosci. *13*: 263-275.

Logan, C., and Nusse, R. (2004). The Wnt signaling pathway in development and disease. Annu. Rev. Cell Dev. Biol. *20*: 781-810.

Loh, Y., Wu, Q., Chew, J., Vega, V., 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.

Lynn, F., Smith, S., Wilson, M., Yang, K., Nekrep, N., and German, M. (2007). Sox9 coordinates a transcriptional network in pancreatic progenitor cells. Proc. Natl. Acad. Sci. *104*: 10500-10505.

Marson, A., Levine, S., Cole, M., Frampton, G., Brambrink, T., Guenther, M., Johnston, W., Wernig, M., Volkert, T., Bartel, D., et al. (submitted). Connecting microRNA genes to the core transcriptional regulatory circuitry of ES cells.

Martinez-Antonio, A., and Collado-Vides, J. (2003). Identifying global regulators in transcriptional regulatory networks in bacteria. Curr. Opin. Microbiol. *6*: 482-489.

McManus, M., and Sharp, P. (2002). Nat Rev Gen *3*: 737-747.

Merrill, B., Pasolli, H., Polak, L., Rendl, M., Garcia-Garcia, M., Anderson, K., and Fuchs, E. (2004). Tcf3: a transcriptional regulator of axis induction in the early embryo. Development *131*: 263-274.

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.

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

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.

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

Nimwegen, E. (2003). Scaling laws in the functional content of genomes. TRENDS Gen. *19*: 479-484.

Niwa, H., Miyazaki, J., and Smith, A. (2000). Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. Nat. Genet. *24*: 372-376.

O'Carroll, D., Erhardt, S., Pagani, M., Barton, S.C., Surani, M.A., and Jenuwein, T. (2001). The polycomb-group gene Ezh2 is required for early mouse development. Mol. Cell. Biol. 21, 4330–4336.

Odom, D., Zizlsperger, N., Gordon, D., Bell, G., Rinaldi, N., Murray, H., Volkert, T., Schreiber, J., Rolfe, P., Gifford, D., et al. (2004). Control of pancreas and liver gene expression by HNF transcription factors. Science *303*: 1378-1381.

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

Olson, E. (2006). Gene regulatory networks in the evolution and development of the heart. Science *313*: 1922-1927.

Orphanides, G., and Reinberg, D. (2002). A unified theory of gene expression. Cell *108*: 439-451.

Park, I., Zhao, R., West, J., Yabuuchi, A., Huo, H., Ince, T., Lerou, P., Lensch, M., and Daley, G. (2008). Reprogramming of human somatic cells to pluripotency with defined factors. Nature *451*: 141-146.

Pasini, D., Bracken, A., Jensen, M., Denchi, E., and Helin, K. (2004). Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity. EMBO J. *23*: 4061-4071.

Pereira, L., Yi, F., and Merrill, B. (2006). Repression of Nanog gene transcription by Tcf3 limits embryonic stem cell self-renewal. Mol. Cell. Biol. *26*: 7479-7491.

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

Rajasekhar, V., and Begemann, M. (2007). Concise review: roles of polycomb group proteins in development and disease: a stem cell perspective. Stem Cells *25*: 2498-2510.

Reya, T., and Clevers, H. (2005). Wnt signaling in stem cells and cancer. Nature *434*: 843-850.

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

Ryffel, G. (2001). Mutations in the human genes encoding the transcription factors of the hepatocyte nuclear factor (HNF)1 and HNF4 families: functional and pathological consequences. J Mol Endocrinol *27*: 11-29.

Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Peralta-Gil, M., Penolaza-Spinola, M., Matrinez-Antonio, A., Karp, P., and Collado-Vides, J. (2006). The comprehensive updated regulatory network of Escherichia coli K-12. BMC Bioinformatics *7*:5.

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

Sato, N., Sanjuan, I., Heke, M., Uchida, M., Naef, F., and Brivanlou, A. (2003). Molecular signature of human embryonic stem cells and its comparison with the mouse. Dev. Biol. *260*: 404-413.

Schöler, H.R., Dressler, G.R., Balling, R., Rohdewohld, H., and Gruss, P. (1990). Oct-4: a germline-specific transcription factor mapping to the mouse t-complex. EMBO J. *9*, 2185–2195.

Sims, R., Belotserkovskaya, R., and Reinberg, D. (2004). Elongation by RNA polymerase II: the short and long of it. Genes & Dev. *18*: 2437-2468.

Sinkkonen, L., Hugenschmidt, T., Berninger, P., Gaidatzis, D., Mohn, F., Artus-Revel, C., Zavolan, M., Svoboda, P., 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.

Stadler, B., and Ruohola-Baker, H. (2008). Small RNAs: keeping stem cells in line. Cell *132*: 563-566.

Struhl, K. (2005). Transcriptional activation: mediator can act after preinitiation complex formation. Mol. Cell *17*: 752-754.

Su, A., Wiltshire, T., Batalov, S., Lapp, H., Ching, K., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl. Acad. Sci. USA *101*:6062-6067.

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.

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

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

Thieffry, D., Huerta, A., Perez-Rueda, E., and Collado-Vides, J. (1998). Bioessays *20*: 433-440.

Tupler, R., Giovanni, P., and Green, M. (2001). Expressing the human genome. Nature *409*: 832-833.

Turner, B. (2002). Cellular memory and the histone code. Cell *111*: 285-291.

Villard, J. (2004). Transcription regulation and human diseases. Swiss Med. Wkly. *134*: 571-579.

Viswanathan, S., Daley, G., and Gregory, R. (2008). Selective blockade of microRNA processing by Lin28. Science *320*: 97-100.

Vogelstein, B., Lane, D., and Levine, A. (2000). Surfing the p53 network. Nature *408*: 307-310.

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

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.

Weintraub, H., Tapscott, S., Davis, R., Thayer, M., Adam, M., Lassar, A., and Miller, A. (1989). Activation of muscle-specific genes in pigment, nerve, fat, liver and fibroblast cell lines by forced expression of MyoD. Proc. Natl. Acad. Sci. *86*: 5434-5438.

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

West, A., and Fraser, P. (2005). Remote control of gene transcription. Human Mol Gen *14*: 101-111.

Winzeler, E., Shoemaker, D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J., Bussey, H., et al. (1999). Funcitonal characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science *285*: 901-906.

Wu, L., and Belasco, J. (2008). Let me count the ways: mechanisms of gene regulation by miRNAs and siRNAs. Mol. Cell *29*: 1-7.

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

Chapter 2

Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells

Published as: Laurie A. Boyer, Tong Ihn Lee, Megan F. Cole, Sarah E. Johnstone, Stuart S. Levine, Jacob P. Zucker, Mathew G. Guenther, Roshan M. Kumar, Heather L. Murray, Richard G. Jenner, David K. Gifford, Douglas A. Melton, Rudolf Jaenisch and Richard A. Young (2005). "Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells." Cell 122: 947-956.

My contribution to this project

The effort to profile key transcription factors in human embryonic stem cells was led by Laurie Boyer. I was responsible for leading the computational analyses for this project, working closely with fellow computational labmate Stuart Levine. We tweaked the error model for use with genome-wide human binding data and designed the gene-calling parameters to use with this new type of data. I also worked with Sarah Johnstone to construct most of the manuscript figures. I also made significant contributions to the conceptual content of the manuscript by uncovering the high overlap between target gene sets and the core interconnected autoregulatory loop.

Summary

The transcription factors OCT4, SOX2, and NANOG have essential roles in early development and are required for the propagation of undifferentiated embryonic stem (ES) cells in culture. To gain insights into transcriptional regulation of human ES cells, we have identified OCT4, SOX2, and NANOG target genes using genome-scale location analysis. We found, surprisingly, that OCT4, SOX2, and NANOG co-occupy a substantial portion of their target genes. These target genes frequently encode transcription factors, many of which are developmentally important homeodomain proteins. Our data also indicate that OCT4, SOX2, and NANOG collaborate to form regulatory circuitry consisting of autoregulatory and feedforward loops. These results provide new insights into the transcriptional regulation of stem cells and reveal how OCT4, SOX2, and NANOG contribute to pluripotency and self-renewal.

Introduction

Mammalian development requires the specification of over 200 unique cell types from a single totipotent cell. Embryonic stem (ES) cells are derived from the inner cell mass (ICM) of the developing blastocyst and can be propagated in culture in an undifferentiated state while maintaining the capacity to generate any cell type in the body. The recent derivation of human ES cells provides a unique opportunity to study early development and is thought to hold great promise for regenerative medicine (Pera and Trounson, 2004; Reubinoff et al., 2000; Thomson et al., 1998). An understanding of the transcriptional regulatory circuitry that is responsible for pluripotency and self-renewal in human ES cells is fundamental to understanding human development and realizing the therapeutic potential of these cells.

Homeodomain transcription factors are evolutionarily conserved and play key roles in cell-fate specification in many organisms (Hombria and Lovegrove, 2003). Two such factors, OCT4/POU5F1 and NANOG, are essential regulators of early development and ES cell identity (Chambers et al., 2003; Hay et al., 2004; Matin et al., 2004; Mitsui et al., 2003; Nichols et al., 1998; Zaehres et al., 2005). Several genetic studies in mouse suggest that these regulators have distinct roles but may function in related pathways to maintain the developmental potential of these cells (Chambers, 2004). For example, disruption of *OCT4* or *NANOG* results in the inappropriate differentiation of ICM and ES cells to trophectoderm and extra-embryonic endoderm, respectively (Chambers et al., 2003; Mitsui et al., 2003; Nichols et al., 1998). However, overexpression of *OCT4* in ES cells leads to a phenotype that is similar to loss of NANOG function (Chambers et al., 2003; Mitsui et al., 2003; Nichols et al., 1998; Niwa et al., 2000). Knowledge of the set of genes regulated by these two transcription factors might reveal why manipulation of OCT4 and NANOG results in these phenotypic consequences.

OCT4 is known to interact with other transcription factors to activate and repress gene expression in mouse ES cells (Pesce and Schöler, 2001). For example, OCT4, a member of the POU (PIT/OCT/UNC) class of homeodomain proteins, can heterodimerize with the HMG-box transcription factor, SOX2, to affect the expression of several genes in mouse ES cells (Botquin et al., 1998; Nishimoto et al., 1999; Yuan et al., 1995). The cooperative interaction of POU homeodomain and HMG factors is thought to be a fundamental mechanism for the developmental control of gene expression (Dailey and Basilico, 2001). The extent to which ES cell gene regulation is accomplished by OCT4 through an OCT4/SOX2 complex and whether NANOG has a role in this process are unknown.

OCT4, SOX2, and NANOG are thought to be central to the transcriptional regulatory hierarchy that specifies ES cell identity because of their unique expression patterns and their essential roles during early development (Avilion et al., 2003; Chambers et al., 2003; Hart et al., 2004; Lee et al., 2004; Mitsui et al., 2003; Nichols et al., 1998; Schöler et al., 1990). Studies in a broad range of eukaryotes have shown that transcriptional regulators that have key roles in cellular processes frequently regulate other regulators associated with that process (Guenther et al., 2005; Lee et al., 2002; Odom et al., 2004). It is likely that the key stem cell regulators bind and regulate genes encoding other transcriptional regulators, which in turn determine the developmental potential of these cells, but we currently lack substantial knowledge of the regulatory

circuitry of ES cells and other vertebrate cells.

To further our understanding of the means by which OCT4, SOX2, and NANOG control the pluripotency and self-renewal of human ES cells, we have used genomescale location analysis (chromatin immunoprecipitation coupled with DNA microarrays) to identify the target genes of all three regulators in vivo. The results reveal that OCT4, SOX2, and NANOG co-occupy the promoters of a large population of genes, that many of these target genes encode developmentally important homeodomain transcription factors, and that these regulators contribute to specialized regulatory circuits in ES cells.

Results and Discussion

OCT4 Promoter Occupancy in Human ES Cells

DNA sequences occupied by OCT4 in human H9 ES cells (NIH code WA09; Supplemental Data) were identified in a replicate set of experiments using chromatin immunoprecipitation (ChIP) combined with DNA microarrays (Figure 1A and Supplemental Data). For this purpose, DNA microarrays were designed that contain 60-mer oligonucleotide probes covering the region from −8 kb to +2 kb relative to the transcript start sites for 17,917 annotated human genes. Although some transcription factors are known to regulate genes from distances greater than 8 kb, 98% of known binding sites for human transcription factors occur within 8 kb of target genes (Figure S1). The sites occupied by OCT4 were identified as peaks of ChIP-enriched DNA that span closely neighboring probes (Figure 1B). OCT4 was associated with 623 (3%) of the promoter regions for known protein-coding genes and 5 (3%) of the promoters for known miRNA genes in human ES cells (Table S2).

Two lines of evidence suggested that this protein- DNA interaction dataset is of high quality. First, the genes occupied by OCT4 in our analysis included many previously identified or supposed target genes in mouse ES cells or genes whose transcripts are highly enriched in ES cells, including *OCT4*, *SOX2*, *NANOG*, *LEFTY2/ EBAF*, *CDX2*, *HAND1*, *DPPA4*, *GJA1/CONNEXIN43*, *FOXO1A*, *CRIPTO/TDGF1*, and *ZIC3* (Abeyta et al., 2004; Brandenberger et al., 2004; Catena et al., 2004; Kuroda et al., 2005; Niwa, 2001; Okumura-Nakanishi et al., 2005; Rodda et al., 2005; Sato et al., 2003; Wei et al., 2005) (Table S2). Second, we have used improved protocols and DNA microarray technology in these experiments (Supplemental Data) that should reduce false positive rates relative to those obtained in previous genome-scale experiments (Odom et al., 2004). By using this new technology with yeast transcription factors, where considerable prior knowledge of transcription factor binding sites has been established, we estimated that this platform has a false positive rate of <1% and a false negative rate of 20% (Supplemental Data).

OCT4, *SOX2*, and *NANOG* Co-Occupy Many Target Genes

We next identified, with location analysis, protein-coding and miRNA genes targeted by the stem cell regulators SOX2 and NANOG. SOX2 and NANOG were found associated with 1271 (7%) and 1687 (9%), respectively, of the promoter regions for known protein-coding genes in human ES cells (Tables S2–S4). It was immediately evident that many of the target genes were shared by OCT4, SOX2, and NANOG (Figure 2A). Examples of protein-coding genes that are co-occupied by the three regulators are shown in Figure 2B (Table S5). Control experiments showed that the set of promoters bound by the cell-cycle transcription factor E2F4 in these human ES cells did not overlap substantially with those bound by the three stem cell regulators (Tables S2 and S6). We found that OCT4, SOX2, and NANOG together occupy at least 353 genes in human ES cells.

Previous studies in murine ES cells have shown that SOX2 and OCT4 can interact to synergistically activate transcription of target genes and that this activity is dependent upon the juxtaposition of OCT4 and SOX2 binding sites (Ambrosetti et al., 1997; Remenyi et al., 2004). Our results revealed that approximately half of the promoter

regions occupied by OCT4 were also bound by SOX2 in human ES cells (Figure 2A; Table S2). It was surprising, however, to find that >90% of promoter regions bound by both OCT4 and SOX2 were also occupied by NANOG. Furthermore, we found that OCT4, SOX2, and NANOG binding sites occurred in close proximity at nearly all of the genes that they cooccupied (Figure 2C). These data suggest that OCT4, SOX2, and NANOG function together to regulate a significant proportion of their target genes in human ES cells.

A class of small noncoding RNAs known as micro- RNAs (miRNA) play vital roles in gene regulation, and recent studies indicate that more than a third of mammalian protein-coding genes are conserved miRNA targets (Bartel, 2004; Lewis et al., 2005). ES cells lacking the machinery that processes miRNA transcripts are unable to differentiate (Kanellopoulou et al., 2005). Moreover, recent evidence indicates that microRNAs play an important role in organismal development through regulation of gene expression (Pasquinelli et al., 2005). OCT4, SOX2, and NANOG were found associated with 14 miRNA genes and co-occupied the promoters of at least two miRNA genes, *mir-137* and *mir-301* (Table 1). Our results suggest that miRNA genes are likely regulated by OCT4, SOX2, and NANOG in human ES cells and are important components of the transcriptional regulatory circuitry in these cells.

ES Cell Transcription Factors Occupy Active and Inactive Genes

OCT4 and SOX2 are known to be involved in both gene activation and repression in vivo (Botquin et al., 1998; Nishimoto et al., 1999; Yuan et al., 1995), so we sought to identify the transcriptional state of genes occupied by the stem cell regulators. To this end, the set of genes bound by OCT4, SOX2, and NANOG were compared to gene expression datasets generated from multiple ES cell lines (Abeyta et al., 2004; Brandenberger et al., 2004; Sato et al., 2003; Wei et al., 2005) to identify transcriptionally active and inactive genes (Table S2). The results showed that one or more of the stem cell transcription factors occupied 1303 actively transcribed genes and 957 inactive genes.

The importance of OCT4, SOX2, and NANOG for early development and ES cell identity led us to focus additional analyses on the set of 353 genes that are cooccupied by these regulators in human ES cells (Table S5). We first identified transcriptionally active genes. Transcripts were consistently detected in ES cells for approximately half of the genes co-bound by OCT4, SOX2, and NANOG. Among these active genes, several encoding transcription factors (e.g., OCT4, SOX2, NANOG, STAT3, ZIC3) and components of the Tgf-β (e.g., TDGF1, LEFTY2/EBAF) and Wnt (e.g., DKK1, FRAT2) signaling pathways were notable targets. Recent studies have shown that Tgf-β and Wnt signaling play a role in pluripotency and self-renewal in both mouse and human ES cells (James et al., 2005; Sato et al., 2004). These observations suggest that OCT4, SOX2, and NANOG promote pluripotency and selfrenewal through positive regulation of their own genes and genes encoding components of these key signaling pathways.

Among transcriptionally inactive genes co-occupied by OCT4, SOX2, and NANOG, we noted a striking enrichment for transcription factor genes ($p < 10-18$; Table S7), many of which have been implicated in developmental processes. These included genes that specify transcription factors important for differentiation into extra-embryonic, endodermal, mesodermal, and ectodermal lineages (e.g., *ESX1l, HOXB1, MEIS1, PAX6, LHX5, LBX1, MYF5, ONECUT1*) (Table S5). Moreover, nearly half of the transcription

Figure 1

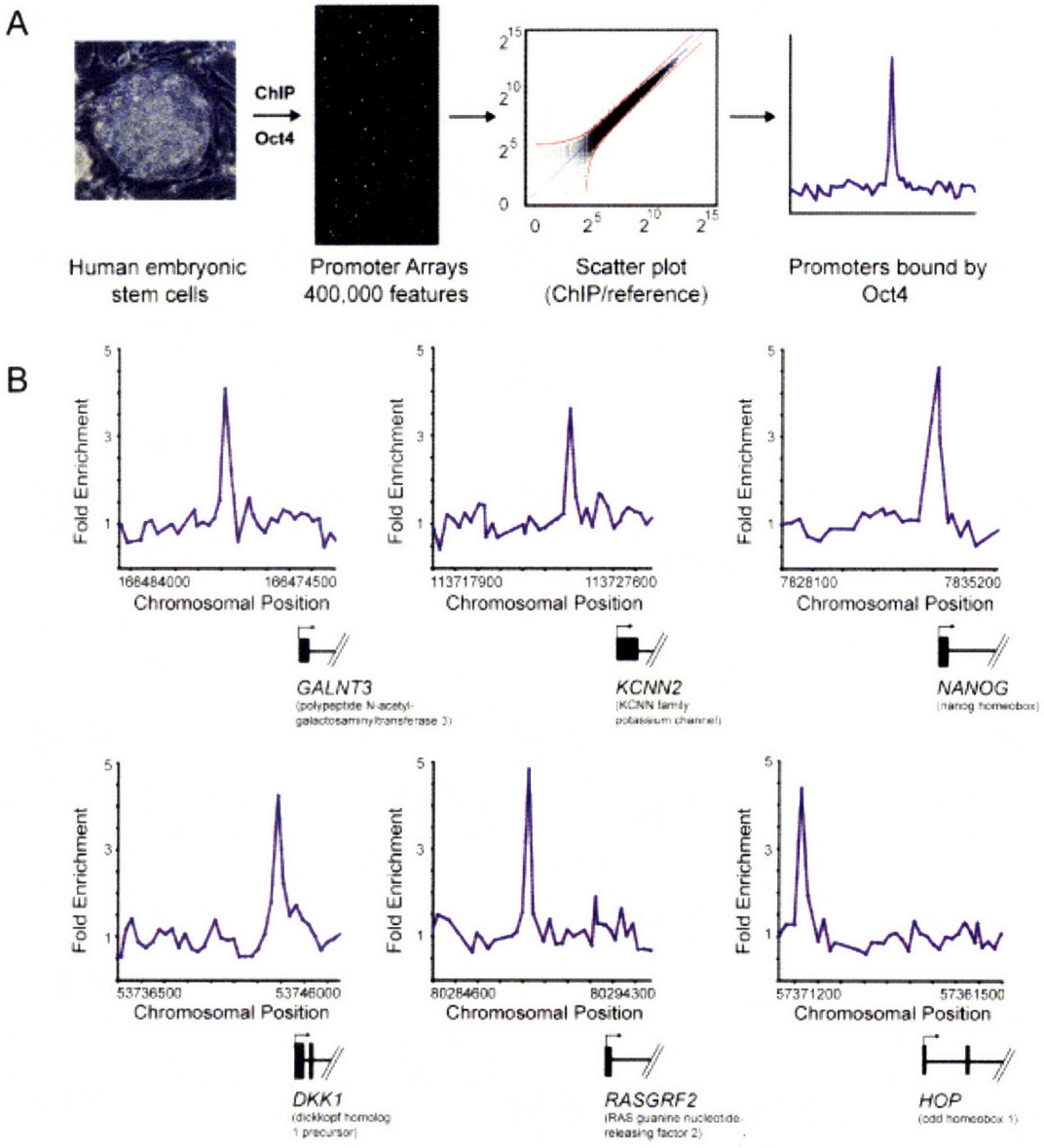

A
ChIP
Oct4
Human embryonic stem cells
Promoter Arrays 400,000 features
Scatter plot (ChIP/reference)
Promoters bound by Oct4
B
Fold Enrichment
Chromosomal Position
GALNT3
KCNN2
NANOG
DKK1
RASGRF2
HOP

Figure 1. Genome-Wide Location Analysis in Human Embryonic Stem Cells
(A) DNA segments bound by transcriptional regulators were identified using chromatin immunoprecipitation (ChIP) and identified with DNA microarrays containing 60-mer oligonucleotide probes covering the region from –8 kb to +2 kb for 17,917 annotated transcription start sites for human genes. ES cell growth and quality control, ChIP protocol, DNA microarray probe design, and data analysis methods are described in detail in Experimental Procedures and Supplemental Data.
(B) Examples of OCT4 bound regions. Plots display unprocessed ChIP-enrichment ratios for all probes within a genomic region. Genes are shown to scale below plots (exons and introns are represented by thick vertical and horizontal lines, respectively), and the genomic region represented is indicated beneath the plot. The transcription start site and transcript direction are denoted by arrows.

Figure 2

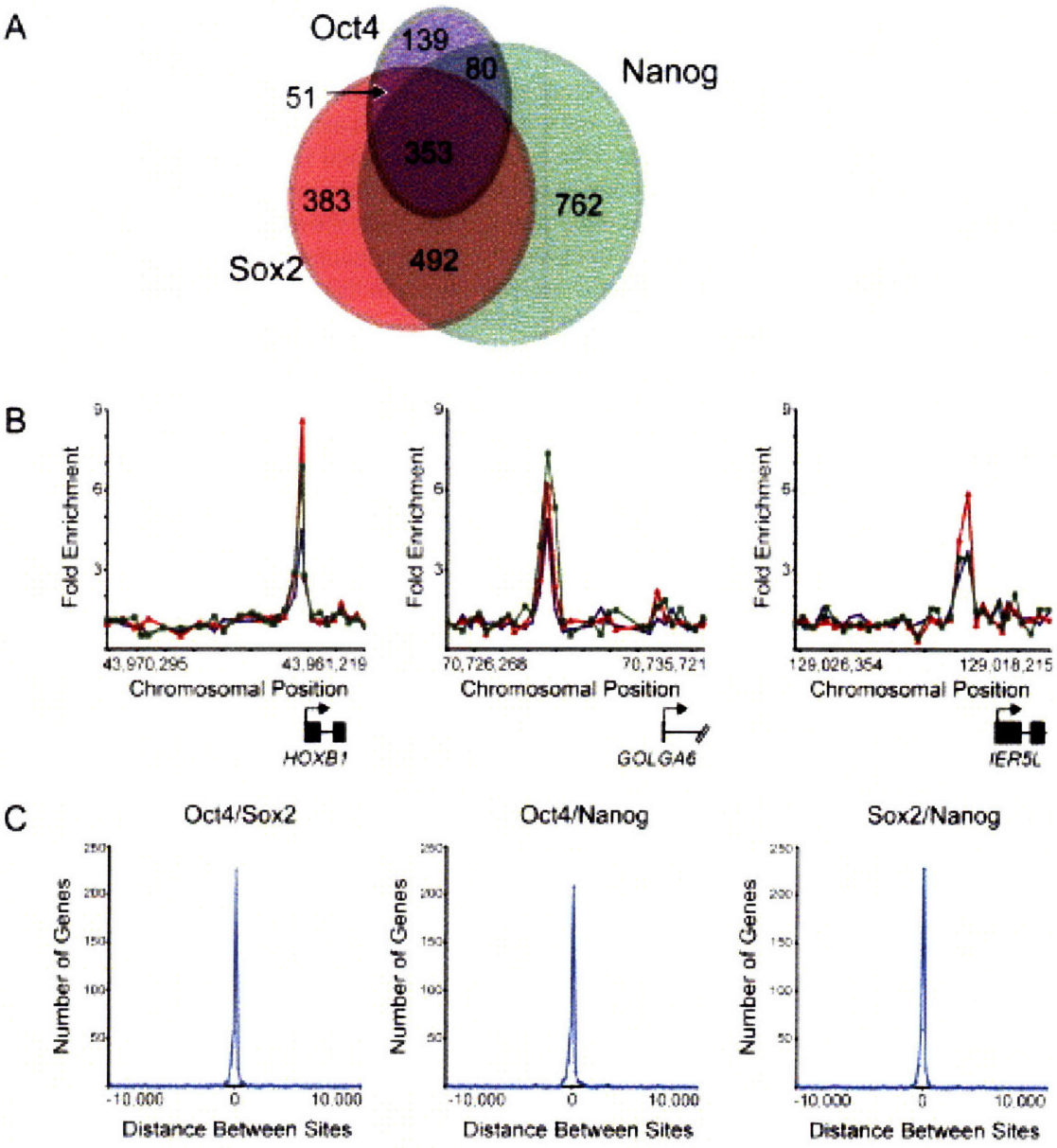

A
Oct4
Nanog
Sox2
139
80
51
353
383
762
492
B
Fold Enrichment
43,970,295
43,961,219
Chromosomal Position
HOXB1
70,726,268
70,735,721
GOLGA6
129,026,354
129,018,215
IER5L
C
Oct4/Sox2
Oct4/Nanog
Sox2/Nanog
Number of Genes
-10,000
0
10,000
Distance Between Sites

Figure 2. OCT4, SOX2, and NANOG Target Genes in Human ES Cells
(A) Venn diagram representing the overlap of OCT4, SOX2, and NANOG promoter bound regions.
(B) Representative examples of protein-coding genes co-occupied by OCT4, SOX2, and NANOG. Plots display unprocessed ChIPenrichment ratios for all probes within a genomic region. Genes are shown to scale relative to their chromosomal position. Exons and introns are represented by thick vertical and horizontal lines, respectively. The start and direction of transcription are denoted by arrows. Green, red, and purple lines represent NANOG, SOX2, and OCT4 bound regions, respectively.
(C) OCT4, SOX2, and NANOG bind in close proximity. The distances between the midpoint of bound regions for pairs of transcription factors was calculated for the 353 regions bound by all three transcription factors. Negative and positive values indicate whether the first factor is upstream or downstream of the second factor in relation to the gene. The frequency of different distances between the bound regions is plotted as a histogram.

Table 1. miRNA Loci near OCT4, SOX2, and NANOG Bound Regions

	Transcription Factor		
miRNA	OCT4	SOX2	NANOG
mir-7-1		+	
mir-10a	+		
mir-22		+	+
mir-32		+	+
mir-128a			+
mir-135b		+	+
mir-137	+	+	+
mir-196a-1			+
mir-196b	+		
mir-204		+	+
mir-205		+	+
mir-301	+	+	+
mir-361			+
mir-448	+		

Proximal binding of OCT4, SOX2, and NANOG to miRNAs from the RFAM database. Transcription factors bound are indicated by a "+."

factor genes that were bound by the three regulators and transcriptionally inactive encoded developmentally important homeodomain proteins (Table 2). These results demonstrate that OCT4, SOX2, and NANOG occupy a set of repressed genes that are key to developmental processes.

To determine which of the OCT4, SOX2, and NANOG bound genes were preferentially expressed in ES cells, we compared expression datasets (Abeyta et al., 2004; Sato et al., 2003) from ES cells and a compendium of differentiated tissues and cell types (Su et al., 2004) (Figure 3; Supplemental Data). It was notable that *DPPA4*, *TDGF1*, *OCT4*, *NANOG*, and *LEFTY2* were at the top of the rank order list of genes that are bound and preferentially expressed in ES cells (Figure 3A). All five of these genes have been implicated in pluripotency (James et al., 2005; Mitsui et al., 2003; Chambers et al., 2003; Nichols et al., 1998; Bortvin et al., 2003). Moreover, several genes that encode developmentally important homeodomain proteins such as *DLX5*, *HOXB1*, *LHX5*, *TITF1*, *LBX1*, and *HOP* were at the bottom of this list, indicating that they are preferentially repressed in ES cells.

The observation that OCT4, SOX2, and NANOG bound to transcriptionally active genes that have roles in pluripotency and transcriptionally inactive genes that promote development suggests that these binding events are regulatory. Two additional lines of evidence indicated that many of the binding events identified in this study contribute to regulation of their target genes. First, some of the genes identified here (e.g., *OCT4*, *SOX2*, and *NANOG*) were previously shown to be regulated by OCT4 and SOX2 in mouse ES cells (Catena et al., 2004; Kuroda et al., 2005, Okumura-Nakanishi et al., 2005; Rodda et al., 2005). Second, we further explored the hypothesis that bound genes are regulated by these transcription factors by taking advantage of the fact that *OCT4* and *NANOG* are expressed in ES cells, but their expression is rapidly downregulated upon differentiation. We compared the expression of OCT4, SOX2, and NANOG occupied genes in human ES cells with expression patterns in 79 differentiated cell types (Su et al., 2004) (Supplemental Data) and focused the analysis on transcription factor genes because these were the dominant functional class targeted by the ES cell regulators (Figure 3B). We expected that for any set of genes, there would be a characteristic change in expression levels between ES cells and differentiated cells. If OCT4, SOX2, and NANOG do not regulate the genes they occupy, then these genes should have the same general expression profile as the control population. We found, however, a significant shift in the distribution of expression changes for genes occupied by OCT4, SOX2, and NANOG (p value < 0.001). Taken together, these data support the model that OCT4, SOX2, and NANOG functionally regulate the genes they occupy and suggest that loss of these regulators upon differentiation results in increased expression of genes necessary for development and reduced expression of a set of genes required for the maintenance of stem cell identity.

Our results suggest that OCT4, SOX2, and NANOG contribute to pluripotency and self-renewal by activating their own genes and genes encoding components of key signaling pathways and by repressing genes that are key to developmental processes. It is presently unclear how the three key regulators can activate some genes and repress others. It is likely that the activity of these key transcription factors is further controlled by additional cofactors, by the precise levels of OCT4, SOX2, and NANOG, and by posttranslational modifications.

Figure 3

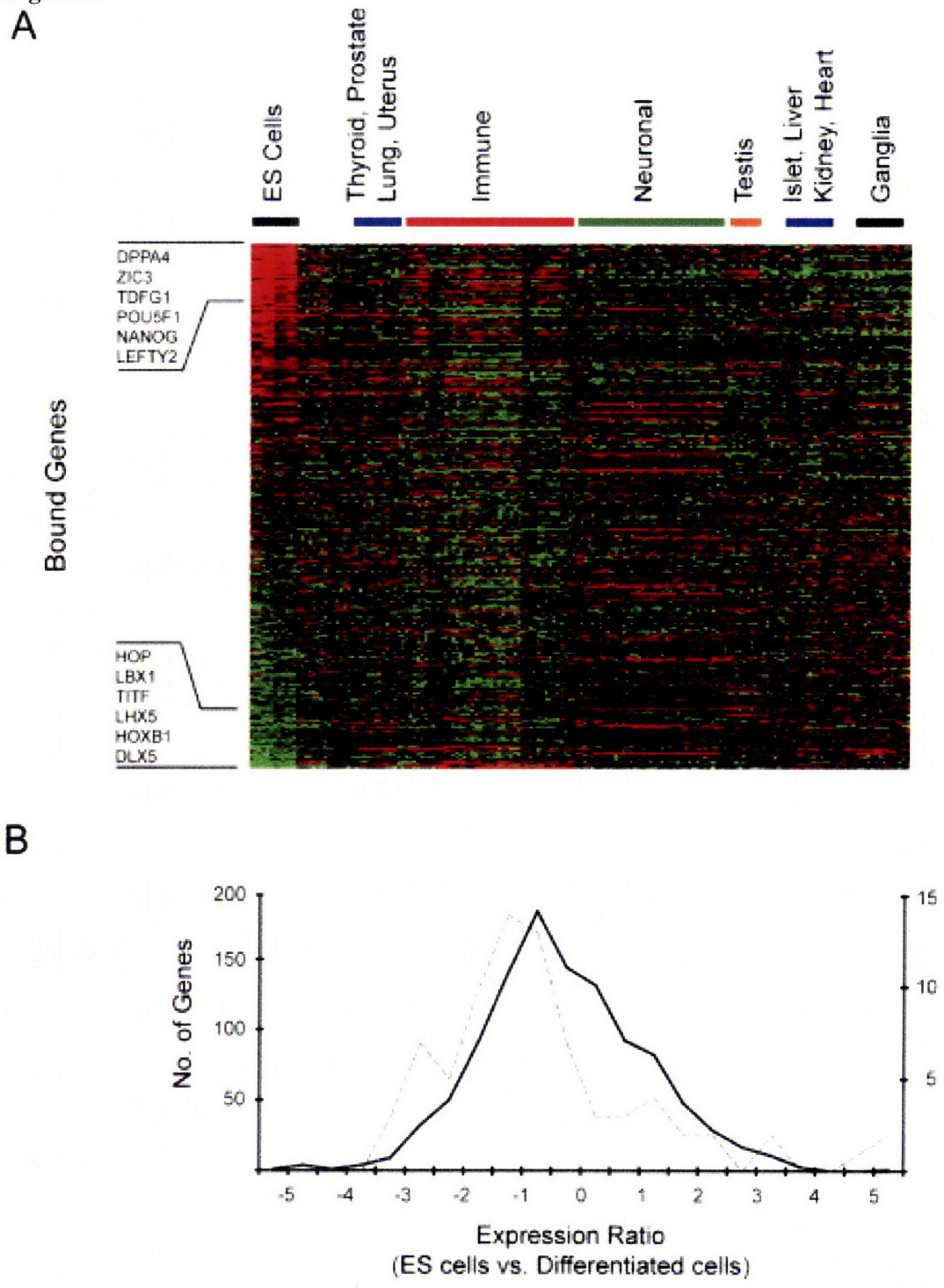

A
ES Cells
Thyroid, Prostate
Lung, Uterus
Immune
Neuronal
Testis
Islet, Liver
Kidney, Heart
Ganglia
Bound Genes
DPPA4
ZIC3
TDFG1
POU5F1
NANOG
LEFTY2
HOP
LBX1
TITF
LHX5
HOXB1
DLX5
B
No. of Genes
200
150
100
50
15
10
5
-5
-4
-3
-2
-1
0
1
2
3
4
5
Expression Ratio
(ES cells vs. Differentiated cells)

Figure 3. Expression of OCT4, SOX2, and NANOG Co-Occupied Genes
(A) Affymetrix expression data for ES cells were compared to a compendium of expression data from 158 experiments representing 79 other differentiated tissues and cell types (Supplemental data). Ratios were generated by comparing gene expression in ES cells to the median level of gene expression across all datasets for each individual gene. Genes were ordered by relative expression in ES cells, and the results were clustered by expression experiment using hierarchical clustering. Each gene is represented as a separate row and individual expression experiments are in separate columns. Red indicates higher expression in ES cells relative to differentiated cells. Green indicates lower expression in ES cells relative to differentiated cells. Examples of bound genes that are at the top and bottom of the rank order list are shown.
(B) Relative levels of gene expression in H9 ES cells compared to differentiated cells were generated and converted to log2 ratios. The distribution of these fold changes was calculated to derive a profile for different sets of genes. Data are shown for the distribution of expression changes between H9 ES cells and differentiated tissues for transcription factor genes that are not occupied by OCT4, SOX2, and NANOG (solid black line) and transcription factor genes occupied by all three (dotted line). The change in relative expression is indicated on the *x* axis, and the numbers of genes in each bin are indicated on the y axes (left axis for unoccupied genes, right axis for occupied genes). The shift in distribution of expression changes for genes occupied by OCT4, SOX2, and NANOG is significant (p value < 0.001 using a two-sampled Kolmogorov-Smirnov test), consistent with the model that OCT4, SOX2, and NANOG are contributing to the regulation of these genes.

Table 2

Table 2. Examples of Inactive Homeodomain Genes Co-occupied by OCT4, SOX2, and NANOG

Gene Symbol	Entrez Gene ID	Gene Name
ATBF1	463	AT binding transcription factor 1
DLX1	1745	distal-less homeobox 1
DLX4	1748	distal-less homeobox 4
DLX5	1749	distal-less homeobox 5
EN1	2019	engrailed homolog 1
ESX1L	80712	extraembryonic, spermatogenesis, homeobox 1-like
GBX2	2637	gastrulation brain homeobox 2
GSC	145258	goosecoid
HOP	84525	homeodomain-only protein
HOXB1	3211	homeobox B1
HOXB3	3213	homeobox B3
HOXC4	3221	homeobox C4
IPF2	3651	insulin promoter factor 2
ISL1	3670	ISL1 transcription factor, LIM/homeodomain (islet-1)
LBX1	10660	transcription factor similar to *D. melanogaster* homeodomain protein lady bird late
LHX2	9355	LIM homeobox 2
LHX5	64211	LIM homeobox 5
MEIS1	4211	myeloid ecotropic viral integration site 1 homolog (mouse)
NKX2-2	4821	NK2 transcription factor related, locus 2 (*Drosophila*)
NKX2-3	159296	NK2 transcription factor related, locus 3 (*Drosophila*)
ONECUT1	3175	one cut domain, family member 1
OTP	23440	orthopedia homolog (*Drosophila*)
OTX1	5013	orthodenticle homolog 1 (*Drosophila*)
PAX6	5080	paired box gene 6
TITF1	7080	thyroid transcription factor 1

Core Transcriptional Regulatory Circuitry in ES Cells

In order to identify regulatory network motifs associated with OCT4, SOX2, and NANOG, we assumed that regulator binding to a gene implies regulatory control and used algorithms that were previously devised to discover such regulatory circuits in yeast (Lee et al., 2002). The simplest units of commonly used transcriptional regulatory network architecture, or network motifs, provide specific regulatory capacities such as positive and negative feedback loops to control the levels of their components (Lee et al., 2002; Milo et al., 2002; Shen-Orr et al., 2002).

Our data indicated that OCT4, SOX2, and NANOG form feedforward loops that involve at least 353 protein coding and 2 miRNA genes (Figure 4A). Feedforwardloop motifs contain a regulator that controls a second regulator and have the additional feature that both regulators bind a set of common target genes. The feedforward loop has multiple regulatory capacities that may be especially useful for stem cells. When both regulators are positive, the feedforward loop can provide consistent activity that is relatively insensitive to transient changes in input (Mangan et al., 2003; Shen-Orr et al., 2002). If the regulators have positive and negative functions, the feedforward loop can act as a switch that enables a rapid response to inputs by providing a timesensitive delay where the downstream regulator acts to counter the effects of the upstream regulator in a delayed fashion (Mangan and Alon, 2003; Mangan et al., 2003). In ES cells, both regulatory capacities could be useful for maintaining the pluripotent state while retaining the ability to react appropriately to differentiation signals. Previous studies have shown that feedforwardloop architecture has been highly favored during the evolution of transcriptional regulatory networks in less complex eukaryotes (Lee et al., 2002; Ma et al., 2004; Milo et al., 2002; Resendis-Antonio et al., 2005; Shen-Orr et al., 2002). Our data suggest that feedforward regulation is an important feature of human ES cells as well.

Our results also showed that OCT4, SOX2, and NANOG together bound to the promoters of their own genes, forming interconnected autoregulatory loops (Figure 4B; see also Figure S2). Transcriptional regulation of OCT4, SOX2, and NANOG by the OCT4-SOX2 complex was recently described in murine ES cells (Catena et al., 2004; Kuroda et al., 2005; Okumura-Nakanishi et al., 2005; Rodda et al., 2005). Our data indicate that this autoregulatory loop is conserved in human ES cells and, more importantly, that NANOG is a component of the regulatory apparatus at these genes. Thus, it is likely that the expression and function of these three key stem cell factors are inextricably linked to one another. Autoregulation is thought to provide several advantages, including reduced response time to environmental stimuli and increased stability of gene expression (McAdams and Arkin, 1997; Rosenfeld et al., 2002; Shen-Orr et al., 2002; Thieffry et al., 1998).

The autoregulatory and feedforward circuitry described here may provide regulatory mechanisms by which stem cell identity can be robustly maintained yet permit cells to respond appropriately to developmental cues. Modifying OCT4 and NANOG levels and function can change the developmental potential of murine ES cells (Chambers et al., 2003; Mitsui et al., 2003; Nichols et al., 1998; Niwa et al., 2000), and this might be interpreted as being a consequence of perturbing independent regulatory pathways under the control of these two regulators. Our results argue that the levels and functions of these key stem cell regulators are tightly linked at both target genes and at their own promoters

and thus provide an additional framework for interpreting the genetic studies. Changes in the relative stoichiometry of these factors would disturb the autoregulatory and feedforward circuitry, producing changes in global gene regulation and thus cell fate.

Expanded Transcriptional Regulatory Circuitry

An initial model for ES cell transcriptional regulatory circuitry was constructed by identifying OCT4, SOX2, and NANOG target genes that encode transcription factors and chromatin regulators and integrating knowledge of the functions of these downstream regulators in both human and mouse based on the available expression studies and literature (Figure 5). The model includes a subset of active and a subset of repressed target genes based on the extensive expression characterization of the 353 co-bound genes as described earlier. The active targets include genes encoding components of chromatin remodeling and histone-modifying complexes (e.g., *SMARCAD1*, *MYST3*, and *SET*), which may have general roles in transcriptional regulation, and genes encoding transcription factors (e.g., *REST*, *SKIL*, *HESX1*, and *STAT3*), which themselves are known to regulate specific genes. For instance, REST has recently been shown to be highly abundant in ES cells and functions in part to repress neuronal specific genes (Ballas et al., 2005). Previous studies have proposed that NANOG may function through the Tgf-β pathway in ES cells (Chambers, 2004). Our model suggests that this occurs through direct regulation of key components of this pathway (e.g., TDFG1, LEFTY2/EBAF) and through regulation of at least one transcription factor, SKIL, which controls the activity of downstream components of this pathway (SMAD2, SMAD4) (He et al., 2003). Our data also reveal that OCT4, SOX2, and NANOG co-occupy STAT3, a key regulator of selfrenewal in mouse ES cells (Chambers, 2004), suggesting that STAT3 may also play a role in human ES cells.

The model described in Figure 5 also depicts a subset of the genes bound by OCT4, SOX2, and NANOG that are inactive and that encode transcription factors that have key roles in differentiation and development. These include regulators with demonstrated roles in development of all embryonic lineages. This initial model for ES cell transcriptional regulatory circuitry is consistent with previous genetic studies in mice that suggest that OCT4 and NANOG maintain pluripotency through repression of differentiation programs (Chambers et al., 2003; Mitsui et al., 2003; Niwa et al., 2000). This model also provides a mechanistic framework for understanding how this is accomplished through regulation of specific sets of genes that control cell-fate specification.

Concluding Remarks

Discovering how gene expression programs are controlled in living cells promises to improve our understanding of cell biology, development, and human health. Identifying the target genes for key transcriptional regulators of human stem cells is a first critical step in the process of understanding these transcriptional regulatory networks and learning how they control cell identity. Mapping OCT4, SOX2, and NANOG to their binding sites within known promoters has revealed that these regulators collaborate to form in ES cells regulatory circuitry consisting of specialized autoregulatory and feedforward loops. Continued advances in our ability to culture and genetically manipulate human ES cells will allow us to test and manipulate this circuitry. Identification of the targets of additional transcription factors and chromatin regulators using the approaches described here should allow investigators to produce a more comprehensive map of transcriptional regulatory circuitry in these cells. Connecting signaling pathways to this circuit map may reveal how these pluripotent cells can be stimulated to differentiate into different cell types or how to reprogram differentiated cells back to a pluripotent state.

Figure 4

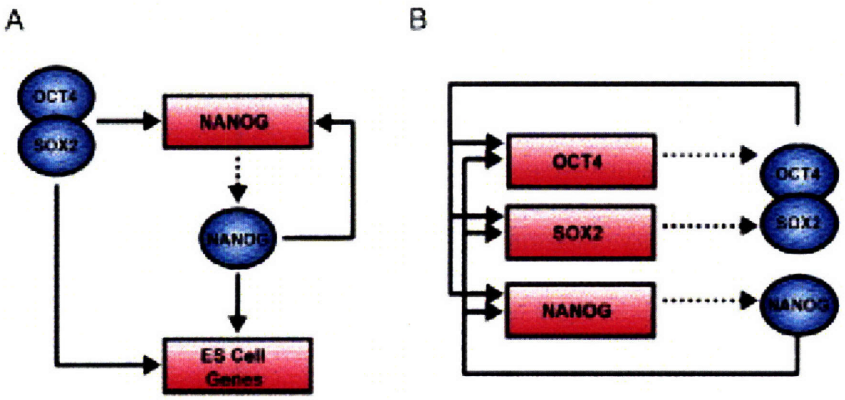

A
B
OCT4
SOX2
NANOG
NANOG
ES Cell Genes
OCT4
SOX2
NANOG
OCT4
SOX2
NANOG

Figure 4. Transcriptional Regulatory Motifs in Human ES Cells
(A) An example of feedforward transcriptional regulatory circuitry in human ES cells. Regulators are represented by blue circles; gene promoters are represented by red rectangles. Binding of a regulator to a promoter is indicated by a solid arrow. Genes encoding regulators are linked to their respective regulators by dashed arrows.
(B) The interconnected autoregulatory loop formed by OCT4, SOX2, and NANOG.

Figure 5

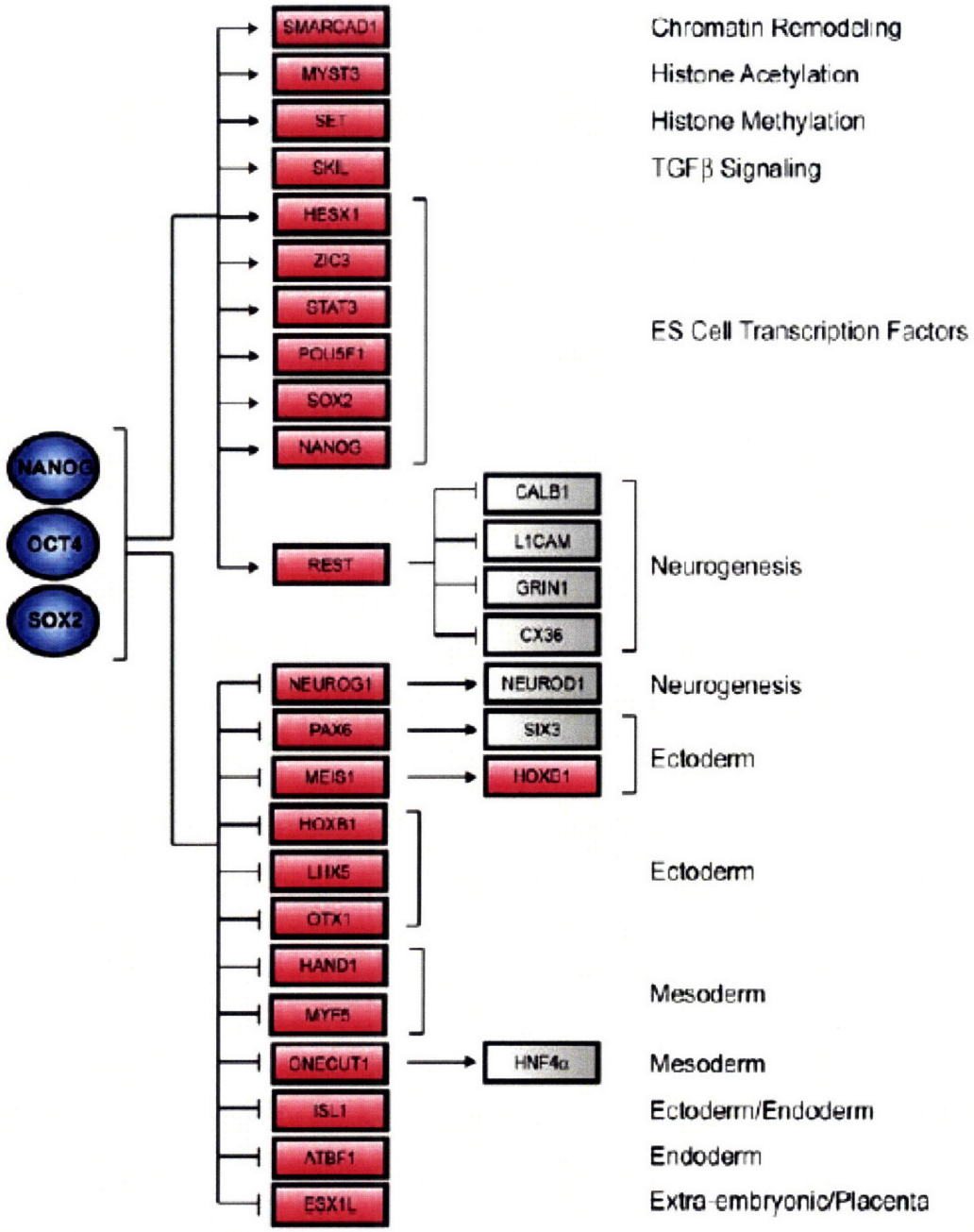

NANOG
OCT4
SOX2
SMARCAD1
MYST3
SET
SKIL
HESX1
ZIC3
STAT3
POU5F1
SOX2
NANOG
REST
CALB1
L1CAM
GRIN1
CX36
NEUROG1
PAX6
MEIS1
NEUROD1
SIX3
HOXB1
HOXB1
LHX5
OTX1
HAND1
MYF5
ONECUT1
HNF4α
ISL1
ATBF1
ESX1L
Chromatin Remodeling
Histone Acetylation
Histone Methylation
TGFβ Signaling
ES Cell Transcription Factors
Neurogenesis
Neurogenesis
Ectoderm
Ectoderm
Mesoderm
Mesoderm
Ectoderm/Endoderm
Endoderm
Extra-embryonic/Placenta

Figure 5. Core Transcriptional Regulatory Network in Human ES Cells
A model for the core transcriptional regulatory network was constructed by identifying OCT4, SOX2, and NANOG target genes that encode transcription factors and chromatin regulators and integrating knowledge of the functions of these downstream regulators based on comparison to multiple expression datasets (Supplemental Data) and to the literature. A subset of active and inactive genes co-occupied by the three factors in human ES cells is shown here. Regulators are represented by blue circles; gene promoters are represented by red rectangles; gray boxes represent putative downstream target genes. Positive regulation was assumed if the target gene was expressed whereas negative regulation was assumed if the target gene was not transcribed.

Experimental Procedures

Growth Conditions for Human Embryonic Stem Cells

Human embryonic stem (ES) cells were obtained from WiCell (Madison, Wisconsin; NIH Code WA09). Detailed protocol information on human ES cell growth conditions and culture reagents are available at http://www.mcb.harvard.edu/melton/hues. Briefly, passage 34 cells were grown in KO-DMEM medium supplemented with serum replacement, basic fibroblast growth factor (bFGF), recombinant human leukemia inhibitory factor (LIF), and a human plasma protein fraction. In order to minimize any MEF contribution in our analysis, H9 cells were cultured on a low density of irradiated murine embryonic fibroblasts (ICR MEFs) resulting in a ratio of approximately >8:1 H9 cell to MEF. The culture of H9 on low-density MEFs had no adverse effects on cell morphology, growth rate, or undifferentiated status as compared to cells grown under typical conditions. In addition, immunohistochemistry for pluripotency markers (e.g., OCT4, SSEA-3) indicated that H9 cells grown on a minimal feeder layer maintained the ability to generate derivates of ectoderm, mesoderm, and endoderm upon differentiation (Figures S3 and S4).

Antibodies

The NANOG (AF1997) and SOX2 (AF2018) antibodies used in this study were immunoaffinity purified against the human proteins and shown to recognize their target proteins in Western blots and by immunocytochemistry (R&D Systems Minneapolis, Minnesota). Multiple OCT4 antibodies directed against different portions of the protein (AF1759 R&D Systems, sc-8628 Santa Cruz, sc-9081 Santa Cruz), some of which were immunoaffinity purified, were used in this study and have been shown to recognize their target protein in Western blots and by immunocytochemistry. The E2F4 antibody used in this study was obtained from Santa Cruz (sc-1082) and has been shown to recognize E2F4-responsive genes identified in previous ChIP studies (Table S2) (Ren et al., 2002; Weinmann et al., 2002).

Chromatin Immunoprecipitation

Protocols describing all materials and methods can be downloaded from http://jura.wi.mit.edu/young/hESRegulation/.

Human embryonic stem cells were grown to a final count of $5 \times 10^7 - 1 \times 10^8$ cells for each location analysis reaction. Cells were chemically crosslinked by the addition of one-tenth volume of fresh 11% formaldehyde solution for 15 min at room temperature. Cells were rinsed twice with 1x PBS and harvested using a silicon scraper and flash frozen in liquid nitrogen and 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 power 7 for 10 x 30 s pulses (90 s pause between pulses) at 4°C while samples were immersed in an ice bath. The resulting wholecell extract was incubated overnight at 4°C with 100 ul of Dynal Protein G magnetic beads that had been preincubated with 10 ug of the appropriate antibody. Beads were washed five times with RIPA buffer and one 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 sonicationstep) 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 and ligated to linker and amplified using a two-stage PCR protocol. Amplified DNA was labeled and purified using Invitrogen Bioprime random primer labeling kits (immunoenriched DNA was labeled with Cy5 fluorophore, whole-cell extract DNA was labeled with Cy3 fluorophore). Labeled DNA was combined (5–6 ug each of immunoenriched and whole-cell extract DNA) and hybridized to arrays in Agilent hybridization chambers for 40 hr at 40°C. Arrays were then washed and scanned (Supplemental Data).

Array Design and Data Extraction

The design of the 10-slide oligo-based promoter arrays used in this study and data extraction methods are described in detail in Supplemental Data. Arrays were manufactured by Agilent Technologies (http://www.agilent.com).

Supplemental Data

Supplemental Data include seven figures, seven tables, and Supplemental text and can be found with this article online at http:// www.cell.com/cgi/content/full/122/6/947/DC1/.

Acknowledgments

We would like to thank Bioinformatics and Research Computing (BaRC) and the Center for Microarray Technology (CMT) at the Whitehead Institute for computational and technical support. We would also like to thank members of the Young lab as well as Chad Cowan and Kevin Eggan for helpful discussions. L.A.B. was supported by NRSA postdoctoral fellowship CA094664, and H.L.M. by NRSA postdoctoral fellowship GM068273. R.M.K. was supported by a fellowship from the American Cancer Society. This work was supported by NHGRI grant HG002668 to D.K.G. and R.A.Y. and NIH grant GM069400 to R.A.Y. T.I.L., D.K.G., and R.A.Y. consult for Agilent Technologies.

References

Abeyta, M.J., Clark, A.T., Rodriguez, R.T., Bodnar, M.S., Pera, R.A., and Firpo, M.T. (2004). Unique gene expression signatures of independently derived human embryonic stem cell lines. Hum. Mol. Genet. *13*, 601–608.

Ambrosetti, D.C., Basilico, C., and Dailey, L. (1997). Synergistic activation of the fibroblast growth factor 4 enhancer by Sox2 and Oct-3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites. Mol. Cell. Biol. *17*, 6321–6329.

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

Ballas, N., Grunseich, C., Lu, D.D., Speh, J.C., and Mandel, G. (2005). REST and its corepressors mediate plasticity of neuronal gene chromatin throughout neurogenesis. Cell *121*, 645–657.

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

Bortvin, A., Eggan, K., Skaletsky, H., Akutsu, H., Berry, D.L., Yanagimachi, R., Page, D.C., and Jaenisch, R. (2003). Incomplete reactivation of Oct4-related genes in mouse embryos cloned from somatic nuclei. Development *130*, 1673–1680.

Botquin, V., Hess, H., Fuhrmann, G., Anastassiadis, C., Gross, M.K., Vriend, G., and Scholer, H.R. (1998). New POU dimer configuration mediates antagonistic control of an osteopontin preimplantation enhancer by Oct-4 and Sox-2. Genes Dev. *12*, 2073–2090.

Brandenberger, R., Khrebtukova, I., Thies, R.S., Miura, T., Jingli, C., Puri, R., Vasicek, T., Lebkowski, J., and Rao, M. (2004). MPSS profiling of human embryonic stem cells. BMC Dev. Biol. *4*, 10–25.

Catena, R., Tiveron, C., Ronchi, A., Porta, S., Ferri, A., Tatangelo, L., Cavallaro, M., Favaro, R., Ottolenghi, S., Reinbold, R., et al. (2004). Conserved POU binding DNA sites in the Sox2 upstream enhancer regulate gene expression in embryonic and neural stem cells. J. Biol. Chem. *279*, 41846–41857.

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

Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., and Smith, A. (2003). Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell *113*, 643–655.

Dailey, L., and Basilico, C. (2001). Coevolution of HMG domains and homeodomains and the generation of transcriptional regulation by Sox/POU complexes. J. Cell. Physiol. *186*, 315–328.

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

Hart, A.H., Hartley, L., Ibrahim, M., and Robb, L. (2004). Identification, cloning and expression analysis of the pluripotency promoting Nanog genes in mouse and human. Dev. Dyn. *230*, 187–198.

Hay, D.C., Sutherland, L., Clark, J., and Burdon, T. (2004). Oct-4 knockdown induces similar patterns of endoderm and trophoblast differentiation markers in human and mouse embryonic stem cells. Stem Cells *22*, 225–235.

He, J., Tegen, S.B., Krawitz, A.R., Martin, G.S., and Luo, K. (2003). The transforming activity of Ski and SnoN is dependent on their ability to repress the activity of Smad proteins. J. Biol. Chem. *278*, 30540–30547.

Hombria, J.C., and Lovegrove, B. (2003). Beyond homeosis–HOX function in morphogenesis and organogenesis. Differentiation *71*, 461–476.

James, D., Levine, A.J., Besser, D., and Hemmati-Brivanlou, A. (2005). TGFbeta/activin/nodal signaling is necessary for the maintenance of pluripotency in human embryonic stem cells. Development *132*, 1273–1282.

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

Kuroda, T., Tada, M., Kubota, H., Kimura, H., Hatano, S.Y., Suemori, H., Nakatsuji, N., and Tada, T. (2005). Octamer and Sox elements are required for transcriptional cis regulation of Nanog gene expression. Mol. Cell. Biol. *25*, 2475–2485.

Lee, J.H., Hart, S.R., and Skalnik, D.G. (2004). Histone deacetylase activity is required for embryonic stem cell differentiation. Genesis *38*, 32–38.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., et al. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science *298*, 799–804.

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.

Ma, H.W., Kumar, B., Ditges, U., Gunzer, F., Buer, J., and Zeng, A.P. (2004). An extended transcriptional regulatory network of Escherichia coli and analysis of its hierarchical structure and network motifs. Nucleic Acids Res. *32*, 6643–6649.

Mangan, S., and Alon, U. (2003). Structure and function of the feedforward loop network motif. Proc. Natl. Acad. Sci. USA *100*, 11980–11985.

Mangan, S., Zaslaver, A., and Alon, U. (2003). The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. J. Mol. Biol. *334*, 197–204.

Matin, M.M., Walsh, J.R., Gokhale, P.J., Draper, J.S., Bahrami, A.R., Morton, I., Moore, H.D., and Andrews, P.W. (2004). Specific knockdown of Oct4 and beta2-microglobulin expression by RNA interference in human embryonic stem cells and embryonic carcinoma cells. Stem Cells *22*, 659–668.

McAdams, H.H., and Arkin, A. (1997). Stochastic mechanisms in gene expression. Proc. Natl. Acad. Sci. USA *94*, 814–819.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. Science *298*, 824–827.

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

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

Nishimoto, M., Fukushima, A., Okuda, A., and Muramatsu, M. (1999). The gene for the embryonic stem cell coactivator UTF1 carries carries a regulatory element which selectively interacts with a complex composed of Oct-3/4 and Sox-2. Mol. Cell. Biol. *19*, 5453–5465.

Niwa, H. (2001). Molecular mechanism to maintain stem cell renewal of ES cells. Cell Struct. Funct. *26*, 137–148.

Niwa, H., Miyazaki, J., and Smith, A.G. (2000). Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or selfrenewal of ES cells. Nat. Genet. *24*, 372–376.

Odom, D.T., Zizlsperger, N., Gordon, D.B., Bell, G.W., Rinaldi, N.J., Murray, H.L., Volkert, T.L., Schreiber, J., Rolfe, P.A., Gifford, D.K., et al. (2004). Control of pancreas and liver gene expression by HNF transcription factors. Science *303*, 1378–1381.

Okumura-Nakanishi, S., Saito, M., Niwa, H., and Ishikawa, F. (2005). Oct-3/4 and Sox2 regulate Oct-3/4 gene in embryonic stem cells. J. Biol. Chem. *280*, 5307–5317.

Pasquinelli, A.E., Hunter, S., and Bracht, J. (2005). MicroRNAs: a developing story. Curr. Opin. Genet. Dev. *15*, 200–205.

Pera, M.F., and Trounson, A.O. (2004). Human embryonic stem cells: prospects for development. Development *131*, 5515–5525.

Pesce, M., and Schöler, H.R. (2001). Oct-4: gatekeeper in the beginnings of mammalian development. Stem Cells *19*, 271–278.

Remenyi, A., Schöler, H.R., and Wilmanns, M. (2004). Combinatorial control of gene expression. Nat. Struct. Mol. Biol. *11*, 812–815.

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.

Resendis-Antonio, O., Freyre-Gonzalez, J.A., Menchaca-Mendez, R., Gutierrez-Rios, R.M., Martinez-Antonio, A., Avila-Sanchez, C., and Collado-Vides, J. (2005). Modular analysis of the transcriptional regulatory network of E. coli. Trends Genet. *21*, 16–20.

Reubinoff, B.E., Pera, M.F., Fong, C.Y., Trounson, A., and Bongso, A. (2000). Embryonic stem cell lines from human blastocysts: somatic differentiation in vitro. Nat. Biotechnol. *18*, 399–404.

Rodda, D.J., Chew, J.-L., Lim, L.-H., Loh, Y.-H., Wang, B., Ng, H.-H., and Robson, P. (2005). Transcriptional regulation of Nanog by Oct4 and Sox2. J. Biol. Chem. *18*, in press.

Rosenfeld, N., Elowitz, M.B., and Alon, U. (2002). Negative autoregulation speeds the response times of transcription networks. J. Mol. Biol. *323*, 785–793.

Sato, N., Sanjuan, I.M., Heke, M., Uchida, M., Naef, F., and Brivanlou, A.H. (2003). Molecular signature of human embryonic stem cells and its comparison with the mouse. Dev. Biol. *260*, 404–413.

Sato, N., Meijer, L., Skaltsounis, L., Greengard, P., and Brivanlou, A.H. (2004). Maintenance of pluripotency in human and mouse embryonic stem cells through activation of Wnt signaling by a pharmacological GSK-3-specific inhibitor. Nat. Med. *10*, 55–63.

Schöler, H.R., Dressler, G.R., Balling, R., Rohdewohld, H., and Gruss, P. (1990). Oct-4: a germline-specific transcription factor mapping to the mouse t-complex. EMBO J. *9*, 2185–2195.

Shen-Orr, S.S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. Nat. Genet. *31*, 64–68.

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. USA *101*, 6062–6067.

Thieffry, D., Salgado, H., Huerta, A.M., and Collado-Vides, J. (1998). Prediction of transcriptional regulatory sites in the complete genome sequence of Escherichia coli K-12. Bioinformatics *14*, 391– 400.

Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S., and Jones, J.M. (1998). Embryonic stem cell lines derived from human blastocysts. Science *282*, 1145–1147.

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

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 immunoprecipiation and CpG island microarray analysis. Genes Dev. *16*, 235–244.

Yuan, H., Corbi, N., Basilico, C., and Dailey, L. (1995). Developmental- specific activity of the FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3. Genes Dev. *9*, 2635–2645.

Zaehres, H., Lensch, M.W., Daheron, L., Stewart, S.A., Itskovitz- Eldor, J., and Daley, G.Q. (2005). High-efficiency RNA interference in human embryonic stem cells. Stem Cells *23*, 299–305.

Accession Numbers

All microarray data from this study are available at ArrayExpress at the EBI (http://www.ebi.ac.uk/arrayexpress) under the accession designation E-WMIT-5.

Chapter 3

Control of Developmental Regulators by Polycomb in Human Embryonic Stem Cells

Published as: Tong Ihn Lee, Richard G. Jenner, Laurie A. Boyer, Matthew G. Guenther, Stuart S. Levine, Roshan M. Kumar, Brett Chevalier, Sarah E. Johnstone, Megan F. Cole, Kyo-ichi Isono, Haruhiko Koseki, Takuya Fuchikami, Kuniya Abe, Heather L. Murray, Jacob P. Zucker, Bingbing Yuan, George W. Bell, Elizabeth Herbolsheimer, Nancy M. Hannett, Kaiming Sun, Duncan T. Odom, Arie P. Otte, Thomas L. Volkert, David P. Bartel, Douglas A. Melton, David K. Gifford, Rudolf Jaenisch, and Richard A. Young (2006). "Control of Developmental Regulators by Polycomb in Human Embryonic Stem Cells." Cell 125: 301-313.

My contribution to this project

The work studying Polycomb was a highly collaborative effort within the Young Lab, involving many lab members. I performed gene ontology analyses and made computational tools aiding data visualization. I also helped the experimental effort of hybing the 115 slide microarray sets. I also made conceptual contributions, particularly involving the overlap with Oct4, Sox2 and Nanog target genes.

Summary

Polycomb group proteins are essential for early development in metazoans, but their contributions to human development are not well understood. We have mapped the Polycomb Repressive Complex 2 (PRC2) subunit SUZ12 across the entire non-repeat portion of the genome in human embryonic stem (ES) cells. We found that SUZ12 is distributed across large portions of over two hundred genes encoding key developmental regulators. These genes are occupied by nucleosomes trimethylated at histone H3K27, are transcriptionally repressed, and contain some of the most highly conserved non-coding elements in the genome. We found that PRC2 target genes are preferentially activated during ES cell differentiation and that the ES cell regulators OCT4, SOX2, and NANOG co-occupy a significant subset of these genes. These results indicate that PRC2 occupies a special set of developmental genes in ES cells that must be repressed to maintain pluripotency and that are poised for activation during ES cell differentiation.

Introduction

Embryonic stem (ES) cells are a unique self-renewing cell type that can give rise to the ectodermal, endodermal, and mesodermal germ layers during embryogenesis. Human ES cells, which can be propagated in culture in an undifferentiated state but selectively induced to differentiate into many specialized cell types, are thought to hold great promise for regenerative medicine (Thomson et al., 1998; Reubinoff et al., 2000; Mayhall et al., 2004; Pera and Trounson, 2004). The gene expression program of ES cells must allow these cells to maintain a pluripotent state but also allow for differentiation into more specialized states when signaled to do so. Learning how this is accomplished may be key to realizing the therapeutic potential of ES cells and further understanding early development.

Among regulators of development, the Polycomb group proteins (PcG) are of special interest. These regulators were first described in Drosophila, where they repress the homeotic genes controlling segment identity in the developing embryo (Lewis, 1978; Denell and Frederick, 1983; Simon et al., 1992; Orlando and Paro, 1995; Pirrotta, 1998; Kennison, 2004). The initial repression of these genes is carried out by DNA binding transcriptional repressors, and PcG proteins modify chromatin to maintain these genes in a repressed state (Duncan, 1986; Bender et al., 1987; Strutt et al., 1997; Horard et al., 2000; Hodgson et al., 2001; Mulholland et al., 2003).

The PcG proteins form multiple Polycomb Repressive Complexes (PRCs), the components of which are conserved from Drosophila to humans (Franke et al., 1992; Shao et al., 1999; Birve et al., 2001; Tie et al., 2001; Cao et al., 2002; Czermin et al., 2002; Kuzmichev et al., 2002; Levine et al., 2002). The PRCs are brought to the site of initial repression and act through epigenetic modification of chromatin structure to promote gene silencing (Pirrotta, 1998; Levine et al., 2004; Lund and van Lohuizen, 2004; Ringrose and Paro, 2004). PRC2 catalyzes histone H3 lysine-27 (H3K27) methylation, and this enzymatic activity is required for PRC2-mediated gene silencing (Cao et al., 2002; Czermin et al., 2002; Kuzmichev et al., 2002; Muller et al., 2002; Kirmizis et al., 2004). H3K27 methylation is thought to provide a binding surface for PRC1, which facilitates oligomerization, condensation of chromatin structure, and inhibition of chromatin remodeling activity in order to maintain silencing (Shao et al., 1999; Francis et al., 2001; Cao et al., 2002; Czermin et al., 2002).

Components of PRC2 are essential for the earliest stages of vertebrate development (Faust et al., 1998; O'Carroll et al., 2001; Pasini et al., 2004). PRC2 and its related complexes, PRC3 and PRC4, contain the core components EZH2, SUZ12, and EED (Kuzmichev et al., 2004; Kuzmichev et al., 2005). EZH2 is a H3K27 methyltransferase, and SUZ12 (Suppressor of zeste 12) is required for this activity (Cao and Zhang, 2004; Pasini et al., 2004). ES cell lines cannot be established from Ezh2-deficient blastocysts (O'Carroll et al., 2001), suggesting that PRC2 is involved in regulating pluripotency and self-renewal. Although the PRCs are known to repress individual HOX genes (van der Lugt et al., 1996; Akasaka et al., 2001; Wang et al., 2002; Cao and Zhang, 2004), it is not clear how these important PcG regulators contribute to early development in vertebrates.

Because the nature of PRC2 target genes in ES cells might reveal why PRC2 is essential for early embryonic development, pluripotency, and self-renewal, we have

mapped the sites occupied by the SUZ12 subunit throughout the genome in human ES cells. This genome- wide map reveals that PRC2 is associated with a remarkable cadre of genes encoding key regulators of developmental processes that are repressed in ES cells. The genes occupied by PRC2 contain nucleosomes that are trimethylated at histone H3 lysine-27 (H3K27me3), a modification catalyzed by PRC2 and associated with the repressed chromatin state. Both PRC2 and nucleosomes with histone H3K27me3 occupy surprisingly large genomic domains around these developmental regulators and are frequently associated with highly conserved noncoding sequence elements previously identified by comparative genomic methods. The transcription factors OCT4, SOX2, and NANOG, which are also key regulators of ES cell pluripotency and self-renewal, occupy a significant subset of these genes. Thus, the model of epigenetic regulation of homeotic genes extends to a large set of developmental regulators whose repression in ES cells appears to be key to pluripotency. We suggest that PRC2 functions in ES cells to repress developmental genes that are preferentially activated during differentiation.

Results and Discussion

Mapping Genome Occupancy in ES Cells

We mapped the location of both RNA polymerase II and the SUZ12 subunit of PRC2 genome-wide in human ES cells (Figure 1). The initiating form of RNA polymerase II was mapped to test the accuracy of the method and provide a reference for comparison with sites occupied by PRC2. The SUZ12 subunit of PRC2 is critical for the function of the complex and was selected for these genomewide experiments. Human ES cells (H9, NIH code WA09) were analyzed by immunohistochemistry for characteristic stem cell markers, tested for their ability to generate cell types from all three germ layers upon differentiation into embryoid bodies, and shown to form teratomas in immunocompromised mice (Supplemental Data; Figures S1–S3).

DNA sequences bound by RNA polymerase II were identified in replicate chromatin-immunoprecipitation (ChIP) experiments using DNA microarrays that contain over 4.6 million unique 60-mer oligonucleotide probes spanning the entire nonrepeat portion of the human genome (Figure 1 and Supplemental Data). To obtain a probabilistic assessment of binding events, an algorithm was implemented that incorporates information from multiple probes representing contiguous regions of the genome, and threshold criteria were established to identify a dataset with minimal false positives and false negatives. RNA polymerase II was associated with the promoters of 7,106 of the approximately 22,500 annotated human genes, indicating that one-third of protein-coding genes are prepared to be transcribed in ES cells. Three lines of evidence suggest this dataset is of high quality. Most of the RNA polymerase II sites (87%) occurred at promoters of known or predicted genes. Transcripts were detected for 88% of the genes bound by RNA polymerase II in previous expression experiments in ES cells. Finally, independent analysis using gene-specific PCR (Supplemental Data) indicated that the frequency of false positives was approximately 4% and the frequency of false negatives was approximately 30% in this dataset. A detailed analysis of the RNA polymerase II dataset, including binding to miRNA genes, can be found in Supplemental Data (Tables S1–S6 and Figures S4 and S5).

The sites occupied by SUZ12 were then mapped throughout the entire nonrepeat genome in H9 ES cells using the same approach described for RNA polymerase II (Figure 1C). SUZ12 was associated with the promoters of 1,893 of the approximately 22,500 annotated human genes, indicating that 8% of protein-coding genes are occupied by SUZ12 in ES cells (Supplemental Data; Tables S7 and S8). Independent site-specific analysis indicated that the frequency of false positives was approximately 3% and the frequency of false negatives was approximately 27% in this dataset.

Comparison of the genes occupied by SUZ12 with those occupied by RNA polymerase II revealed that the two sets were largely exclusive (Figure 1D; Supplemental Data; Table S8). There were, however, genes where SUZ12 and RNA polymerase II cooccupied promoters. At these genes, PRC complexes may fail to block assembly of the preinitiation complex (Dellino et al., 2004), consistent with the observation that Polycomb group proteins can associate with components of the general transcription apparatus (Breiling et al., 2001; Saurin et al., 2001). The vast majority of SUZ12 bound sites were found at gene promoters (Figure 1E). Ninety-five percent of the SUZ12 bound regions were found within 1 kb of known or predicted transcription start sites

(Supplemental Data and Table S7). This suggests that SUZ12 functions in human ES cells primarily at promoters rather than at distal regulatory elements. It is interesting that 40% of all SUZ12 bound regions are within 1 Kb of CpG islands (Table S7), given the recent discovery of a mechanistic link between PcG proteins and DNA methyltransferoses (Vire et al., 2006).

Global Transcriptional Repression by PRC2

PRC2 is composed of three core subunits, SUZ12, EED, and EZH2, and has been shown to mediate histone H3K27 methylation at specific genes in vivo. To confirm that SUZ12 is associated with active PRC2 at target genes, we used chromatin immunoprecipitation with antibodies against EED and the histone H3K27me3 mark and analyzed the results with promoter microarrays. We found that EED and the histone H3K27me3 mark cooccurred with SUZ12 at most genes using a high-confidence binding threshold (Figure 2). The false negative rates of thresholded data can lead to an underestimate of the similarity between different datasets. Plotting raw enrichment ratios for genes associated with SUZ12, EED, or H3K27me3 demonstrates that SUZ12 binding represents PRC2 binding at almost all target genes (Figure S6).

Genetic and biochemical studies at selected genes indicate that PRC2-mediated H3K27 methylation represses gene expression, but it has not been established if it acts as a repressor genome-wide. If genes occupied by SUZ12 are repressed by PRC2, then transcripts from these genes should generally be present at lower levels in ES cells than in differentiated cell types. To test this prediction, we compared the expression levels of PRC2-occupied genes in four different ES cell lines with the expression level of these genes in 79 differentiated human cell and tissue types (Sato et al., 2003; Abeyta et al., 2004; Su et al., 2004). We found that PRC2 occupied genes were generally underexpressed in ES cells relative to other cell types (Figure 2C). A small fraction of the genes occupied by PRC2 were relatively overexpressed in ES cells (Figure 2C); these tended to show less extensive SUZ12 occupancy and were more likely to be cooccupied by RNA polymerase II (Supplemental Data). These results are consistent with the model that PRC2-mediated histone H3K27 methylation promotes gene silencing at the majority of its target genes throughout the genome in ES cells.

Key Developmental Regulators Are Targets of PRC2

Examination of the targets of SUZ12 revealed that they were remarkably enriched for genes that control development and transcription (Figure 3) and that SUZ12 tended to occupy large domains at these genes (Figure 4). Although only 8% of all annotated genes were occupied by SUZ12, 50% of those encoding transcription factors associated with developmental processes were occupied by SUZ12. By comparison, RNA polymerase II preferentially occupied genes involved in a broad spectrum of cell proliferation functions such as nucleic acid metabolism, protein synthesis, and cell cycle (Figure 3A and examples in Figure 1B; Supplemental Data; Table S10).

It was striking that SUZ12 occupied many families of genes that control development and transcription (Figures 3B and S7 and Table S11). These included 39 of 40 of the homeotic genes found in the HOX clusters and the majority of homeodomain genes. SUZ12 bound homeodomain genes included almost all members of the DLX, IRX, LHX, and PAX gene families, which regulate early developmental steps in

Figure 1

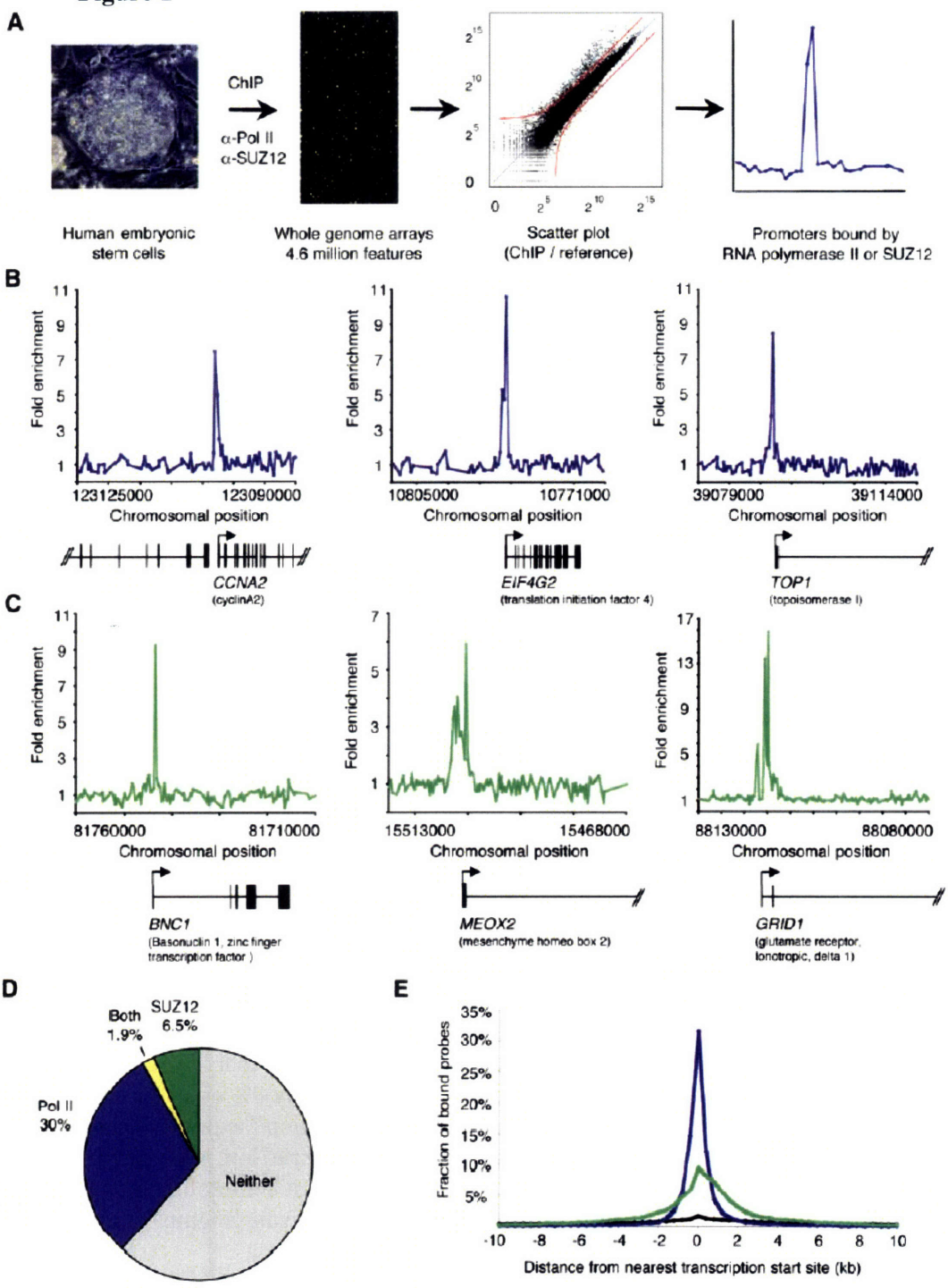

A
ChIP
α-Pol II
α-SUZ12
Human embryonic stem cells
Whole genome arrays 4.6 million features
Scatter plot (ChIP / reference)
Promoters bound by RNA polymerase II or SUZ12
B
Fold enrichment
Chromosomal position
123125000
123090000
CCNA2
(cyclinA2)
10805000
10771000
EIF4G2
(translation initiation factor 4)
39079000
39114000
TOP1
(topoisomerase I)
C
81760000
81710000
BNC1
(Basonuclin 1, zinc finger transcription factor)
15513000
15468000
MEOX2
(mesenchyme homeo box 2)
88130000
88080000
GRID1
(glutamate receptor, ionotropic, delta 1)
D
Both 1.9%
SUZ12 6.5%
Pol II 30%
Neither
E
Fraction of bound probes
Distance from nearest transcription start site (kb)

Figure 1. Genome-Wide ChIP-Chip in Human Embryonic Stem Cells
(A) DNA segments bound by the initiation form of RNA polymerase II or SUZ12 were isolated using chromatin-immunoprecipitation (ChIP) and identified with DNA microarrays containing over 4.6 million unique 60-mer oligonucleotide probes spanning the entire nonrepeat portion of the human genome. ES cell growth and quality control, the antibodies, ChIP protocol, DNA microarray probe design, and data analysis methods are described in detail in Supplemental Data.
(B) Examples of RNA polymerase II ChIP signals from genome-wide ChIP-Chip. The plots show unprocessed enrichment ratios (blue) for all probes within a genomic region (ChIP versus whole genomic DNA). Chromosomal positions are from NCBI build 35 of the human genome. Genes are shown to scale below plots (exons are represented by vertical bars). The start and direction of transcription are noted by arrows.
(C) Examples of SUZ12 ChIP signals from genome-wide ChIP-Chip. The plots show unprocessed enrichment ratios (green) for all probes within a genomic region (ChIP versus whole genomic DNA). Chromosomal positions, genes, and notations are as described in (B).
(D) Chart showing percentage of all annotated genes bound by RNA polymerase II (blue), SUZ12 (green), both (yellow), or neither (gray).
(E) Distribution of the distance between bound probes and the closest transcription start sites from RefSeq, Ensembl, MGC, UCSC Known Genes and H-Inv databases for SUZ12 (green line), and RNA polymerase II (blue line). The number of bound probes is given as the percentage of total probes and is calculated for 400 bp intervals from the start site. The null-distribution of the distance between all probes and the closest transcription are shown as a black line.

Figure 2

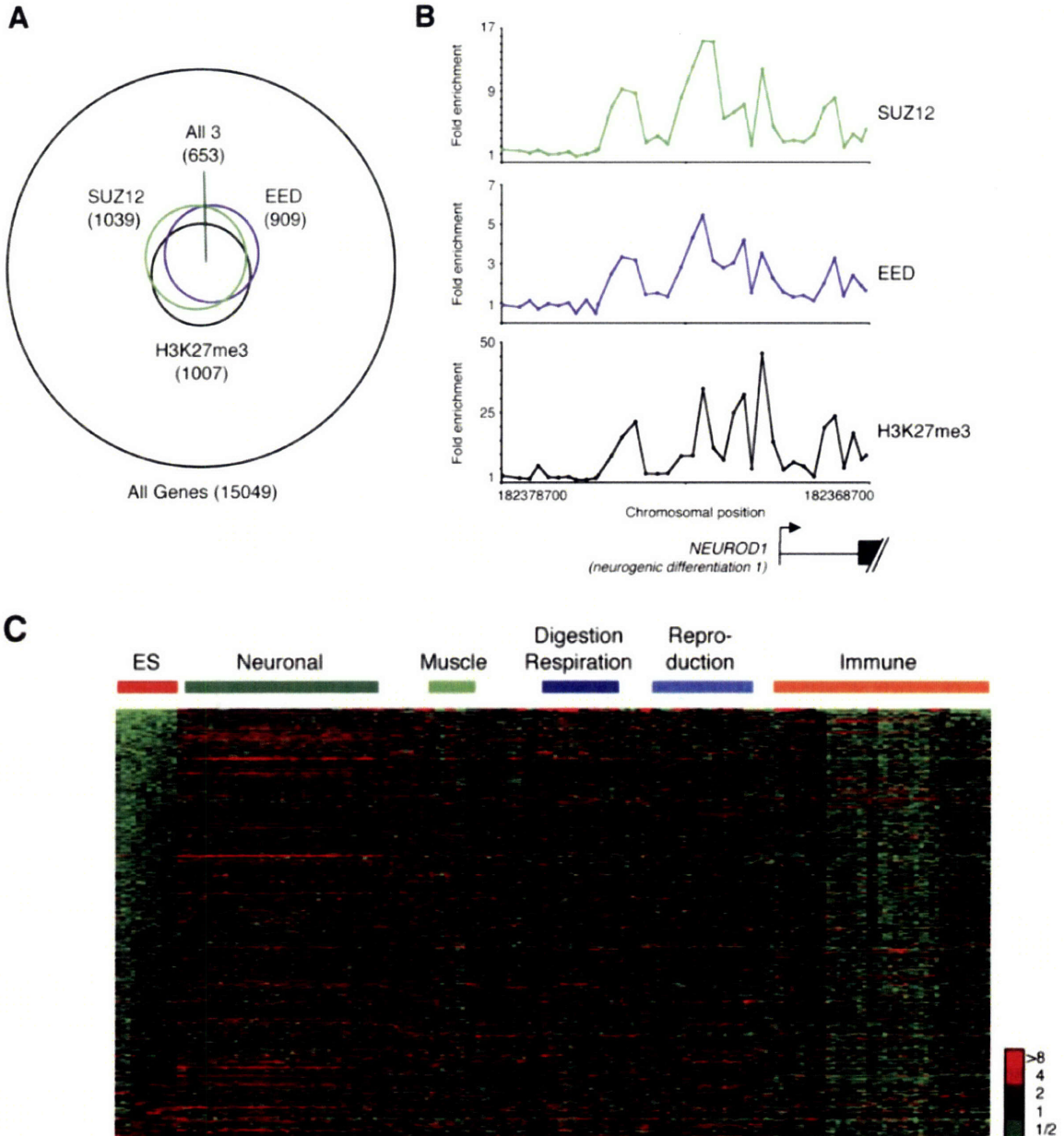

A
All 3
(653)
SUZ12
(1039)
EED
(909)
H3K27me3
(1007)
All Genes (15049)
B
Fold enrichment
17
9
1
SUZ12
Fold enrichment
7
5
3
1
EED
Fold enrichment
50
25
1
H3K27me3
182378700
182368700
Chromosomal position
NEUROD1
(neurogenic differentiation 1)
C
ES
Neuronal
Muscle
Digestion
Respiration
Repro-
duction
Immune
>8
4
2
1
1/2
1/4
<1/8

Figure 2. SUZ12 Is Associated with EED, histone H3K27me3 Modification, and Transcriptional Repression in ES Cells

(A) Venn diagram showing the overlap of genes bound by SUZ12 at high-confidence, genes bound by EED at high-confidence, and genes trimethylated at H3K27 at highconfidence. The data are from promoter microarrays that contain probes tiling 8 kb and +2 kb around transcription start. 72% of the genes bound by SUZ12 at high-confidence are also bound by EED at high-confidence; others are bound by EED at lower confidence (Figure S6).

(B) SUZ12 (top), EED (middle), and H3K27me3 (bottom) occupancy at NEUROD1. The plots show unprocessed enrichment ratios for all probes within this genomic region (SUZ12 ChIP versus whole genomic DNA, EED ChIP versus whole genomic DNA, and H3K27me3 ChIP versus total H3 ChIP). Chromosomal positions are from NCBI build 35 of the human genome. NEUROD1is shown to scale below plots (exons are represented by vertical bars). The start and direction of transcription are noted by arrows.

(C) Relative expression levels of 604 genes occupied by PRC2 and trimethylated at H3K27 in ES cells. Comparisons were made across four ES cell lines and 79 differentiated cell types. Each row corresponds to a single gene that is bound by SUZ12, associated with EED and H3K27me3, and for which Affymetrix expression data are available. Each column corresponds to a single expression microarray. ES cells are in the following order: H1, H9, HSF6, HSF1. 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 according to the scale on the right. Cell types are grouped by tissue or organ function, and genes are ranked according the significance of their relative level of gene expression in ES cells.

Figure 3

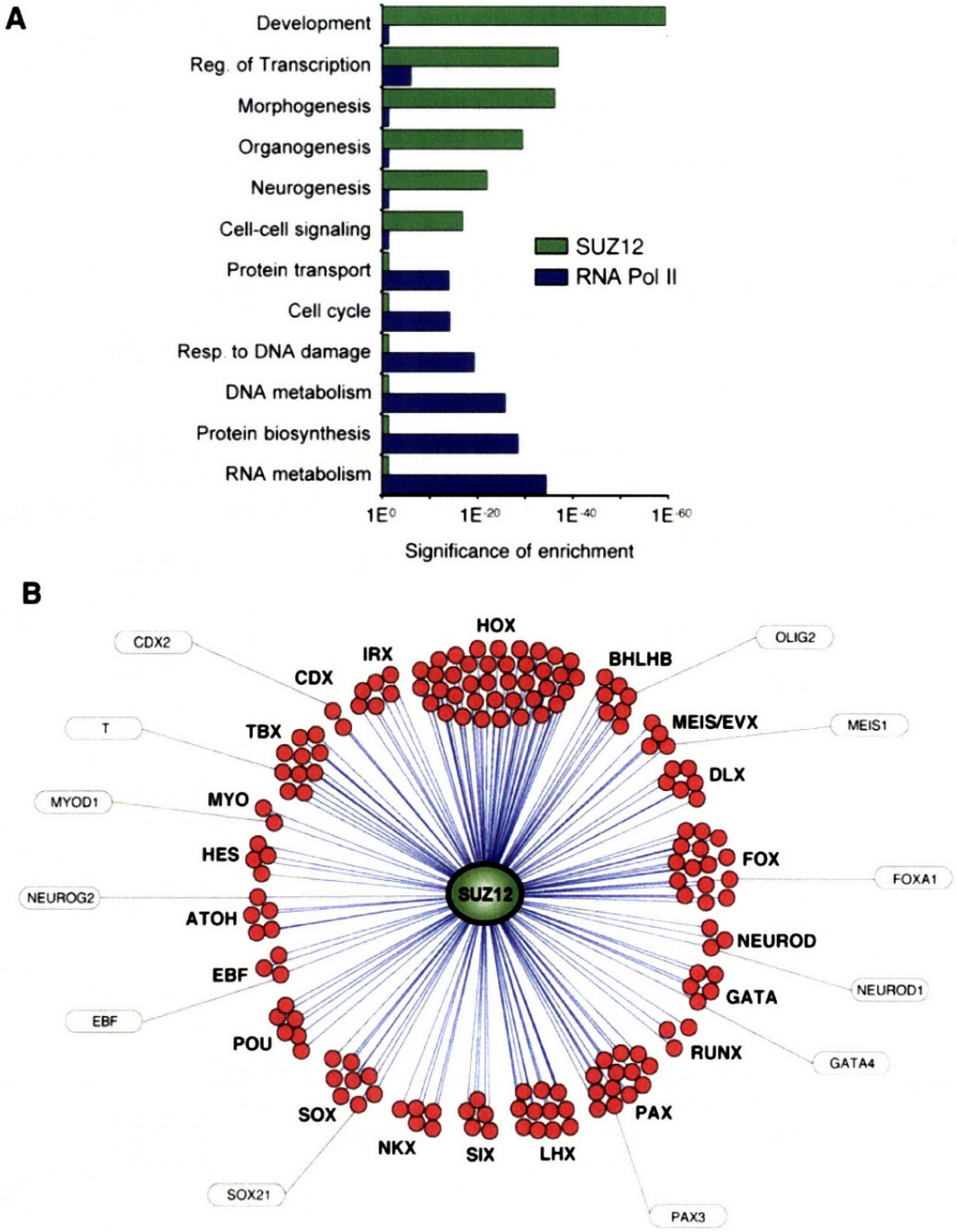

A
Development
Reg. of Transcription
Morphogenesis
Organogenesis
Neurogenesis
Cell-cell signaling
Protein transport
Cell cycle
Resp. to DNA damage
DNA metabolism
Protein biosynthesis
RNA metabolism
SUZ12
RNA Pol II
1E0
1E-20
1E-40
1E-60
Significance of enrichment
B
SUZ12
HOX
IRX
CDX
TBX
MYO
HES
ATOH
EBF
POU
SOX
NKX
SIX
LHX
PAX
RUNX
GATA
NEUROD
FOX
MEIS/EVX
DLX
BHLHB
CDX2
T
MYOD1
NEUROG2
EBF
SOX21
OLIG2
MEIS1
FOXA1
NEUROD1
GATA4
PAX3

Figure 3. Cellular Functions of Genes Occupied by SUZ12
(A) Genes bound by SUZ12 or RNA polymerase II were compared to biological process gene ontology categories; highly represented categories are shown. Ontology terms are shown on the y axis; p-values for the significance of enrichment are graphed along the x axis (SUZ12 in green, RNA polymerase II in blue).
(B) Selected examples of developmental transcription factor families bound by SUZ12. SUZ12 is represented by the green oval; individual transcription factors are represented by circles and grouped by family as indicated. Examples of transcription factors with defined roles in development are labeled. Transcription factor families include homeobox protein (HOX), basic helix-loop-helix domain containing, class B (BHLHB), HOX cofactors (MEIS/EVX), distal-less homeobox (DLX), Forkhead box (FOX), NEUROD, GATA binding protein (GATA), runt related transcription factor (RUNX), paired box and paired-like (PAX), LIM homeobox (LHX), sine oculis homeobox homolog (SIX), NK transcription factor related (NKX), SRY box (SOX), POU domain containing, classes 3 and 4 (POU), early B-cell factor (EBF), atonal homolog (ATOH), hairy and enhancer of split protein (HES), myogenic basic domain (MYO), T-box (TBX), caudal type homeobox (CDX), and iroquois homeobox protein (IRX).

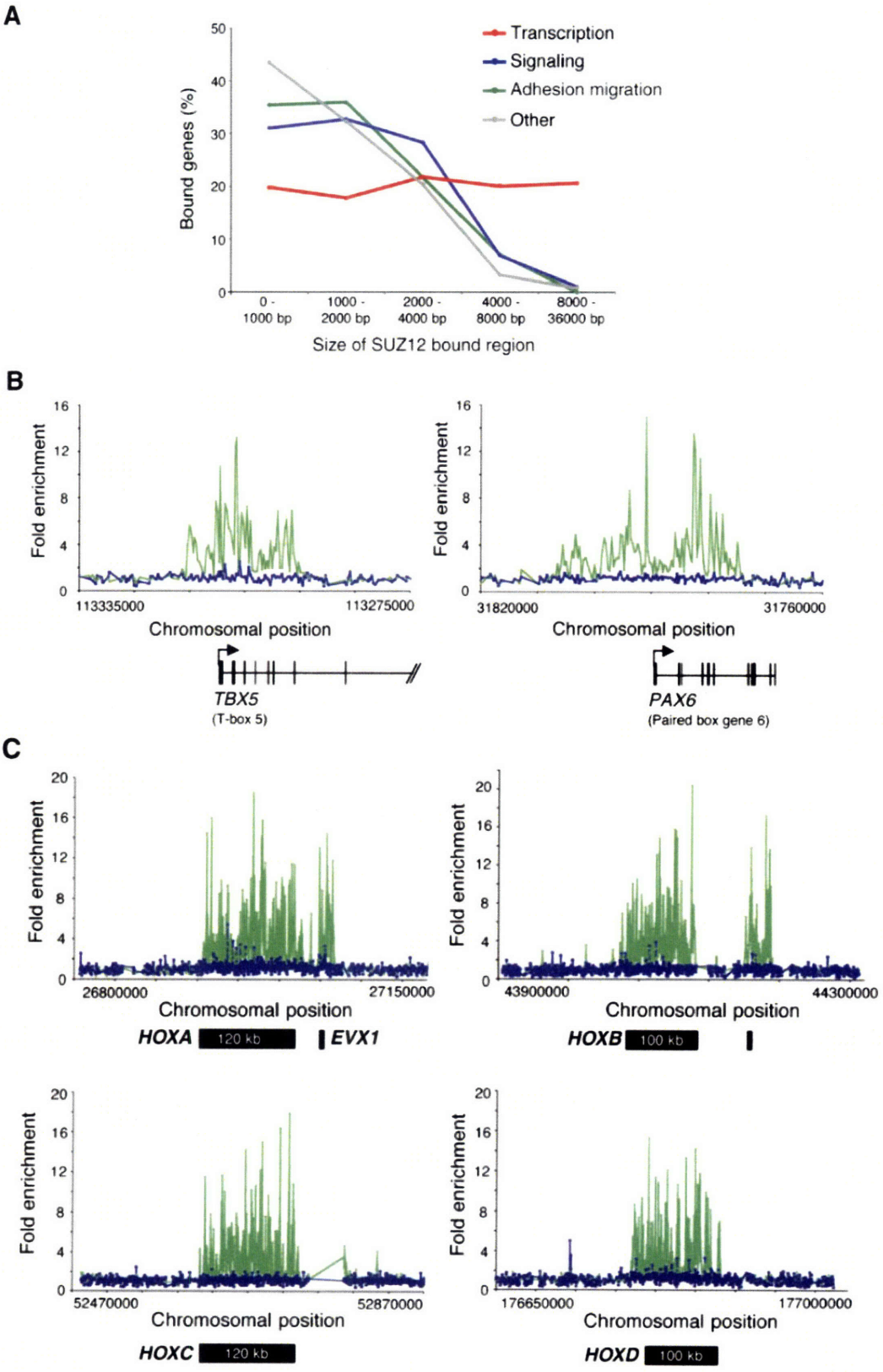

A
Transcription
Signaling
Adhesion migration
Other
Bound genes (%)
50
40
30
20
10
0
0 - 1000 bp
1000 - 2000 bp
2000 - 4000 bp
4000 - 8000 bp
8000 - 36000 bp
Size of SUZ12 bound region
B
Fold enrichment
16
12
8
4
0
113335000
113275000
Chromosomal position
TBX5
(T-box 5)
31820000
31760000
Chromosomal position
PAX6
(Paired box gene 6)
C
Fold enrichment
20
16
12
8
4
0
26800000
27150000
Chromosomal position
HOXA
120 kb
EVX1
43900000
44300000
Chromosomal position
HOXB
100 kb
52470000
52870000
Chromosomal position
HOXC
120 kb
176650000
177000000
Chromosomal position
HOXD
100 kb

Figure 4. SUZ12 Occupies Large Portions of Genes Encoding Transcription Factors with Roles in Development

(A) The fraction of SUZ12 target genes associated with different sizes of binding domains. Genes are grouped into four categories according to their function: Signaling, Adhesion/ migration, Transcription, and Other.

(B) Examples of SUZ12 (green) and RNA polymerase II (blue) binding at the genes encoding developmental regulators TBX5 and PAX6. The plots show unprocessed enrichment ratios for all probes within a genomic region (ChIP versus whole genomic DNA). Genes are shown to scale below plots (exons are represented by vertical bars). The start and direction of transcription are noted by arrows.

(C) Binding profiles of SUZ12 (green) and RNA polymerase II (blue) across 500 kb regions encompassing HOX clusters A–D. Unprocessed enrichment ratios for all probes within a genomic region are shown (ChIP versus whole genomic DNA). Approximate HOX cluster region sizes are indicated within black bars.

neurogenesis, hematopoiesis, axial patterning, tissue patterning, organogenesis, and cellfate specification. SUZ12 also occupied promoters for large subsets of the FOX, SOX, and TBX gene families. The forkhead family of FOX genes is involved in axial patterning and tissue development from all three germ layers (Lehmann et al., 2003). Mutations in members of the SOX gene family alter cell-fate specification and differentiation and are linked to several developmental diseases (Schepers et al., 2002). The TBX family of genes regulates a wide variety of developmental processes such as gastrulation, early pattern formation, organogenesis, and limb formation (Showell et al., 2004). Thus, the genes preferentially bound by SUZ12 have functions that, when expressed, promote differentiation. This is likely to explain, at least in part, why PRC2 is essential for early development and ES cell pluripotency.

A remarkable feature of PRC2 binding at most genes encoding developmental regulators was the extensive span over which the regulator occupied the locus (Figures 4, S8, and S9). For the majority (72%) of bound sites across the genome, SUZ12 occupied a small region of the promoter similar in size to regions bound by RNA polymerase II (Figure 1). For the remaining bound regions, SUZ12 occupancy encompassed large domains spanning 2–35 kb and extending from the promoter into the gene. A large portion of genes encoding developmental regulators (72%) exhibited these extended regions of SUZ12 binding. In some cases, binding encompassed multiple contiguous genes. For instance, SUZ12 binding extended 100 kb across the entire HOXA, HOXB, HOXC, and HOXD clusters but did not bind to adjacent genomic sequences, yielding a highly defined spatial pattern (Figure 4B). In contrast, clusters of unrelated genes, such as the interleukin 1-b cluster, were not similarly bound by SUZ12. Thus, genes encoding developmental regulators showed an unusual tendency to be occupied by PRC2 over much or all of their transcribed regions.

PRC2 and Highly Conserved Elements

Previous studies have noted that many highly conserved noncoding elements of vertebrate genomes are associated with genes encoding developmental regulators (Bejerano et al., 2004; Siepel et al., 2005; Woolfe et al., 2005). Given SUZ12's strong association with this class of genes, we investigated the possibility that SUZ12 bound regions are associated with these highly conserved elements. Inspection of individual genes suggested that SUZ12 occupancy was associated with regions of sequence conservation (Figure 5A). Eight percent of the approximately 1,400 highly conserved noncoding DNA elements described by Woolfe and colleagues (Woolfe et al., 2005) were found to be associated with the SUZ12 bound developmental regulators (p-value 10 14). Using entries from the PhastCons database of conserved elements (Siepel et al., 2005), we found that SUZ12 occupancy of highly conserved elements was highly significant (using highly conserved elements with a LoD conservation score of 100 or better, the p-value for significances was less than 10^{-85}). Since PRC2 has not been shown to directly bind DNA sequences, we expect that specific DNA binding proteins occupy the highly conserved DNA sequences and may associate with PRC2, which spreads and occupies adjacent chromatin. Thus, the peaks of SUZ12 occupancy might not be expected to precisely collocate with the highly conserved elements, even if these elements are associated with PRC2 recruitment.

Remarkably, the degree of the association between SUZ12 binding and conserved

sequences increases when considering sequences with an increasing degree of conservation (Figure 5B). By comparison, RNA polymerase II showed no such enrichment. These results suggest that the subset of highly conserved noncoding elements at genes encoding developmental regulators may be associated with PcG-mediated silencing of these regulators.

Signaling Genes Are among PRC2 Targets

The targets of SUZ12 were also enriched for genes that encode components of signaling pathways (Figure 3A and Table S12). There is evidence that transforming growth factor-b (TGFb), bone morphogenic protein (BMP), wingless-type MMTV integration site (Wnt), and fibroblast growth factor (FGF) signaling pathways, which are required for gastrulation and lineage differentiation in the embryo, are also essential for self-renewal and differentiation of ES cells in culture (Loebel et al., 2003; Molofsky et al., 2004). SUZ12 generally occupied the promoters of multiple components of these pathways, but it occupied larger domains within a group of signaling genes that contained highly conserved elements. This group contained members of the Wnt family (WNT1, WNT2, WNT6) as well as components of the TGFb superfamily (BMP2, GDF6). Recent studies have shown that Wnt signaling plays a role in pluripotency and self-renewal in both mouse and human ES cells (Sato et al., 2004), and our results suggest that it is important to maintain specific family members in a repressed state in ES cells.

Activation of PRC2 Target Genes during Differentiation

PRC2 is associated with an important set of developmental regulators that must be silent in ES cells but activated during differentiation. This observation suggests that PRC2 ultimately functions to repress occupied genes in ES cells and that these genes may be especially poised for transcriptional activation during ES cell differentiation. We reasoned that if this model is correct, genes bound by SUZ12 should be preferentially activated upon ES cell differentiation or in cells that lack SUZ12. Furthermore, in differentiated cells, SUZ12 might continue to be observed at silent genes but must be removed from genes whose expression is essential for that cell type.

We first examined gene expression in ES cells stimulated to undergo differentiation (Sato et al., 2003). We found that genes occupied by SUZ12 were more likely to be activated during ES cell differentiation than genes that were not occupied by SUZ12 (Figure 6A; Supplemental Data; Table S13), indicating that SUZ12-occupied genes show preferential activation during differentiation under these conditions. Thirty-six percent of genes bound by SUZ12 showed greater than 2-fold increases in expression during ES cell differentiation, whereas only 16% of genes not bound by SUZ12 showed such an increase. This effect was particularly striking at the set of developmental regulators (Figure 6B). SUZ12 occupied most (83%) of the developmental regulators that were induced more than 10-fold during ES cell differentiation.

We next examined the expression of SUZ12 target genes in Suz12-deficient cell lines derived from homozygous mutant blastocysts (Supplemental Data). We reasoned that genes bound by SUZ12 in human ES cells have orthologs in mice that should be upregulated in Suz12-deficient mouse cells, although we expected the overlap in these sets of genes to be imperfect because of potential differences between human and mouse ES cells, the possible repression of PRC2 target genes by additional mechanisms, and

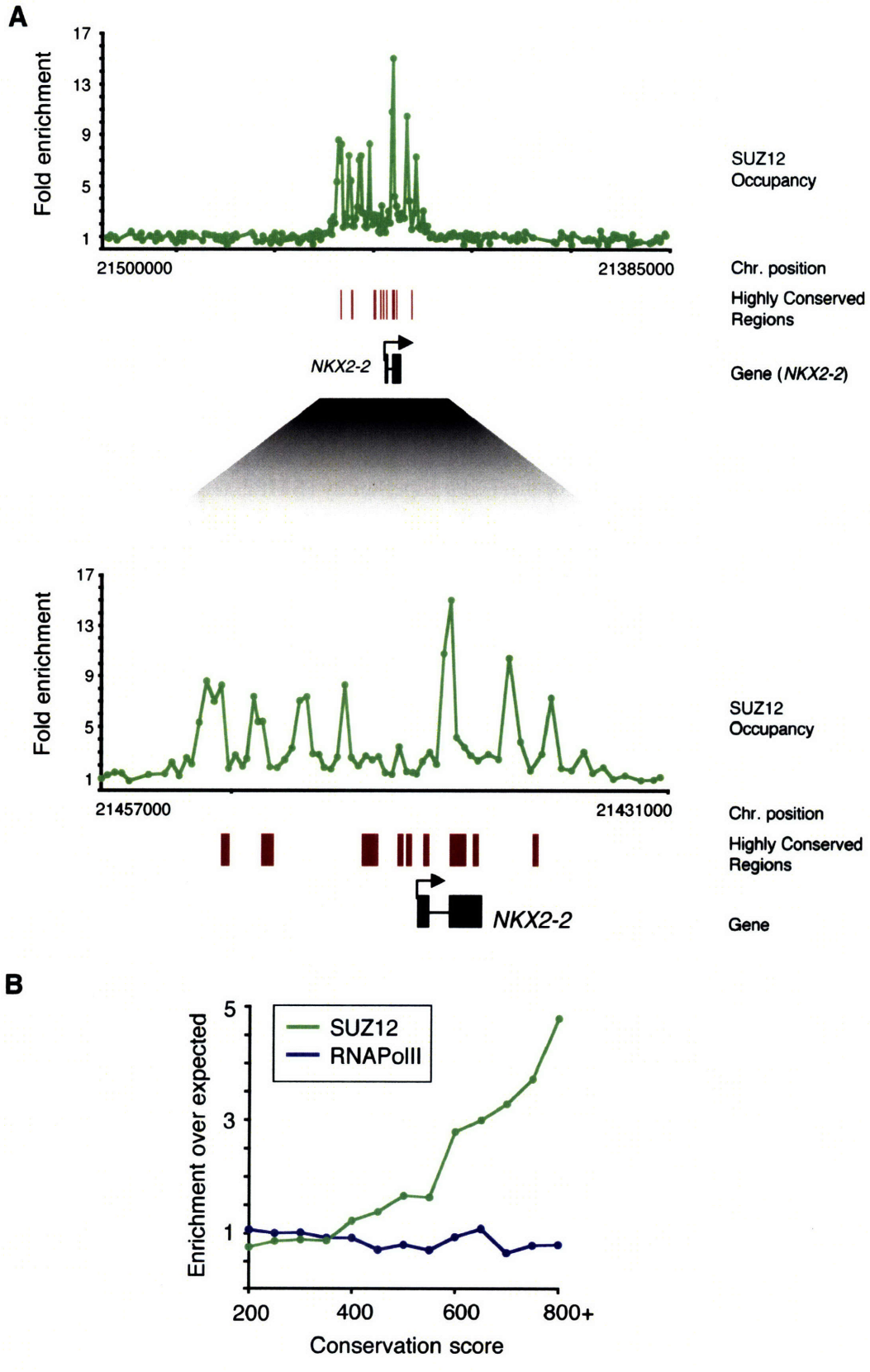

A
Fold enrichment
17
13
9
5
1
21500000
21385000
SUZ12
Occupancy
Chr. position
Highly Conserved
Regions
NKX2-2
Gene (NKX2-2)
Fold enrichment
17
13
9
5
1
21457000
21431000
SUZ12
Occupancy
Chr. position
Highly Conserved
Regions
NKX2-2
Gene
B
Enrichment over expected
5
3
1
SUZ12
RNAPolII
200
400
600
800+
Conservation score

Figure 5. SUZ12 Binding Is Associated with Highly Conserved Regions
(A) SUZ12 occupancy (green) and conserved elements are shown at NKX2-2 and adjacent genomic regions. The plots show unprocessed enrichment ratios for all probes within this genomic region (SUZ12 ChIP versus whole genomic DNA). Conserved elements (red) with LoD scores > 160 derived from the PhastCons program (Siepel et al., 2005) are shown to scale above the plot. Genes are shown to scale below plots (exons are represented by vertical bars). A higher resolution view is also shown below.
(B) Enrichment of conserved noncoding elements within SUZ12 (green) and RNA polymerase II (blue) bound regions. The maximum nonexonic PhastCons conservation score was determined for each bound region. For comparison, the same parameter was determined using a randomized set of genomic regions with the same size distribution. The graph displays the ratio of the number of bound regions with that score versus the number of randomized genomic regions with that score.

Figure 6

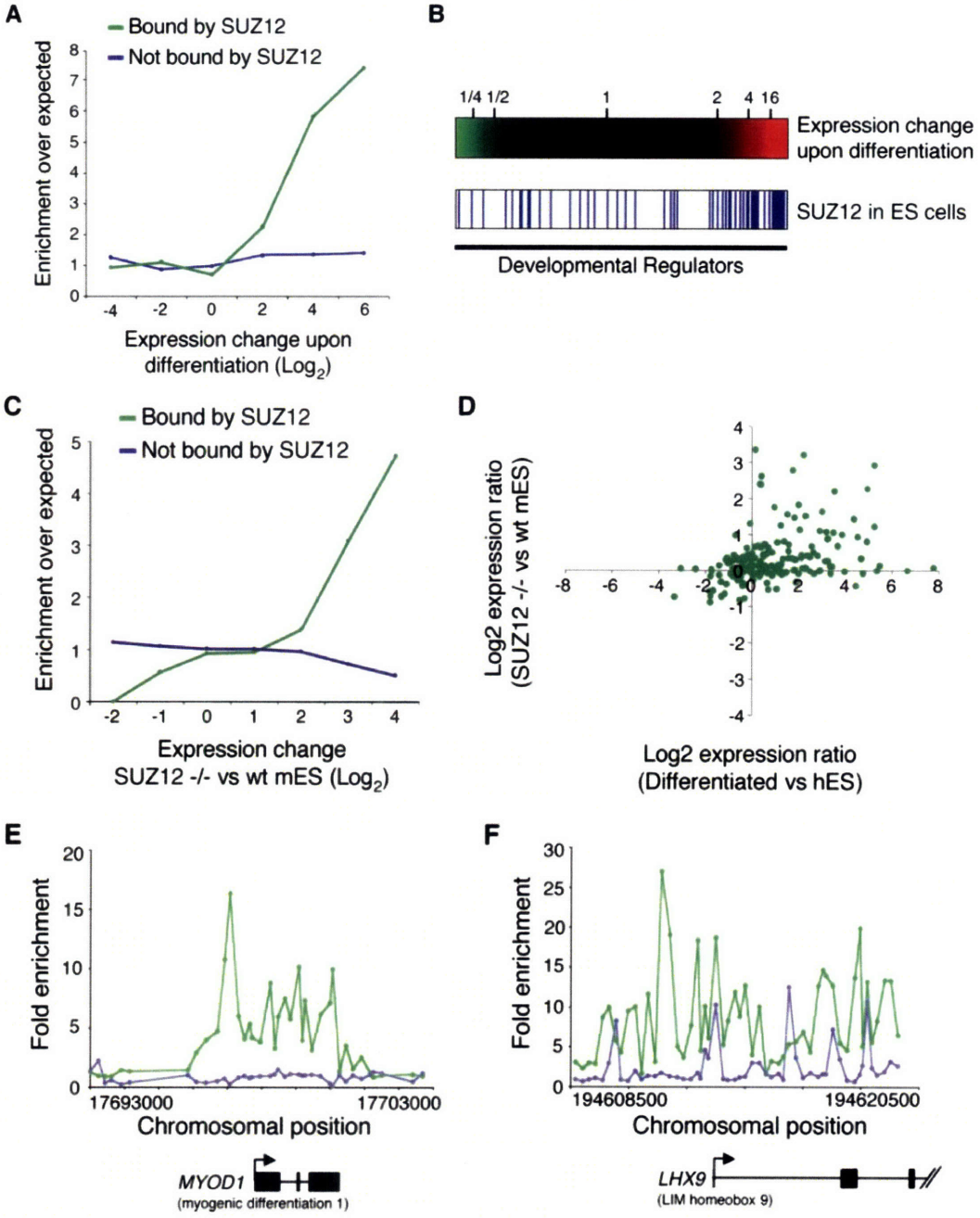

A
Bound by SUZ12
Not bound by SUZ12
Enrichment over expected
Expression change upon differentiation (Log2)
B
1/4 1/2
1
2
4 16
Expression change upon differentiation
SUZ12 in ES cells
Developmental Regulators
C
Bound by SUZ12
Not bound by SUZ12
Enrichment over expected
Expression change SUZ12 -/- vs wt mES (Log2)
D
Log2 expression ratio (SUZ12 -/- vs wt mES)
Log2 expression ratio (Differentiated vs hES)
E
Fold enrichment
17693000
17703000
Chromosomal position
MYOD1
(myogenic differentiation 1)
F
Fold enrichment
194608500
194620500
Chromosomal position
LHX9
(LIM homeobox 9)

Figure 6. Preferential Activation of PRC2 Target Genes during ES Cell Differentiation
(A) Fold enrichment in the number of genes induced or repressed during ES cell differentiation. The change in gene expression is given as the log(2) transformed ratio of the signals in differentiated H1 cells versus pluripotent H1 cells and is binned into six groups. The upper limit of each bin is indicated on the x axis. The two lines show genes transcriptionally inactive in ES cells (absence of RNA polymerase II) and bound by SUZ12 (green) and genes transcriptionally inactive in ES cells and repressed by other means (blue). In both cases, fold enrichment is calculated against the total population of genes and normalized for the number of genes present in each group.
(B) Expression changes of genes encoding developmental regulators during ES cell differentiation. Expression ratio (differentiated/pluripotent) is represented by color, with shades of red indicating upregulation and shades of green downregulation according to the scale shown above. Genes are ordered according to change in gene expression, with genes exhibiting higher expression in pluripotent ES cells to the left and genes exhibiting higher expression in differentiated cells to the right. Genes bound by SUZ12 in undifferentiated ES cells are indicated by blue lines in the lower panel.
(C) Fold enrichment in the number of genes induced or repressed in SUZ12-deficient mouse cells. The change in gene expression is given as the log(2) transformed ratio of the signals in Suz12-deficient cells versus wild-type ES cells. The two lines show genes transcriptionally inactive in human ES cells (absence of RNA polymerase II) and bound by SUZ12 (green) and genes transcriptionally inactive in human ES cells and repressed by other means (blue). In both cases, fold enrichment is calculated against the total population of genes.
(D) Gene expression ratios (log base 2) of Suz12 target genes in differentiated human H1 ES cells relative to pluripotent H1 ES cells (x axis) and in Suz12-deficient mouse cells relative to wild-type mouse ES cells (y axis). Upper right quadrant: genes upregulated during human ES cell differentiation and in Suz12-deficient mouse cells; lower right: genes upregulated during ES cell differentiation and downregulated in Suz12-deficient cells; lower left: genes downregulated during ES cell differentiation and in Suz12-deficient cells; upper left: genes downregulated during ES cell differentiation and upregulated in Suz12-deficient cells.
(E) SUZ12 binding profiles across the gene encoding muscle regulator MYOD1 in H9 human ES cells (green) and primary human skeletal myotubes (gray). The plots show unprocessed enrichment ratios for all probes within a genomic region (ChIP versus whole genomic DNA). Genes are shown to scale below plots (exons are represented by vertical bars). The start and direction of transcription are noted by arrows.
(F) Suz12 binding profiles across the gene encoding LHX9 in H9 human ES cells (green) and primary human skeletal myotubes (gray). The plots show unprocessed enrichment ratios for all probes within a genomic region (ChIP versus whole genomic DNA). Genes are shown to scale below plots (exons are represented by vertical bars). The start and direction of transcription are noted by arrows.

pleiotropic effects of the Suz12 knockout on genes downstream of Suz12-target genes. Differences in gene expression between Suz12 homozygous mutant cells and wild-type ES cells were measured using gene expression microarrays and the human SUZ12 binding data mapped to orthologous mouse genes using HomoloGene (www.ncbi.nlm.nih.gov/HomoloGene). We found that a significant portion of mouse genes whose counterparts were bound by SUZ12 in human ES cells were upregulated in Suz12-deficient mouse cells (70 of 346 genes, p = 6 10 4); these genes are listed in Table S14. Orthologs of genes occupied by SUZ12 in human ES cells were more likely to be activated and less likely to be repressed in Suz12-deficient mouse cells than orthologs of genes not occupied by SUZ12 (Figure 6C). Furthermore, we found that orthologs of Suz12 target genes that were induced upon human ES cell differentiation were generally also induced upon loss of Suz12 in mouse cells (Figure 6D). Genes that were activated during ES cell differentiation and in Suz12-deficient cells included those encoding transcriptional regulators (GATA2, GATA3, GATA6, HAND1, MEIS2, and SOX17) signaling proteins (WNT5A, DKK1, DKK2, EFNA1, EFNB1, EPHA4, and EPHB3) and the cell-cycle inhibitor CDKN1A. These data indicate that Suz12 is necessary to fully repress the genes that are occupied by PRC2 in wild-type ES cells and have since been confirmed with binding data and knockout studies of a second PRC subunit in mouse (Boyer et al., 2006).

If PRC2 functions to repress genes in ES cells that are activated during differentiation, then in differentiated tissues SUZ12 occupancy should be diminished at genes encoding developmental regulators that have a role in specifying the identity of that tissue, similar to results seen with Ezh2 at specific genes in mouse (Caretti et al., 2004). To test this, we designed an array focused on the promoters of developmental regulators and used ChIP-Chip to investigate SUZ12 occupancy at these promoters in primary differentiated muscle cells. The results demonstrated that genes encoding key regulators of muscle differentiation, including MYOD1, displayed greatly diminished SUZ12 occupancy when compared to ES cells (Figure 6E). MYOD1 is a master regulator for muscle differentiation (Tapscott, 2005), and the gene encoding this transcription factor displayed no significant SUZ12 occupancy when compared to the levels of SUZ12 occupancy observed in ES cells. Genes encoding other transcriptional regulators that play a central role in muscle development, such as PAX3 and PAX7 (Brand- Saberi, 2005), showed reduced levels of SUZ12 occupancy in muscle cells relative to ES cells (Supplemental Data and Figure S11). In contrast, other developmental regulators important for differentiation of nonmuscle tissues remained occupied by SUZ12 in differentiated muscle cells (Figure 6F and Table S15). These data support a model where PRC2 binding in ES cells represses key developmental regulators that are later expressed during differentiation.

Targets of PRC2 Are Shared with Key ES Cell Regulators

The transcription factors OCT4, SOX2, and NANOG have essential roles in early development and are required for the propagation of undifferentiated ES cells in culture (Nichols et al., 1998; Avilion et al., 2003; Chambers et al., 2003; Mitsui et al., 2003). We recently reported that these transcription factors occupied promoters for many important developmental regulators in human ES cells (Boyer et al., 2005). This led us to compare the set of genes encoding developmental regulators and occupied by

OCT4, SOX2, and NANOG with those occupied by PRC2 (Figure 7 and Supplemental Data). We found that each of the three DNA binding transcription factors occupied approximately one-third of the PRC2-occupied genes that encode developmental transcription factors (Figure 7A; Supplemental Data; Table S11). Remarkably, we found that the subset of genes encoding developmental regulators that were occupied by OCT4, SOX2, and NANOG and repressed in the regulatory circuitry highlighted in Boyer et al. were almost all occupied by PRC2 (Figure 7B). These included genes for transcription factors known to be important for differentiation into extraembryonic, endodermal, mesodermal, and ectodermal lineages (e.g., ESX1L, ONECUT1, HAND1, HOXB1). As expected, active genes encoding ES cell transcription factors (e.g., ZIC3, STAT3, OCT4, NANOG) were occupied by OCT4, SOX2, NANOG, and RNA polymerase II but not by PRC2 (Figure 7B).

The observation that OCT4, SOX2, and NANOG are bound to a significant subset of developmental genes occupied by PRC2 supports a link between repression of developmental regulators and stem cell pluripotency. Like PRC2, OCT4 and NANOG have been shown to be important for early development and ES cell identity. It is possible, therefore, that inappropriate regulation of developmental regulators that are common targets of OCT4, NANOG, and PRC2 contributes to the inability to establish ES cell lines in OCT4, NANOG, and EZH2 mutants (Nichols et al., 1998; O'Carroll et al., 2001; Chambers et al., 2003; Mitsui et al., 2003).

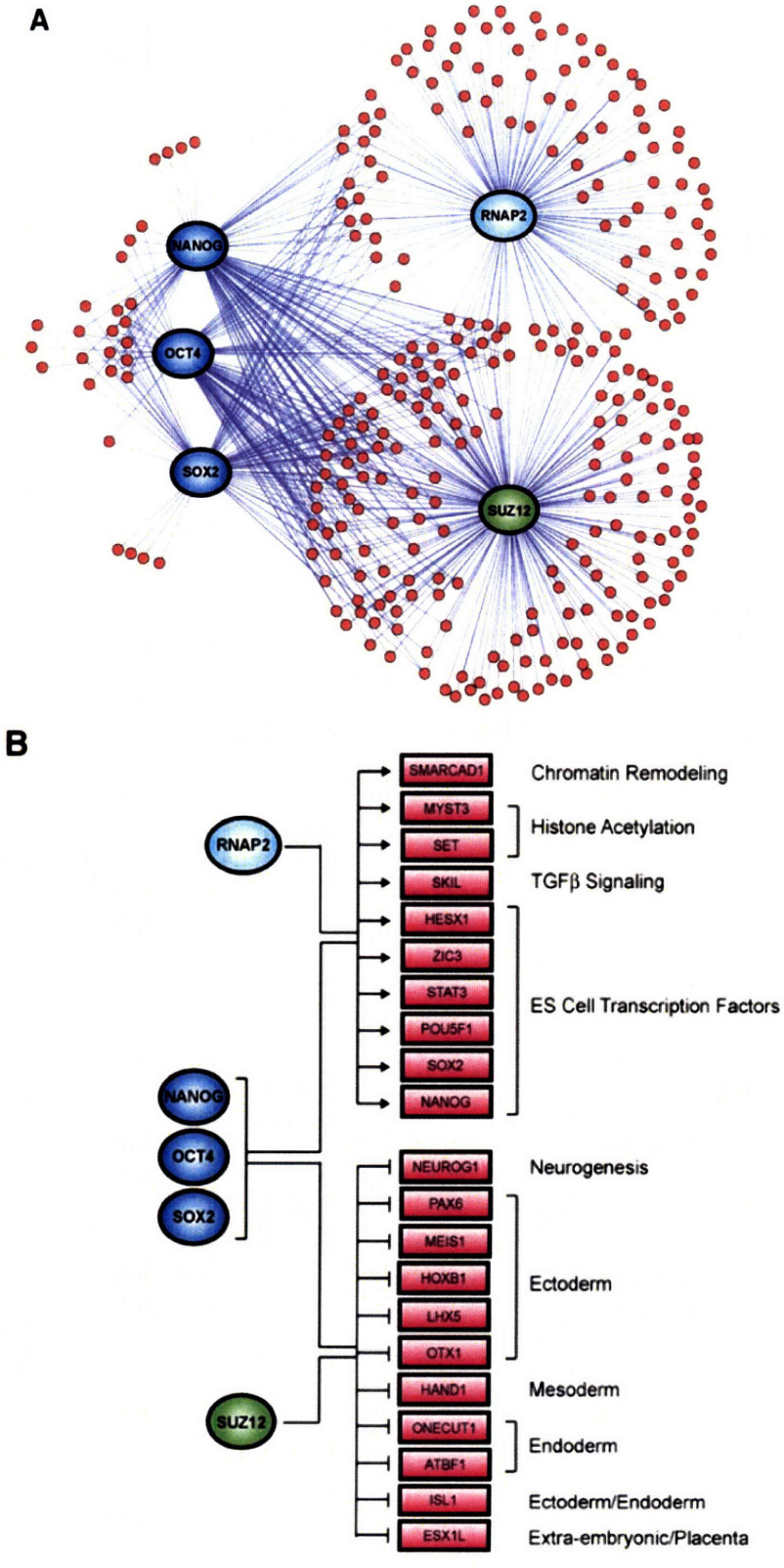

A
NANOG
OCT4
SOX2
RNAP2
SUZ12
B
RNAP2
NANOG
OCT4
SOX2
SUZ12
SMARCAD1
MYST3
SET
SKIL
HESX1
ZIC3
STAT3
POU5F1
SOX2
NANOG
NEUROG1
PAX6
MEIS1
HOXB1
LHX5
OTX1
HAND1
ONECUT1
ATBF1
ISL1
ESX1L
Chromatin Remodeling
Histone Acetylation
TGFβ Signaling
ES Cell Transcription Factors
Neurogenesis
Ectoderm
Mesoderm
Endoderm
Ectoderm/Endoderm
Extra-embryonic/Placenta

Figure 7. SUZ12 Is Localized to Genes also Bound by ES Cell Transcriptional Regulators
(A) Transcriptional regulatory network model of developmental regulators governed by OCT4, SOX2, NANOG, RNA polymerase II, and SUZ12 in human ES cells. The ES cell transcription factors each bound to approximately one-third of the PRC2-occupied, developmental transcription factor genes. Developmental regulators were selected based on gene ontology. Regulators are represented by dark blue circles; RNA polymerase II is represented by a light blue circle; SUZ12 is represented by a green circle; gene promoters for developmental regulators are represented by small red circles.
(B) SUZ12 occupies a set of repressed developmental regulators also bound by OCT4, SOX2, and NANOG in human ES cells. Genes annotated as bound by OCT4, SOX2, and NANOG previously and identified as active or repressed based on expression data (Boyer et al., 2005) were tested to see if they were bound by SUZ12 or RNA polymerase II. Ten of eleven previously identified active genes were found to be bound by RNA polymerase II at known promoters, while eleven of twelve previously identified repressed genes were bound by SUZ12. Regulators are represented by dark blue circles, RNA polymerase II by a light blue circle, and SUZ12 by a green circle. Gene promoters are represented by red rectangles.

Concluding Remarks

We have mapped the sites occupied by SUZ12 throughout the genome to gain insights into how PRC2 contributes to pluripotency in human embryonic stem cells. ES cells proliferate in an undifferentiated state yet remain poised to respond to development cues. Genes encoding the transcriptional regulators that promote differentiation must therefore be repressed in ES cells but activated upon receiving signals to differentiate. We found that PRC2 occupies large domains at genes encoding a key set of repressed developmental regulators that are preferentially activated upon cellular differentiation, thus implicating this complex directly in the maintenance of the pluripotent state.

Transcription factors and chromatin regulators contribute to the transcriptional regulatory circuitry responsible for pluripotency and self-renewal in human ES cells. Understanding this circuitry is fundamental to understanding human development and realizing the therapeutic potential of these cells. In this context, we find it exciting that the outlines of the core transcriptional regulatory circuitry of human ES cells are emerging. The transcription factors OCT4, SOX2, and NANOG are associated with actively transcribed genes that contribute to growth and self-renewal (Boyer et al., 2005). These factors also occupy genes encoding key developmental regulators that are transcriptionally repressed, due at least in part to their association with PRC2 and nucleosomes modified at histone H3K27me3. Further study of transcription factors and chromatin regulators genome-wide will allow investigators to produce a more comprehensive map of transcriptional regulatory circuitry in ES cells and to test models that emerge from the circuitry. This information may provide insights into approaches by which pluripotent cells can be stimulated to differentiate into different cell types.

Experimental Procedures

Cells and Cell Culture

Human H9 ES cells (WiCell, Madison, WI) were cultured as described (Boyer et al., 2005). Primary human skeletal muscle cells were obtained from Cell Applications (San Diego, CA) and expanded and differentiated into myotubes according to the supplier's protocols. Suz12 / mouse cell lines were derived from blastocysts from crosses between heterozygous Suz12 mutant animals, as described in Supplemental Data.

Chromatin Immunoprecipitation and DNA Microarray Analysis

ChIP was combined with DNA microarray analysis as described (Boyer et al., 2005). The antibodies used here were specific for hypophosphorylated RNA polymerase II (8WG16) (Thompson et al., 1989), SUZ12 (Upstate, 07-379), EED (Hamer et al., 2002), H3K27me3 (Abcam, AB6002), and total histone H3 (Abcam, AB1791). The design of the oligo-based arrays, which were manufactured by Agilent Technologies, is described in detail in Supplemental Data. A whole-chip error model was used to calculate confidence values from the enrichment ratio and the signal intensity of each probe (probe p-value) and of each set of three neighboring probes (probe-set p-value). Probe-sets with significant probe-set p-values ($p < 0.001$) and significant individual probe p-values were judged to be bound (see Supplemental Data for additional information). Bound regions were assigned to genes if they were within 1 kb of the transcription start site from one of five genomic databases; RefSeq, MGC, Ensembl, UCSC Known Gene, or H-Inv. All microarray data is available at ArrayExpress under the accession designation E-WMIT-7.

Gene Expression Analysis

Gene expression data were collated from H1 ES cells (Sato et al., 2003), H9, HSF1, and HSF6 ES cells (Abeyta et al., 2004), and 79 differentiated human cell and tissue types (Su et al., 2004) and analyzed as described in detail in Supplemental Data. Replicate gene expression data was obtained for wild-type mouse ES cells and Suz12- deficient cells using Agilent Mouse Development arrays and were analyzed as described in Supplemental Data.

Supplemental Data

Supplemental Data include fifteen figures, fifteen tables, Experimental Procedures, and References and can be found with this article online at http://www.cell.com/cgi/content/full/125/2/301/DC1/.

Acknowledgments

We thank Elizabeth Jacobsen for technical assistance and Robert Brady for help with array design. L.A.B. and H.L.M. were supported by NRSA postdoctoral fellowships. M.G.G. is an Amgen Fellow of LSRF. R.M.K. was supported by a fellowship from the ACS. D.T.O. was supported by NIH award DK070813. This work was supported by NIH grants HG002668 and GM069400. T.L., T.L.V., D.K.G., and R.A.Y. consult for Agilent Technologies.

References

Abeyta, M.J., Clark, A.T., Rodriguez, R.T., Bodnar, M.S., Pera, R.A., and Firpo, M.T. (2004). Unique gene expression signatures of independently- derived human embryonic stem cell lines. Hum. Mol. Genet. 13, 601–608.

Akasaka, T., van Lohuizen, M., van der Lugt, N., Mizutani-Koseki, Y., Kanno, M., Taniguchi,M., Vidal,M., Alkema, M., Berns, A., and Koseki, H. (2001). Mice doubly deficient for the Polycomb Group genes Mel18 and Bmi1 reveal synergy and requirement for maintenance but not initiation of Hox gene expression. Development 128, 1587–1597.

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

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved elements in the human genome. Science 304, 1321–1325.

Bender, M., Turner, F.R., and Kaufman, T.C. (1987). A development genetic analysis of the gene regulator of postbithorax in Drosophila melanogaster. Dev. Biol. 119, 418–432.

Birve, A., Sengupta, A.K., Beuchle, D., Larsson, J., Kennison, J.A., Rasmuson-Lestander, A., and Muller, J. (2001). Su(z)12, a novel Drosophila Polycomb group gene that is conserved in vertebrates and plants. Development 128, 3371–3379.

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., Otte, A.P., Vidal,M., Gifford, D.K., Young, R.A., and Jaenisch, R. (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature, in press.

Brand-Saberi, B. (2005). Genetic and epigenetic control of skeletal muscle development. Ann. Anat. 187, 199–207.

Breiling, A., Turner, B.M., Bianchi, M.E., and Orlando, V. (2001). General transcription factors bind promoters repressed by Polycomb group proteins. Nature 412, 651–655.

Cao, R., Wang, L.,Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R.S., and Zhang, Y. (2002). Role of histone H3 lysine 27 methylation in Polycomb-group silencing. Science 298, 1039–1043.

Cao, R., and Zhang, Y. (2004). SUZ12 is required for both the histone methyltransferase

activity and the silencing function of the EED-EZH2 complex. Mol. Cell 15, 57–67.

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

Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., and Smith, A. (2003). Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell 113, 643–655.

Czermin, B., Melfi, R., McCabe, D., Seitz, V., Imhof, A., and Pirrotta, V. (2002). Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. Cell 111, 185–196.

Dellino, G.I., Schwartz, Y.B., Farkas, G., McCabe, D., Elgin, S.C., and Pirrotta, V. (2004). Polycomb silencing blocks transcription initiation. Mol. Cell 13, 887–893.

Denell, R.E., and Frederick, R.D. (1983). Homoeosis in Drosophila: a description of the Polycomb lethal syndrome. Dev. Biol. 97, 34–47.

Duncan, I. (1986). Control of bithorax complex functions by the segmentation gene fushi tarazu of D. melanogaster. Cell 47, 297–309.

Faust, C., Lawson, K.A., Schork, N.J., Thiel, B., and Magnuson, T. (1998). The Polycomb-group gene eed is required for normal morphogenetic movements during gastrulation in the mouse embryo. Development 125, 4495–4506.

Francis, N.J., Saurin, A.J., Shao, Z., and Kingston, R.E. (2001). Reconstitution of a functional core polycomb repressive complex. Mol. Cell 8, 545–556.

Franke, A., DeCamillis, M., Zink, D., Cheng, N., Brock, H.W., and Paro, R. (1992). Polycomb and polyhomeotic are constituents of a multimeric protein complex in chromatin of Drosophila melanogaster. EMBO J. 11, 2941–2950.

Hamer, K.M., Sewalt, R.G., den Blaauwen, J.L., Hendrix, T., Satijn, D.P., and Otte, A.P. (2002). A panel of monoclonal antibodies against human polycomb group proteins. Hybrid. Hybridomics 21, 245–252.

Hodgson, J.W., Argiropoulos, B., and Brock, H.W. (2001). Site-specific recognition of a 70-base-pair element containing d(GA)(n) repeats mediates bithoraxoid polycomb group response element-dependent silencing. Mol. Cell. Biol. 21, 4528–4543.

Horard, B., Tatout, C., Poux, S., and Pirrotta, V. (2000). Structure of a polycomb response element and in vitro binding of polycomb group complexes containing GAGA factor. Mol. Cell. Biol. 20, 3187–3197.

Kennison, J.A. (2004). Introduction to Trx-G and Pc-G genes. Methods Enzymol. 377, 61–70.

Kirmizis, A., Bartley, S.M., Kuzmichev, A., Margueron, R., Reinberg, D., Green, R., and Farnham, P.J. (2004). Silencing of human polycomb target genes is associated with methylation of histone H3 Lys 27. Genes Dev. 18, 1592–1605.

Kuzmichev, A., Nishioka, K., Erdjument-Bromage, H., Tempst, P., and Reinberg, D. (2002). Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. Genes Dev. 16, 2893–2905.

Kuzmichev, A., Jenuwein, T., Tempst, P., and Reinberg, D. (2004). Different EZH2-containing complexes target methylation of histone H1 or nucleosomal histone H3. Mol. Cell 14, 183–193.

Kuzmichev, A., Margueron, R., Vaquero, A., Preissner, T.S., Scher, M., Kirmizis, A., Ouyang, X., Brockdorff, N., Abate-Shen, C., Farnham, P., and Reinberg, D. (2005). Composition and histone substrates of polycomb repressive group complexes change during cellular differentiation. Proc. Natl. Acad. Sci. USA 102, 1859–1864.

Lehmann, O.J., Sowden, J.C., Carlsson, P., Jordan, T., and Bhattacharya, S.S. (2003). Fox's in development and disease. Trends Genet. 19, 339–344.

Levine, S.S., Weiss, A., Erdjument-Bromage, H., Shao, Z., Tempst, P., and Kingston, R.E. (2002). The core of the polycomb repressive complex is compositionally and functionally conserved in flies and humans. Mol. Cell. Biol. 22, 6070–6078.

Levine, S.S., King, I.F., and Kingston, R.E. (2004). Division of labor in polycomb group repression. Trends Biochem. Sci. 29, 478–485.

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

Loebel, D.A., Watson, C.M., De Young, R.A., and Tam, P.P. (2003). Lineage choice and differentiation in mouse embryos and embryonic stem cells. Dev. Biol. 264, 1–14.

Lund, A.H., and van Lohuizen, M. (2004). Polycomb complexes and silencing mechanisms. Curr. Opin. Cell Biol. 16, 239–246.

Mayhall, E.A., Paffett-Lugassy, N., and Zon, L.I. (2004). The clinical potential of stem cells. Curr. Opin. Cell Biol. 16, 713–720.

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

Molofsky, A.V., Pardal, R., and Morrison, S.J. (2004). Diverse mechanisms regulate stem cell self-renewal. Curr. Opin. Cell Biol. 16, 700– 707.

Mulholland, N.M., King, I.F., and Kingston, R.E. (2003). Regulation of Polycomb group complexes by the sequence-specific DNA binding proteins Zeste and GAGA. Genes Dev. 17, 2741–2746.

Muller, J., Hart, C.M., Francis, N.J., Vargas, M.L., Sengupta, A., Wild, B.,Miller, E.L., O'Connor,M.B., Kingston, R.E., and Simon, J.A. (2002). Histone methyltransferase activity of a Drosophila Polycomb group repressor complex. Cell 111, 197–208.

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

O'Carroll, D., Erhardt, S., Pagani, M., Barton, S.C., Surani, M.A., and Jenuwein, T. (2001). The polycomb-group gene Ezh2 is required for early mouse development. Mol. Cell. Biol. 21, 4330–4336.

Orlando, V., and Paro, R. (1995). Chromatin multiprotein complexes involved in the maintenance of transcription patterns. Curr. Opin. Genet. Dev. 5, 174–179.

Pasini, D., Bracken, A.P., Jensen, M.R., Denchi, E.L., and Helin, K. (2004). Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity. EMBO J. 23, 4061–4071.

Pera, M.F., and Trounson, A.O. (2004). Human embryonic stem cells: prospects for development. Development 131, 5515–5525.

Pirrotta, V. (1998). Polycombing the genome: PcG, trxG, and chromatin silencing. Cell 93, 333–336.

Reubinoff, B.E., Pera, M.F., Fong, C.Y., Trounson, A., and Bongso, A. (2000). Embryonic stem cell lines from human blastocysts: somatic differentiation in vitro. Nat. Biotechnol. 18, 399–404.

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

Sato, N., Sanjuan, I.M., Heke, M., Uchida, M., Naef, F., and Brivanlou, A.H. (2003). Molecular signature of human embryonic stem cells and its comparison with the mouse. Dev. Biol. 260, 404–413.

Sato, N., Meijer, L., Skaltsounis, L., Greengard, P., and Brivanlou, A.H. (2004). Maintenance of pluripotency in human and mouse embryonic stem cells through

activation of Wnt signaling by a pharmacological GSK-3-specific inhibitor. Nat. Med. 10, 55–63.

Saurin, A.J., Shao, Z., Erdjument-Bromage, H., Tempst, P., and Kingston, R.E. (2001). A Drosophila Polycomb group complex includes Zeste and dTAFII proteins. Nature 412, 655–660.

Schepers, G.E., Teasdale, R.D., and Koopman, P. (2002). Twenty pairs of sox: extent, homology, and nomenclature of the mouse and human sox transcription factor gene families. Dev. Cell 3, 167–170.

Shao, Z., Raible, F., Mollaaghababa, R., Guyon, J.R., Wu, C.T., Bender, W., and Kingston, R.E. (1999). Stabilization of chromatin structure by PRC1, a Polycomb complex. Cell 98, 37–46.

Showell, C., Binder, O., and Conlon, F.L. (2004). T-box genes in early embryogenesis. Dev. Dyn. 229, 201–218.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15, 1034–1050.

Simon, J., Chiang, A., and Bender, W. (1992). Ten different Polycomb group genes are required for spatial control of the abdA and AbdB homeotic products. Development 114, 493–505.

Strutt, H., Cavalli, G., and Paro, R. (1997). Co-localization of Polycomb protein and GAGA factor on regulatory elements responsible for the maintenance of homeotic gene expression. EMBO J. 16, 3621–3632.

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. USA 101, 6062–6067.

Tapscott, S.J. (2005). The circuitry of a master switch: Myod and the regulation of skeletal muscle gene transcription. Development 132, 2685–2695.

Thompson, N.E., Steinberg, T.H., Aronson, D.B., and Burgess, R.R. (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, 11511–11520.

Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S., and Jones, J.M. (1998). Embryonic stem cell lines derived from human blastocysts. Science 282, 1145– 1147.

Tie, F., Furuyama, T., Prasad-Sinha, J., Jane, E., and Harte, P.J. (2001). The Drosophila Polycomb Group proteins ESC and E(Z) are present in a complex containing the histone-binding protein p55 and the histone deacetylase RPD3. Development 128, 275–286.

van der Lugt, N.M., Alkema, M., Berns, A., and Deschamps, J. (1996). The Polycomb-group homolog Bmi-1 is a regulator of murine Hox gene expression. Mech. Dev. 58, 153–164.

Vire, E., Brenner, C., Deplus, R., Blanchon, L., Fraga, M., Didelot, C., Morey, L., Van Eynde, A., Bernard, D., Vanderwinden, J.M., Bollen, M., Esteller, M., Di Croce, L., de Launoit, Y., and Fuks, F. (2006). The Polycomb group protein EZH2 directly controls DNA methylation. Nature 439, 871–874.

Wang, J., Mager, J., Schnedier, E., and Magnuson, T. (2002). The mouse PcG gene eed is required for Hox gene repression and extraembryonic development. Mamm. Genome 13, 493–503.

Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K., et al. (2005). Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. 3, e7.

Chapter 4

Tcf3 is an Integral Component of the Core Regulatory Circuitry of Embryonic Stem Cells

Published as: Megan F. Cole, Sarah E. Johnstone, Jamie J. Newman, Michael H. Kagey, and Richard A. Young (2008). "Tcf3 is an Integral Component of the Core Regulatory Circuitry of Embryonic Stem Cells." Genes Dev. 22: 746-755.

My contribution to this project

I initiated the effort within the Young Lab to study signaling pathways in embryonic stem cells, identifying key pathways to study and general experimental designs to be used. Along with labmates Sarah Johnstone and Jamie Newman, we led the study of Tcf3 and the Wnt pathway in embryonic stem cells. We additionally spearheaded the establishment of embryonic stem cell tissue culture within the Young Lab, including the foundation of a new tissue culture facility for the lab. I helped work out tissue culture conditions, performed all the ChIP-chip experiments and did the bulk of the computational analyses for this project. I also led the conceptual and experimental approaches taken for this work.

Abstract

Embryonic stem cells have a unique regulatory circuitry, largely controlled by the transcription factors Oct4, Sox2 and Nanog, which generates a gene expression program necessary for pluripotency and self-renewal (Boyer et al. 2005; Loh et al. 2006; Chambers et al. 2003; Mitsui et al. 2003; Nichols et al. 1998). How external signals connect to this regulatory circuitry to influence embryonic stem cell fate is not known. We report here that a terminal component of the canonical Wnt pathway in embryonic stem cells, the transcription factor Tcf3, co-occupies promoters throughout the genome in association with the pluripotency regulators Oct4 and Nanog. Thus Tcf3 is an integral component of the core regulatory circuitry of ES cells, which includes an autoregulatory loop involving the pluripotency regulators. Both *Tcf3* depletion and Wnt pathway activation cause increased expression of Oct4, Nanog and other pluripotency factors and produce ES cells that are refractory to differentiation. Our results suggest that the Wnt pathway, through Tcf3, brings developmental signals directly to the core regulatory circuitry of ES cells to influence the balance between pluripotency and differentiation.

Introduction

Embryonic stem (ES) cells provide a unique opportunity to study early development and hold great promise for regenerative medicine (Thomson et al. 1998; Reubinoff et al. 2000; Pera and Trounson 2004). ES cells are derived from the inner cell mass of the developing blastocyst and can be propagated in culture in an undifferentiated state while maintaining the capacity to generate any cell type in the body. Discovering how signaling pathways and transcriptional regulatory circuitry contribute to self-renewal and pluripotency is essential for understanding early development and realizing the therapeutic potential of ES cells.

A model for the core transcriptional regulatory circuitry of ES cells has emerged from studying the target genes of the ES cell transcription factors Oct4, Sox2 and Nanog (Boyer et al. 2005; Loh et al. 2006). These master regulators occupy the promoters of active genes encoding transcription factors, signal transduction components and chromatin modifying enzymes that promote ES cell selfrenewal. They also occupy the promoters of a large set of developmental transcription factors that are silent in ES cells, but whose expression is associated with lineage commitment and cellular differentiation. Polycomb Repressive Complexes co-occupy the genes encoding these developmental transcription factors to help maintain a silent transcriptional state in ES cells (Boyer et al. 2006; Lee et al. 2006; Wilkinson et al. 2006; Rajaskhar and Begemann 2007; Stock et al. 2007).

External signals can promote ES cell pluripotency or cause these cells to differentiate, but precisely how these pathways are connected to the ES cell regulatory network has not been determined. These signals are produced by the stem cell niche in the developing blastocyst or, for cultured ES cells, can be produced by added factors or serum to maintain stem cell identity or promote differentiation. Recent studies have demonstrated the importance of several signaling pathways in maintaining or modifying ES cell state, including the Activin/Nodal, Notch, BMP4 and Wnt pathways (Rao et al. 2004; Kristensen et al. 2005; Friel et al. 2005; Boiani and Scholer 2005; Valdimarsdottir and Mummery 2005; Dreesen and Brivanlou 2007; Pan and Thomson 2007). By understanding how these signaling pathways influence the gene expression program of ES cells, it should be possible to discover how they contribute to embryonic stem cell identity or promote specific differentiation programs.

The Wnt/β-catenin signaling pathway has multiple roles in embryonic stem cell biology, development and disease (Logan and Nusse 2004; Reya and Clevers 2005; Clevers 2006). Several studies have shown that activation of the Wnt pathway can cause ES cells to remain pluripotent under conditions that induce differentiation (Kielman et al. 2002; Sato et al. 2004; Singla et al. 2006; Hao et al. 2006; Ogawa et al. 2006; Miyabashi et al. 2007; Takao et al. 2007), while other studies have shown that the Wnt pathway has an important role in directing differentiation of ES cells (Otero et al. 2004; Lindsley et al. 2006). Recent studies have shown that T Cell Factor-3 (Tcf3), a terminal component of the Wnt pathway, acts to repress the *Nanog* gene in ES cells (Pereira et al. 2006), providing an important clue for at least one mechanism by which the Wnt pathway regulates stem cell state. Nonetheless, we have an incomplete understanding of how the pathway exerts its effects, in part because few target genes have been identified for its terminal components in ES cells.

Stimulation of the canonical Wnt signaling pathway causes the transcriptional co-activator β-catenin to translocate to the nucleus, where it interacts with constitutively DNA-bound Tcf/Lef proteins to activate target genes (Behrens et al. 1996; Brantjes et al. 2001; Cadigan 2002). Tcf3, a member of the Tcf/Lef family, is highly expressed in murine embryonic stem (mES) cells and is critical for early embryonic development (Korinek et al. 1998; Merrill et al. 2004; Pereira et al. 2006). To determine how the Wnt pathway is connected to the gene expression program of ES cells, we have determined the genome-wide binding profile of Tcf3 and examined how perturbations of the pathway affect the gene expression program. Remarkably, the genome-wide data reveal that Tcf3 co-occupies the ES cell genome with the pluripotency transcription factors Oct4 and Nanog. These and other results reveal that the Wnt pathway brings developmental signals directly to the core regulatory circuitry of ES cells, which consists of the pluripotency transcription factors and Tcf3, together with their mutual target genes.

Results

Identification of Tcf3 Binding Sites Genome-wide

To determine how the Wnt pathway regulates the gene expression program of murine embryonic stem cells, we first identified genes occupied by Tcf3. Murine embryonic stem cells were grown under standard conditions (Supplemental Fig. S1) and DNA sequences occupied by Tcf3 were identified using chromatin immunoprecipitation (ChIP) combined with DNA microarrays (ChIP-Chip). For this purpose, DNA microarrays were designed with 60-mer oligonucleotide probes tiling the entire non-repeat portion of the mouse genome. The results revealed that Tcf3 occupies over 1000 murine promoters (Supplemental Table S1), including those of the known Wnt targets Axin2 and Myc (Fig. 1A)(He et al. 1998; Yan et al. 2001; Jho et al. 2002).

Tcf3 Co-occupies the Genome with ES Cell Master Regulators

Inspection of the genes occupied by Tcf3 revealed a large set that were previously shown to be bound by the homeodomain transcription factor Oct4 (Boyer et al. 2005; Loh et al. 2006), which is an essential regulator of early development and ES cell identity (Nichols et al. 1998; Hay et al. 2004). To examine the overlap of gene targets more precisely, we carried out ChIP-Chip experiments with antibodies directed against Oct4 in mES cells and used the same genome-wide microarray platform employed in the Tcf3 experiment. Remarkably, the binding profiles of Tcf3 and Oct4 revealed that they bind the same genomic regions and display identical spatial distribution patterns with regards to transcription start sites (Fig. 1B; Supplemental Fig. S2). These results identified a set of 1224 genes that are co-occupied by Tcf3 and Oct4 at high confidence (Supplemental Table S1) and suggested that the Wnt pathway connects directly to genes regulated by Oct4 through Tcf3.

Previous studies in human embryonic stem cells have shown that Oct4 shares target genes with the transcription factors Nanog and Sox2 (Boyer et al. 2005), suggesting that Tcf3-occupied genes in murine ES cells should also be occupied by Nanog and Sox2. Additional genome-wide ChIP-Chip experiments with antibodies directed against Nanog revealed that it does indeed bind the same sites occupied by Oct4 and Tcf3 (Fig. 1B,C and 2, Supplemental Fig. S2). The fact that Oct4 and Sox2 form heterodimers in ES cells (Dailey and Bascilico 2001; Okumura-Nakanishi et al. 2005) and frequently co-occupy promoters in human ES cells (Boyer et al. 2005) makes it likely that Tcf3 co-occupies much of the genome with Oct4, Nanog and Sox2.

The observation that Tcf3 co-occupies much of the genome with the ES cell pluripotency transcription factors has a number of implications for the regulatory circuitry of these cells. Tcf3 binds its own promoter as well as the promoters of genes encoding Oct4, Sox2 and Nanog (Fig. 2). Thus Tcf3 is an integral component of an interconnected autoregulatory loop, where all four transcription factors together occupy each of their own promoters (Fig. 3A). This feature of ES cell regulatory circuitry was previously described for Oct4, Sox2 and Nanog alone (Boyer et al. 2005) and has been postulated to be a common regulatory motif for master regulators of cell state (Chambers et al. 2003; Okumura- Nakanishi et al. 2005; Rodda et al. 2005; Odom et al. 2004; Odom et al. 2006). Autoregulation is thought to provide several advantages to the control of cell state, including reduced response time to environmental stimuli and increased

stability of gene expression (McAdams et al. 1997; Rosenfeld et al. 2002; Shenn- Orr et al. 2002; Thieffry et al. 1998). It is also notable that Tcf3 and the pluripotency transcription factors together occupy genes encoding many Wnt pathway components (Supplemental Fig. S3), suggesting that this transcription factor regulates much of its own signaling pathway apparatus together with the pluripotency factors.

A model for the core regulatory circuitry of ES cells has been proposed in which the genes bound by the master regulators Oct4, Sox2 and Nanog fall into two classes: transcriptionally active genes encoding transcription factors, signaling components and other products that support the stem cell state, and transcriptionally inactive genes, consisting mostly of developmental regulators, where Polycomb is bound and RNA polymerase II is recruited, but transcription is stalled (Boyer et al. 2005; Boyer et al. 2006; Lee et al. 2006; Guenther et al. 2007; Stock et al. 2007; Zeitlinger et al. 2007). Our results reveal that Tcf3, together with the pluripotency regulators, is associated with both classes of genes, and thus provide a modified model of the core regulatory circuitry of ES cells (Fig. 3B). The association of Tcf3 with the set of genes encoding key transcription factors, signaling pathway components, and developmental regulators suggests that the Wnt signaling pathway contributes to the regulation of these genes, thereby impacting embryonic stem cell pluripotency and selfrenewal.

Expression Analysis of Tcf3 Knockdown in mES Cells

Genes bound by Tcf/Lef proteins are thought to be repressed in the absence of Wnt/β-catenin signaling and to be activated upon Wnt pathway stimulation (Behrens et al. 1996; Brantjes et al. 2001; Miyabayashi et al. 2007; Daniels et al. 2005; Cavallo et al. 1998). Murine ES cells have low endogenous Wnt activity in standard culture conditions and the Wnt pathway can be further stimulated in culture (Dravid et al. 2005; Yamaguchi et al. 2005; Lindsley et al. 2006; Ogawa K et al. 2006; Anton et al. 2007; Takao et al. 2007) (Supplemental Fig. S4). Thus it is unclear whether Tcf3-occupied genes are being repressed or activated at the low level of Wnt activity characteristic of standard ES cell culture conditions. To investigate whether the effect of Tcf3 occupancy is to repress or to activate genes, RNAi constructs were used to deplete *Tcf3* mRNA in mES cells in two independent experiments (Supplemental Fig. S5) and changes in global mRNA levels were assayed with DNA microarrays (Fig. 4A). The ~3.5% of mouse genes whose mRNA levels changed by at least two-fold were significantly enriched for Tcf3 targets relative to genes whose expression was unaltered by the *Tcf3* knockdown (p value $< 2 \times 10^{-10}$; Supplemental Fig. S6; Supplemental Table S2). The genes whose expression increased upon loss of *Tcf3* included those encoding the master regulators Oct4, Sox2 and Nanog, other genes involved in pluripotency such as Lefty2 and Nodal, and the Wnt pathway component Dkk1 (Fig. 4A). The fact that upregulated genes are strongly enriched for Tcf3 binding suggests that Tcf3 mainly acts to repress genes. Upon loss of *Tcf3*, target genes are no longer repressed and can now be activated by other factors (such as Oct4, Sox2 and Nanog) present at their promoters.

While expression of Tcf3 target genes was often up-regulated upon loss of *Tcf3*, the expression of a substantial number of Tcf3-bound genes remained unchanged, and a relatively small number of Tcf3-bound genes showed reduced expression (Fig. 4A). Nearly half of the genes occupied by Tcf3, Oct4 and Nanog are co-occupied by Polycomb Repressive Complexes (Boyer et al. 2006; Lee et al. 2006; Wilkinson et al.

Figure 1

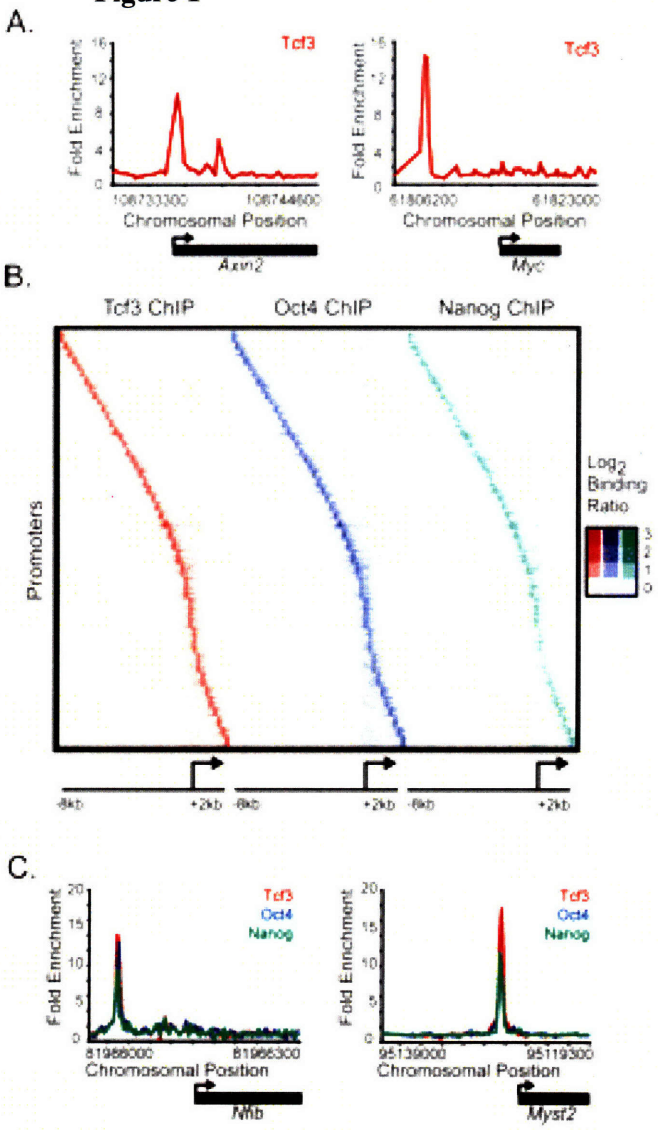

A.
Tcf3
Fold Enrichment
Chromosomal Position
Axin2
Myc
B.
Tcf3 ChIP
Oct4 ChIP
Nanog ChIP
Promoters
Log2 Binding Ratio
-8kb
+2kb
C.
Tcf3
Oct4
Nanog
Fold Enrichment
Chromosomal Position
Nfib
Myst2

Figure 1. Tcf3, Oct4, and Nanog co-occupy the genome in mouse ES cells.
(*A*) Tcf3 binds to known target genes. Examples of previously known Tcf3-bound regions are displayed as unprocessed ChIP-enrichment ratios for all probes within the chromosomal region indicated *below* the plot. The gene is depicted *below* the plot, and the TSS and direction are denoted by an arrow.
(*B*) Tcf3, Oct4, and Nanog display nearly identical binding profiles. Analysis of ChIP–chip data from genes bound by Tcf3, Oct4, or Nanog reveals that the three factors bind to similar genomic regions at all promoters. Regions from –8 kb to +2 kb around each TSS were divided into bins of 250 bp. The raw enrichment ratio for the probe closest to the center of the bin was used. If there was no probe within 250 bp of the bin center then no value was assigned. For genes with multiple promoters, each promoter was used for analysis. The analysis was performed on 3764 genes, which represents 4086 promoters. Promoters are organized according to the distance between the maximum Tcf3-binding ratio and the TSS.
(*C*) Tcf3, Oct4, and Nanog bind in close proximity at target genes. Plots display unprocessed ChIP-enrichment ratios for all probes within the chromosomal region indicated *below* the plot. The gene is depicted *below* the plot, and the TSS and direction are denoted by an arrow.

Figure 2

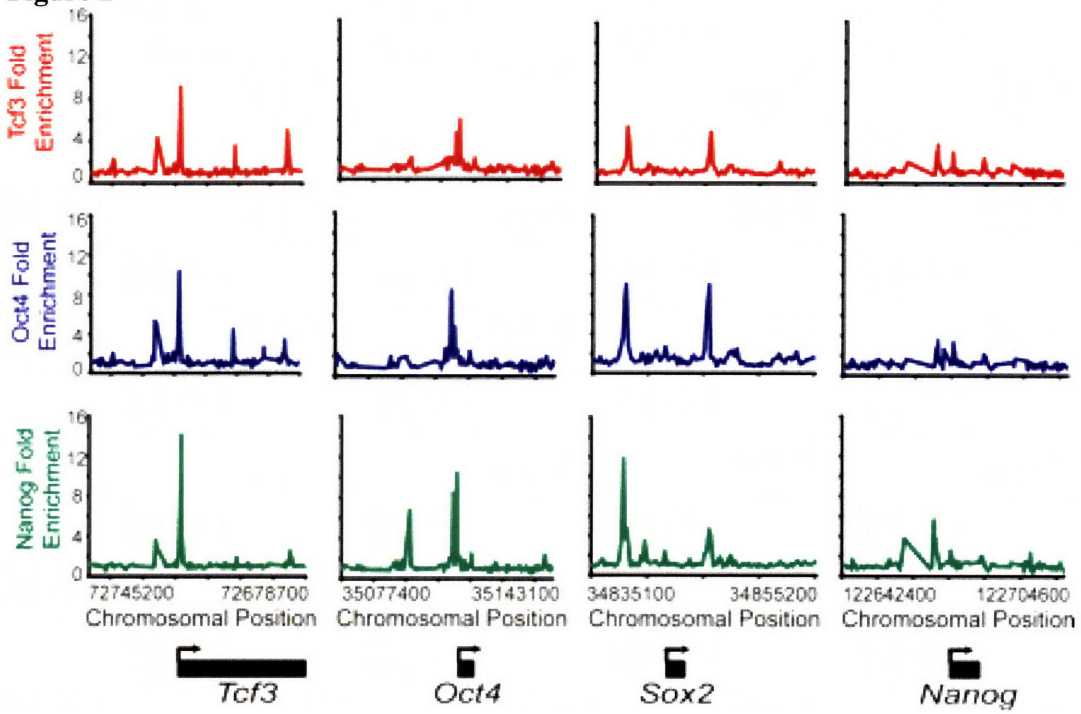

Tcf3 Fold Enrichment
Oct4 Fold Enrichment
Nanog Fold Enrichment
16
12
8
4
0
72745200
72678700
35077400
35143100
34835100
34855200
122642400
122704600
Chromosomal Position
Tcf3
Oct4
Sox2
Nanog

Figure 2. Tcf3, Oct4, and Nanog bind the promoters of *Tcf3*, *Oct4*, *Sox2*, and *Nanog*. Plots display unprocessed ChIP-enrichment ratios for all probes within the chromosomal region indicated *below* the plot. The gene is depicted *below* the plot, and the TSS and direction are denoted by an arrow.

Figure 3

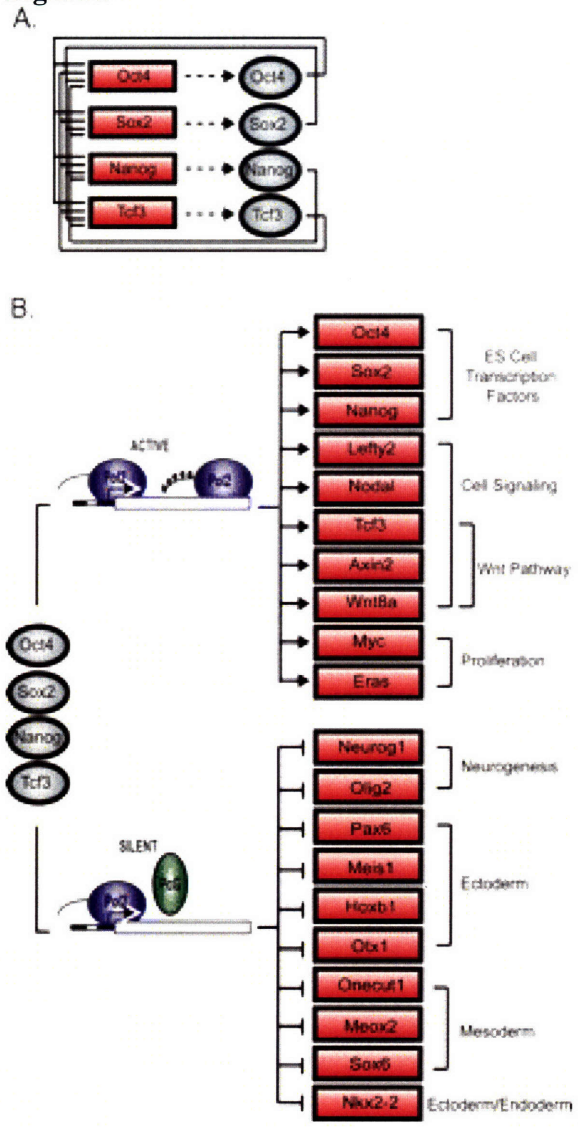

A.
Oct4
Sox2
Nanog
Tcf3
B.
ACTIVE
SILENT
Oct4
Sox2
Nanog
Tcf3
Oct4
Sox2
Nanog
ES Cell Transcription Factors
Lefty2
Nodal
Cell Signaling
Tcf3
Axin2
Wnt8a
Wnt Pathway
Myc
Eras
Proliferation
Neurog1
Olig2
Neurogenesis
Pax6
Meis1
Hoxb1
Otx1
Ectoderm
Onecut1
Mesp2
Sox6
Mesoderm
Nkx2-2
Ectoderm/Endoderm

Figure 3. Tcf3 is an integral component of the core regulatory circuitry of ES cells.
(*A*) Tcf3 forms an interconnected autoregulatory loop with Oct4, Sox2, and Nanog. Proteins are represented by ovals and genes are indicated by rectangles.
(*B*) Model showing a key portion of the regulatory circuitry of mES cells where Oct4, Sox2, Nanog, and Tcf3 occupy both active and silent genes. The evidence that Oct4, Nanog, and Tcf3 occupy these genes is described here; Sox2 occupancy is inferred from previous studies in human ES cells (Boyer et al. 2005). Evidence that the transcriptionally silent genes are occupied by Polycomb Repressive Complexes is from Boyer et al. (2006), and unpublished data and that these genes have stalled RNA polymerases is from Guenther et al. (2007) and Stock et al. (2007). Proteins are represented by ovals and genes are indicated by rectangles.

Figure 4

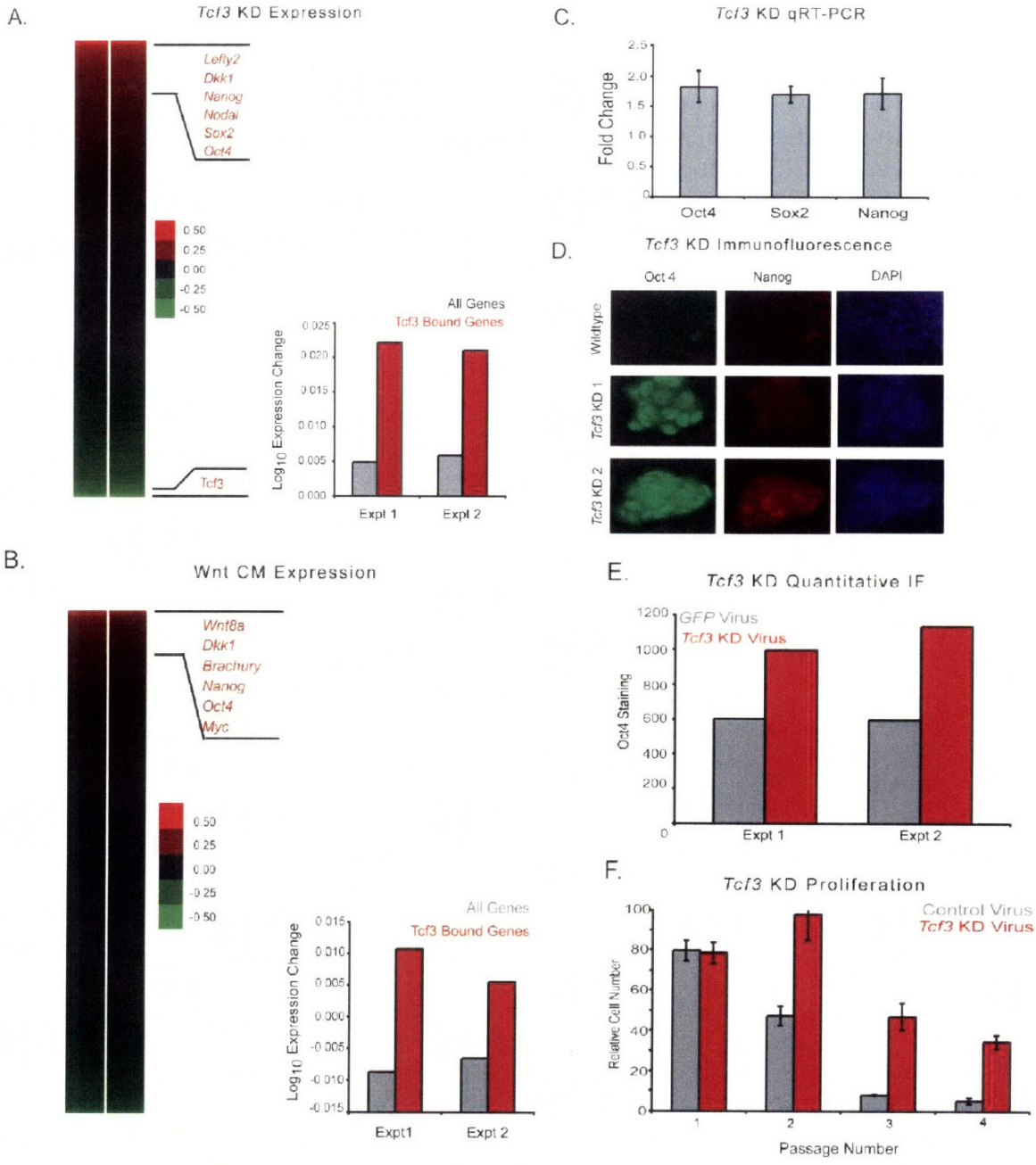

A.
Tcf3 KD Expression
Lefty2
Dkk1
Nanog
Nodal
Sox2
Oct4
Tcf3
0.50
0.25
0.00
-0.25
-0.50
All Genes
Tcf3 Bound Genes
Log10 Expression Change
Expt 1
Expt 2
B.
Wnt CM Expression
Wnt8a
Dkk1
Brachury
Nanog
Oct4
Myc
0.50
0.25
0.00
-0.25
-0.50
All Genes
Tcf3 Bound Genes
Log10 Expression Change
Expt1
Expt 2
C.
Tcf3 KD qRT-PCR
Fold Change
Oct4
Sox2
Nanog
D.
Tcf3 KD Immunofluorescence
Oct 4
Nanog
DAPI
Wildtype
Tcf3 KD 1
Tcf3 KD 2
E.
Tcf3 KD Quantitative IF
GFP Virus
Tcf3 KD Virus
Oct4 Staining
Expt 1
Expt 2
F.
Tcf3 KD Proliferation
Control Virus
Tcf3 KD Virus
Relative Cell Number
Passage Number

Figure 4. Knockdown of *Tcf3* and activation of the Wnt pathway in mES cells reveal a role for Tcf3 in repression of target genes and a role in regulating pluripotency.
(*A*) *Tcf3* knockdown results in up-regulation of target genes. The effect of *Tcf3* knockdown on gene expression was measured by hybridization of labeled RNA prepared from *Tcf3* knockdown cells against RNA prepared from cells infected with nonsilencing control lentivirus at 48 h post-infection. A heat map of biological replicate data sets of expression changes was generated where genes are ordered according to average expression change. Tcf3 target genes have a higher average expression change than the average for all genes upon knockdown of *Tcf3*.
(*B*) Wnt CM results in up-regulation of Tcf3 target genes. The effect of Wnt activation on gene expression was measured by hybridization of labeled RNA prepared from mES cells grown in Wnt CM against RNA prepared from cells grown in mock CM. A heat map of biological replicate data sets of expression change upon addition of Wnt CM where genes are ordered according to average expression change of replicates. Tcf3 target genes have a higher average expression change than the average for all genes upon addition of Wnt CM.
(*C*) *Tcf3* knockdown results in increased expression of *Oct4*, *Sox2*, and *Nanog*. Real-time PCR demonstrates that *Oct4*, *Sox*, and *Nanog* have increased expression upon knockdown of *Tcf3*. Values are normalized to *Gapdh* transcript levels, and fold change is relative to cells transfected with a nonsilencing hairpin.
(*D*) *Tcf3* knockdown results in increased staining for Oct4 and Nanog. Immunofluorecence was performed on mES cells grown one passage off of feeders that were either infected with *Tcf3* knockdown lentivirus or infected with nonsilencing control lentivirus. Cells were fixed with 4% paraformaldehyde 96 h post-infection. Cells were stained with Oct4, Nanog, and DAPI. Images for Oct4 and Nanog staining were taken at 40°— magnification and an exposure time of 300 msec. *Tcf3* KD 1 and KD 2 represent different knockdown hairpin constructs. *Tcf3* KD 2 is the virus also used in *A*, *C*, *E*, and *F*.
(E) *Tcf3* knockdown results in a significant increase of Oct4 staining. Quantification of Oct4 staining was performed in cells infected with *Tcf3* or *Gfp* knockdown virus.
(*F*) *Tcf3* knockdown cells proliferate over more passages in the absence of LIF. Relative cell numbers of ES cells transfected with *Tcf3* or control virus through multiple passages off of feeders in the presence or absence of LIF. Identical cell numbers were initially plated, and cells were split 1:12 every 2–3 d. Cells were counted at each passage and values for cells grown in the absence of LIF were normalized to cells grown in the presence of LIF.

2006; Rajaskhar and Begemann 2007), and their transcriptional state would not be expected to change as Polycomb would prevent elongation of transcripts at these genes (Stock et al. 2007). Indeed, we find that expression of genes occupied by Tcf3 and Polycomb do not show a significant expression change upon loss of *Tcf3* (p value > 0.4). There were some Tcf3 target genes whose expression was down-regulated upon loss of *Tcf3*; because mES cells have a low level of Wnt pathway activation, it is possible that sufficient β-catenin enters the nucleus in order to associate with and activate this subset of genes. Indeed, we find that some amount of β-catenin does associate with Tcf3 and Oct4 as β-catenin can be detected in crosslinked chromatin extracts immunoprecipitated for either Tcf3 or Oct4 (Supplemental Fig. S7). It is also possible that the loss of expression of this set of Tcf3 target genes is a secondary consequence of the knockdown. The repressive activity of Tcf3 appears to be its dominant function for most genes under these conditions, as the set of Tcf3 bound genes were found to have a significantly higher increase in expression upon knockdown compared to all genes (Fig. 4A; p value < 7×10^{-5}).

Expression Analysis of Wnt Pathway Activation in mES Cells

We next studied the effect of increased stimulation of the Wnt pathway on Tcf3 target genes in murine ES cells. Cells were treated with Wnt3a conditioned media in two independent experiments, and changes in global mRNA levels were assayed with DNA microarrays (Fig. 4B). The <1% of mouse genes whose mRNA levels changed by at least two-fold in the Wnt treated cells were significantly enriched for Tcf3 targets relative to genes whose expression was unaltered by the addition of Wnt (p value < 1.5 x 10-5; Supplemental Fig. S8; Supplemental Table S3). The genes whose expression most increased encode the pluripotency factors Oct4 and Nanog, Wnt pathway components such as Wnt8a and Dkk1, and known Wnt targets such as Brachury (Fig. 4B). These results are consistent with a model where Tcf3 acts to partially repress many of its target genes under standard mES cell culture conditions, yet contributes to increased expression of its target genes under conditions of increased Wnt stimulation. We would therefore expect a correlation between genes upregulated upon loss of Tcf3 and genes up-regulated upon Wnt stimulation. Indeed, we do find these gene sets to be significantly correlated (p value < 1×10^{-8}; Supplemental Fig. S9). Although a significant portion of Tcf3 target genes undergo expression changes upon Wnt stimulation, it is possible that a second class of Tcf3 target genes are regulated independently of Wnt signaling and therefore are uninfluenced by changes in pathway activation (Yi and Merrill 2007). In fact, several studies have shown a β-catenin independent role for Tcf3 (Kim et al. 2000; Merrill et al. 2001; Roel et al. 2002). It should also be noted that ES cells express other mammalian Tcf/Lef proteins and that these factors may also mediate the functional consequences of Wnt signaling (Pereira et al., 2006).

Influence of Tcf3 and Wnt on Pluripotency Regulators and ES Cell State

Evidence that Tcf3 is an integral component of the core transcriptional circuitry of ES cells that functions to partially repress transcription of pluripotency genes led us to examine whether *Tcf3* knockdown enhances features of ES cells associated with pluripotency and self-renewal. Quantitative real-time PCR analysis demonstrated that *Tcf3* knockdown in mES cells results in higher transcript levels for the pluripotency

genes *Oct4*, *Sox2* and *Nanog* (Fig. 4C). Upregulation of *Nanog* upon *Tcf3* depletion confirms a previous report that Tcf3 acts to repress this gene under normal ES cell growth conditions (Pereira et al. 2006). Thus the results of the *Tcf3* knockdown experiment indicate that under normal conditions Tcf3 functions to reduce expression of the three pluripotency regulators.

We next measured the levels of Oct4 and Nanog proteins in ES cells subjected to *Tcf3* knockdown. The results of immunofluorescence experiments show that there are substantial increases in the levels of Oct4 and Nanog transcription factors in the nucleus of such cells (Fig. 4D). There is a significant increase of Oct4 in *Tcf3* knockdown cells compared to control cells based on quantitative measurements of staining intensity using Cellomics software (Fig. 4E). Remarkably, *Tcf3* knockdown mES cells display enhanced proliferation and Oct4 staining in the absence of feeders and LIF compared to control cells, supporting previous results (Fig. 4F; Supplemental Fig. S10)(Pereira et al. 2006). Previous studies have demonstrated that activation of the Wnt/β-catenin pathway can have similar effects on ES cell pluripotency (Sato et al. 2003; Singla et al. 2006; Hao et al. 2006) and we also find that cells treated with Wnt conditioned media show increased staining of Oct4 (Supplemental Fig. S11). The observation that *Tcf3* knockdown and Wnt stimulation have similar functional consequences is consistent with the expression data described above for ES cells subjected to *Tcf3* knockdown and ES cells treated with Wnt3a CM. These studies demonstrate the functional importance of Tcf3 occupancy and Wnt pathway activation for a subset of target genes that includes the pluripotency regulators.

Discussion

It is fundamentally important to determine how signaling pathways control ES cell pluripotency and differentiation and how these pathways connect to discrete sets of target genes to affect such states. We have found that a terminal component of the Wnt signaling pathway, the transcription factor Tcf3, is physically associated with the same genomic sites as the pluripotency regulators Oct4 and Nanog in murine embryonic stem cells. This result reveals that the Wnt pathway is physically connected to the core regulatory circuitry of these cells. This core circuitry consists of two key features: an interconnected autoregulatory loop and the set of target genes that are mutually bound by the pluripotency transcription factors and Tcf3.

The genome-wide datasets we report here enhance our knowledge of the targets of Oct4, Nanog and Tcf3. These new datasets were generated using the same protocols and genome-wide tiling microarrays in ES cells grown under identical conditions, allowing more reliable conclusions about the overlap of these factors throughout the genome; previous datasets for these factors came from different murine ES cells grown in different settings, using different chromatin IP analysis platforms, and these data were not always genome-wide (Boyer et al. 2005; Boyer et al. 2006; Loh et al. 2006). The new data reveal, for example, the remarkable extent to which Oct4 and Nanog binding overlap throughout the ES cell genome and the striking association of Tcf3 with those sites (Fig. 1B). The new data also provide a revised model for the core regulatory circuitry of murine ES cells, which incorporates Tcf3 and high confidence target genes of key ES cell regulators (Fig. 3).

The revised model of core regulatory circuitry extends our knowledge of how extracellular signals from the Wnt pathway contribute to stem cell state. Pereira et al. (2006) demonstrated that Tcf3 binds the *Nanog* promoter and represses its mRNA expression in mES cells. Our data confirm Tcf3 binding and function at *Nanog* and extend our knowledge of Tcf3 targets to the other well-characterized pluripotency regulators Oct4 and Sox2, as well as most of their target genes. Pereira et al. (2006) proposed a model wherein Tcf3-mediated control of Nanog levels allows stem cells to balance the creation of lineage-committed and undifferentiated cells. Our results also support this model, but argue that Tcf3 contributes to the balance through its functions in the core regulatory circuitry described here.

Our results suggest that the Wnt pathway, through Tcf3, influences the balance between pluripotency and differentiation in ES cells, as modeled in Figure 5. Under standard culture conditions, where there is a low-level of Wnt activation, ES cells are poised between the pluripotent state and any of a number of differentiated states. It is well established that Oct4, Sox2 and Nanog act to promote the pluripotent state, as depicted in the model where the influence of these factors is shown by an arrow. Under standard culture conditions, Tcf3 may exist in an activating or repressive complex, but is predominantly in a repressive complex promoting differentiation. The loss of Tcf3 in *Tcf3* knockdown cells, would, in this model, favor pluripotency. Wnt stimulation converts the repressive complex to an activating complex and thus promotes pluripotency. Our results suggest that the Wnt pathway, through Tcf3, influences the balance between pluripotency and differentiation by bringing developmental signals directly to the core regulatory circuitry of ES cells. The observation that the Wnt pathway

can be manipulated to affect the balance between pluripotency and differentiation suggests that perturbation of this pathway may impact the efficiency of reprogramming somatic cells into pluripotent stem cells.

Figure 5

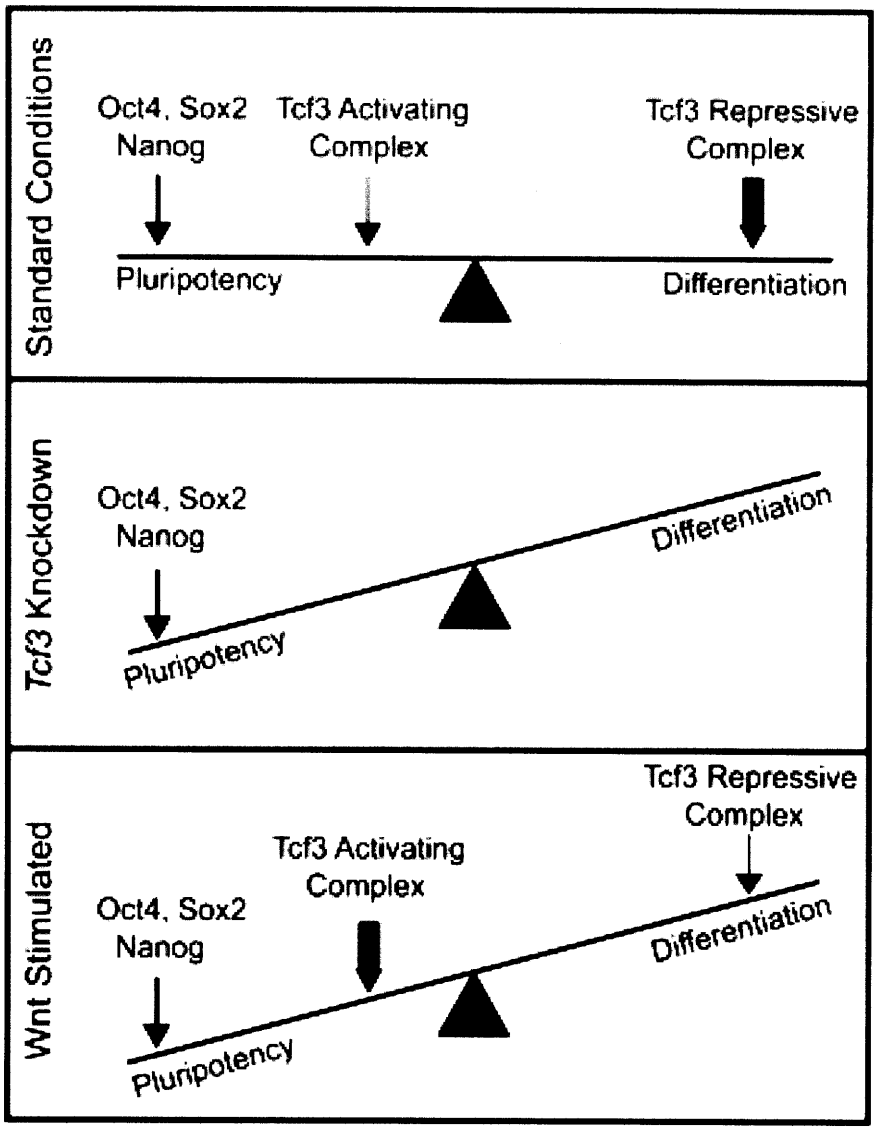

Standard Conditions
Oct4, Sox2
Nanog
Tcf3 Activating
Complex
Tcf3 Repressive
Complex
Pluripotency
Differentiation
Tcf3 Knockdown
Oct4, Sox2
Nanog
Differentiation
Pluripotency
Wnt Stimulated
Tcf3 Repressive
Complex
Tcf3 Activating
Complex
Oct4, Sox2
Nanog
Differentiation
Pluripotency

Figure 5. Model depicting the influence of Wnt pathway components on pluripotency and differentiation in ES cells.
ES cells are poised between the pluripotent state and any of a number of differentiated states. Oct4, Sox2, and Nanog act to promote the pluripotent state (depicted by an arrow). Tcf3 can exist in an activating complex with -catenin or a repressive complex with Groucho (Reya and Clevers 2005). Under standard growth conditions, the Wnt pathway is only active at low levels (Supplemental Fig. S4; Dravid et al. 2005; Yamaguchi et al. 2005; Lindsley et al. 2006; Ogawa et al. 2006; Anton et al. 2007; Takao et al. 2007). Therefore, Tcf3 is mainly in a repressive complex promoting differentiation (depicted by a thick arrow), although some Tcf3 associates with -catenin to activate target genes and promote pluripotency (depicted by a thin arrow). In *Tcf3* knockdown cells, there is no influence from Tcf3 on cell state. Thus, the balance is tipped toward maintaining pluripotency. Upon Wnt stimulation, the balance again tips toward maintaining pluripotency as more Tcf3 associates with -catenin in an activating complex (depicted by a thick arrow). This model is not meant to imply that Wnt or Tcf3 are themselves pluripotency factors, but rather that they can influence cell state in the presence of other pluripotency factors, such as Oct4, Sox2, and Nanog.

Materials and methods

Mouse embryonic stem cell culture conditions

V6.5 murine ES cells were grown on irradiated murine embryonic fibroblasts (MEFs) unless otherwise stated. Cells were grown under mES cell conditions as previously described in Boyer et al. (2005). Briefly, cells were grown on 0.2% gelatinized tissue culture plates in DMEM-KO (Invitrogen 10829-018) supplemented with 15 % fetal bovine serum (Hyclone, Characterized SH3007103), 1000 Units/mL leukemia inhibitory factor (LIF) (ESGRO ESG1106), 100μM nonessential amino acids (Invitrogen 11140-050), 2mM L-glutamine (Invitrogen 25030-081),100 Units/mL pennicillin and 100 μg/mL streptomycin (Invitrogen 15140-122), and 8 nL/ml 2-mercaptoethanol (Sigma M7522).

Genome-wide location analysis

Chromatin immunoprecipitation protocol

Protocols describing ChIP methods were downloaded from http://jura.wi.mit.edu/young_public/hESregulation/ChIP.html with slight modifications. Briefly, 10^8 mES cells were grown for one passage off of feeders and then crosslinked using formaldehyde. Cells were resuspended, lysed in lysis buffer and sonicated to solubilize and sheer crosslinked DNA. Triton X-100 and SDS were added to the lysate after sonication to final concentrations of 1% and 0.1% respectively. The resulting whole cell extract was incubated at 4°C overnight with 100 uL of Dynal Protein G magnetic beads that had been preincubated with 10 ug of the appropriate antibody overnight. After 16-18 hours, beads were washed with the following 4 buffers for 4 minutes per buffer: low salt buffer (20mM Tris pH 8.1, 150mM NaCl, 2mM EDTA, 1% Triton X- 100, 0.1% SDS), high salt buffer (20mM Tris pH 8.1, 500mM NaCl, 2mM EDTA, 1% Triton X-100, 0.1% SDS), LiCl buffer (10mM Tris pH 8.1, 250mM LiCl, 1mM EDTA, 1% deoxycholate, 1% NP-40), and TE+ 50mM NaCl. Bound complexes were eluted from the beads in elution buffer by heating at 65°C with occasional vortexing, and crosslinks were reversed by overnight incubation at 65°C.

ChIP Antibodies

Cell extracts were immunoprecipitated using antibodies against Tcf3 (Santa Cruz sc-8635), Oct4 (Santa Cruz sc-8628) or Nanog (Bethyl Labs bl1662).

Array Design

The murine 244k whole genome array was purchased from Agilent Technology (www.agilent.com). The array consists of 25 slides each containing ~244,000 60mer oligos (slide ID 15310-3, 15317, 15319-21, 15323, 15325, 15327-30, 15332-7, 15339-41, 15343-44) covering the entire non-repeat portion of the mouse genome at a density of about 1 oligo per 250bp.

Data Normalization and Analysis

Data normalization and analyses were performed as previously described in Boyer et al. (2005).

Tcf3 Knockdown

Lentiviral Production

Lentivirus was produced according to Open Biosystems Trans-lentiviral™ shRNA

Packaging System (TLP4614). The shRNA constructs targeting murine *Tcf3* were designed using an siRNA rules based algorithm consisting of sequence, specificity and position scoring for optimal hairpins that consist of a 21 base stem and a 6 base loop (RMM4534-NM-009332). Five hairpin constructs were used to produce virus targeting *Tcf3*. A negative control virus was made from the pLKO.1 empty vector (RHS4080).

Lentiviral Infection of mES Cells

Murine V6.5 ES cells were plated at approximately 30% confluence on the day of infection. Cells were seeded in 2x mES media with 6 ug/ml of polybrene (Sigma H9268-10G) and *Tcf3* knockdown or control (pLKO.1) virus was immediately added. After 24 hours, infection media was removed and replaced with mES media with 2 ug/ml of Puromycin (Sigma P8833). RNA was harvested at 48 hours after infection.

Knockdown Efficiency

Knockdown efficiency was measured using real-time PCR to measure levels of *Tcf3* mRNA (Supplemental Fig. S5).

RNA Isolation, Real-time PCR and Analysis of Transcript Levels

To determine transcript levels by RT-PCR, RNA was isolated from approximately $10^6 - 10^7$ mES cells using TRIzol reagent following the protocol for cells grown in monolayer (Invitrogen 15596-026). Samples were treated with Dnase I (Invitrogen 18068-015) and cDNA was prepared using SuperScript III reverse transcriptase kit (Invitrogen 180808-051) using oligo dT primed first strand synthesis. Real-time PCR was carried out on the 7000 ABI Detection System using Taqman probes for the housekeeping gene *Gapdh* (Applied Biosystems Mm99999915_g1) as a control and genes of interest (Applied Biosystems; *Tcf3* Mm00493456_m1, *Oct4* Mm00658129_gH, *Sox2* Mm00488369_s1, *Nanog* Mm02384862_g1).

Expression Arrays

Genomic expression analysis was measured using Agilent Whole Mouse Genome Microarrays (Agilent G4122F). 2 ug of RNA was labeled for each sample using the Two-color Low RNA Input Linear Amplification Kit PLUS (Agilent 5188-5340). RNA from the treated sample (either *Tcf3* KD cells or cells treated with Wnt3a conditioned media) were labeled with Cy5 and RNA from control cells (infected with empty-vector virus or a mock conditioned media control, respectively) were labeled with Cy3. Labeled cRNA was hybridized overnight at 65°C. Slides were washed according to the Agilent protocol and scanned on an Agilent DNA microarray scanner BA. Data was analyzed using Agilent Feature Extraction Version 9.5.3 with default settings recommended by Agilent. Flagged and low-intensity spots were then removed and spots representing a single gene were averaged.

Wnt Pathway Activation

Wnt pathway activity in mES cells was stimulated using Wnt3a conditioned media (ATCC CRL-2647) and mock conditioned media (ATCC CRL-2648) was used as a control. Preparation of conditioned medias was performed as per protocol provided with the cells. Conditioned media was diluted with mES media at a ratio of 1:1.

Immunohistochemical Analysis

Mouse ES cells were crosslinked for 10 minutes at room temperature with 4% paraformaldehyde. Cells were permeabilized with 0.2% Triton X-100 for 10 minutes and stained for Oct4 (Santa Cruz, sc-5279; 1:200 dilution), Nanog (Abcam, ab1603; 1:250 dilution), and DAPI Nucleic Acid Stain (Invitrogen D1306; 1:10000 dilution) overnight at 4°C. After several washes cells were incubated for 2 hours at room temperature with goat-anti mouse conjugated Alexa Fluor 488 (Invitrogen 1:200 dilution) or goat-anti rabbit conjugated Alexa Fluor 568 (Invitrogen 1:200 dilution).

Quantitative Image Acquisition and Data Analysis

Image acquisition and data analysis was performed essentially as described in Moffat et al. (2006). Five days post infection cells were fixed and stained with Oct4 and Hoechst 33342 (1:1000 dilution). Stained cells were imaged on an Arrayscan HCS Reader (Cellomics) using the standard acquisition camera mode (10x objective, 9 fields). Hoechst was used as the focus channel and intra-well focusing was done every 3 fields. The Apotome feature was applied to acquire all images. Objects selected for analysis were identified based on the Hoechst staining intensity using the Target Activation Protocol and the Isodata Threshold method. Parameters were established requiring that individual objects pass an intensity and size threshold. The Object Segmentation Assay Parameter was adjusted for maximal resolution. Following object selection the average Oct4 intensity was determined and then a mean value for each well was calculated. All wells used for subsequent analysis contained at least 5000 selected objects.

Supplemental Data

Supplemental Data include nine figures, three tables, and Supplemental text and can be found with this article online at http://www.genesdev.org/cgi/content/full/22/6/746/DC1.

Accession Numbers

All microarray data from this study are available at ArrayExpress at the EBI (http://www.ebi.ac.uk/arrayexpress) under the accession designation E-TABM-409.

Acknowledgements

We thank Stuart Levine, Alex Marson, Martin Aryee and Sumeet Gupta for experimental and analytical support, Warren Whyte for the *Gfp* lentivirus vector, Roshan Kumar for knockdown and microarray advice, Jennifer Love for microarray advice, Laurie Boyer and Mathias Pawlak for cell culture advice and Tony Lee, Scott McCuine, Brett Chevalier and Rudolph Jaenisch for helpful discussions. Images for immunofluorescence were collected using the W.M. Keck Foundation Biological Imagining Facility at the Whitehead Institute and Whitehead-MIT Bioimaging Center. The SSEA-1 monoclonal antibody developed by D. Solter and B.B. Knowles was obtained from the Developmental Studies Hybridoma Bank developed under the auspices of the NICGH and maintained by the University of Iowa, Department of Biological Sciences, Iowa City, IA 52242. This work was supported by grants from the NIH and The Whitehead Institute. SJ was supported by an NSF Predoctoral Training Fellowship and MK was supported by an NIH NIGMS Postdoctural Fellowship.

References

Anton, R., Kestler, H.A., and Kuhl, M. 2007. beta-Catenin signaling contributes to stemness and regulates early differentiation in murine embryonic stem cells. *FEBS Lett* **581:** 5247-5254.

Behrens, J., von Kries, J.P., Kuhl, M., Bruhn, L., Wedlich, D., Grosschedl, R., and Birchmeier, W. 1996. Functional interaction of beta-catenin with the transcription factor LEF-1. *Nature* **382:** 638-642.

Boiani, M., and Scholer, H.R. 2005. Regulatory networks in embryo-derived pluripotent stem cells. *Nat Rev Mol Cell Biol* **6:** 872-884.

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.

Brantjes, H., Roose, J., van De Wetering, M., and Clevers, H. 2001. All Tcf HMG box transcription factors interact with Groucho-related co-repressors. *Nucleic Acids Res* **29:** 1410-1419.

Cadigan, K.M. 2002. Wnt signaling--20 years and counting. *Trends Genet* **18:** 340-342.

Cavallo, R.A., Cox, R.T., Moline, M.M., Roose, J., Polevoy, G.A., Clevers, H., Peifer, M., and Bejsovec, A. 1998. Drosophila Tcf and Groucho interact to repress Wingless signalling activity. *Nature* **395:** 604-608.

Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., and Smith, A. 2003. Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell* **113:** 643-655.

Clevers, H. 2006. Wnt/beta-catenin signaling in development and disease. *Cell* **127:** 469-480.

Dailey, L., and Basilico, C. 2001. Coevolution of HMG domains and homeodomains and the generation of transcriptional regulation by Sox/POU complexes. *J Cell Physiol* **186:** 315-328.

Daniels, D.L., and Weis, W.I. 2005. Beta-catenin directly displaces

Groucho/TLE repressors from Tcf/Lef in Wnt-mediated transcription activation. *Nat Struct Mol Biol* **12:** 364-371.

Dravid, G., Ye, Z., Hammond, H., Chen, G., Pyle, A., Donovan, P., Yu, X., and Cheng, L. 2005. Defining the role of Wnt/beta-catenin signaling in the survival, proliferation, and self-renewal of human embryonic stem cells. *Stem Cells* **23**: 1489-501.

Dreesen, O., and Brivanlou, A.H. 2007. Signaling pathways in cancer and embryonic stem cells. *Stem Cell Rev* **3:** 7-17.

Friel, R., van der Sar, S., and Mee, P.J. 2005. Embryonic stem cells: understanding their history, cell biology and signalling. *Adv Drug Deliv Rev* **57:** 1894-1903.

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.

Hao, J., Li, T.G., Qi, X., Zhao, D.F., and Zhao, G.Q. 2006. WNT/beta-catenin pathway up-regulates Stat3 and converges on LIF to prevent differentiation of mouse embryonic stem cells. *Dev Biol* **290:** 81-91.

Hay, D.C., Sutherland, L., Clark, J., and Burdon, T. 2004. Oct-4 knockdown induces similar patterns of endoderm and trophoblast differentiation markers in human and mouse embryonic stem cells. *Stem Cells* **22:** 225- 235.

He, T.C., Sparks, A.B., Rago, C., Hermeking, H., Zawel, L., da Costa, L.T., Morin, P.J., Vogelstein, B., and Kinzler, K.W. 1998. Identification of c-MYC as a target of the APC pathway. *Science* **281:** 1509-1512.

Jho, E.H., Zhang, T., Domon, C., Joo, C.K., Freund, J.N., and Costantini, F. 2002. Wnt/beta-catenin/Tcf signaling induces the transcription of Axin2, a negative regulator of the signaling pathway. *Mol Cell Biol* **22:** 1172-1183.

Kim CH, Oda T, Itoh M, Jiang D, Artinger KB, Chandrasekharappa SC, Driever W, Chitnis AB. 2000. Repressor activity of Headless/Tcf3 is essential for vertebrate head formation. *Nature* **407**: 913-6.

Kielman, M.F., Rindapaa, M., Gaspar, C., van Poppel, N., Breukel, C., van Leeuwen, S., Taketo, M.M., Roberts, S., Smits, R., and Fodde, R. 2002. Apc modulates embryonic stem-cell differentiation by controlling the dosage of beta-catenin signaling. *Nat Genet* **32:** 594-605.

Korinek, V., Barker, N., Willert, K., Molenaar, M., Roose, J., Wagenaar, G., Markman, M., Lamers, W., Destree, O., and Clevers, H. 1998. Two members of

the Tcf family implicated in Wnt/beta-catenin signaling during embryogenesis in the mouse. *Mol Cell Biol* **18:** 1248-1256.

Kristensen, D.M., Kalisz, M., and Nielsen, J.H. 2005. Cytokine signalling in embryonic stem cells. *Apmis* **113:** 756-772.

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.

Lindsley, R.C., Gill, J.G., Kyba, M., Murphy, T.L., and Murphy, K.M. 2006. Canonical Wnt signaling is required for development of embryonic stem cell-derived mesoderm. *Development* **133:** 3787-3796.

Logan, C.Y., and Nusse, R. 2004. The Wnt signaling pathway in development and disease. *Annu Rev Cell Dev Biol* **20:** 781-810.

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.

McAdams, H.H., and Arkin, A. 1997. Stochastic mechanisms in gene expression. *Proc Natl Acad Sci U S A* **94:** 814-819. Merrill, B.J., Gat, U., DasGupta, R., and Fuchs, E. 2001. Lef-1 and Tcf-3 transcription factors mediate tissue specific Wnt signaling during Xenopus development. *Genes & Development* **15**: 1688-1705.

Merrill, B.J., Pasolli, H.A., Polak, L., Rendl, M., Garcia-Garcia, M.J., Anderson, K.V., and Fuchs, E. 2004. Tcf3: a transcriptional regulator of axis induction in the early embryo. *Development* **131:** 263-274.

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

Miyabayashi, T., Teo, J.L., Yamamoto, M., McMillan, M., Nguyen, C., and Kahn, M. 2007. Wnt/beta-catenin/CBP signaling maintains long-term murine embryonic stem cell pluripotency. *Proc Natl Acad Sci U S A* **104:** 5668- 5673.

Moffat, J., Grueneberg, D.A., Yang, X., Kim, S.Y., Kloepfer, A.M., Hinkle, G., Piqani, B., Eisenhaure, T.M., Luo, B., Grenier, J.K., et al. 2006. A lentiviral RNAi library for human and mouse genes applied to an arrayed viral highcontent screen. *Cell* **124:** 1283-1298.

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

Odom, D.T., Dowell, R.D., Jacobsen, E.S., Nekludova, L., Rolfe, P.A., Danford, T.W., Gifford, D.K., Fraenkel, E., Bell, G.I., and Young, R.A. 2006. Core transcriptional regulatory circuitry in human hepatocytes. *Mol Syst Biol* **2:** 2006 0017.

Odom, D.T., Zizlsperger, N., Gordon, D.B., Bell, G.W., Rinaldi, N.J., Murray, H.L., Volkert, T.L., Schreiber, J., Rolfe, P.A., Gifford, D.K., et al. 2004. Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303:** 1378-1381.

Ogawa, K., Nishinakamura, R., Iwamatsu, Y., Shimosato, D., and Niwa, H. 2006. Synergistic action of Wnt and LIF in maintaining pluripotency of mouse ES cells. *Biochem Biophys Res Commun* **343:** 159-166.

Okumura-Nakanishi, S., Saito, M., Niwa, H., and Ishikawa, F. 2005. Oct-3/4 and Sox2 regulate Oct-3/4 gene in embryonic stem cells. *J Biol Chem* **280:** 5307-5317.

Otero, J.J., Fu, W., Kan, L., Cuadra, A.E., and Kessler, J.A. 2004. Beta-catenin signaling is required for neural differentiation of embryonic stem cells. *Development* **131:** 3545-3557.

Pan, G., and Thomson, J.A. 2007. Nanog and transcriptional networks in embryonic stem cell pluripotency. *Cell Res* **17:** 42-49.

Pera, M.F., and Trounson, A.O. 2004. Human embryonic stem cells: prospects for development. *Development* **131:** 5515-5525.

Pereira, L., Yi, F., and Merrill, B.J. 2006. Repression of Nanog gene transcription by Tcf3 limits embryonic stem cell self-renewal. *Mol Cell Biol* **26:** 7479- 7491.

Rajasekhar, V.K., and Begemann, M. 2007. Concise review: roles of polycomb group proteins in development and disease: a stem cell perspective. *Stem Cells* **25:** 2498-2510.

Rao, M. 2004. Conserved and divergent paths that regulate self-renewal in mouse and human embryonic stem cells. *Dev Biol* **275:** 269-286.

Reubinoff, B.E., Pera, M.F., Fong, C.Y., Trounson, A., and Bongso, A. 2000.

Embryonic stem cell lines from human blastocysts: somatic differentiation in vitro. *Nat Biotechnol* **18:** 399-404.

Reya, T., and Clevers, H. 2005. Wnt signalling in stem cells and cancer. *Nature* **434:** 843-850.

Rodda, D.J., Chew, J.L., Lim, L.H., Loh, Y.H., Wang, B., Ng, H.H., and Robson, P. 2005. Transcriptional regulation of nanog by OCT4 and SOX2. *J Biol Chem* **280:** 24731-24737.

Roël G., Hamilton F.S., Gent Y., Bain A.A., Destrée O., and Hoppler S. 2002. Lef-1 and Tcf-3 transcription factors mediate tissue-specific Wnt signaling during Xenopus development. *Current Bio* **12**: 1941-1945.

Rosenfeld, N., Elowitz, M.B., and Alon, U. 2002. Negative autoregulation speeds the response times of transcription networks. *J Mol Biol* **323:** 785-793.

Sato, N., Meijer, L., Skaltsounis, L., Greengard, P., and Brivanlou, A.H. 2004. Maintenance of pluripotency in human and mouse embryonic stem cells through activation of Wnt signaling by a pharmacological GSK-3-specific inhibitor. *Nat Med* **10:** 55-63.

Shen-Orr, S.S., Milo, R., Mangan, S., and Alon, U. 2002. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet* **31:** 64-68.

Singla, D.K., Schneider, D.J., LeWinter, M.M., and Sobel, B.E. 2006. wnt3a but not wnt11 supports self-renewal of embryonic stem cells. *Biochem Biophys Res Commun* **345:** 789-795.

Stock, J.K., Giadrossi, S., Casanova, M., Brookes, E., Vidal, M., Koseki, H., Brockdorff, N., Fisher, A.G., and Pombo, A. 2007. Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. *Nat Cell Biol* **9:** 1428-1435.

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

Thieffry, D., Salgado, H., Huerta, A.M., and Collado-Vides, J. 1998. Prediction of transcriptional regulatory sites in the complete genome sequence of Escherichia coli K-12. *Bioinformatics* **14:** 391-400.

Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S., and Jones, J.M. 1998. Embryonic stem cell lines derived from human blastocysts. *Science* **282:** 1145-1147.

Valdimarsdottir, G., and Mummery, C. 2005. Functions of the TGFbeta superfamily in human embryonic stem cells. *Apmis* **113:** 773-789.

Wilkinson, F.H., Park, K., and Atchison, M.L. 2006. Polycomb recruitment to DNA in vivo by the YY1 REPO domain. *Proc Natl Acad Sci U S A* **103:** 19296- 19301.

Yamaguchi, Y., Ogura, S., Ishida, M., Karasawa, M., and Takada, S. 2005. Gene trap screening as an effective approach for identifcation of Wnt-responsive genes in the mouse embryo. *Dev Dyn* **233**: 484-95.

Yan, D., Wiesmann, M., Rohan, M., Chan, V., Jefferson, A.B., Guo, L., Sakamoto, D., Caothien, R.H., Fuller, J.H., Reinhard, C., et al. 2001. Elevated expression of axin2 and hnkd mRNA provides evidence that Wnt/beta -catenin signaling is activated in human colon tumors. *Proc Natl Acad Sci U S A* **98:** 14973-14978.

Yi, F., and Merrill, B.J. 2007. Stem cells and TCF proteins: a role for betacatenin-- independent functions. *Stem Cell Rev* **3:** 39-48.

Zeitlinger, J., Stark, A., Kellis, M., Hong, J.W., Nechaev, S., Adelman, K., Levine, M., and Young, R.A. 2007. RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat Genet* **39:** 1512-1516.

Chapter 5

Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells

Submitted to Cell 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

My contribution to this project

The miRNA project was led by Alex Marson and Stuart Levine from the Young Lab. I generated the location analysis data for Oct4, Sox2, Nanog and Tcf3. I also worked with a group of Young Lab and Microarray Facility members to establish a working protocol for ChIP-seq using Solexa sequencing equipment.

Summary

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, Supplementary Figures S1 and S2, Supplementary 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 (Supplementary 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 (Supplementary Figure S3, Supplementary 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 (Supplementary 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 Supplementary Figure S4).

High-confidence promoters were identified for over 80% of miRNAs in both mouse and human (Figure 2, Supplementary Figure S4 and Supplementary 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) (Supplementary 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 Supplementary Figure S5 and Supplementary 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 Supplementary 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 Supplementary 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 Supplementary 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 (Supplementary 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 Supplementary Tables S7 and S8). Transcription factor occupancy does not necessarily mean that the adjacent gene is regulated by that factor; conserved transcription factor occupancy, however, has been shown to occur preferentially at genes that are regulated by that factor (Odom et al., 2007). 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

Figure 1

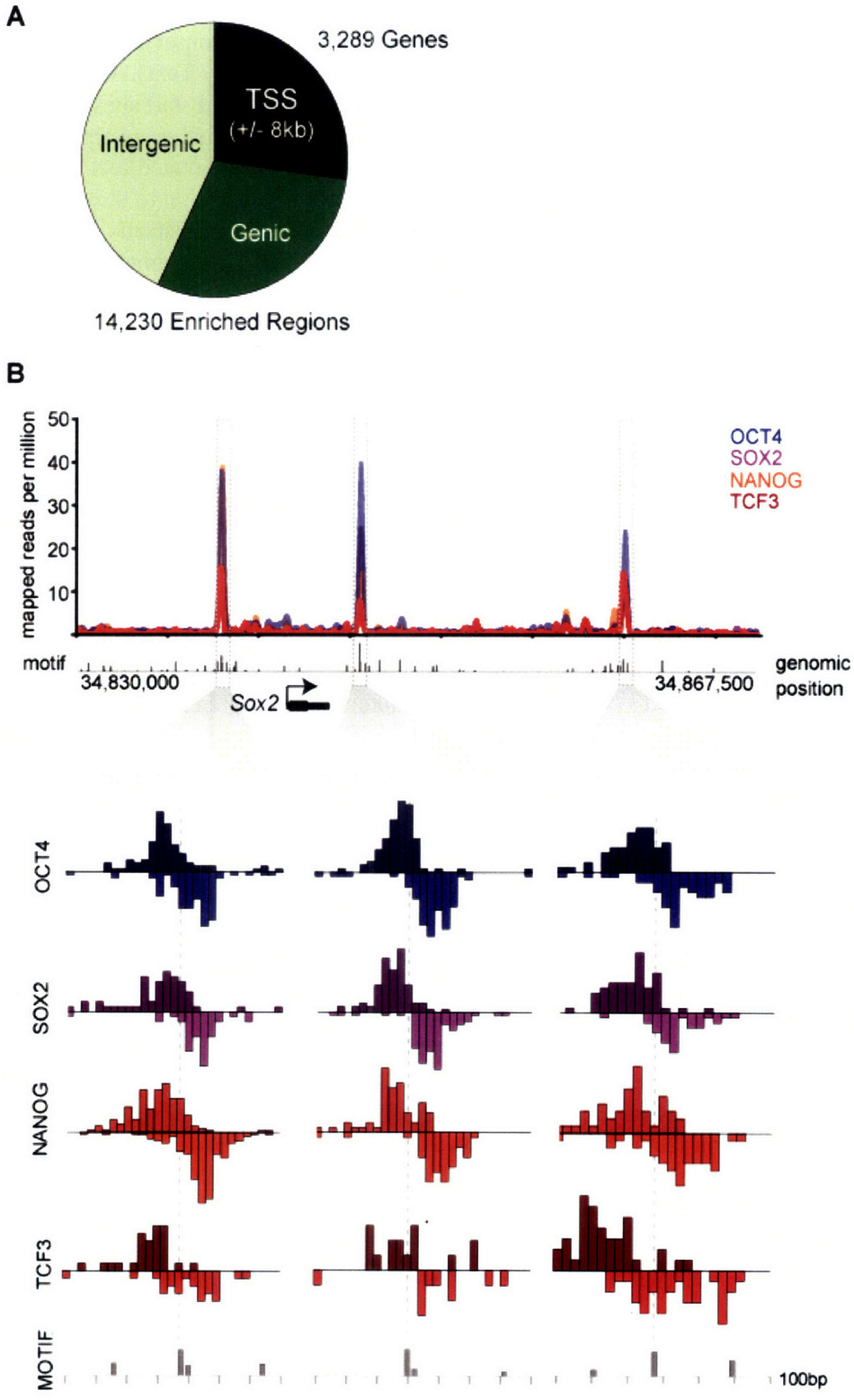

A
3,289 Genes
TSS
(+/- 8kb)
Intergenic
Genic
14,230 Enriched Regions
B
mapped reads per million
50
40
30
20
10
OCT4
SOX2
NANOG
TCF3
motif
34,830,000
34,867,500
genomic
position
Sox2
OCT4
SOX2
NANOG
TCF3
MOTIF
100bp

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

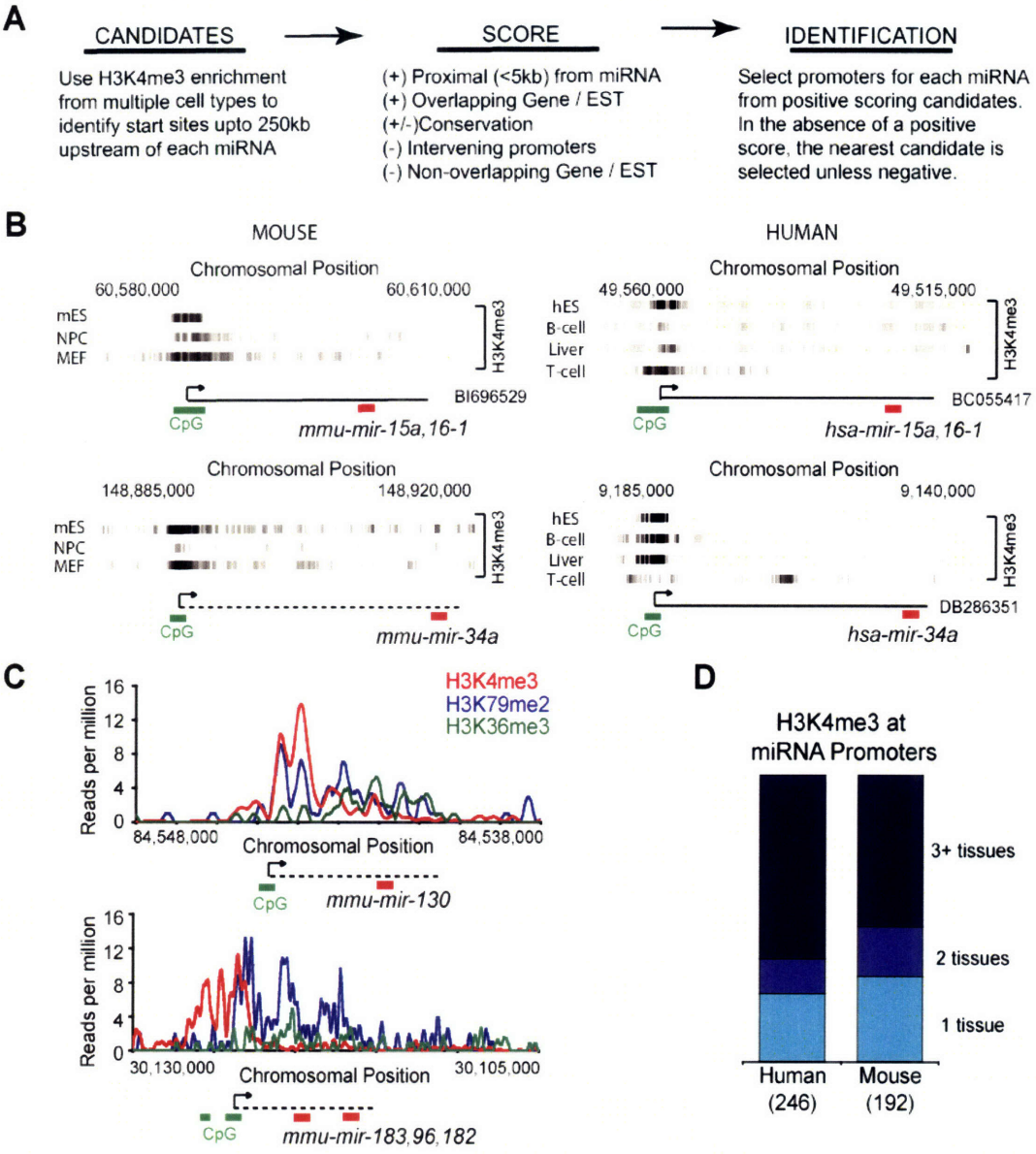

A
CANDIDATES
Use H3K4me3 enrichment from multiple cell types to identify start sites upto 250kb upstream of each miRNA
SCORE
(+) Proximal (<5kb) from miRNA
(+) Overlapping Gene / EST
(+/-)Conservation
(-) Intervening promoters
(-) Non-overlapping Gene / EST
IDENTIFICATION
Select promoters for each miRNA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

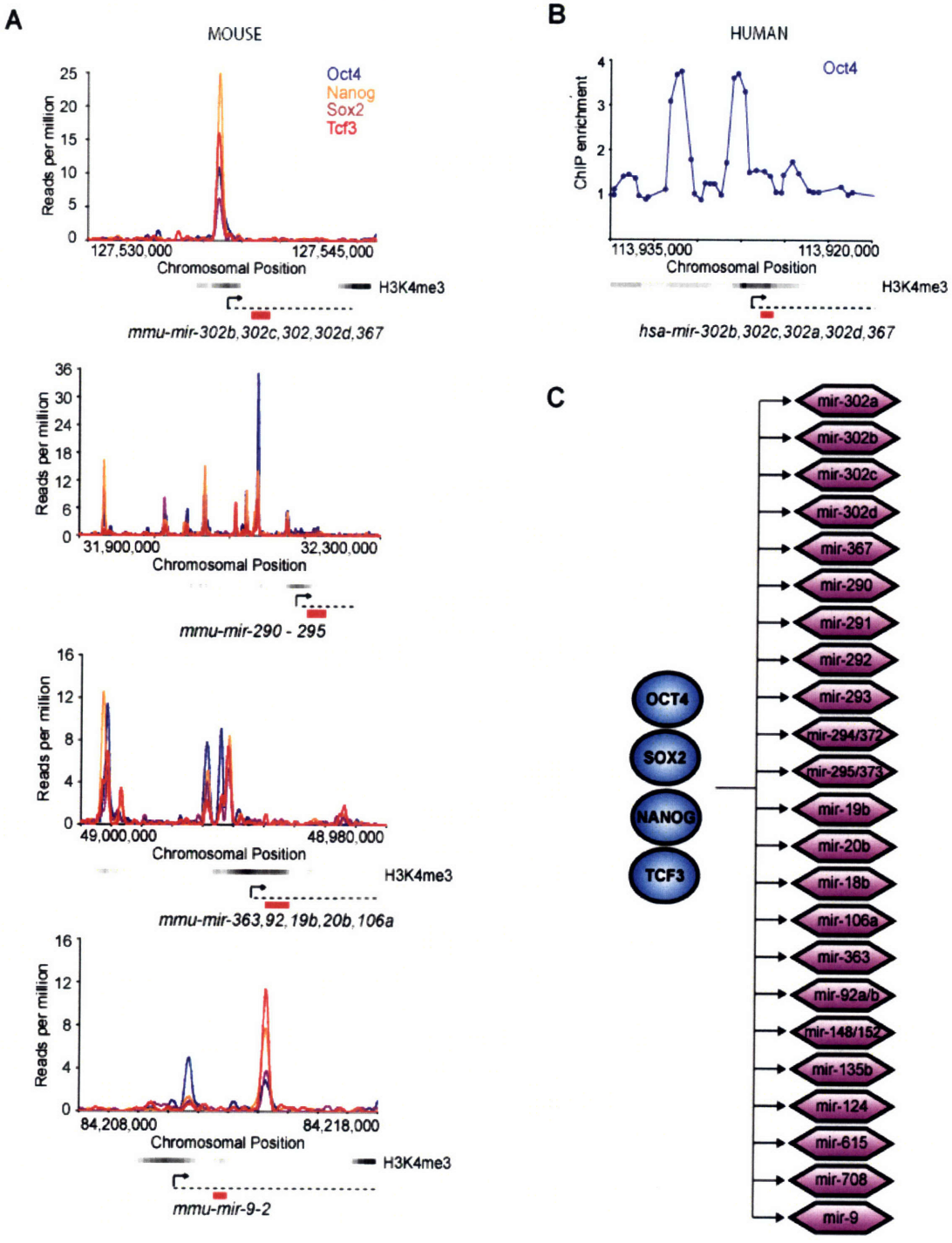

A
MOUSE
Oct4
Nanog
Sox2
Tcf3
Reads per million
Chromosomal Position
127,530,000
127,545,000
H3K4me3
mmu-mir-302b,302c,302,302d,367
31,900,000
32,300,000
mmu-mir-290 - 295
49,000,000
48,980,000
mmu-mir-363,92,19b,20b,106a
84,208,000
84,218,000
mmu-mir-9-2
B
HUMAN
Oct4
ChIP enrichment
113,935,000
113,920,000
hsa-mir-302b,302c,302a,302d,367
C
OCT4
SOX2
NANOG
TCF3
mir-302a
mir-302b
mir-302c
mir-302d
mir-367
mir-290
mir-291
mir-292
mir-293
mir-294/372
mir-295/373
mir-19b
mir-20b
mir-18b
mir-106a
mir-363
mir-92a/b
mir-148/152
mir-135b
mir-124
mir-615
mir-708
mir-9

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).

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 (Supplementary 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 occupies 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 Supplementary 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 Supplementary 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 (Supplementary 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 (Supplementary 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 (Supplementary 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).

Figure 4

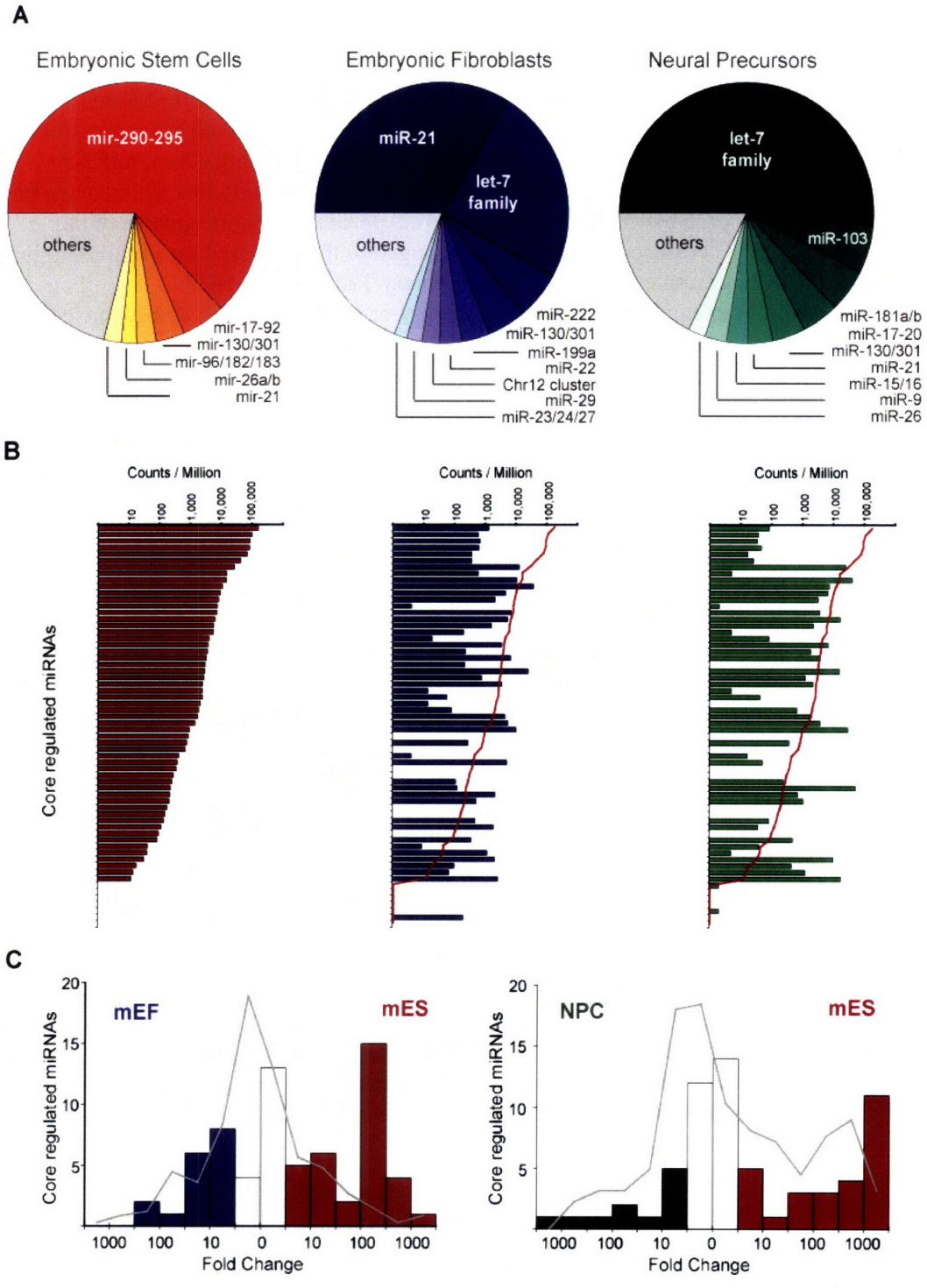

A
Embryonic Stem Cells
mir-290-295
others
mir-17-92
mir-130/301
mir-96/182/183
mir-26a/b
mir-21
Embryonic Fibroblasts
miR-21
let-7 family
others
miR-222
miR-130/301
miR-199a
miR-22
Chr12 cluster
miR-29
miR-23/24/27
Neural Precursors
let-7 family
others
miR-103
miR-181a/b
miR-17-20
miR-130/301
miR-21
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 Supplementary 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

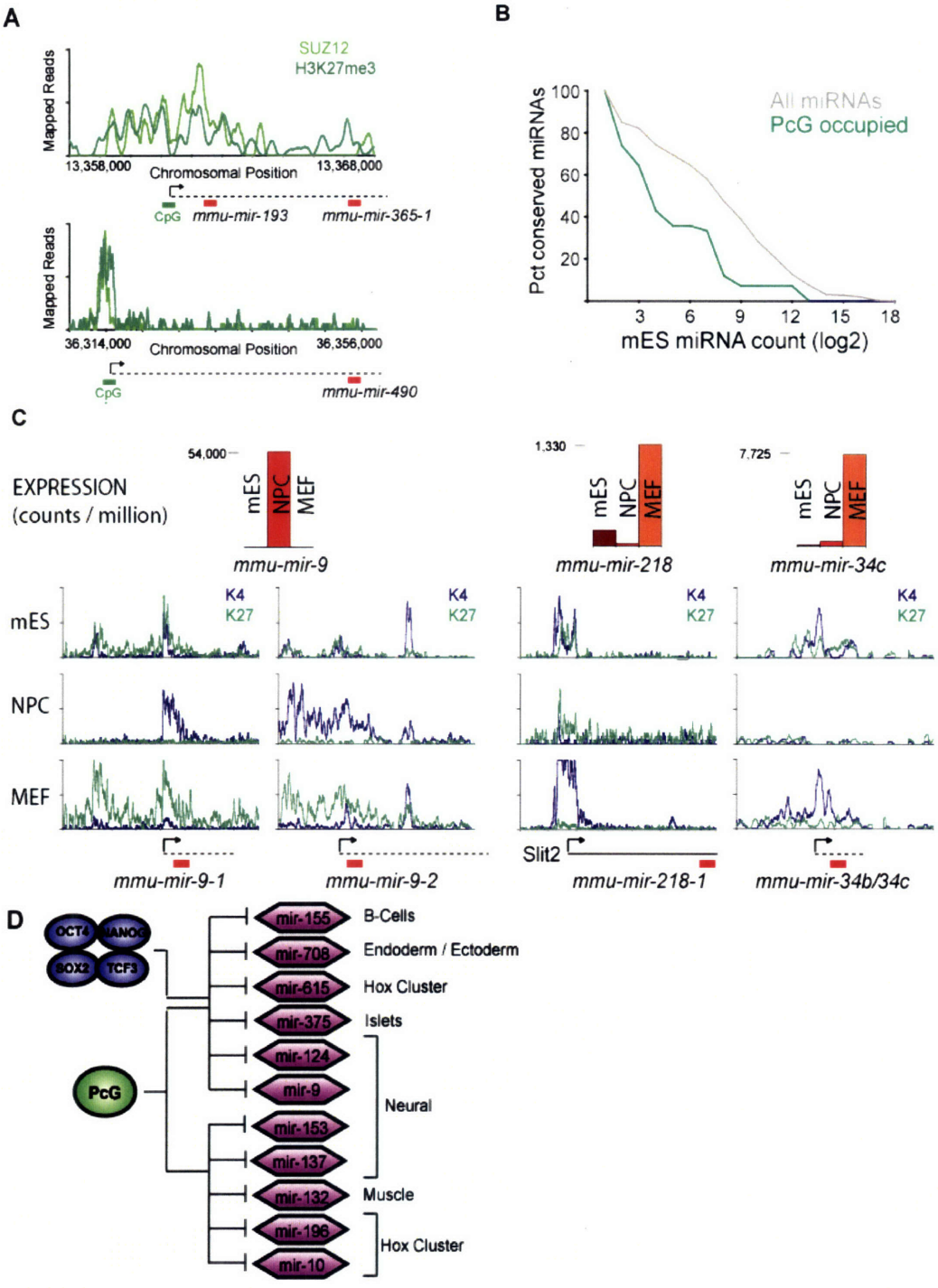

A
SUZ12
H3K27me3
Mapped Reads
13,358,000
Chromosomal Position
13,368,000
CpG
mmu-mir-193
mmu-mir-365-1
Mapped Reads
36,314,000
Chromosomal Position
36,356,000
CpG
mmu-mir-490
B
All miRNAs
PcG occupied
Pct conserved miRNAs
100
80
60
40
20
3
6
9
12
15
18
mES miRNA count (log2)
C
EXPRESSION
(counts / million)
54,000
mES
NPC
MEF
mmu-mir-9
1,330
mES
NPC
MEF
mmu-mir-218
7,725
mES
NPC
MEF
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.

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; Calabrese et al., 2008). 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.

Figure 6

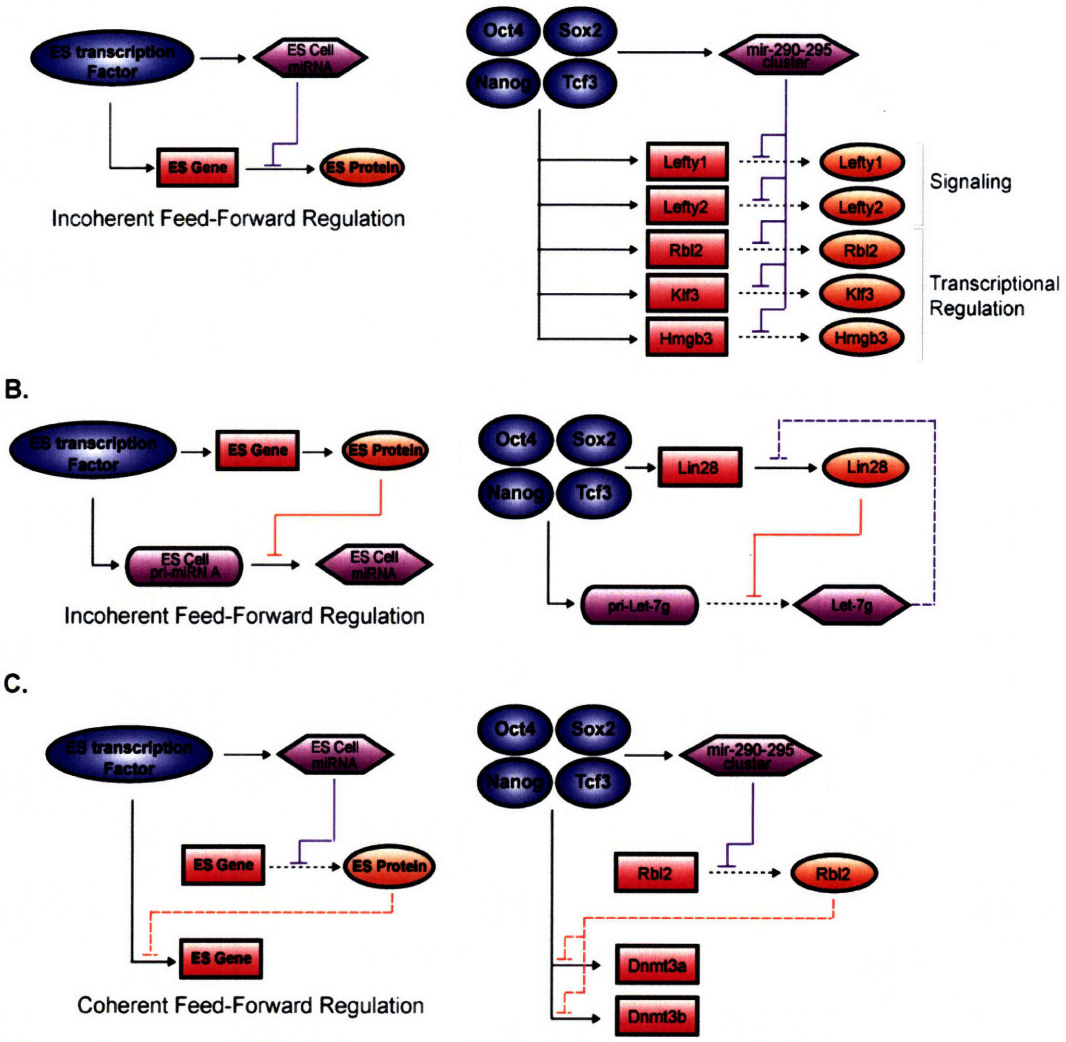

A.
ES transcription Factor
ES Cell miRNA
ES Gene
ES Protein
Incoherent Feed-Forward Regulation
Oct4
Sox2
Nanog
Tcf3
mir-290-295 cluster
Lefty1
Lefty2
Rbl2
Klf3
Hmgb3
Signaling
Transcriptional Regulation
B.
ES transcription Factor
ES Gene
ES Protein
ES Cell pri-miRN A
ES Cell miRNA
Incoherent Feed-Forward 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

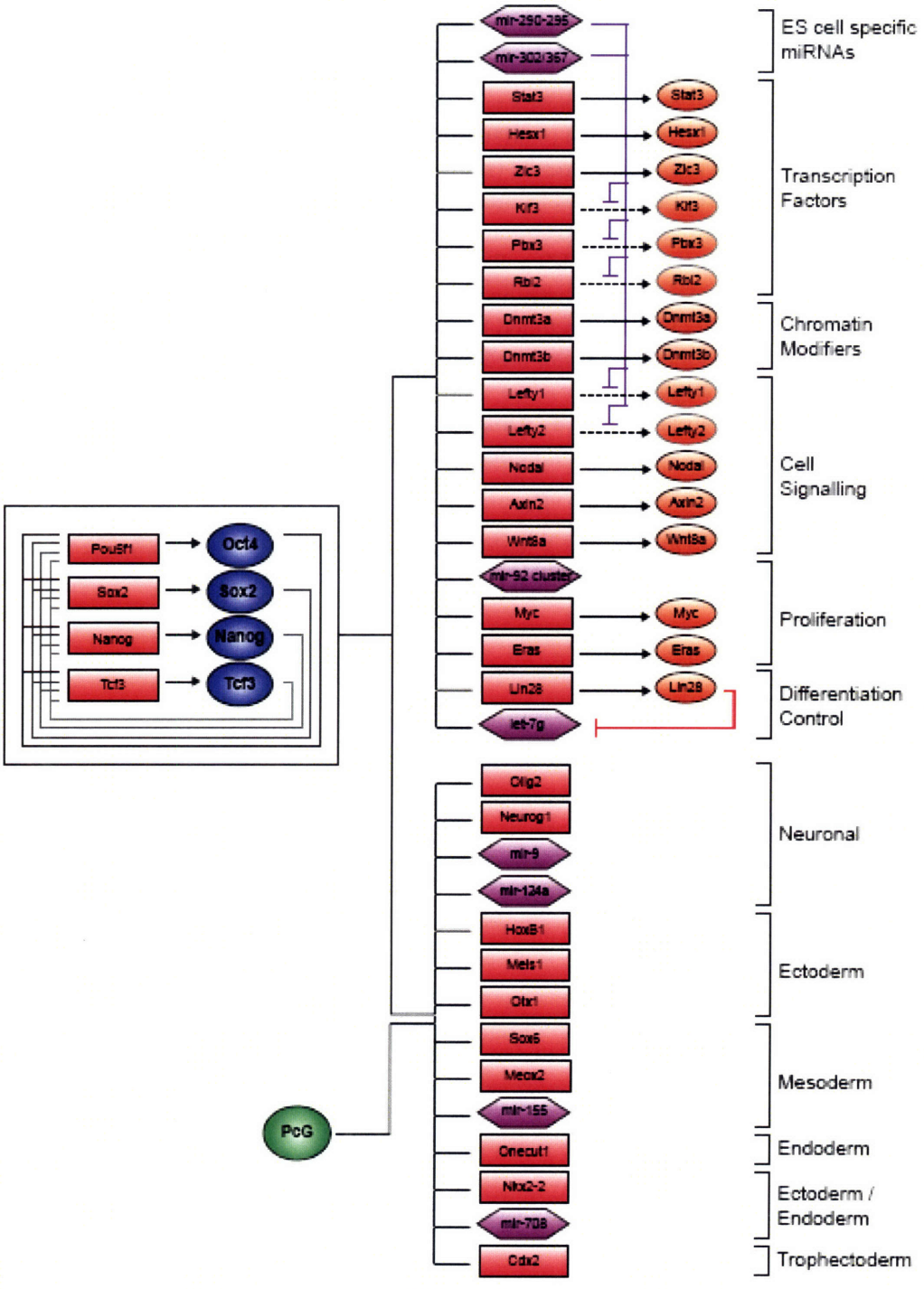

mir-290-295
mir-302/367
Stat3
Hesx1
Zic3
Klf3
Pbx3
Rbl2
Dnmt3a
Dnmt3b
Lefty1
Lefty2
Nodal
Axin2
Wnt8a
mir-92 cluster
Myc
Eras
Lin28
let-7g
Olig2
Neurog1
mir-9
mir-124a
HoxB1
Meis1
Otx1
Sox6
Mecx2
mir-155
Cnecut1
Nkx2-2
mir-708
Cdx2
Pou5f1
Sox2
Nanog
Tcf3
Oct4
Sox2
Nanog
Tcf3
PcG
ES cell specific miRNAs
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.

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 Supplementary Information.

Cell Culture

V6.5 (C57BL/6-129) murine ES cells were grown under typical ES conditions (see Supplementary 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 Supplementary 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 Supplementary Information). Mouse embryonic fibroblasts were prepared and cultured from DR-4 strain mice as previously described (See Supplementary Information).

Antibodies and ChIP assays

Detailed descriptions of antibodies, antibody specificity and ChIP methods used in this study are provided in Supplementary 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 (GUUCAGAGUUCUACAGUCCGACGAUC) was ligated with 4 RNA Ligase.

RT-PCR (Superscript II, Invitrogen) was performed with 5' primer (CAAGCAGAAGACGGCATA). Splicing of overlapping ends PCR (SOEPCR) was performed (Phusion, NEB) with 5' primer and 3' PCR primer (AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA), 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.

Odom, D. T., Dowell, R. D., Jacobsen, E. S., Gordon, W., Danford, T. W., MacIsaac, K. D., Rolfe, P. A., Conboy, C. M., Gifford, D. K., and Fraenkel, E. (2007). Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nat Genet *39*, 730-732.

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, Q., Li, W., Liu, X., Caroll, J., Janne, O., Keeton, E., Chinnaiyan, A., Pienta, K., and Brown, M. (2007). A hierarchical network of transcription factors governs androgen receptor-dependent prostate cancer growth. Mol. Cell *27*: 380-392.

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.

Chapter 6

Concluding Remarks

Concluding Remarks

The work presented in this thesis has piece by piece begun to dissect the network of embryonic stem cells by first mapping the role of key transcription factors and then incorporating inputs from chromatin modifiers, signaling pathways and miRNAs into the network diagram. Together this work has greatly improved our understanding both of how the ES cell state is regulated as well as general properties of vertebrate networks. The approaches taken in these studies to examine the ES cell network can be applied to and guide studies in other vertebrate cell types. These studies have also uncovered or highlighted interesting topics that require further exploration. Here I discuss possible future directions that stem from my thesis work and conclude with a brief description of some recent work that highlights the importance of understanding vertebrate networks and their manipulation.

Transcription Factors

While a handful of key ES cell transcription factors, including Oct4, Sox2 and Nanog, have been incorporated into the map of the ES cell network, there are hundreds of other transcription factors in ES cells that affect gene expression and should be added to the network diagram. Several of these factors have been implicated as playing an important role in ES cells (Hanna et al., 2002; Lim et al., 2007). In order to both understand the role of these other transcription factors in the network and to understand how transcription factors work in combination at target genes to specify expression levels it is important to incorporate inputs from these other factors in the network map. Expanding the network will allow us to identify network hubs, examine the combinatorial nature of gene regulation in vertebrates and may uncover other important network themes.

Binding data of Oct4, Sox2, Nanog and Tcf3 revealed that they bind in extremely close proximity to each other on DNA. This proximity of binding suggests the possibility that these factors may be physically interacting with each other or even forming a complex on chromatin. It would be interesting to perform biochemical studies to examine the possible interactions between these factors and the functional significance of these interactions.

Studies of vertebrate networks are still in the early stage of data generation but as more data is collected it will be important to develop algorithms that make sense of and interpret the data. For example, as we uncover the set of regulators controlling target gene expression it may be possible to determine the logic created by the combination of regulators at a promoter that controls gene expression. It will be interesting to examine whether a 'transcription factor code' can be deduced that could predict transcription levels simply from the set of transcription factors regulating a gene.

Chromatin Regulators

The discovery that Polycomb complexes silence developmental regulators controlling other cell lineages seems to be a fundamental mechanism maintaining cell state. However, as Polycomb complexes do not directly bind DNA it will be important to identify how Polycomb proteins are recruited to the set of silent developmental regulators in ES and other cells. One hypothesis is that non-coding RNAs direct Polycomb to this

set of genes as some Polycomb proteins bind RNA (Bernstein et al., 2006; Zhang et al., 2004).

During differentiation the repression of the subset of developmental regulators necessary for the particular cell lineage being adopted must be relieved without removing repression of other developmental regulators. The removal of Polycomb complexes from these regulators being activated must be a key mechanisms allowing and directing differentiation. Understanding how Polycomb is selectively removed from a particular subset of developmental regulators is therefore an important question that should be explored in future studies.

Studies of chromatin marks revealed that a large number of genes in ES and other cell types have stalled transcription. Given this, it is clear that transcription elongation is a highly regulated process that is likely to play a key role in controlling proper gene expression. It will be important to further explore both the mechanisms regulating transcription elongation as well as the role that this type of regulation plays in the transcriptional regulatory network controlling cell state.

Signaling Pathways

Although input from the Wnt pathway into the ES cell network has begun to be mapped through its terminal component Tcf3, its role in the network that is mediated by the other three Tcf proteins must also be explored. It is not clear how much of Wnt signaling is mediated by Tcf3 versus the other Tcf proteins, which are also expressed in ES cells. Although these proteins can function redundantly they are also known in some instances to have independent roles (Liu et al., 2005). It would be very interesting to examine the role of the other Tcf proteins in the ES cell network and to examine their relative use in mediating Wnt signaling. It would be particularly interesting to profile the role of Tcf4 as this factor is up-regulated upon differentiation and so may mediate the role of Wnt in the process of differentiation.

While the role of the Wnt pathway in maintaining the ES cell state has begun to be teased apart there are several other pathways known to play an important role in maintaining ES cells whose inputs into the network have not yet been examined (Valdimarsdottir and Mummery, 2005; Xu et al., 2005). It will be interesting to determine whether these other pathways also control the ES cell state by directly connecting to the core regulators and their target genes as seems to be the case for the Wnt pathway. It is also known that the effects of signaling pathways can be affected by the status of other pathways and so incorporating inputs from multiple pathways may lead to a better understanding for how they act in a combinatorial manner to determine cell state.

Currently, the only signaling pathway whose input into the network has been studied is the Wnt pathway, which is believed to help maintain the ES cell network. There are several signaling pathways, however, which are known to be involved in the process of ES cell differentiation (Lowell et al., 2006; Ying et al., 2003). It would be very interesting to examine how these pathways initiate changes in the network to direct differentiation down particular lineages.

Non-coding RNAs

Chapter 5 describes an initial study of non-coding RNAs in ES cells that examines the regulation of miRNA genes. The set of genes regulated by these miRNA genes, however, is largely unknown. In order to fully incorporate these miRNA genes into the ES cell network it is necessary to identify their target genes. This is perhaps especially important for the set of miRNA genes that are uniquely expressed in ES cells and so are likely to play an important role in the ES cell network.

While miRNA genes have begun to be incorporated into the map of the ES cell network, there are other classes of non-coding RNAs that play a role in regulating gene expression whose role in the ES cell network is entirely unknown. The incorporation of other types of non-coding RNAs into the network would incorporate additional layers of regulation and deepen our understanding of how multiple levels of regulation act together to specify gene expression.

Medical Applications

A chief motivation for the study and mapping of vertebrate regulatory networks is the belief that a better understanding can guide efforts in reprogramming networks and cell state. Scientists have recently demonstrated the ability to reprogram fibroblast cells into ES cells by forced expression of several ectopic genes (Jaenisch and Young, 2008; Okita et al., 2007; Park et al., 2008; Takahashi and Yamanaka, 2006; Takahashi et al., 2007; Wernig et al., 2007; Yu et al., 2007). This technique has been applied to medical treatments in model organisms. For example, scientists have taken skin cells from mice with sickle cell anemia or Parkinson's disease, reprogrammed these cells to ES cells, differentiated these cells to a useful cell type, injected them into diseased mice and been able to alleviate the disease symptoms (Hanna et al., 2007; Wernig et al., 2008). These studies demonstrate the enormous medical promise offered by reprogramming cell states.

In order to understand the process of cellular reprogramming it will be necessary to study transcriptional regulatory networks as they transition from one cell state to another. The study of dynamic network changes is experimentally quite challenging, however these studies should nonetheless be attempted as they are important both to understanding the creation of cells to be used for medical treatments and to understanding the dynamic nature of networks.

Current reprogramming techniques rely on the expression of ectopic genes to 'jump start' the core network of the desired cell type. This genetic manipulation in order to reprogram cells limits their use in medical treatments as genetic perturbations can lead to increased risk of diseases such as cancer. During development cells do not undergo genetic changes but instead rely on signaling pathways to induce network modifications. Through better understanding how cells use signaling pathways to initialize network changes scientists may be able to leverage this knowledge to manipulate cell networks without the use of ectopic genes. For this reason it will be important for future studies to aggressively pursue the study of signaling pathways and natural network manipulations.

While our understanding of the ES cell network and its manipulation is quickly developing there are many other medically relevant cell types and networks that need to be examined. In order to develop methods to manipulate a patient's cells into the needed cell type it will likely be important to understand the network governing the desired cell

type. Additionally, many diseases are caused by dysregulation of a cell's network and so by understanding the errors in the network scientists may be able to develop methods to correct these errors. It is therefore imperative to begin dissecting the networks of other medically relevant cell types.

Summary

The work presented in this thesis together begin to form a coherent picture of the regulatory network governing the embryonic stem cell state. Transcriptional regulatory networks are quite complex, especially in higher organisms, and involve many layers of regulation that we are only beginning to understand. Each study presented here attacked the network from a distinct angle but nonetheless function together to create a unified picture of the control of ES cells and of vertebrate networks in general. As these studies ventured into the relatively new field of vertebrate regulatory networks they both uncovered many important themes and highlighted the necessity of numerous future studies. Given the developmental and medical relevance of the field it seems clear that it will receive much attention, progress quickly, and yield many important insights.

References

Bernstein, E., Duncan, E., Masui, O., Gil, J., Heard, E., and Allis, C. (2006). Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin. Mol. Cell Biol. *26*: 2560-2569.

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

Hanna, J., Wernig, M., Markoulaki, S., Sun, C., Meissner, A., Cassady, J., Beard, C., Brambrink, T., Wu, L., Townes, T., and Jaenisch, R. (2007). Treatment of sickle cell anemia mouse model with iPS cells generated from autologous skin. Science *318*: 1920-1923.

Hanna, L., Foreman, R., Tarasenko, I., Kessler, D., and Labosky, P. (2002). Requirement for Foxd3 in maintaining pluripotent cells of the early mouse embryo. Genes Dev. *16*: 2650-2661.

Lim, L., Loh, Y., Zhang, W., Li, Y., Chen, X., Wang, Y., Bakre, M., Ng, H., and Stanton, L. (2007). Zic3 is required for maintenance of pluripotency in embryonic stem cells. Mol. Biol. Cell *18*: 1348-1358.

Liu, F., van den Broek, O., Destree, O., and Hoppler, S. (2005). Distinct roles for Xenopus Tcf.Lef genes in mediating specific responses to Wnt/B-catenin signaling in mesoderm development. Development *132*: 5375-5385.

Lowell, S., Benchoua, A., Heavey, B., and Smith, A. G. (2006). Notch promotes neural lineage entry by pluripotent embryonic stem cells. PLoS Biol *4*, e121.

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

Park, I., Zhao, R., West, J., Yabuuchi, A., Huo, H., Ince, T., Lerou, P., Lensch, M., and Daley, G. (2008). Reprogramming of human somatic cells to pluripotency with defined factors. Nature *451*: 141-146.

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

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

Valdimarsdottir, G., and Mummery, C. (2005). Functions of the TGFbeta superfamily in human embryonic stem cells. APMIS *113*: 773-789.

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

Wernig, M., Zhao, J., Pruszak, J., Hedlund, E., Fu, D., Soldner, F., Broccoli, V., Constantine-Paton, M., Isacson, O., and Jaenisch, R. (2008). Neurons derived from reprogrammed fibroblasts functionally integrate into the fetal brain and improve symptoms of rats with parkinson's disease. Proc. Natl. Acad. Sci. *105*: 5856-5861.

Xu, C., Rosler, E., Jiang, J., Lebkowski, J., Gold, J., O'Sullivan, C., Delavan-Boorsma, K., Mok, M., Bronstein, A., and Carpenter, M. (2005). Basic fibroblast growth factor supports undifferentiated human embryonic stem cell growth without conditioned medium. Stem Cells *23*: 315-323.

Ying, Q. L., Stavridis, M., Griffiths, D., Li, M., and Smith, A. (2003). Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. Nat Biotechnol *21*, 183-186.

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

Zhang, H., Christoforou, A., Aravind, L., Emmons, S., van den Heuvel, S., and Haber, D. (2004). The C. elegans polycomb gene SOP-2 encodes an RNA binding protein. Mol. Cell *14*: 841-847.

Appendix A

Genome-wide Map of Nucleosome Acetylation and Methylation in Yeast

Published as: Dmitry K. Pokholok, Christopher T. Harbison, Stuart Levine, Megan Cole, Nancy M. Hannett, Tong Ihn Lee, George W. Bell, Kimberly Walker, P. Alex Rolfe, Elizabeth Herbolsheimer, Julia Zeitlinger, Fran Lewitter, David K. Gifford, and Richard A. Young. (2005). "Genome-wide Map of Nucleosome Acetylation and Methylation in Yeast." Cell 122: 517-527.

My contribution to this project

This project was led by Dmitri Pokholok and Chris Harbison. I worked with Stuart Levine and Elizabeth Herbolsheimer to develop an error model and gene calling algorithm for use with genome-wide binding data. I also played a large part in determining the proper experimental and computational normalizations for the histone mark ChIPs.

Summary

Eukaryotic genomes are packaged into nucleosomes whose position and chemical modification state can profoundly influence regulation of gene expression. We profiled nucleosome modifications across the yeast genome using chromatin immunoprecipitation coupled with DNA microarrays to produce high-resolution genome-wide maps of histone acetylation and methylation. These maps take into account changes in nucleosome occupancy at actively transcribed genes and, in doing so, revise previous assessments of the modifications associated with gene expression. Both acetylation and methylation of histones are associated with transcriptional activity, but the former occurs predominantly at the beginning of genes, whereas the latter can occur throughout transcribed regions. Most notably, specific methylation events are associated with the beginning, middle, and end of actively transcribed genes. These maps provide the foundation for further understanding the roles of chromatin in gene expression and genome maintenance.

Introduction

The genomes of eukaryotes are packaged into chromatin, the fundamental unit of which is the nucleosome. Nucleosomes consist of approximately 146 base pairs of DNA wrapped around a histone octamer (Luger et al., 1997). The histone components of nucleosomes and additional chromatin proteins can interact to form higherorder chromosomal structures. Nucleosomes are thus critical to the organization and maintenance of genetic material, and their position and modification state can profoundly influence genetic activities such as regulation of gene expression (Kouzarides, 2002; Narlikar et al., 2002).

Several recent studies have explored the relative occupancy and modification state of nucleosomes across the yeast genome by using chromatin immunoprecipitation of histones and DNA-microarray analysis, a technique known as "genome-wide location analysis" or "ChIP-chip" (Bernstein et al., 2002, 2004; Kurdistani et al., 2004; Lee et al., 2004; Robyr et al., 2002; Santos- Rosa et al., 2002). The results of these studies, which used DNA microarrays that probed a single site in the intergenic or transcribed portions of the genome, suggested that the intergenic regions of *S. cerevisiae* are less densely occupied by nucleosomes than transcribed regions (Lee et al., 2004) and that nucleosome density in both regions inversely correlates with transcription rates (Bernstein et al., 2004; Lee et al., 2004). Similar studies have explored the relationship between transcription and genome occupancy by chromatin regulators (Humphrey et al., 2004; Kurdistani et al., 2002; Lieb et al., 2001; Ng et al., 2002b, 2003b; Robert et al., 2004; Robyr et al., 2002) or transcription and modification of nucleosomes (Kurdistani et al., 2004; Robyr et al., 2002; Roh et al., 2004). Some conclusions from these studies are difficult to reconcile with one another; for example, the histone acetyltransferases Gcn5 and Esa1 appear to be recruited to genes upon transcriptional activation (Robert et al., 2004), but acetylation of the amino acid residues they target in histones does not appear to be strongly associated with transcriptional activity (Kurdistani et al., 2004).

To more accurately define nucleosomal occupancy and modification and their relationships to transcription, we have investigated genome-wide chromatin structure at considerably higher resolution than that afforded by the experimental designs that have been used thus far. Such designs have typically used DNA microarrays with a single feature to capture signals from each intergenic or transcribed region (Bernstein et al., 2004; Kurdistani et al., 2004; Lee et al., 2004). We describe here profiling of protein-DNA occupancy at high resolution and accuracy using improved protocols and DNA microarrays that tile the yeast genome. The results we obtain with this approach provide a more complete picture of nucleosome occupancy and the nature of modifications associated with transcriptional activity, resolve discrepancies in previous reports, and allow us to produce the first high-resolution maps of histone acetylation and methylation in the yeast genome.

Results and Discussion

High-Resolution Genome-wide ChIP-Chip

To increase the resolution and accuracy of genomewide location analysis, we designed a DNA microarray that contains over 40,000 probes for the yeast genome and developed hybridization methods that maximized signal-to-noise ratios on this array (see Experimental Procedures). To test whether these modifications improved the resolution and accuracy of genome-wide binding analysis, we explored the genome-wide occupancy of Gcn4, a transcriptional regulator of amino acid-biosynthetic genes with a well-characterized DNA binding specificity (Hope and Struhl, 1985; Oliphant et al., 1989), crystal structure (Ellenberger et al., 1992; O'Shea et al., 1991), and previously identified target genes (Arndt and Fink, 1986; Natarajan et al., 2001). The binding data for individual target genes are shown in Figure 1A and in Figure S1 in the Supplemental Data available with this article online. For those regions for which there is strong evidence for Gcn4 binding, we found that the peak of Gcn4 binding occurred directly over that binding site.

To test the accuracy of the new method, we identified a test set of 84 genes most likely to be targeted by Gcn4 in vivo (Figure 1B, Table S1, Experimental Procedures) and a set of 945 genes least likely to be targeted by Gcn4; the selection criteria for these sets of genes is described in Experimental Procedures. Based on these positive and negative Gcn4 targets, analysis of Gcn4 binding with the new method suggests a falsepositive rate of less than 1% and a false-negative rate of ~25%, corresponding with a total of 210 genes whose promoters are bound within the optimal p-value threshold of 6×10^{-6} (Experimental Procedures). These results demonstrate that the new array and protocol modifications provide substantially higher resolution and accuracy than our previous method using self-printed arrays (Harbison et al., 2004; Lee et al., 2002).

Global Nucleosome Occupancy

The improved accuracy and resolution of this ChIP-Chip method was used to investigate nucleosome occupancy and modification throughout the yeast genome. When we examined histone occupancy with antibodies against core histone H3 or histone H4, using genomic DNA as the reference channel, we found a relatively high density of nucleosomes over transcribed regions and a lower density over intergenic regions (Figures 1C and 1D). Figure 1C shows a stereotypical example of histone occupancy at a portion of chromosome XV. Figure 1D presents composite profiles of histone H3 and H4 for 5324 genes aligned according to the location of translation initiation and termination sites. There was an ~20% reduction in histone occupancy in intergenic sequences relative to genic sequences for the average gene. These results are consistent with previous observations (Lee et al., 2004) and suggest that the majority of yeast genes have higher nucleosome density over transcribed regions relative to intergenic regions.

We were surprised to find that differential enrichment of intergenic and genic regions also occurred in control experiments lacking antibody (compare Figure 2A and Figure 1D). Results similar to those in Figure 2A were obtained in control experiments when ChIPs were performed with antibodies directed against nonhistone proteins (data not shown). Others have noted that different relative levels of intergenic and genic DNA are recovered using various extraction strategies (Nagy et al., 2003), but control data of

Figure 1

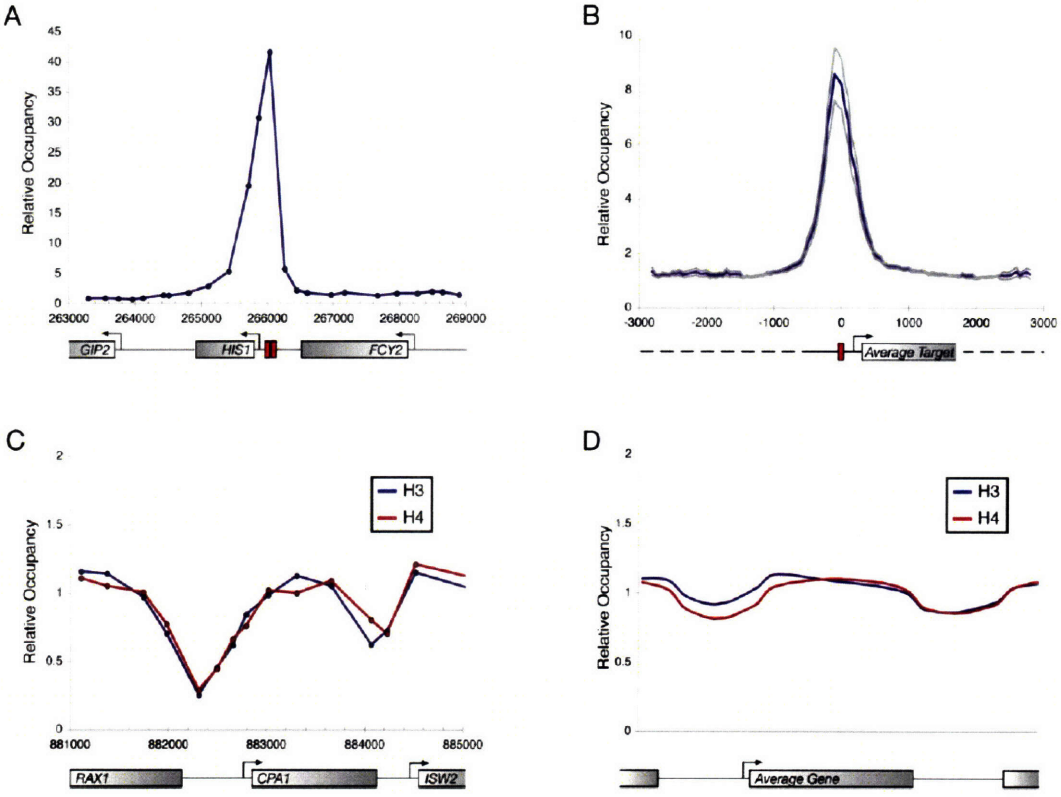

A
Relative Occupancy
45
40
35
30
25
20
15
10
5
0
263000
264000
265000
266000
267000
268000
269000
GIP2
HIS1
FCY2
B
Relative Occupancy
10
8
6
4
2
0
-3000
-2000
-1000
0
1000
2000
3000
Average Target
C
Relative Occupancy
2
1.5
1
0.5
0
H3
H4
881000
882000
883000
884000
885000
RAX1
CPA1
ISW2
D
Relative Occupancy
2
1.5
1
0.5
0
H3
H4
Average Gene

Figure 1. Nucleosome Occupancy across the Yeast Genome with High-Resolution Genome-wide Location Analysis

(A) Occupancy of the *HIS1* promoter by Gcn4. The genomic positions of probe regions are arrayed along the x axis, with the ratio of enrichment of Gcn4 for probes along the y axis. ORFs are depicted as gray rectangles, and arrows indicate the direction of transcription. Red boxes represent sequence matches to the Gcn4 binding specificity within promoter regions.

(B) Composite profile of Gcn4 binding at the set of 84 high-confidence Gcn4 target genes. Promoter and downstream regions were aligned with each other according to the position of a sequence match to the Gcn4 binding specificity. Aligned probes were then assigned to 50 bp segment bins, and an average of the corresponding enrichment ratio was calculated. Standard error of the mean is shown in gray. Genetic elements are depicted as in Figure 1A, except that dashed lines represent sites including both ORFs and intergenic regions.

(C) Nucleosome occupancy at the promoter of *CPA1*, a gene encoding an amino acid biosynthetic enzyme. The genomic positions of probe regions are arrayed along the x axis, with the ratio of enrichment of histone H3 (blue) or H4 (red) for probes along the y axis. ORFs are depicted as gray rectangles, and arrows indicate the direction of transcription.

(D) A composite profile of histone occupancy at 5324 genes. The ends of ORFs were defined at fixed points according to the position of translational start and stop sites. The length of the ORF was then subdivided into 40 regions of equal length, and probes were assigned according to their nearest corresponding relative position. Probes in promoter regions were similarly assigned following subdivision into 20 regions. The average histone H3 (blue) or H4 (red) enrichment for each subdivided bin is plotted.

Figure 2

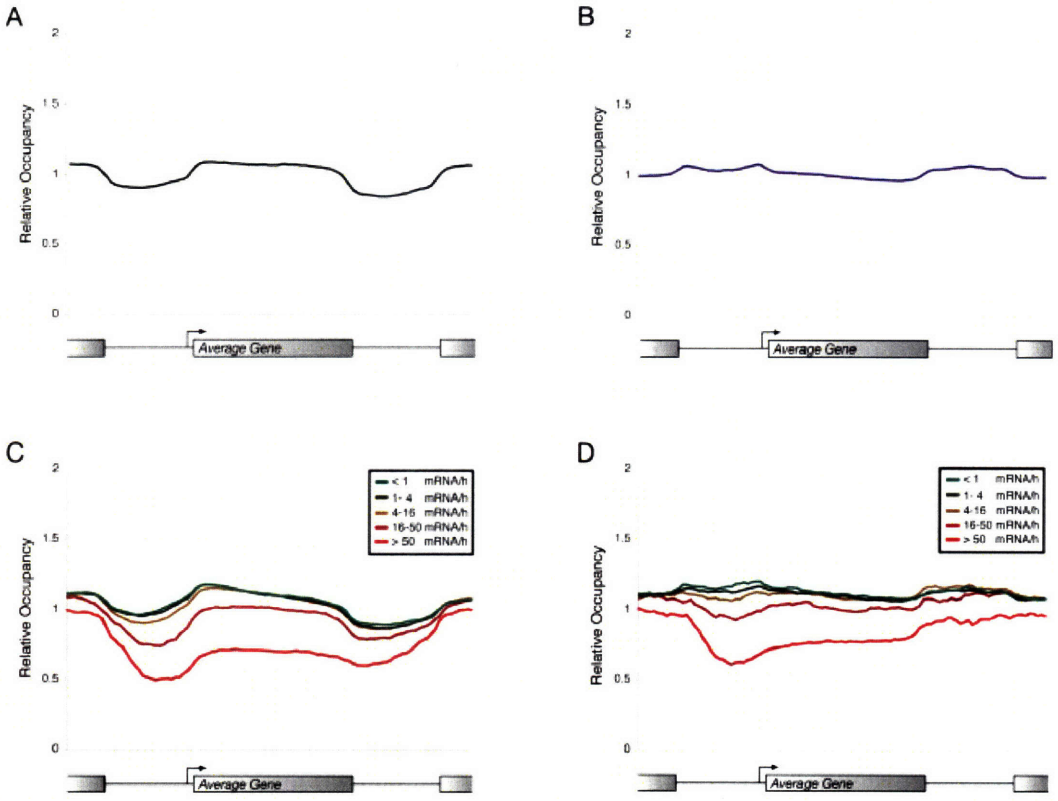

A
B
C
D
Relative Occupancy
2
1.5
1
0.5
0
Average Gene
< 1 mRNA/h
1- 4 mRNA/h
4-16 mRNA/h
16-50 mRNA/h
> 50 mRNA/h

Figure 2. Comparisons of Histone Profiles
(A) A composite profile of enrichment in a control experiment. The profile is created as in Figure 1D, except that enrichment is measured from a mock immunoprecipitation, in which no antibody has been included (Experimental Procedures).
(B) A composite profile of histone occupancy normalized to a control. The profile is created as in Figure 1D, except that enrichment from H3 immunoprecipitation is normalized to enrichment from mock immunoprecipitations.
(C) A composite profile of histone occupancy according to transcriptional activity. All genes for which data were available (Holstege et al., 1998) were divided into five classes according to their transcriptional rate. Composite data were computed for H3 enrichment as in Figure 1D.
(D) A composite profile of normalized histone occupancy according to transcriptional activity. The composite profile is created as in Figure 2C, except that enrichment from H3 immunoprecipitation is normalized to enrichment from mock immunoprecipitations.

this type have not yet been used to normalize the results of histone ChIP studies (Bernstein et al., 2004, 2005; Kurdistani et al., 2004; Lee et al., 2004). When these control experiments were used to normalize the histone H3 data, we found that there were not substantial differences in the relative levels of intergenic versus genic DNA at the average gene (Figure 2B). Nonetheless, approximately 40% of yeast promoters do have lower levels of histones than their downstream transcribed regions, even after the normalization by control experiments (Figure S2), and we show below that these are associated with transcribed genes.

To examine the relationship between gene expression and nucleosome occupancy, we assigned genes into five different classes depending on their transcriptional rate (Holstege et al., 1998) and created a composite histone H3 profile for each class (Figures 2C and 2D). The composite histone profile in Figure 2C was generated by using whole genomic DNA in the reference channel, and that in Figure 2D was generated by normalizing to a no-antibody control ChIP. The results in both profiles confirm that nucleosome occupancy at both promoter and transcribed regions inversely correlates with gene activity, in agreement with previous gene-specific and genome-wide studies (Bernstein et al., 2004; Boeger et al., 2003; Lee et al., 2004; Reinke and Horz, 2003). The results shown in Figure 2D also suggest that nucleosome occupancy is reduced maximally at the promoters of active genes. In contrast, the promoters of transcriptionally inactive genes are as densely populated with nucleosomes as genic regions.

If gene activation leads to reduced nucleosome occupancy, then dynamic activation of specific genes should cause reduced histone levels at these newly transcribed genes. To test this notion, we performed ChIP-Chip with histone antibodies on cells before and after exposure to oxidative stress (Causton et al., 2001). At genes known to be activated by oxidative stress (e.g., *HSP30* and *HSP82*), nucleosome occupancy dropped substantially (Figure S3). These results confirm that gene activation leads to reduced nucleosome density in both promoter and transcribed regions, with the greatest effect occurring at the promoter.

Histone Acetylation

The histone acetylases Gcn5 and Esa1 are generally recruited to the promoter regions of active genes (Robert et al., 2004), and thus we would expect that the amino acid residues that are substrates of these HATs would be preferentially acetylated at active genes. A recent genome-wide study, however, reported little correlation between transcriptional activity and acetylation of the histone H3 and H4 amino acid residues targeted by Gcn5 and Esa1 (Kurdistani et al., 2004). To understand the source of these discrepancies, we used the new methods to investigate selected histone modifications genome-wide.

Histone H3 lysine 9 acetylation (H3K9ac) and histone H3 lysine 14 acetylation (H3K14ac) are among the modifications catalyzed by Gcn5 (Kuo et al., 1996; Utley et al., 1998; Zhang et al., 1998). We used ChIP-Chip to measure the levels of H3K9ac relative to the levels of core histone H3 genome-wide. The results show that acetylation of histone H3 at lysine 9 peaks at the predicted transcriptional start sites of active genes (Figure 3A) and that this modification correlates with transcription rates genome-wide (Figure 3B). We also found that acetylation of histone H3 at lysine 14 peaks over the start sites of active genes (Figure 3C) and correlates with transcription rates genome-wide

(Figure 3D). We conclude that there is a positive association between Gcn5, the modifications known to be catalyzed by Gcn5, and transcriptional activity (Figure 3 and Figure S4A).

Four lysine residues of histone H4 are acetylated by Esa1, an acetyltransferase associated with the NuA4 complex (Allard et al., 1999; Clarke et al., 1999; Vogelauer et al., 2000). We measured the levels of hyperacetylated histone H4 relative to core histone genome-wide using ChIP-Chip with an antibody that recognizes histone H4 acetylated at lysines 5, 8, 12, and 16 (H4K5ac8ac12ac16ac). The results showed that H4 hyperacetylation peaks over the start sites of active genes (Figure 3E) and correlates with transcription rates (Figure 3F), although the association is not as strong as that observed for H3K9ac and H3K14ac. Our analysis cannot exclude the possibility that acetylation of individual lysine residues in the N-terminal tail of histone H4 might correlate differently with transcriptional activity. Nonetheless, our data reveal a positive, albeit modest, correlation between Esa1 occupancy, the modifications known to be catalyzed by this enzyme, and transcriptional activity (Figure 3 and Figure S4B).

To ascertain whether dynamic gene activation leads to the expected increase in histone acetylation at sites catalyzed by Gcn5 and Esa1, we performed ChIP-Chip with the relevant histone antibody on cells before and after exposure to oxidative stress. The results confirm that gene activation leads to increased histone acetylation at sites catalyzed by Gcn5 and Esa1 in the promoter and transcribed regions of activated genes (Figure S5).

In general, we find that histones with the acetylated residues studied here are enriched predominantly at promoter regions and transcriptional start sites of active genes and that enrichment drops substantially across the ORFs (Figure 3 and Figure S5). This is consistent with the model that transcriptional activators generally recruit Gcn5 and Esa1 to promoters of genes upon their activation (Robert et al., 2004) and with the idea that the two HATs acetylate local nucleosomes when recruited to these genes. Our conclusion that there is a strong correlation between transcriptional activity and acetylation of the histone H3 and H4 amino acid residues targeted by Gcn5 and Esa1 is in contrast to that of Kurdistani et al. (2004). This discrepancy is most likely due to differences in the material used in the control channel in the ChIP-Chip procedure. The experiments described here compare ChIP with a histone-modification antibody to a control ChIP with a core histone antibody. The experiments reported in Kurdistani et al. (2004) used whole genomic DNA in the reference channel. We found we could replicate the results in Kurdistani et al. (2004) if we used whole genomic DNA as a reference in ChIP-Chip experiments (Figure S6), but for reasons described above, this method of normalization is inappropriate.

Histone Methylation

Methylation of histones in *S. cerevisiae* is carried out by three known histone methyltransferases, which are capable of covalently modifying specific lysine residues in histone H3 with up to three methyl groups (Peterson and Laniel, 2004). Since characterization of specific methylation marks has been studied primarily at the level of individual genes (Bannister et al., 2005; Bernstein et al., 2002; Krogan et al., 2003b; Ng et al., 2003a; Xiao et al., 2003), we sought to systematically profile mono-, di-, and trimethylated residues at K4, K36, and K79 of histone H3 in nucleosomes associated with genomic DNA.

Figure 3

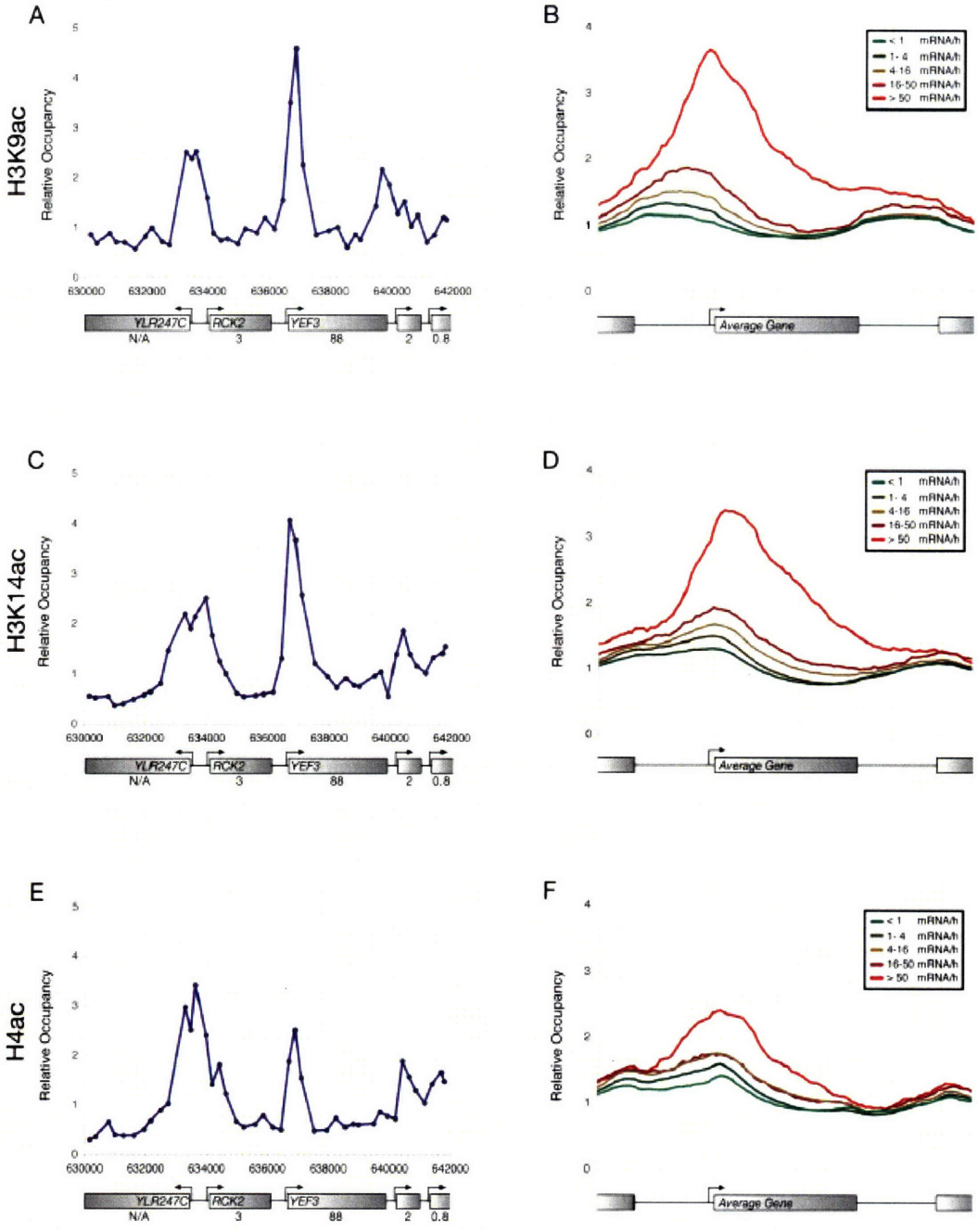

A
H3K9ac
Relative Occupancy
B
Relative Occupancy
< 1 mRNA/h
1 - 4 mRNA/h
4-16 mRNA/h
16-50 mRNA/h
> 50 mRNA/h
Average Gene
C
H3K14ac
Relative Occupancy
D
Relative Occupancy
E
H4ac
Relative Occupancy
F
Relative Occupancy
630000
632000
634000
636000
638000
640000
642000
YLR247C
RCK2
YEF3
N/A
3
88
2
0.8

Figure 3. Nucleosome Acetylation Generally Correlates with Transcriptional Activity
(A) Acetylation of H3K9 at a locus on chromosome XII. Enrichment is depicted as in Figure 1. The number beneath each gene represents the transcriptional frequency of the corresponding ORF (Holstege et al., 1998) in mRNA/hr.
(B) Composite profile of acetylation of H3K9 across the average gene. Composite profiles of acetylation according to transcriptional-frequency class are shown as in Figure 2.
(C) Acetylation of H3K14 at a locus on chromosome XII. Enrichment is depicted as in Figure 1. The number beneath each gene represents the transcriptional frequency of the corresponding ORF (Holstege et al., 1998) in mRNA/hr.
(D) Composite profile of acetylation of H3K14 across the average gene. Composite profiles of acetylation according to transcriptional frequency class are shown as in Figure 2.
(E) Hyperacetylation of H4 at a locus on chromosome XII. Enrichment is depicted as in Figure 1. The number beneath each gene represents the transcriptional frequency of the corresponding ORF (Holstege et al., 1998) in mRNA/hr.
(F) Composite profile of hyperacetylation of H4 across the average gene. Composite profiles of acetylation according to transcriptional frequency class are shown as in Figure 2.

The histone methyltransferase Set1 has previously been shown to be recruited to the 5' end of actively transcribed genes where it is responsible for histone H3K4 methylation (Bernstein et al., 2002; Briggs et al., 2001; Krogan et al., 2003a; Ng et al., 2003b; Santos-Rosa et al., 2002). We measured histone H3K4 trimethylation (H3K4me3) using ChIP-Chip and found that the results confirm previous studies (Santos-Rosa et al., 2002) and provide a higher-resolution picture of H3K4 trimethylation across the yeast genome (Figures 4A and 4B). Peaks of histone H3K4 trimethylation occurred at the beginning of actively transcribed genes, and there was a positive correlation between this modification and transcription rates (Figures 4A and 4B).

We also investigated the profiles of mono- and dimethylated histone H3K4-containing nucleosomes and found that they exhibit a pattern distinct from that observed for trimethylated histone H3K4 (Figure S7). While trimethylated H3K4 peaks at the beginning of the transcribed portions of genes, dimethylated H3K4 (H3K4me2) is most enriched in the middle of genes, and monomethylated H3K4 (H3K4me) is found predominantly at the end of genes.

We measured genome-wide the relative levels of H3K36 trimethylation, which is catalyzed by Set2, a factor associated with the later stages of transcriptional elongation (Strahl et al., 2002). In contrast to the pattern observed with H3K4 trimethylated histones, we found that trimethylated H3K36 (H3K36me3) was enriched throughout the coding region, peaking near the 3# ends of transcription units (Figures 4C and 4D). H3K36 trimethylation also correlated with transcriptional activity. These results are consistent with the model that Set2 is recruited by the transcription elongation apparatus and that it methylates local nucleosomes during active transcription.

The Dot1 histone methyltransferase modifies histone H3 lysine 79 (H3K79), which occurs within the core domain of histone H3 (Feng et al., 2002; Ng et al., 2002a, 2003a). Methylation of this residue is estimated to occur in ~90% of all histones and is associated with telomeric silencing control in yeast (Ng et al., 2003a; van Leeuwen et al., 2002), but global ChIP profiles of dimethylated H3K79 in *Drosophila* (Schubeler et al., 2004) have linked this modification to active transcription. We investigated the genomic profile of H3K79 trimethylation in yeast (H3K79me3) and found that histones with this modification are enriched within the transcribed regions of genes (Figures 4E and 4F). Most genes appeared to have nucleosomes modified at H3K79; there was little correlation between the relative levels of H3K79 trimethylation at genes and transcriptional activity (Figure 4F).

Global Map of Histone Marks

We recently mapped the locations of conserved transcription-factor binding sites throughout the yeast genome (Harbison et al., 2004). We used the results described here to generate a complementary genome-wide map of nucleosome occupancy and histone modifications that includes results for eight sets of histone modifications (H3K9ac, H3K14ac, H4K5ac8ac12ac16ac, H3K4me1, H3K4me2, H3K4me3, H3K36me3, and H3K79me3). A portion of this map is shown in Figure 5, and a browsable form of the complete yeast-genome chromatin map is available at the authors' website (http://web.wi.mit.edu/young/nucleosome).

Concluding Remarks

It is well established that nucleosomes play fundamentally important roles in the organization and maintenance of the genome. Nucleosome modifications have been shown to be associated with transcriptional regulation at well-studied genes, and models have emerged that connect regulation of gene expression to histone modification by specific chromatin regulators (Cosma et al., 1999; Gregory et al., 1999; Kuo et al., 1998; Reinke and Horz, 2003). We have carried out a systematic genome- wide analysis of nucleosome acetylation and methylation at sufficient resolution to determine whether models that connect regulation of gene expression to histone modification (Deckert and Struhl, 2001; Reid et al., 2000; Reinke et al., 2001; Suka et al., 2002) apply to gene regulation throughout the yeast genome.

The results described here, taken together with recent discoveries, are consistent with the following general model connecting gene expression to histone modification. Transcriptional activation by DNA binding regulators generally involves recruitment of Gcn5 and Esa1 to promoters, where these HATs acetylate specific residues on histones H3 and H4 at local nucleosomes (Figure 3) (Bhaumik and Green, 2001; Cosma et al., 1999; Larschan and Winston, 2001; Reid et al., 2000). We were able to find few exceptions to this general rule, where only one or the other HAT acetylates its target residues at the promoters of actively transcribed genes (data not shown). Active transcription is characteristically accompanied by histone H3K4 trimethylation by Set1 at the beginning of genes (Figure 4B) and by H3K4 dimethylation and monomethylation at nucleosomes positioned further downstream in the transcription unit (Figure S7) (Krogan et al., 2003a; Ng et al., 2003b). As the transcription apparatus proceeds down the transcription unit, increasing levels of histone H3K36 trimethylation are observed at most active genes, catalyzed by Set2 (Figure 4D) (Krogan et al., 2003b; Strahl et al., 2002). Histone H3K79me3, which is catalyzed by Dot1 (Feng et al., 2002; Ng et al., 2003a; van Leeuwen et al., 2002), is enriched within genes, but, unlike the other modifications studied here, this enrichment is not clearly associated with active transcription (Figure 4F). Correlations between transcriptional activity and histone occupancy or modification at intergenic and transcribed regions are summarized in Table S2.

The genome-wide maps of histone occupancy and modification described here should provide investigators with information useful for further exploring the histone code and its implications for gene regulation and chromosome organization and maintenance. We expect that the approaches used here to map histone occupancy and modification in yeast can also be used to gain insights into the linkage between gene expression and histone modification across the genome in higher eukaryotes.

Figure 4

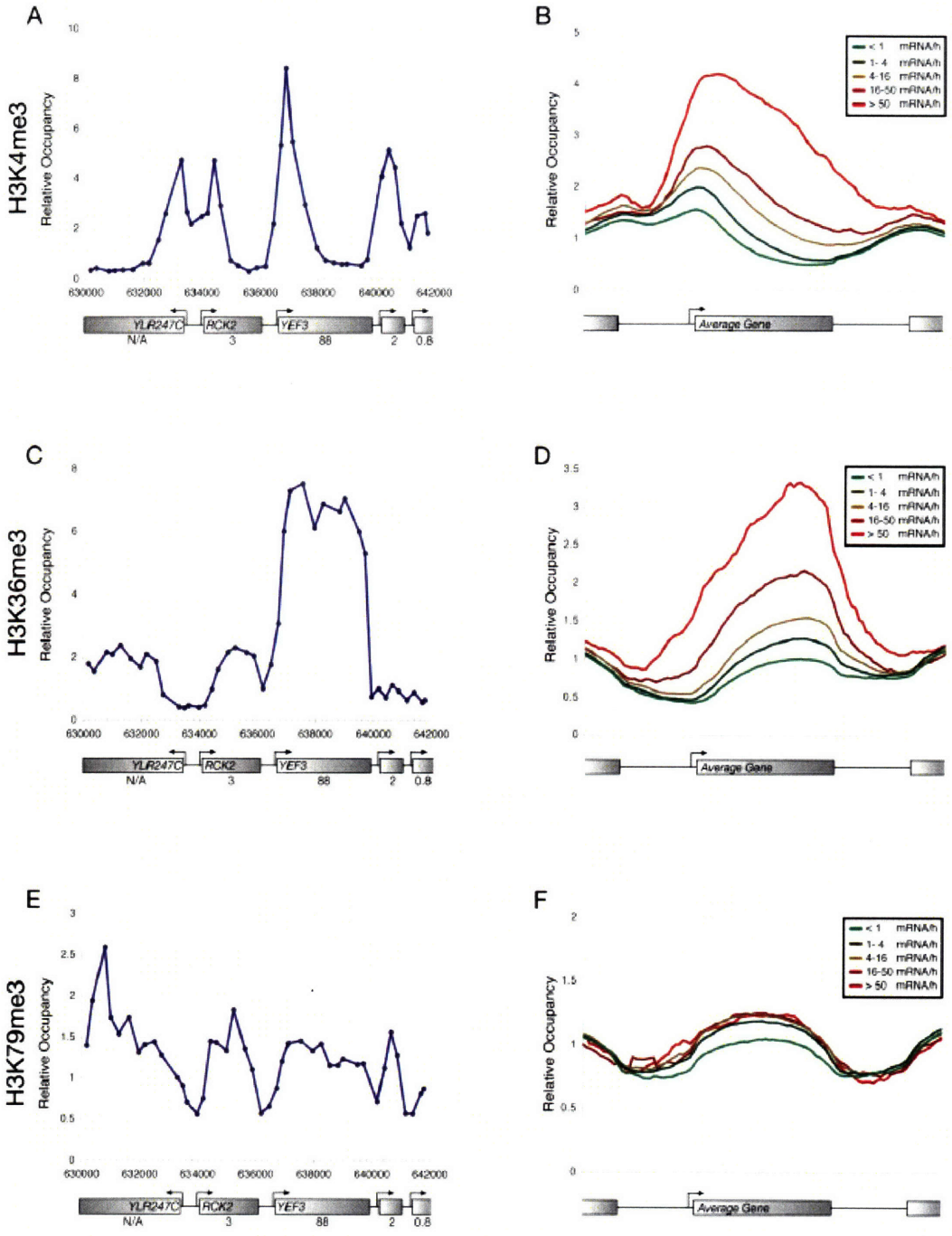

A
H3K4me3
Relative Occupancy
B
Relative Occupancy
< 1 mRNA/h
1- 4 mRNA/h
4-16 mRNA/h
16-50 mRNA/h
> 50 mRNA/h
Average Gene
C
H3K36me3
Relative Occupancy
D
Relative Occupancy
Average Gene
E
H3K79me3
Relative Occupancy
F
Relative Occupancy
Average Gene
630000 632000 634000 636000 638000 640000 642000
YLR247C RCK2 YEF3
N/A 3 88 2 0.8

Figure 4. Nucleosome Methylation Generally Correlates with Transcriptional Activity
(A) Trimethylation of H3K4 at a locus on chromosome XII. Enrichment is depicted as in Figure 1. The number beneath each gene represents the transcriptional frequency of the corresponding ORF (Holstege et al., 1998) in mRNA/hr.
(B) Composite profile of trimethylation of H3K4 across the average gene. Composite profiles of methylation according to transcriptional frequency class are shown as in Figure 2.
(C) Trimethylation of H3K36 at a locus on chromosome XII. Enrichment is depicted as in Figure 1. The number beneath each gene represents the transcriptional frequency of the corresponding ORF (Holstege et al., 1998) in mRNA/hr.
(D) Composite profile of trimethylation of H3K36 across the average gene. Composite profiles of methylation according to transcriptional frequency class are shown as in Figure 2.
(E) Trimethylation of H3K79 at a locus on chromosome XII. Enrichment is depicted as in Figure 1. The number beneath each gene represents the transcriptional frequency of the corresponding ORF (Holstege et al., 1998) in mRNA/hr.
(F) Composite profile of trimethylation of H3K79 across the average gene. Composite profiles of methylation according to transcriptional frequency class are shown as in Figure 2.

Figure 5

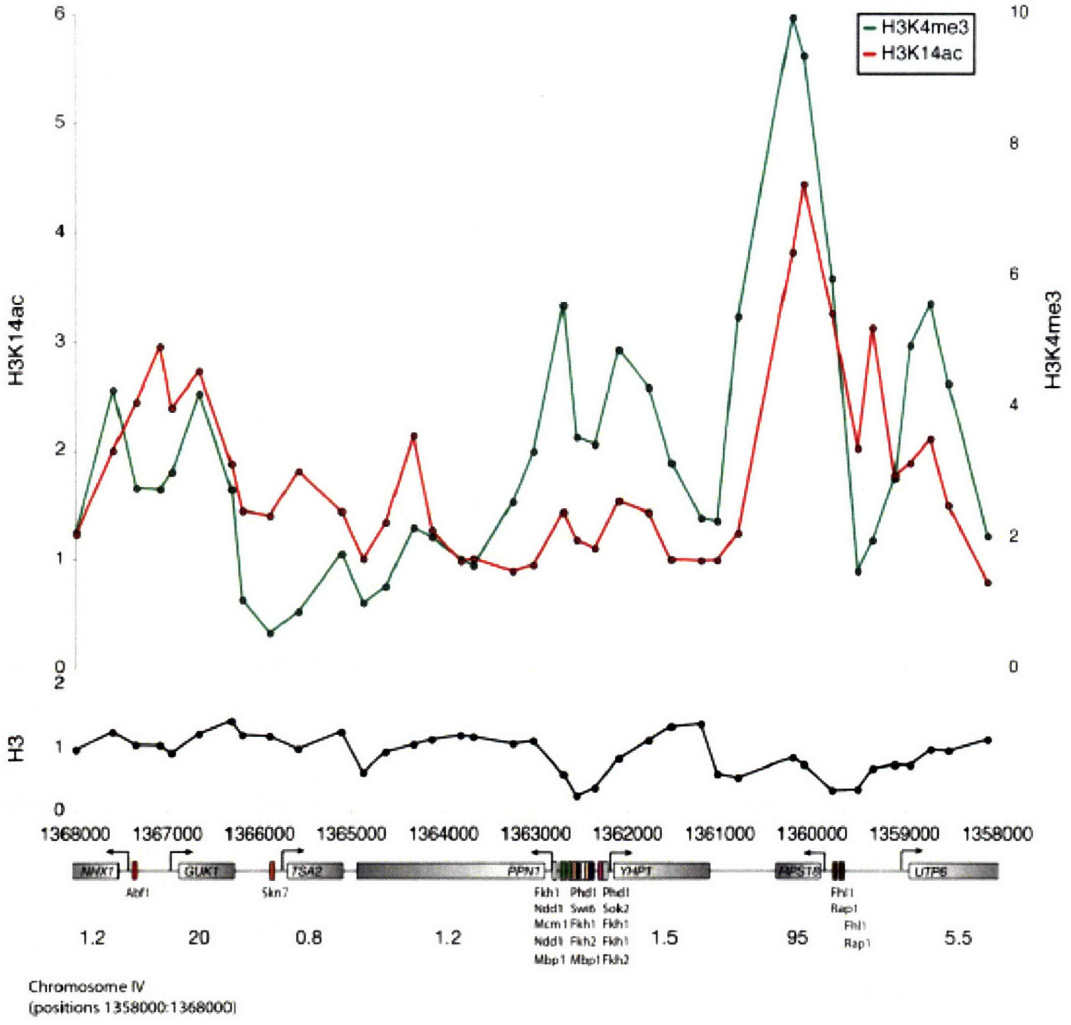

H3K4me3
H3K14ac
H3K14ac
H3K4me3
H3
1368000
1367000
1366000
1365000
1364000
1363000
1362000
1361000
1360000
1359000
1358000
NWX1
GUK1
TSA2
PPN1
YHP1
RPS18
UTP6
Abf1
Skn7
Fkh1 Phd1 Phd1
Ndd1 Swi6 Sok2
Mcm1 Fkh1 Fkh1
Ndd1 Fkh2 Fkh1
Mbp1 Mbp1 Fkh2
Fhl1
Rap1
Fhl1
Rap1
1.2
20
0.8
1.2
1.5
95
5.5
Chromosome IV
(positions 1358000:1368000)

Figure 5. A High-Resolution Genome-wide Map of Nucleosome States
A map for a region on chromosome IV is depicted as in Figure 1. Conserved binding sites for transcriptional regulators (Harbison et al., 2004) are depicted as colored boxes. Numbers beneath genes represent transcriptional activity (mRNA/hr). Enrichment values from acetylated H3K14, trimethylated H3K4, and histone H3 are depicted in red, green, and black, respectively.

Experimental Procedures

A detailed protocol of the ChIP-chip procedure, the microarray data described herein, and a description of the error model used for data analysis are available at the authors' website (http://web.wi.mit.edu/young/nucleosome).

Array Design

The Agilent DNA microarray used here has 44,290 features consisting of 60-mer oligonulceotide probes. The array covers 12 Mb of the yeast genome (85%), excluding highly repetitive regions, with an average probe density of 266 bp. Intergenic regions are represented by 14,256 probes, and ORFs are represented by 27,185 probes. The remaining 2,849 features included blank spots and controls.

Epitope Tagging, Antibodies, and Strains

Transcriptional and chromatin regulators were tagged at the C terminus with a 9-copy myc epitope. The sequence encoding the myc epitope was introduced into the endogenous gene immediately upstream of the stop codon. Specific oligonucleotides were used to generate PCR products from plasmids described by Cosma et al. (1999). The resulting PCR products were transformed into a W303 yeast to generate the tagged strains by one-step genomic integration. Clones were selected for growth on the appropriate selective media plates, and the insertion was confirmed by PCR. The expression of the epitope-tagged protein was confirmed by Western blotting using an anti-Myc (9E11). The antibodies used in this study are listed in Table 1.

Table 1. Antibodies Used in This Study

Specificity	Supplier	Catalog #
Anti-Histone H3	Abcam	ab1791
Anti-Histone H4	Abcam	ab10156
Anti-H3K9ac	Upstate Biotechnology	06-942
Anti-H3K14ac	Upstate Biotechnology	06-911
Anti-H4ac	Upstate Biotechnology	06-866
Anti-H3K4me3	Abcam	ab8580
Anti-H3K4me2	Abcam	ab7766
Anti-H3K4me1	Abcam	ab8895
Anti-H3K36me3	Abcam	ab9050
Anti-H3K79me3	Abcam	ab2621
Anti-myc	Taconic Biotechnology	9E11
Rabbit IgG	Upstate Biotechnology	12-370

Chromatin Immunoprecipitation and Genome-wide ChIP-Chip

Chromatin immunoprecipitation and genome-wide location analysis were performed as described previously (Ren et al., 2000), except that the crosslinking time was reduced to 30 min at room temperature, the order of Proteinase K and RNase treatment was reversed, and high-resolution oligonucleotide arrays (Agilent Technologies) were used for hybridizations. Briefly, yeast cells were grown in at least two independent cultures in

rich medium. Response to hydrogen peroxide was induced by adding hydrogen peroxide to the cell cultures grown at mid-log phase in YPD medium at 30°C to final concentration of 0.4 mM for 20 min. Cultures were treated with formaldehyde (1%) for 30 min, and cells were collected by centrifugation, washed with ice-cold TBS, and disrupted by vortexing in lysis buffer in the presence of glass beads. The chromatin was sonicated to yield an average DNA fragment of 500 bp. The DNA fragments crosslinked to the proteins were enriched by immunoprecipitation with specific antibodies. After reversal of the crosslinks and purification, the immunoprecipitated and input DNA was labeled by ligation-mediated PCR with Cy5 and Cy3 fluorescent dyes, respectively. Both pools of labeled DNA were hybridized to a single DNA microarray (described above). Images of Cy5 and Cy3 fluorescence intensities were generated by scanning array using GenePix 5000 scanner and were analyzed with GenePix Pro 5.1 software. Experiments were carried out at least in duplicate.

We and other groups have noted that there can be modest differences in the relative levels of intergenic and genic yeast DNA that are recovered during phenol extraction (Nagy et al., 2003). Experimental analysis indicates that this is not due to differences in our ability to detect intergenic and genic DNA on the DNA microarrays. Our experiments also indicate that this observation is not due to artifacts due to differential labeling of DNA. Others have speculated that differential recovery is due to contaminating nucleases that might preferentially digest intergenic DNA (Nagy et al., 2003). It is also possible that there are intrinsic differences in susceptibility to shearing by sonication in intergenic and genic DNA.

Mock-Immunoprecipitation Normalization

Control immunoprecipitations were performed as above with two exceptions. In one case, an antibody with no specificity to histones (rabbit IgG) was substituted for the H3- or H4-specific antibody. In the second case, no antibody was added during the overnight incubation with magnetic beads. For histone H3 and H4 (data not shown), data were normalized relative to the "no antibody" control. Distributions of relative occupancy at ORF and intergenic regions by histone H3 are depicted in Figure S8. Following normalization by this method, the standard deviation is 0.38, and we used two-sampled t tests to determine the likelihood of differences occurring by chance in both the original and controlled experiments (Figure S8).

Data Analysis

Genome-wide location data were subjected to quality-control filters and median normalized, and the weighted average ratio of immunoprecipitated to control DNA was determined for each spot across all replicates. A confidence value (p value) for single probes and an averaged confidence value for neighboring probes were calculated.

A binding cutoff for Gcn4 was determined by comparing maximum IP/WCE ratios to a high-likelihood positive list and a highlikelihood negative list using ROC curve analysis. A positive list of 84 genes (Table S1) was selected on the basis of previous highconfidence binding data (p % 0.001) (Harbison et al., 2004), the presence of a perfect or near perfect Gcn4 consensus binding site (TGASTCA) in the region of –400 bp to +50 bp, and a greater than 2-fold change in steady-state mRNA levels dependent on Gcn4 when shifted to amino acid starvation medium (Natarajan et al., 2001). The

negative list of 945 genes not transcribed from divergent intergenic regions was selected by weak binding (p R 0.1), absence of a motif near the presumed start site, and a less than 60% change in steady-state mRNA levels in response to shift to amino acid starvation. Each gene was scored based on the minimum p value found in the region −250 bp to +50 bp from the UAS using the higher of the single and averaged confidence score. Optimal parameters were determined by maximizing the absolute difference in identified genes in both the positive list and the negative list using the Statistics-ROC package for Perl.

Nucleosome-depleted promoters were defined as intergenic regions upstream of protein-coding genes for which unmodified histone H3 or H4 enrichment met the following criterion: the enrichment of any probe within the intergenic region was less than the average ratio of enrichment at two neighboring ORFs.

Throughout the text, histone H3 occupancy is often referred to as nucleosome occupancy. There are two reasons to believe that the results with H3 likely reflect nucleosome occupancy and not nucleosomes that are missing H3 specifically. First, we obtained similar results in independent experiments with H3 and H4. Second, previous in vitro studies suggest that it is H2A-H2B dimmers (and not H3 or H4) that preferentially dissociate from nucleosomes during transcription (Kireeva et al., 2002).

Supplemental Data

Supplemental Data include two tables and eight figures and can be found with this article online at http://www.cell.com/cgi/content/full/122/4/517/DC1/.

Acknowledgments

We thank E. Fraenkel for helpful discussions and D. Reynolds for technical assistance. This work was supported by NHGRI grant HG002668 and NIH grant GM069676. T.I.L., D.K.G., and R.A.Y. consult for Agilent Technologies.

References

Allard, S., Utley, R.T., Savard, J., Clarke, A., Grant, P., Brandl, C.J., Pillus, L., Workman, J.L., and Cote, J. (1999). NuA4, an essential transcription adaptor/histone H4 acetyltransferase complex containing Esa1p and the ATM-related cofactor Tra1p. EMBO J. *18*, 5108–5119.

Arndt, K., and Fink, G.R. (1986). GCN4 protein, a positive transcription factor in yeast, binds general control promoters at all 5# TGACTC 3# sequences. Proc. Natl. Acad. Sci. USA *83*, 8516–8520.

Bannister, A.J., Schneider, R., Myers, F.A., Thorne, A.W., Crane-Robinson, C., and Kouzarides, T. (2005). Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. J. Biol. Chem. *280*, 17732–17736.

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. USA *99*, 8695–8700.

Bernstein, B.E., Liu, C.L., Humphrey, E.L., Perlstein, E.O., and Schreiber, S.L. (2004). Global nucleosome occupancy in yeast. Genome Biol. *5*, R62. Published online August 20, 2004.. 10.1186/gb-2004-5-9-r62

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas, E.J., 3rd, Gingeras, T.R., et al. (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. Cell *120*, 169–181.

Bhaumik, S.R., and Green, M.R. (2001). SAGA is an essential in vivo target of the yeast acidic activator Gal4p. Genes Dev. *15*, 1935–1945.

Boeger, H., Griesenbeck, J., Strattan, J.S., and Kornberg, R.D. (2003). Nucleosomes unfold completely at a transcriptionally active promoter. Mol. Cell *11*, 1587–1598.

Briggs, S.D., Bryk, M., Strahl, B.D., Cheung, W.L., Davie, J.K., Dent, S.Y., Winston, F., and Allis, C.D. (2001). Histone H3 lysine 4 methylation is mediated by Set1 and required for cell growth and rDNA silencing in Saccharomyces cerevisiae. Genes Dev. *15*, 3286–3295.

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

Clarke, A.S., Lowell, J.E., Jacobson, S.J., and Pillus, L. (1999). Esa1p is an essential histone acetyltransferase required for cell cycle progression. Mol. Cell. Biol. *19*, 2515–2526.

Cosma, M.P., Tanaka, T., and Nasmyth, K. (1999). Ordered recruitment of transcription and chromatin remodeling factors to a cell cycle- and developmentally regulated promoter. Cell *97*, 299–311.

Deckert, J., and Struhl, K. (2001). Histone acetylation at promoters is differentially affected by specific activators and repressors. Mol. Cell. Biol. *21*, 2726–2735.

Ellenberger, T.E., Brandl, C.J., Struhl, K., and Harrison, S.C. (1992). The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted alpha helices: crystal structure of the protein-DNA complex. Cell *71*, 1223–1237.

Feng, Q., Wang, H., Ng, H.H., Erdjument-Bromage, H., Tempst, P., Struhl, K., and Zhang, Y. (2002). Methylation of H3-lysine 79 is mediated by a new family of HMTases without a SET domain. Curr. Biol. *12*, 1052–1058.

Gregory, P.D., Schmid, A., Zavari, M., Munsterkotter, M., and Horz, W. (1999). Chromatin remodelling at the PHO8 promoter requires SWI-SNF and SAGA at a step subsequent to activator binding. EMBO J. *18*, 6407–6414.

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.

Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S., and Young, R.A. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. Cell *95*, 717–728.

Hope, I.A., and Struhl, K. (1985). GCN4 protein, synthesized in vitro, binds HIS3 regulatory sequences: implications for general control of amino acid biosynthetic genes in yeast. Cell *43*, 177–188.

Humphrey, E.L., Shamji, A.F., Bernstein, B.E., and Schreiber, S.L. (2004). Rpd3p relocation mediates a transcriptional response to rapamycin in yeast. Chem. Biol. *11*, 295–299.

Kireeva, M.L., Walter, W., Tchernajenko, V., Bondarenko, V., Kashlev, M., and Studitsky, V.M. (2002). Nucleosome remodeling induced by RNA polymerase II: loss of the H2A/H2B dimer during transcription. Mol. Cell *9*, 541–552.

Kouzarides, T. (2002). Histone methylation in transcriptional control. Curr. Opin. Genet. Dev. *12*, 198–209.

Krogan, N.J., Dover, J., Wood, A., Schneider, J., Heidt, J., Boateng, M.A., Dean, K., Ryan, O.W., Golshani, A., Johnston, M., et al. (2003a). The Paf1 complex is required for histone H3 methylation by COMPASS and Dot1p: linking transcriptional elongation to

histone methylation. Mol. Cell *11*, 721–729.

Krogan, N.J., Kim, M., Tong, A., Golshani, A., Cagney, G., Canadien, V., Richards, D.P., Beattie, B.K., Emili, A., Boone, C., et al. (2003b). Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II. Mol. Cell. Biol. *23*, 4207–4218.

Kuo, M.H., Brownell, J.E., Sobel, R.E., Ranalli, T.A., Cook, R.G., Edmondson, D.G., Roth, S.Y., and Allis, C.D. (1996). Transcriptionlinked acetylation by Gcn5p of histones H3 and H4 at specific lysines. Nature *383*, 269–272.

Kuo, M.H., Zhou, J., Jambeck, P., Churchill, M.E., and Allis, C.D. (1998). Histone acetyltransferase activity of yeast Gcn5p is required for the activation of target genes in vivo. Genes Dev. *12*, 627–639.

Kurdistani, S.K., Robyr, D., Tavazoie, S., and Grunstein, M. (2002). Genome-wide binding map of the histone deacetylase Rpd3 in yeast. Nat. Genet. *31*, 248–254.

Kurdistani, S.K., Tavazoie, S., and Grunstein, M. (2004). Mapping global histone acetylation patterns to gene expression. Cell *117*, 721–733.

Larschan, E., and Winston, F. (2001). The S. cerevisiae SAGA complex functions in vivo as a coactivator for transcriptional activation by Gal4. Genes Dev. *15*, 1946–1956.

Lee, C.K., Shibata, Y., Rao, B., Strahl, B.D., and Lieb, J.D. (2004). Evidence for nucleosome depletion at active regulatory regions genome- wide. Nat. Genet. *36*, 900–905.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., et al. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science *298*, 799–804.

Lieb, J.D., Liu, X., Botstein, D., and Brown, P.O. (2001). Promoterspecific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat. Genet. *28*, 327–334.

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

Nagy, P.L., Cleary, M.L., Brown, P.O., and Lieb, J.D. (2003). Genomewide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin. Proc. Natl. Acad. Sci. USA *100*, 6364–6369.

Narlikar, G.J., Fan, H.Y., and Kingston, R.E. (2002). Cooperation between complexes

that regulate chromatin structure and transcription. Cell *108*, 475–487.

Natarajan, K., Meyer, M.R., Jackson, B.M., Slade, D., Roberts, C., Hinnebusch, A.G., and Marton, M.J. (2001). Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast. Mol. Cell. Biol. *21*, 4347–4368.

Ng, H.H., Feng, Q., Wang, H., Erdjument-Bromage, H., Tempst, P., Zhang, Y., and Struhl, K. (2002a). Lysine methylation within the globular domain of histone H3 by Dot1 is important for telomeric silencing and Sir protein association. Genes Dev. *16*, 1518–1527.

Ng, H.H., Robert, F., Young, R.A., and Struhl, K. (2002b). Genomewide location and regulated recruitment of the RSC nucleosomeremodeling complex. Genes Dev. *16*, 806–819.

Ng, H.H., Ciccone, D.N., Morshead, K.B., Oettinger, M.A., and Struhl, K. (2003a). Lysine-79 of histone H3 is hypomethylated at silenced loci in yeast and mammalian cells: a potential mechanism for position-effect variegation. Proc. Natl. Acad. Sci. USA *100*, 1820–1825.

Ng, H.H., Robert, F., Young, R.A., and Struhl, K. (2003b). 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.

O'Shea, E.K., Klemm, J.D., Kim, P.S., and Alber, T. (1991). X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. Science *254*, 539–544.

Oliphant, A.R., Brandl, C.J., and Struhl, K. (1989). Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. Mol. Cell. Biol. *9*, 2944–2949.

Peterson, C.L., and Laniel, M.A. (2004). Histones and histone modifications. Curr. Biol. *14*, R546–R551. Reid, J.L., Iyer, V.R., Brown, P.O., and Struhl, K. (2000). Coordinate regulation of yeast ribosomal protein genes is associated with targeted recruitment of Esa1 histone acetylase. Mol. Cell *6*, 1297–1307.

Reinke, H., and Horz, W. (2003). Histones are first hyperacetylated and then lose contact with the activated PHO5 promoter. Mol. Cell *11*, 1599–1607.

Reinke, H., Gregory, P.D., and Horz, W. (2001). A transient histone hyperacetylation signal marks nucleosomes for remodeling at the PHO8 promoter in vivo. Mol. Cell *7*, 529–538.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., et al. (2000). Genome-wide location and function of DNA binding proteins. Science *290*, 2306–2309.

Robert, F., Pokholok, D.K., Hannett, N.M., Rinaldi, N.J., Chandy, M., Rolfe, A., Workman, J.L., Gifford, D.K., and Young, R.A. (2004). Global position and recruitment of HATs and HDACs in the yeast genome. Mol. Cell *16*, 199–209.

Robyr, D., Suka, Y., Xenarios, I., Kurdistani, S.K., Wang, A., Suka, N., and Grunstein, M. (2002). Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases. Cell *109*, 437–446.

Roh, T.Y., Ngau, W.C., Cui, K., Landsman, D., and Zhao, K. (2004). High-resolution genome-wide mapping of histone modifications. Nat. Biotechnol. *22*, 1013–1016.

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.

Schubeler, D., MacAlpine, D.M., Scalzo, D., Wirbelauer, C., Kooperberg, C., van Leeuwen, F., Gottschling, D.E., O'Neill, L.P., Turner, B.M., Delrow, J., et al. (2004). The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. Genes Dev. *18*, 1263–1271.

Strahl, B.D., Grant, P.A., Briggs, S.D., Sun, Z.W., Bone, J.R., Caldwell, J.A., Mollah, S., Cook, R.G., Shabanowitz, J., Hunt, D.F., and Allis, C.D. (2002). Set2 is a nucleosomal histone H3-selective methyltransferase that mediates transcriptional repression. Mol. Cell. Biol. *22*, 1298–1306.

Suka, N., Luo, K., and Grunstein, M. (2002). Sir2p and Sas2p opposingly regulate acetylation of yeast histone H4 lysine16 and spreading of heterochromatin. Nat. Genet. *32*, 378–383.

Utley, R.T., Ikeda, K., Grant, P.A., Cote, J., Steger, D.J., Eberharter, A., John, S., and Workman, J.L. (1998). Transcriptional activators direct histone acetyltransferase complexes to nucleosomes. Nature *394*, 498–502.

van Leeuwen, F., Gafken, P.R., and Gottschling, D.E. (2002). Dot1p modulates silencing in yeast by methylation of the nucleosome core. Cell *109*, 745–756.

Vogelauer, M., Wu, J., Suka, N., and Grunstein, M. (2000). Global histone acetylation and deacetylation in yeast. Nature *408*, 495–498.

Xiao, T., Hall, H., Kizer, K.O., Shibata, Y., Hall, M.C., Borchers, C.H., and Strahl, B.D. (2003). Phosphorylation of RNA polymerase II CTD regulates H3 methylation in yeast. Genes Dev. *17*, 654–663.

Zhang, W., Bone, J.R., Edmondson, D.G., Turner, B.M., and Roth, S.Y. (1998). Essential and redundant functions of histone acetylation revealed by mutation of target lysines and

loss of the Gcn5p acetyltransferase. EMBO J. *17*, 3155–3167.

Accession Numbers

The microarray data described herein are available at ArrayExpress (http://www.ebi.ac.uk/arrayexpress) under the accession number E-WMIT-3.

Appendix B

Supplementary Material for Chapter 5

Table of Contents

Index of Supplementary Files

The following files contain data formatted for upload into the UCSC genome browser (Kent et al., 2002). To upload the files, first copy the files onto a computer with internet access. Then use a web browser to go to http://genome.ucsc.edu/cgi-bin/hgCustom?hgsid=105256378 for mouse and http://genome.ucsc.edu/cgi-bin/hgCustom?hgsid=104842340 for human. In the "Paste URLs or Data" section, select "Browse..." on the right of the screen. Use the pop-up window to select the copied files, then select "Submit". The upload process may take some time.

mouse_miRNA_track.mm8.bed – Map of predicted miRNA genes in mouse. Transcripts with EST or gene evidence are shown as black lines. Presumed transcripts are shown as grey lines. Positions of the mature miRNAs are annotated as thicker lines.

human_miRNA_track.hg17.bed – Map of predicted miRNA genes in human. Transcripts with EST or gene evidence are shown as black lines. Presumed transcripts are shown as grey lines. Positions of the mature miRNAs are annotated as thicker lines.

mES_regulator_ChIPseq.mm8.WIG.gz – ChIP-seq data for Oct4, Sox2, Nanog and Tcf3 in mES cells. Top track for each data set illustrates the normalized number of reads assigned to each 25bp bin. Bars in the second track identify regions of the genome enriched at $p < 10^{-9}$.

mES_chromatin_ChIPseq.mm8.WIG.gz – ChIP-seq data for H3K4me3, H3K79me2, H3K36me3 and Suz12 in mES cells. Top track for each data set illustrates the normalized number of reads assigned to each 25bp bin. Bars in the second track identify regions of the genome enriched at $p < 10^{-9}$. Enriched regions were not identified for H3K79me2 and H3K36me3

Supplemental References

Supplementary Tables and Files are available from Young Lab.

Supplementary Methods and Discussion

Growth Conditions and Quality Control for Human Embryonic Stem Cells

Human embryonic stem (ES) 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 Embryonic Stem 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 are listed 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 are listed 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 are listed 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 are listed 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 are listed 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 Preperation 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 $\sim 10^8$ 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 not all egions 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 [red line]).

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 (Poissoin: $p < 10^{-9}$). 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.

Curve fitting was performed using a least squares model by 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)were 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 Preperation 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 labelled and purified using Bioprime random primer labeling kits (Invitrogen): immunoenriched DNA was labeled with Cy5 fluorophore, whole cell extract DNA was labelled with Cy3 fluorophore.

Labelled 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[20] 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[16] 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 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 labelled 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, mouse embryonic fibroblasts, and induced pluripotent stem cells.

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, □-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 (Supplementary 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 are 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.

Supplemental Figures

Figure S1

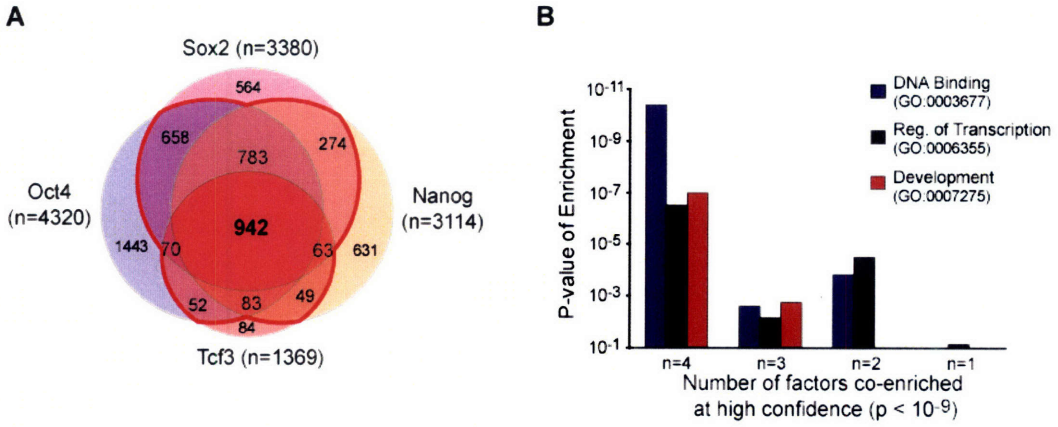

A
Sox2 (n=3380)
564
658
274
783
Oct4
(n=4320)
Nanog
(n=3114)
942
1443
70
63
631
52
83
49
84
Tcf3 (n=1369)
B
DNA Binding
(GO:0003677)
Reg. of Transcription
(GO:0006355)
Development
(GO:0007275)
P-value of Enrichment
10-11
10-9
10-7
10-5
10-3
10-1
n=4
n=3
n=2
n=1
Number of factors co-enriched
at high confidence (p < 10-9)

Figure S1. 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 S2

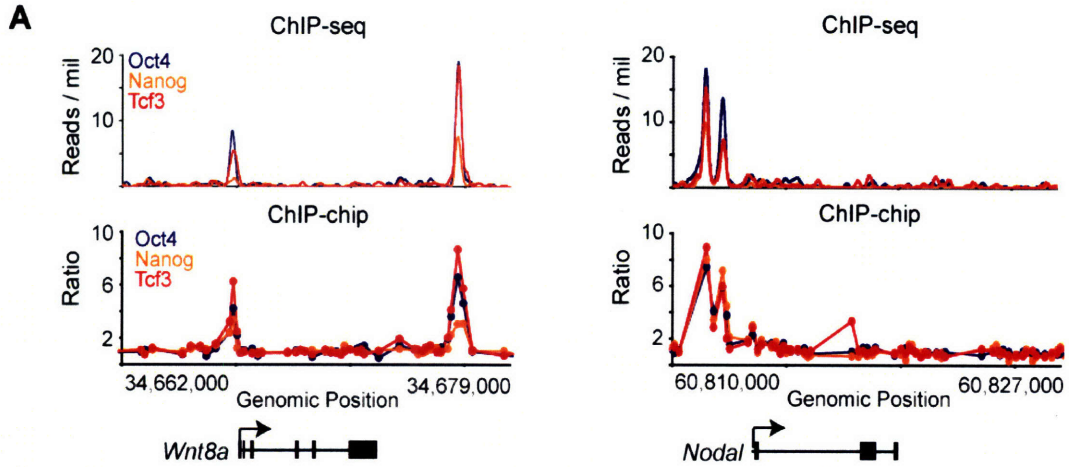

A
ChIP-seq
Reads / mil
20
10
Oct4
Nanog
Tcf3
ChIP-chip
Ratio
10
6
2
34,662,000
Genomic Position
34,679,000
Wnt8a
ChIP-seq
Reads / mil
20
10
ChIP-chip
Ratio
10
6
2
60,810,000
Genomic Position
60,827,000
Nodal

Figure S2. 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.

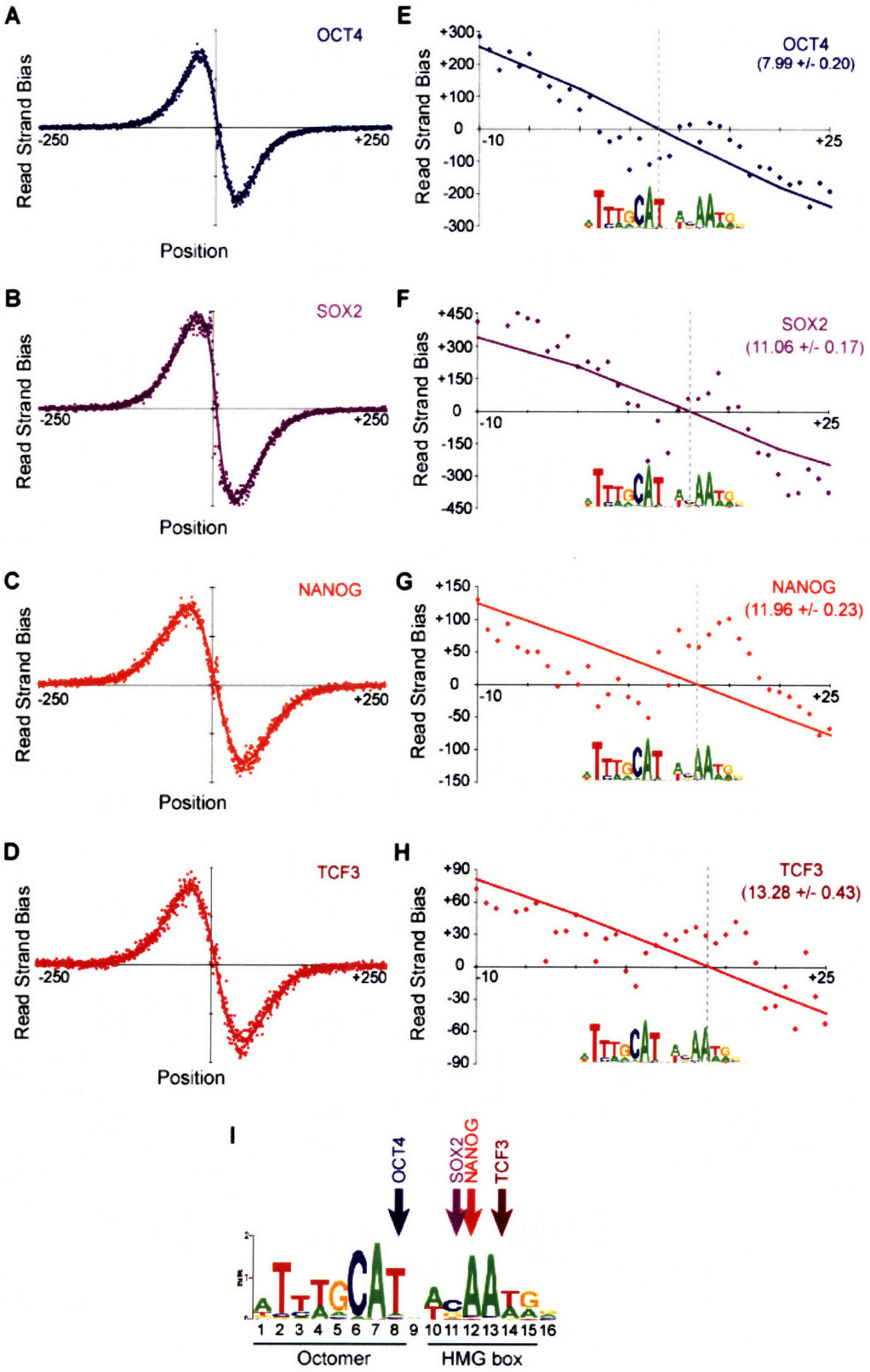

A
OCT4
Read Strand Bias
-250
+250
Position
B
SOX2
Read Strand Bias
-250
+250
Position
C
NANOG
Read Strand Bias
-250
+250
Position
D
TCF3
Read Strand Bias
-250
+250
Position
E
OCT4
(7.99 +/- 0.20)
Read Strand Bias
+300
+200
+100
0
-100
-200
-300
-10
+25
F
SOX2
(11.06 +/- 0.17)
Read Strand Bias
+450
+300
+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
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Octomer
HMG box

Figure S3. 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 S4

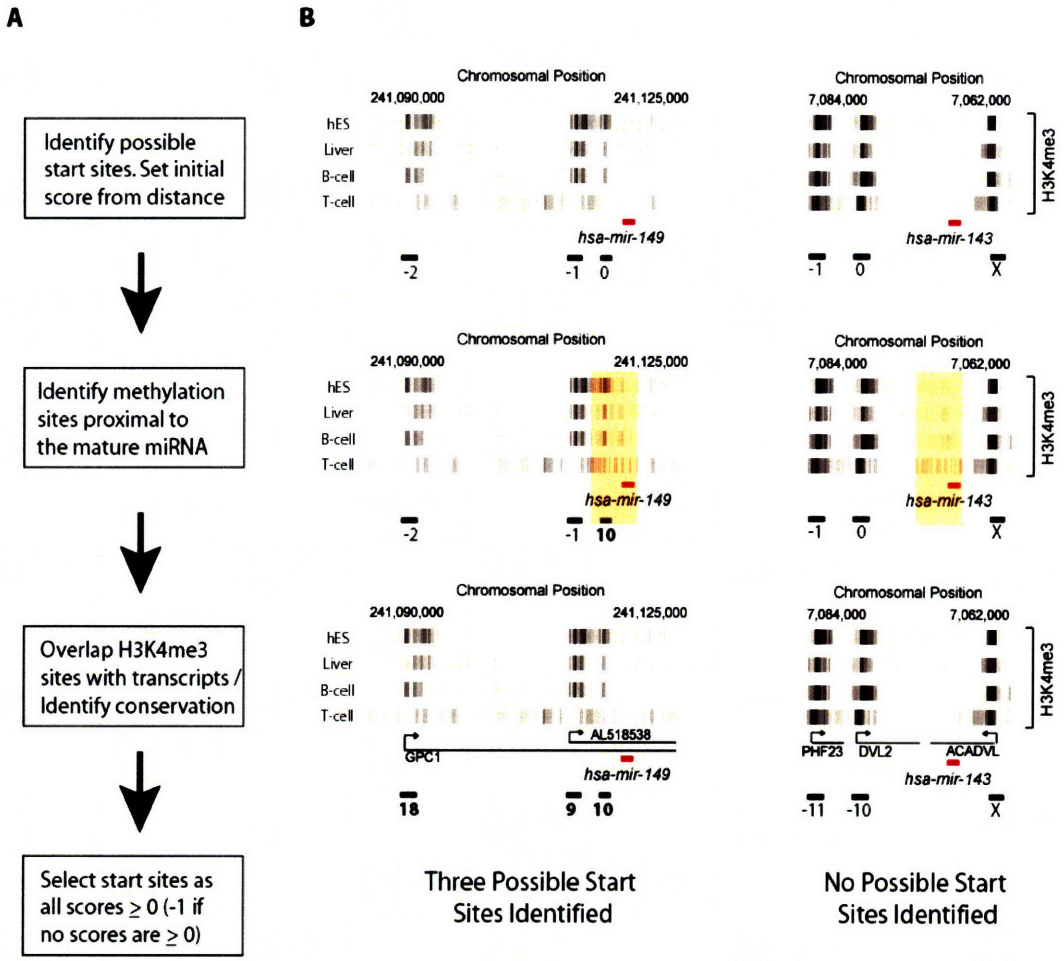

A
Identify possible start sites. Set initial score from distance
Identify methylation sites proximal to the mature miRNA
Overlap H3K4me3 sites with transcripts / Identify conservation
Select start sites as all scores ≥ 0 (-1 if no scores are ≥ 0)
B
Chromosomal Position
241,090,000
241,125,000
hES
Liver
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
-1
X
PHF23
DVL2
ACADVL
-11
-10
No Possible Start Sites Identified

Figure S4. 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[1,2]. 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 S5

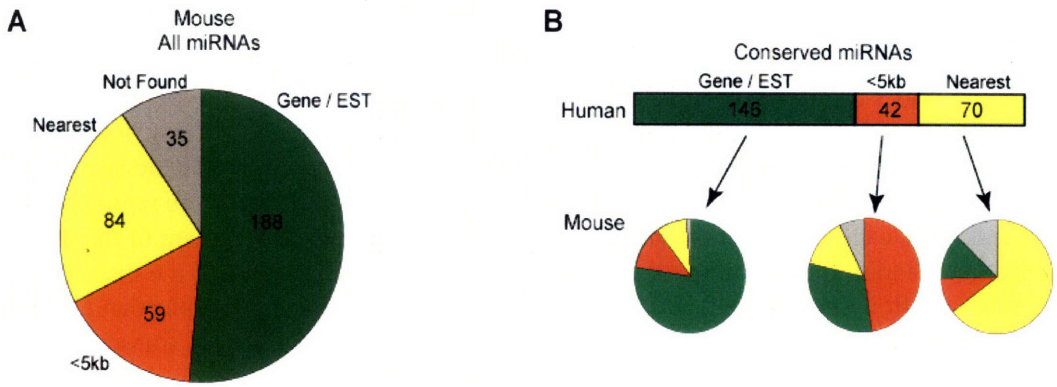

A
Mouse
All miRNAs
Not Found
Gene / EST
Nearest
35
84
188
59
<5kb
B
Conserved miRNAs
Gene / EST
<5kb
Nearest
Human
146
42
70
Mouse

Figure S5. 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 S6

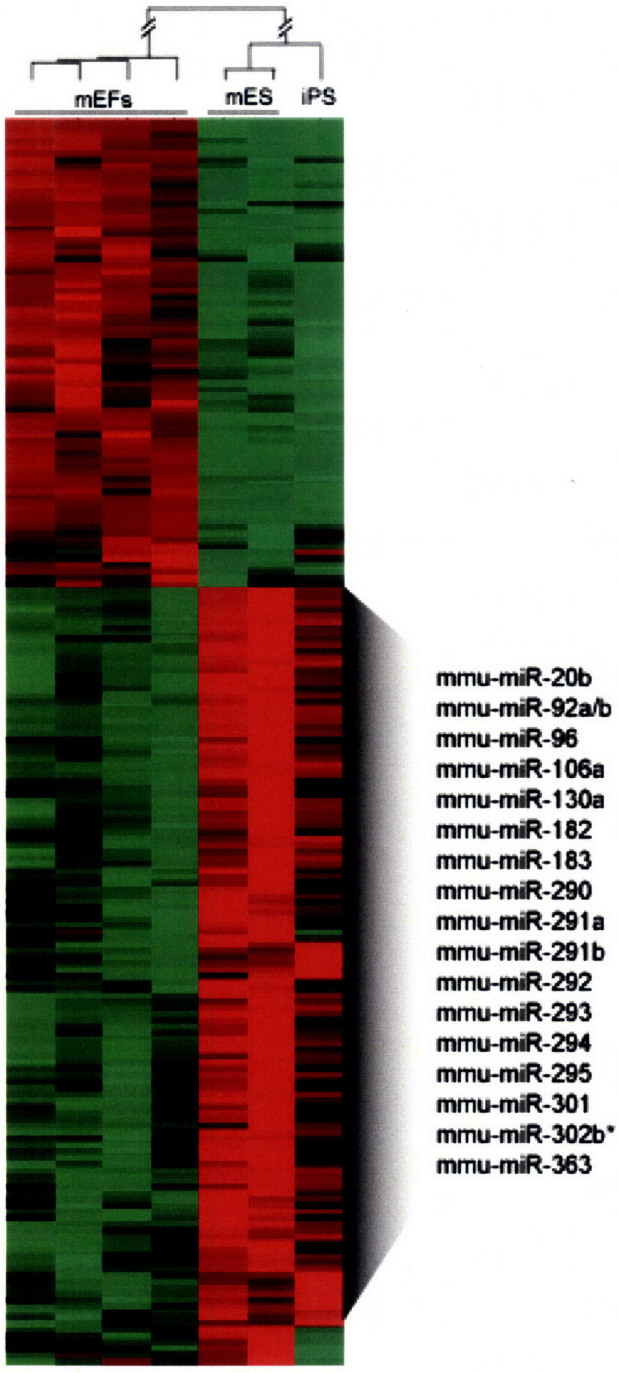

mEFs
mES
iPS
mmu-miR-20b
mmu-miR-92a/b
mmu-miR-96
mmu-miR-106a
mmu-miR-130a
mmu-miR-182
mmu-miR-183
mmu-miR-290
mmu-miR-291a
mmu-miR-291b
mmu-miR-292
mmu-miR-293
mmu-miR-294
mmu-miR-295
mmu-miR-301
mmu-miR-302b*
mmu-miR-363

Figure S6. miRNA genes occupied by the core transcriptional regulators in ES cells are expressed in induced Pluripotent Stem (iPS) cells.
miRNAs were purified from MEFs (lanes 1-3), mES cells (lane 4,5) and iPS cells (lane6) and hybridized to LNA miRNA arrays. Differentially expressed probes enriched in either mEFs or mES cells are shown (FDR < 10%, see supplemental text, iPS cells were not used to determinedifferential expression). miRNA probes were Z-score normalized, and cell types were clustered hierarchically (top). Probes associated with active miRNAs occupied at their promoters by Oct4/Sox2/Nanog/Tcf3 are listed to the right.

Figure S7

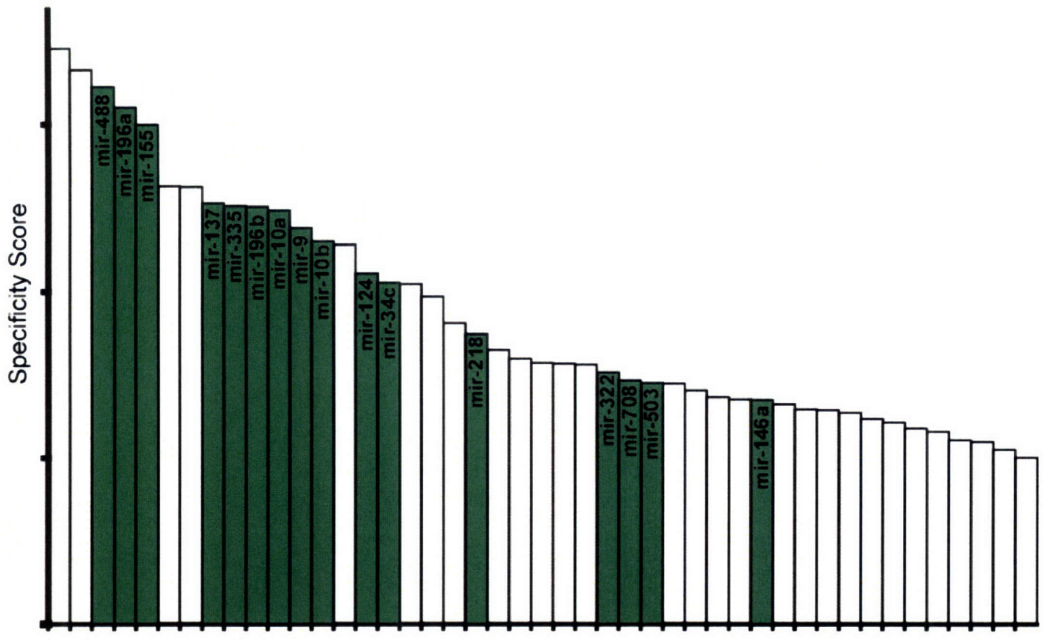

Specificity Score
mir-488
mir-196a
mir-155
mir-137
mir-335
mir-196b
mir-10a
mir-9
mir-10b
mir-124
mir-34c
mir-218
mir-322
mir-708
mir-503
mir-146a

Figure S7. 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).

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.