

UCLA

UCLA Previously Published Works

Title

Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells

Permalink

<https://escholarship.org/uc/item/7q11r5qw>

Journal

BMC Genomics, 10(1)

ISSN

1471-2164

Authors

Mason, Mike J
Fan, Guoping
Plath, Kathrin
et al.

Publication Date

2009-07-20

DOI

<http://dx.doi.org/10.1186/1471-2164-10-327>

Peer reviewed

Research article

Open Access

Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells

Mike J Mason¹, Guoping Fan², Kathrin Plath³, Qing Zhou^{*1} and Steve Horvath^{*2,4}

Address: ¹Statistics, University of California, Los Angeles, CA, 90095, USA, ²Human Genetics, David Geffen School of Medicine, Los Angeles, CA, 90095, USA, ³Biological Chemistry, University of California, Los Angeles, CA, 90095, USA and ⁴Biostatistics, School of Public Health, University of California, Los Angeles, CA, 90095, USA

Email: Mike J Mason - mjrmason@stat.ucla.edu; Guoping Fan - gfan@mednet.ucla.edu; Kathrin Plath - kplath@mednet.ucla.edu; Qing Zhou* - zhou@stat.ucla.edu; Steve Horvath* - shorvath@mednet.ucla.edu

* Corresponding authors

Published: 20 July 2009

Received: 17 February 2009

BMC Genomics 2009, **10**:327 doi:10.1186/1471-2164-10-327

Accepted: 20 July 2009

This article is available from: <http://www.biomedcentral.com/1471-2164/10/327>

© 2009 Mason et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Recent work has revealed that a core group of transcription factors (TFs) regulates the key characteristics of embryonic stem (ES) cells: pluripotency and self-renewal. Current efforts focus on identifying genes that play important roles in maintaining pluripotency and self-renewal in ES cells and aim to understand the interactions among these genes. To that end, we investigated the use of unsigned and signed network analysis to identify pluripotency and differentiation related genes.

Results: We show that signed networks provide a better systems level understanding of the regulatory mechanisms of ES cells than unsigned networks, using two independent murine ES cell expression data sets. Specifically, using signed weighted gene co-expression network analysis (WGCNA), we found a pluripotency module and a differentiation module, which are not identified in unsigned networks. We confirmed the importance of these modules by incorporating genome-wide TF binding data for key ES cell regulators. Interestingly, we find that the pluripotency module is enriched with genes related to DNA damage repair and mitochondrial function in addition to transcriptional regulation. Using a connectivity measure of module membership, we not only identify known regulators of ES cells but also show that Mrpl15, Msh6, Nrfl, Nup133, Ppif, Rbpj, Sh3gl2, and Zfp39, among other genes, have important roles in maintaining ES cell pluripotency and self-renewal. We also report highly significant relationships between module membership and epigenetic modifications (histone modifications and promoter CpG methylation status), which are known to play a role in controlling gene expression during ES cell self-renewal and differentiation.

Conclusion: Our systems biologic re-analysis of gene expression, transcription factor binding, epigenetic and gene ontology data provides a novel integrative view of ES cell biology.

Background

Embryonic stem (ES) cells have two important characteristics: pluripotency, the ability to differentiate into any

type of cell in the body, and self-renewal, the ability to replicate indefinitely. As such, they have tremendous therapeutic potential for regenerative medicine [1,2]. Current

work focuses on understanding and extending the network of genes that controls these key characteristics [3-16]. These efforts identified ES cell-specific transcription factors (TFs) that are differentially expressed between ES cells and differentiated cells (fibroblasts). Several studies have identified the targets of these TFs and the mechanism by which they regulate them [4,8,17]. Highly differentially expressed TFs (Oct4, Sox2, c-Myc, and Klf4) have been found capable of reprogramming fibroblasts to a pluripotent state [3].

While standard differential expression analysis techniques have led to remarkable discoveries they ignore the strong correlations that may exist between gene expression profiles. As a consequence, the user of a standard marginal analysis can drown in information but starve in knowledge. This is especially true when considering ES cells where many genes change expression during differentiation. For example, in a data set from Zhou *et al* 2007, which we consider below, more than 6200 genes were highly differentially expressed (Student t-test p-value smaller than the very stringent threshold of 10^{-6}). It is difficult to further prioritize these genes and to learn the underlying biological pathways. In contrast, co-expression networks, also referred to as 'association,' 'correlation,' or 'influence' networks [18-22], realize that genes can be highly correlated and thus can be grouped into large clusters (co-expression modules). For example, our network analysis of the same data organizes the genes into only 8 large modules. Next our module-centric analysis focuses on understanding the modules and their key regulators. Since it applies significance testing to the level of modules, co-expression network analysis may greatly alleviate the multiple testing problem that plagues standard gene-centric methods [23]. Gene co-expression network methods have been successfully applied in a variety of different settings [18,19,21,22,24-32].

In this article, we demonstrate that a co-expression network analysis of stem cell data sets provides novel biological insights that cannot be found using conventional techniques. Using external data (including gene ontology, TF binding data, epigenetic regulators), we also contrast the performance of signed and unsigned network construction methods. We find that signed co-expression network analysis performs best in this stem cell application. We identify pluripotency and differentiation related co-expression modules and novel ES cell regulators.

Results and discussion

Constructing Signed Co-expression Networks

We first define a gene co-expression similarity measure which is used to define the network. We denote the gene co-expression similarity measure of a pair of genes i and j by s_{ij} . Many co-expression studies use the absolute value of

the Pearson correlation as an unsigned co-expression similarity measure,

$$s_{ij}^{unsigned} = |cor(x_i, x_j)|, \quad (1)$$

where gene expression profiles x_i and x_j consist of the expression of genes i and j across multiple microarray samples. However, using the absolute value of the correlation may obfuscate biologically relevant information, since no distinction is made between gene repression and activation. In contrast, in signed networks the similarity between genes reflects the sign of the correlation of their expression profiles. To define a signed co-expression measure between gene expression profiles x_i and x_j , we use a simple transformation of the correlation:

$$s_{ij}^{signed} = \frac{1 + cor(x_i, x_j)}{2}. \quad (2)$$

As the unsigned measure $s_{ij}^{unsigned}$, the signed similarity

s_{ij}^{signed} takes on a value between 0 and 1. Note that the unsigned similarity between two oppositely expressed genes ($cor(x_i, x_j) = -1$) equals 1 while it equals 0 for the signed similarity. Similarly, while the unsigned co-expression measure of two genes with zero correlation remains zero, the signed similarity equals 0.5.

Next, an adjacency matrix (network), $A = [a_{ij}]$, is used to quantify how strongly genes are connected to one another. A is defined by thresholding the co-expression similarity matrix $S = [s_{ij}]$. 'Hard' thresholding (dichotomizing) the similarity measure S results in an unweighted gene co-expression network. Specifically an unweighted network adjacency is defined to be 1 if $s_{ij} > \tau$ and 0 otherwise, i.e. two genes are considered connected if their similarity measure is above a given threshold τ , and are considered separated otherwise.

Because hard thresholding encodes gene connections in a binary fashion, it can be sensitive to the choice of the threshold and result in the loss of co-expression information [19]. The continuous nature of the co-expression information can be preserved by employing soft thresholding, which results in a weighted network. Specifically, we use a continuous measure to assess their connection strength:

$$a_{ij} = s_{ij}^{\beta}, \quad (3)$$

where the power β is the thresholding parameter. As a default we use $\beta = 6$ and $\beta = 12$ for unsigned and signed networks, respectively. Alternatively, β can be chosen using the scale-free topology criterion [19]. Since $\log(a_{ij}) =$

$\beta \times \log(s_{ij})$, the weighted network adjacency is linearly related to the co-expression similarity on a logarithmic scale. Figure 1 shows the resulting adjacencies after applying the co-expression similarity measures and thresholding. Note that a high power β transforms high similarities into high adjacencies, while pushing low similarities towards 0. Since this soft-thresholding procedure leads to weighted adjacency matrix, the ensuing analysis is referred to as weighted gene co-expression network analysis or WGCNA [19,23,33,34].

A major step in our module centric analysis is to cluster genes into network modules using a network proximity measure. Roughly speaking, a pair of genes has a high proximity if it is closely interconnected. We will use the convention that the maximal proximity between two genes is 1 and the minimum proximity is 0. Specifically, we define the proximity as the topological overlap measure (TOM) [35-37] which can also be defined for weighted networks [19]. The TOM combines the adjacency of two genes and the connection strengths these two genes share with other "third party" genes (see equation 6 in the Methods section and Additional File 1). The TOM is a highly robust measure of network interconnectedness (proximity). This proximity is used as input of average linkage hierarchical clustering. Modules are defined as

branches of the resulting cluster tree [38]. This module detection procedure has been used in many applications [23,25-30,32,39,40] and a comparison to alternative procedures is beyond the scope of this article.

We find it convenient to summarize the gene expression profiles of a given module with the module eigengene, which can be considered as the best summary of the standardized module expression data [33,41]. The module eigengene of a given module is defined as the first principal component of the standardized expression profiles (see equation 8 in the Methods section).

Quantifying Module Membership

To identify possible regulators within a given module, we looked for highly connected intramodular hub genes, i.e. genes that have strong connections within the module. In our effort to find these genes, we examined two types of connectivity measures, which can be applied relative to any module q . The first connectivity measure is intramodular connectivity $k_i^{(q)}$ defined as

$$k_i^{(q)} = a_{i1} + a_{i2} + \dots + a_{in^{(q)}} \quad (4)$$

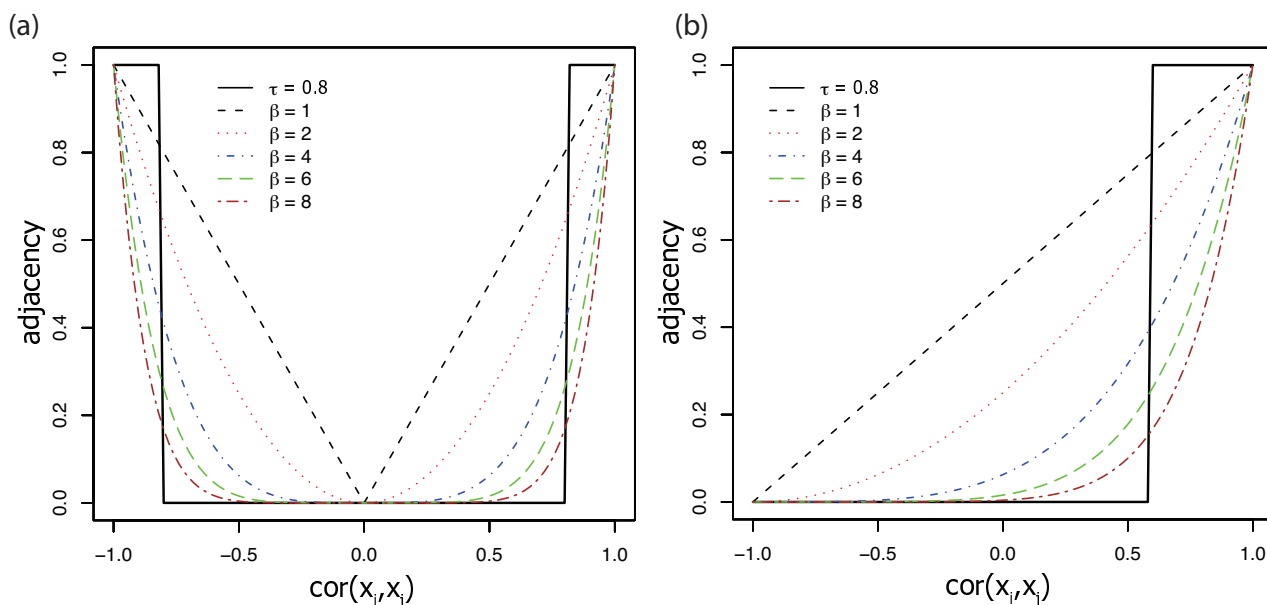


Figure 1
Network Connection Strength Versus Expression Correlation. Network adjacency (y-axis) versus correlation (x-axis) for an unweighted network (black step function with $\tau = 0.8$) and weighted networks (dashed lines corresponding to different powers, β) in an unsigned network (a) and a signed network (b). Note that $\text{cor}(x_i, x_j) = -1$ leads to adjacency = 0 in the signed network. The weighted network preserves the continuous nature of the co-expression information while an unweighted network dichotomizes the correlation.

where $n^{(q)}$ is the number of genes in the q^{th} module. In the case of an unweighted network, $k_i^{(q)}$ simply counts the number of connections to gene i within the q^{th} module. Intramodular connectivity can be interpreted as a measure of module membership: the higher the intramodular connectivity, the more centrally located the gene is in the module and the more certain is its membership with regard to this module. In signed networks, these highly connected hub genes may up-regulate adjacent genes since they are positively correlated with them, while in unsigned networks they may activate or repress their neighboring genes.

The second connectivity measure is the module eigengene based connectivity, $k_{ME,i}^{(q)}$ (also known as the signed module membership measure [33]), defined as

$$k_{ME,i}^{(q)} = \text{cor}(x_i, E^{(q)}), \quad (5)$$

where $E^{(q)}$ is the eigengene of the q^{th} module (see equations 9 and 10 in the Methods section) and x_i is the expression profile of the gene i . We denote modules by colors. For example, $k_{ME,i}^{\text{blue}}$ denotes the module membership measure of the i -th gene with regard to the blue module.

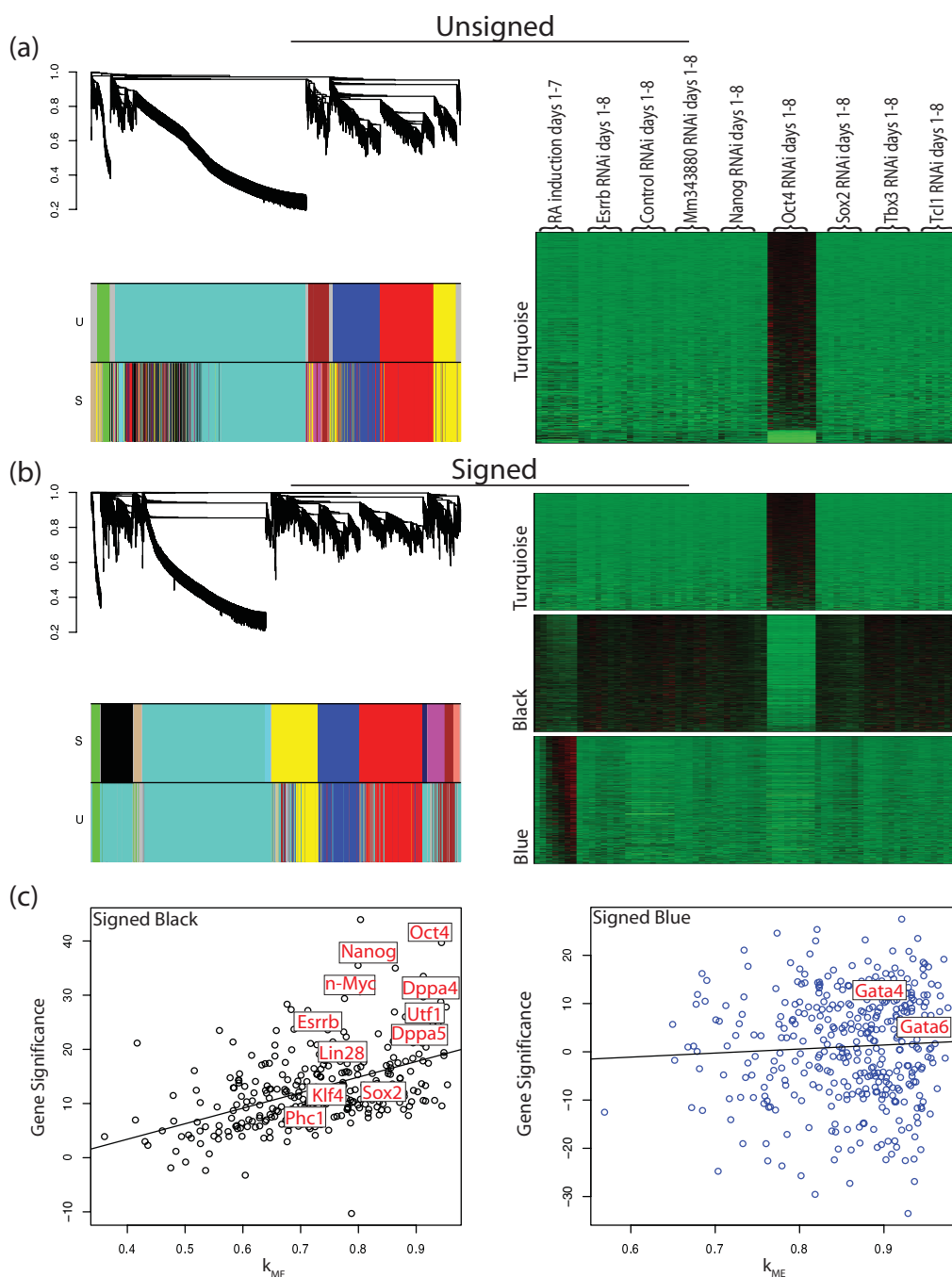
Module eigengene based connectivity has several advantages over intramodular connectivity: first, it is naturally scaled to take on values between -1 and 1; second, one can use a correlation test to calculate a corresponding p-value for a gene's module membership; third it can be used in signed networks to identify genes that are anti-correlated with a given module eigengene (i.e. they may repress genes in the module), and fourth, k_{ME} can be computed for any gene on the array (not just genes used in the network construction). In practice, we found that intramodular and module eigengene based connectivity are highly correlated (Additional File 2). A priori, the connectivity measures defined in equations 4 and 5 are quite different. But we show in the Methods section that a simple theoretical relationship between them can be derived in the context of a signed co-expression module. Due to its advantages, we used the module eigengene based connectivity $k_{ME,i}^{(q)}$ as the measure of module membership in our applications.

Signed WGCNA Identifies Pluripotency Related Modules in Ivanova et al (2006) Data Set

We generated unsigned and signed co-expression networks to analyze over 17,000 genes measured across 70

expression arrays from data published in Ivanova *et al* (2006) [4]. This data set contains expression profiles of ES cells individually depleted for the transcription factors Oct4, Nanog, Sox2, Esrrb, and Tbx3 by RNA interference (RNAi). The data set also includes expression profiles for RNAi knock downs of Tcf1 a co-activator of AKT kinase, and an EST (Mm343880), along with expression profiles of control ES cells carrying an empty RNAi vector and of ES cells differentiated by retinoic acid (RA). Each of these treatments was sampled over approximately eight days. To compare the performances of unsigned and signed WGCNA in identifying gene groups that are important for the regulation of the pluripotent state, we defined gene modules in unsigned and signed networks and assessed module function and importance by determining gene ontology terms associated with each module and examining module membership of genes known to play a role in ES cells. In addition, we analyzed how genes of a given module are bound by chromatin regulators or pluripotency TFs by incorporating independent promoter binding information.

Figure 2a shows the dendrogram of the unsigned network for the Ivanova *et al* data set. Modules were found by cutting branches of the cluster tree (dendrogram), using the dynamic tree cut library in R [38]. Modules are indicated by the color bands below the dendrogram. Genes that do not clearly belong to a branch are colored grey. To compare modules in signed and unsigned networks, we show two color bands: the top color band shows the genes colored by module membership in the unsigned network (corresponding to the dendrogram), while the bottom color band shows genes colored by module membership in the signed network. Similarly, Figure 2b displays the dendrogram of the signed network with the top color band showing genes colored by module membership in the signed network and the bottom color band showing genes colored by module membership in the unsigned network. These figures show that while some large modules (turquoise, yellow, red, and blue) are preserved in both networks, the signed network has two distinct small modules (black and tan) hidden within the unsigned turquoise module. The black and tan modules from the signed network are scattered throughout the unsigned network's turquoise module and cannot be detected since there is no branch of the dendrogram corresponding directly to these modules (i.e. regardless of the tree cutting algorithm employed, these modules would not be found in the unsigned network, data not shown). Figure 2a also shows a heatmap of the expression profiles of genes in the turquoise module from the unsigned network. Genes in this module exhibit a positive (red) or negative (green) change in expression upon knock down of the master regulator Oct4, which is not surprising given that Oct4 RNAi has been shown to cause a distinct differentiation pattern from other TF RNAi's [4,42]. Heat maps are similarly

**Figure 2**

Unsigned and Signed Mouse ES cell Networks in Ivanova et al. **a, left**, Dendrogram of the unsigned network of the Ivanova et al (2006) data set with color bands below indicating module membership for the unsigned network (U) and the signed network (S). **a, right**, heat map for visualizing standardized gene expressions (rows) across samples (columns) for genes in the turquoise module in the unsigned network. **b, left**, Dendrogram of the signed network of the Ivanova et al (2006) data set with color bands below indicating module membership for the signed network (S) and the unsigned network (U). **b, right**, heat map of expression profiles across samples for genes in the turquoise, black, and blue modules in the signed network. Note, modules are not scaled to reflect the number of genes in each module. **c**, scatter plot of module membership, k_{ME} (x-axis) plotted against gene significance, GS (y-axis) for the black and blue modules in the signed network with known ES cell regulators and differentiation genes labelled.

shown in Figure 2b for the turquoise, black, and blue modules from the signed network. In contrast to the unsigned network, each module contains genes with similar expression profiles. Here the turquoise module is made of genes that are activated when Oct4 is knocked down, while the black module contains genes that are repressed when Oct4 is knocked down and are down-regulated during retinoic acid (RA) induced differentiation. The blue module contains genes that are activated upon differentiation. This analysis shows that signed WGCNA identifies modules with more specific expression patterns than unsigned WGCNA.

Functional Enrichment with Regard to Known ES Cell Related Genes

Next we used external data to further study the gene modules defined by the networks and reveal their functional roles. We used two different strategies for this evaluation: first we assigned transcription factors and other regulators with known roles in pluripotency, self-renewal or differentiation to modules [4,8] and second, we incorporated genome-wide binding data for transcription factors and other regulators implicated in ES cell regulation in order to determine if these modules contain genes that are directly controlled by ES cell related TFs or differentiation suppressors [5,6,16].

Many genes known to maintain the pluripotent state of ES cells are found in the black module in the signed network. We defined a measure of gene significance (GS) as the t-statistic from the paired Student's t-test of expression in control RNAi samples and ES cell samples with RNAi knock down of Oct4 (paired by day of treatment). Figure 2c shows GS plotted against its module eigengene based connectivity, k_{ME} , in the black and blue modules of the signed network with marker genes labeled. Since the signed module membership k_{ME} is defined as the correlation between a gene expression profile and the module eigengene, its values lie between -1 and 1 with values near 1 signifying strong module membership to the corresponding signed module. Figure 2c shows a strong linear relationship between k_{ME} and GS in the black module (correlation = 0.5, p-value = 6.5e-13). As expected, most of the genes whose RNAi knock down induced ES cell differentiation in Ivanova *et al* [4] belong to the black module (Oct4, Nanog, Sox2, Esrrb, and Dppa4, Fisher's exact test p-value = 3.2×10^{-5}). Oct4's high connectivity ($k_{ME}^{black} = 0.94$) makes it a hub gene in the black module, consistent with its known role as a master regulator of the pluripotent state. Furthermore, many genes that are known to be highly expressed in ES cells are also in the

black module (e.g. Klf4, Utf1, and Phc1). Klf4 is one of the four TFs that can reprogram differentiated cells into a pluripotency state [3]. Utf1 interacts with Oct4, affects chromatin regulation in ES cells, and has recently been shown to improve reprogramming efficiency [43-45]. Phc1 is a Polycomb Group (PcG) protein. PcG proteins repress genes that become active upon differentiation of ES cells by mediating histone H3 lysine 27 tri-methylation and histone H2a ubiquitination [6]. The blue module contains Gata6 and Gata4, which are both highly connected ($k_{ME}^{blue} = 0.93$ and 0.88, respectively). These TFs are markers of ES cell differentiation, particularly into endoderm. Below we provide further evidence that the black and blue modules are related to pluripotency and differentiation respectively.

Module Enrichment with Regard to Known ES Cell Regulators

We incorporated genome-wide binding data for TFs (Oct4, Sox2, Nanog, Stat3, Smad1, cMyc, nMyc, Zfx, E2f1) and other regulators (Suz12) implicated in the maintenance of pluripotency and self-renewal, which were obtained by chromatin immunoprecipitation (ChIP) and massive parallel sequencing (ChIP-seq) by Chen *et al* (2008) [16]. Oct4, Sox2, Nanog, Smad1, and Stat3 are referred to as the Oct4 group of TFs, as they have been shown to often co-bind genomic regions; cMyc, nMyc, E2f1, and Zfx are referred to as the cMyc group of TFs because they also co-bind genomic regions [16]. Together TFs in the Oct4 and cMyc group are thought to activate expression of genes involved in pluripotency and self-renewal. Suz12, is a subunit of the histone H3K27 methyltransferase PcG protein complex, which represses genes that are activated upon differentiation [6,16].

To determine if binding by the two TF groups and Suz12 occurs more often in certain modules we computed binding enrichment for each module. Enrichment is defined as the odds ratio, that is the probability of a gene being bound by a particular TF or TF complex for genes in a given module divided by the probability of being bound for genes not located in the module. Table 1 shows module enrichment of genes bound by TFs in the Oct4 group, the cMyc group, and Suz12 for modules in the unsigned and signed networks. A gene is called bound by the Oct4 or cMyc groups if it is bound by at least 4 of the 5 and 3 of the 4 TFs in each TF group, respectively (similar results are found when using 3 of 5 and 2 of 4, data not shown). In agreement with the notion that the black module found in the signed network contains genes that are implicated in pluripotency, this module is strongly enriched with genes bound by TFs in the Oct4 and cMyc groups and under-enriched for binding by Suz12. Specifically, the

Table 1: Transcription Factor Binding in Ivanova et al Networks

Module	No. of genes	Unsigned Network Oct4Complex	cMycComplex	Suz12
blue	899	0.863, 0.232	1.09, 0.0398	0.956, 0.366
brown	1241	0.862, 0.179	0.757, 3.09e-09	1.92, 2.5e-21
green	1556	0.825, 0.0832	0.811, 1.78e-07	1.13, 0.0496
grey	7407	1.16, 2.37e-05	1.18, 4.62e-45	0.866, 2.17e-07
red	1673	1.1, 0.169	1.68, 6.75e-74	0.281, 7.61e-28
turquoise	5721	0.898, 0.0268	0.801, 4.3e-36	1.2, 6.96e-10
yellow	1053	1.14, 0.15	1.46, 5.74e-22	0.458, 7.39e-10
Signed Network				
black	941	2.9, 7.25e-24	1.83, 4.8e-58	0.288, 1.95e-15
blue	220	0.608, 0.165	0.488, 1.59e-07	1.93, 4.42e-05
brown	1090	0.736, 0.0402	0.754, 2.04e-08	1.99, 3.27e-21
cyan	581	0.783, 0.174	0.931, 0.15	1.78, 2.58e-08
green	1538	0.731, 0.0138	0.778, 1.14e-09	1.12, 0.0534
grey	4428	0.949, 0.224	0.853, 2.38e-14	1.13, 0.000604
magenta	514	1.09, 0.286	1.11, 0.0548	0.824, 0.122
midnightblue	649	1.61, 0.00131	1.89, 2.87e-45	0.181, 6.99e-15
pink	2242	1.4, 4.41e-05	1.85, 8.63e-160	0.341, 3.32e-31
red	1270	1.01, 0.429	1.56, 1.19e-37	0.371, 8.07e-16
salmon	650	1.03, 0.388	1.19, 0.00083	0.67, 0.00378
tan	445	2.16, 4.62e-06	1.71, 4.65e-21	0.37, 3.14e-06
turquoise	5205	0.679, 2.41e-09	0.749, 3.51e-49	1.24, 2.71e-11
yellow	1042	1.05, 0.326	1.35, 1.9e-13	0.633, 4.69e-05

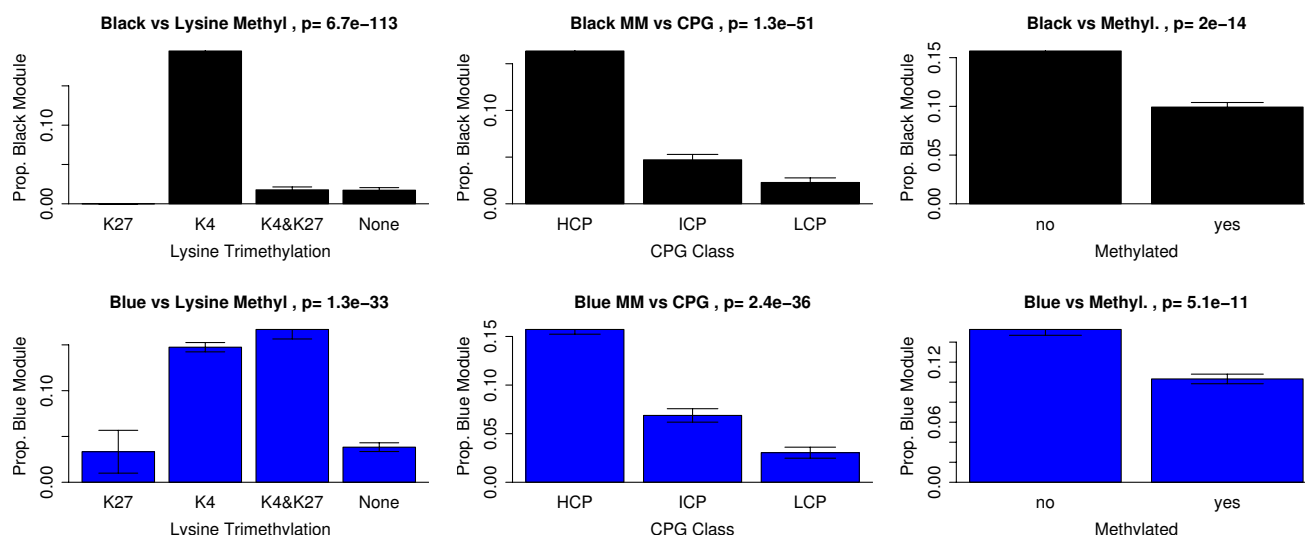
Enrichment for binding by the Oct4 and cMyc TF Groups and Suz12.

Each enrichment score (before the comma) is followed by its corresponding p-value (after the comma) calculated from the hyper-geometric distribution. P-values are uncorrected for multiple comparisons.

proportion of genes bound by the Oct4 group in the black module is almost twice the proportion of genes bound in the general population (1.94). Similarly genes in the black module are almost twice as likely to be bound by TFs in the cMyc group (1.87) and are almost three times less likely to be bound by Suz12 (0.344). These enrichments further support the idea that the black module is a pluripotency and self-renewal module. Other modules, like the blue, brown, and cyan, are enriched for Suz12 bound genes and under-enriched for genes bound by the two TF groups. The blue module is the most significantly enriched for Suz12 binding further supporting the idea that genes in this module are involved in ES cell differentiation. For modules preserved in both unsigned and signed networks, Table 1 shows that enrichment values are generally more significant in the signed network. Similar enrichments can be seen for Oct4 and Nanog binding from Loh *et al* (2006) and Polycomb Group (PcG) protein binding from Boyer *et al* (2006) (Additional File 3). The incorporation of binding data suggests that signed WGCNA better separates genes into modules based on function and regulation. Specifically, the black module can be considered a pluripotency/self-renewal module, while the blue, brown, and cyan modules can be considered differentiation modules in the signed network. It is important to reiterate that the pluripotency/self-renewal black module was only found using signed WGCNA.

Epigenetic Regulation and Module Membership

Recent studies suggest that chromatin structure and epigenetic modifications, like histone modification and DNA methylation, play a role in controlling gene expression during ES cell self-renewal and differentiation [46-51]. For example, gene repression by the PcG protein complex via histone H3 lysine 27 trimethylation (H3K27me3) is required for ES cell self-renewal and pluripotency [6,52]. To understand how epigenetic variables contribute to the regulation of ES cells we studied the relationship of the pluripotency and differentiation modules with ES cell H3K4 and H3K27 trimethylation, DNA methylation, and CpG promoter content from previously published data sets [50,51]. We related the epigenetic variables to module membership in the black (Figure 3, top row) and blue module (Figure 3, bottom row). Specifically, we determined what proportion of genes with a given epigenetic mark (H3K4me3 for example) are also in the top 1000 genes with the highest k_{ME}^{black} (or k_{ME}^{blue}). Below we show that our findings are highly robust with respect to the number of selected module genes (see also Additional File 4). H3K4me3 is associated with gene activation, whereas H3K27me3 is known to silence genes. Figure 3 shows that

**Figure 3**

Relating Module Membership to Epigenetic Regulation. The top 1000 genes with highest module membership in the black module (top row) and blue module (bottom row) are related to 3 epigenetic variables (corresponding to the 3 columns). The y-axis reports the proportion of top 1000 genes that are known to belong to the group of genes defined on the x-axis. Histone H3K4me3 trimethylation status is abbreviated K4, H3K27me3 trimethylation status is abbreviated by K27. Promoters that are both H3K4 and H3K27 trimethylated in ES cells (denoted K4&K27) are thought to poise key developmental genes for activation upon differentiation [50,51]. Note that genes with promoter CpG methylation are significantly ($p = 2.0 \times 10^{-14}$) under-enriched with respect to the top 1000 black module genes.

genes with H3K4 trimethylation in ES cells contain significantly more black module genes than genes without this marker (Kruskal Wallis test p -value = 6.7×10^{-113}) which may reflect the active role the pluripotency module plays in ES cells. Interestingly, genes with H3K4 trimethylation or bivalent methylation contain significantly ($p = 1.3 \times 10^{-33}$) more blue module genes than other gene classifications. Promoters that are both H3K4 and H3K27 trimethylated in ES cells (referred to as bivalent promoters) are thought to poise key developmental genes for activation upon differentiation [50].

Mammalian gene promoters are known to fall into one of at least two major classes: 1) CpG-rich promoters are associated with both ubiquitously expressed 'housekeeping' genes, and genes with more complex expression patterns, particularly those expressed during embryonic development and 2) CpG-poor promoters are generally associated with highly tissue-specific genes. To understand the role of CpG content in our modules we analyzed three CpG content classifications from Mikkelsen *et al* (2007): high (denoted HCP), low (LCP), and intermediate (ICP). Figure 3 shows that HCP genes contain significantly more black module genes ($p = 1.3 \times 10^{-51}$) and significantly more blue module genes ($p = 2.4 \times 10^{-36}$) than ICP or LCP genes. The LCPs are known to have a very different trimethylation pattern than the HCPs. Few (6.5%) of LCPs

have significant H3K4me3 in ES cells and virtually none have H3K27me3. HCPs and LCPs are subject to distinct modes of regulation. In ES cells, all HCPs seem to be targets of trithorax group activity, and may therefore drive transcription unless actively repressed by PcG proteins. In contrast, LCPs seem to be inactive by default, independent of repression by PcG proteins, and may instead be selectively activated by cell-type- or tissue-specific factors [50].

Figure 3 also shows promoter CpG methylation in relation to module membership. DNA methylation in mammalian cells plays multiple roles in cell physiology, including genome stability, repression of endogenous retroviral elements, genomic imprinting. Levels of DNA methylation are dynamically regulated during embryogenesis but less is known about the role DNA methylation play in gene expression and maintenance of pluripotency in ES cells [51]. Figure 3 shows that methylated genes are significantly under-enriched for black module ($p = 2.0 \times 10^{-14}$) and significantly under-enriched for blue module genes ($p = 5.1 \times 10^{-11}$). In Additional File 5, we present the data used for cross-referencing module membership to epigenetic regulators.

Variance in k_{ME} Explained by Epigenetic Variables

The above results demonstrate highly significant relationships between module membership and epigenetic varia-

bles. In the following, we probe deeper and determine the proportion of variance in module membership that can be explained by the epigenetic variables. Using analysis of variance, we can determine what proportion of the variation in eigengene-based connectivity k_{ME} can be explained by the different epigenetic variables. As can be seen from Table 2, the epigenetic variables explain only 8.3% of the variation in k_{ME}^{black} and 4.2% of the variation in k_{ME}^{blue} . For k_{ME}^{black} , histone trimethylation status ($p < 2.2 \times 10^{-16}$, prop. of variance explained = 6.7%) and cMyc complex binding ($p < 2.2 \times 10^{-16}$, prop. of variance explained = 1.5%) explain most of the variation. For k_{ME}^{blue} , histone trimethylation status ($p < 2.2 \times 10^{-16}$, prop. of variance explained = 3.4%) and CPG class ($p = 6.0 \times 10^{-10}$, prop. of variance explained = 0.5%) explain most of the variation. In summary, we find highly significant but relatively weak relationships between module membership and epigenetic variables.

Signed WGCNA Identifies a Pluripotency Module in data from Zhou et al (2007)

To further investigate WGCNA's ability to discover functionally important groups of genes, we turned to an independent data set from Zhou et al (2007) [8]. In this study, ES cells were removed from feeder cells and leukemia inhibitory factor (LIF) to induce differentiation. During the course of differentiation, cells were separated based on expression of an Oct4 green fluorescent protein (GFP)

reporter gene. Multiple samples were taken from undifferentiated ES cells and cells sorted at days 2, 4, 8, and 15 for high and low Oct4 expression. As before, we first identified gene modules via signed and unsigned methods and then related module membership to external data. In the following we show that a pluripotency/self-renewal and a differentiation module can be found in this new data set. For consistency between data sets, we have colored these modules black and blue, respectively.

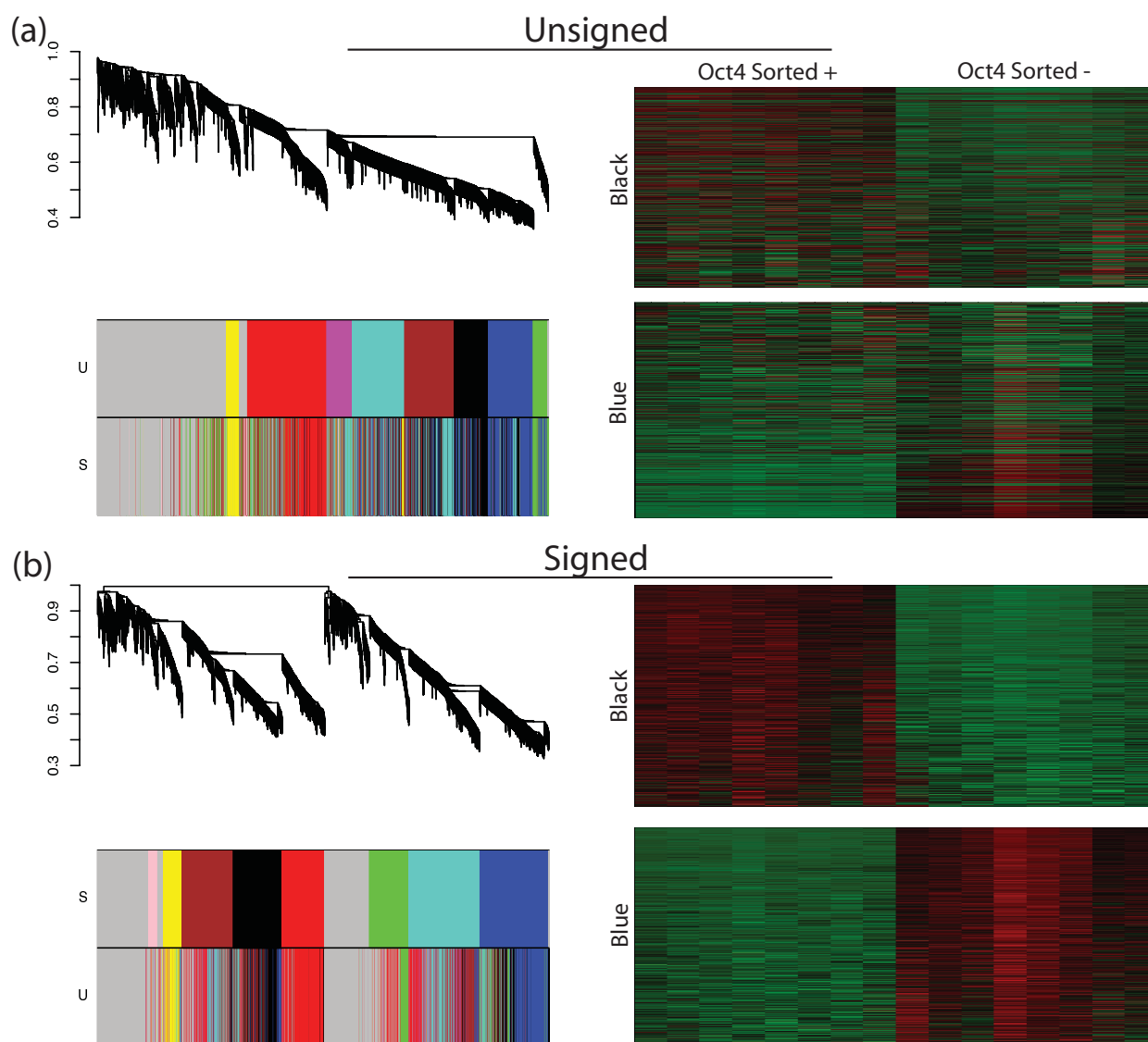
Cluster Tree Comparison of Unsigned and Signed Networks

Similar to Figure 2, Figure 4 shows the dendrograms of the unsigned and signed networks with color bands indicating module membership in the unsigned and signed networks. The heat maps of the expression profiles of genes in the blue and black modules appear the same across the unsigned and signed networks. The dendrograms, however, reveal network differences. Because module genes roughly stay together, we note that the red, black, blue, and yellow modules are partially preserved in the signed network. However, the black and blue modules are co-located on the same large branch in the unsigned network (Figure 4a) but are separated onto two distinct branches in the signed network (Figure 4b), suggesting a change in network topology. Therefore, depending on branch cutting methods used, the blue and black modules could easily be merged into a larger module in the unsigned network. Indeed, the structure of the dendrogram in Figure 4a suggests that the black and blue modules (along with the turquoise, magenta, and brown) may comprise one large module. For illustrative purposes, we have sepa-

Table 2: Module Membership Versus Epigenetic Variables

Source of Variation in kME Source	Degrees Of Freedom	kMEblack, Total Prop Var Explained = 8.3%			kMEblue, Total Prop Var Explained = 4.2%		
		Sums of Sq	Prop. Of Total Var	p-value (F test)	Sums of Sq	Prop. Of Total Var	p-value (F-test)
Histone Trimethylation (K4, K27, K4&K27, none)	3	170.91	0.067	< 2.2E-16	30.49	0.034	< 2.2E-16
cMyc Complex	1	39.21	0.015	< 2.2E-16	1.48	0.002	2.6E-04
Oct4 Complex	1	8.62	0.003	8.0E-08	0.79	0.001	7.5E-03
CPG class (HCP, ICP, LCP)	2	4.56	0.002	4.9E-04	4.71	0.005	6.0E-10
PcG Bound	1	0.88	0.000	8.7E-02	0.19	0.000	1.9E-01
CpG Methylated	1	0.04	0.000	7.1E-01	0.58	0.001	2.2E-02
Residual Error	7846	2353.8	0.917		868.36	0.958	
Total Variation	7855	2567.11			906.6		

This analysis of variance table reports which epigenetic variables and TF binding data have a significant effect on k_{ME}^{black} (columns on the left hand side) and k_{ME}^{blue} (columns on the right hand side). For each variable (source of variation), the columns report the degrees of freedom, the sums of squares, the proportion of total variance explained by the variable and the corresponding F-test p-value. Note that histone trimethylation status is the most significant source of variation for both k_{ME}^{black} and k_{ME}^{blue} .

**Figure 4**

Unsigned and Signed Networks of the Zhou et al ES Expression Data. **a, left,** Dendrogram of the unsigned network of the Zhou et al (2007) data with color bands below indicating module membership for the unsigned network (U) and the signed network (S). **a, right,** A heat map shows microarray expression profiles across samples for genes in the blue and black modules in the unsigned network. **b, left,** Dendrogram of the signed network of the Zhou et al (2007) data with color bands below indicating module membership for the signed network (S) and the unsigned network (U). **b, right,** heatmap of expression profiles (rows) across samples (columns) for genes in the blue and black modules in the signed network.

rated them into smaller modules. In the signed network, on the other hand, the black and blue modules will remain separate regardless of the branch cutting technique employed.

Module Membership in Unsigned and Signed Networks

Figure 5 shows the module eigengene based connectivity, k_{ME} , of genes in the black and blue modules plotted against their fold change between samples with high and

low Oct4 expression as defined in Zhou et al (2007). Figures 5a and 5b show that the *unsigned* black and blue modules each contain genes from both the *signed* black and blue modules. However, these genes are clearly separated by fold change and connectivity. Furthermore, both these unsigned modules contain two types of genes, pluripotency/self-renewal related genes and differentiation related genes. For example, the unsigned black module has pluripotency TFs Oct4, Nanog, and Esrrb, while

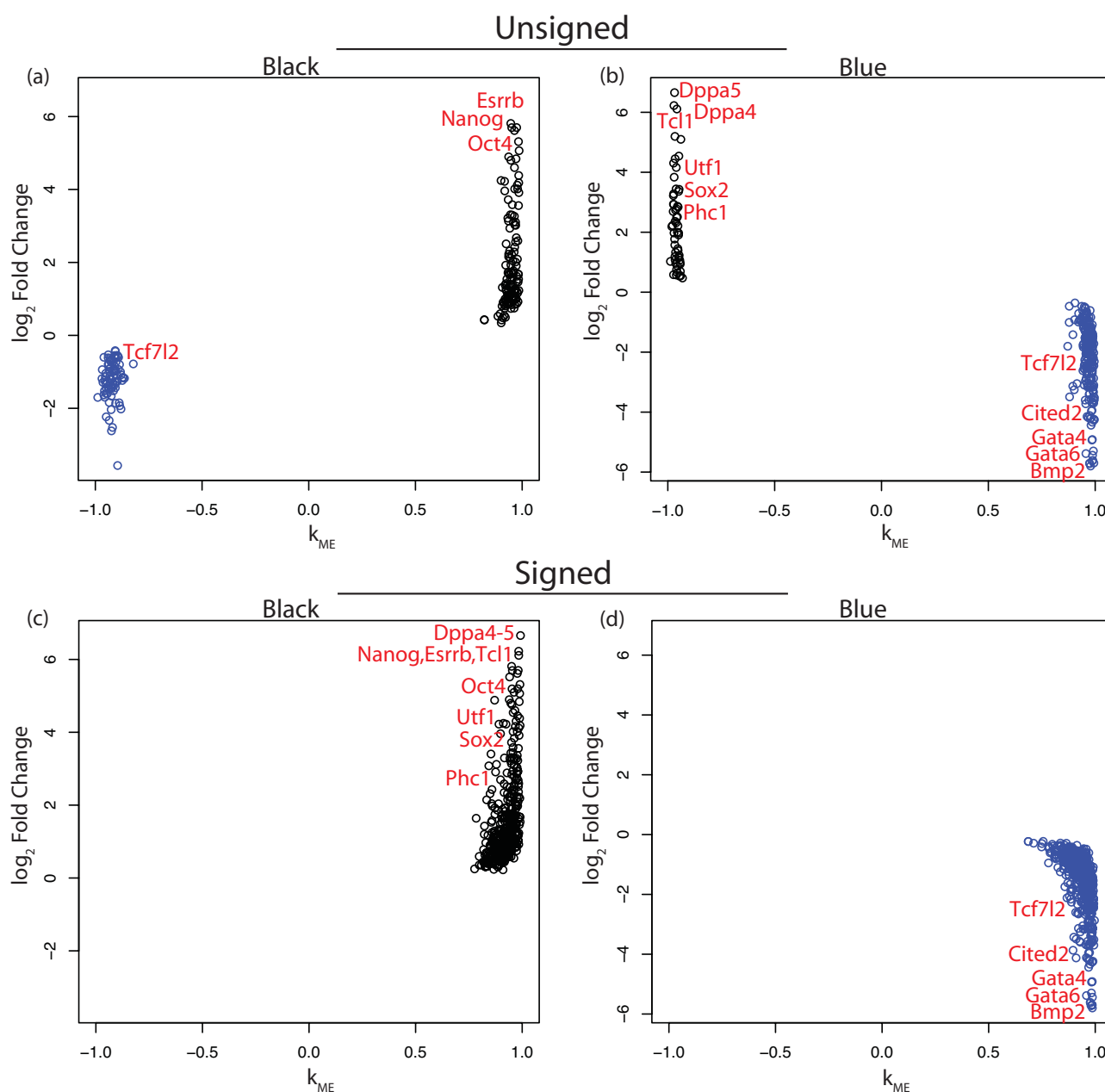


Figure 5
Expression Changes Versus Module Membership in the Black and Blue Modules (Zhou et al). Module membership, k_{ME} , is plotted against $\log_2$ expression fold change (FC) for the black (a) and blue (b) modules in the *unsigned* network of the Zhou et al (2007) data. FC is the ratio between the average expression in Oct4 positive samples and Oct4 negative microarray samples. Known ES cell regulators are labeled. Genes are colored by module membership in the *signed* network. c and d are analogous to a and b but module membership is with regard to the *signed* black and blue modules.

also having Tcf7l2, which is known to play a role in differentiation. The unsigned blue module has pluripotency genes Sox2, Dppa4, Dppa5, Utf1, and Phc1 along with differentiation genes Gata4, Gata6, Cited2, and a second probe for Tcf7l2. Figures 5c and 5d show the black and blue modules in the signed network. These modules con-

tain genes that not only have consistent fold changes and connectivities but also have consistent functional roles. Here the black module contains genes expressed similarly to Oct4 including (Nanog, Sox2, Dppa4, Dppa5, Utf1, and Phc1). Many of the genes found to induce differentiation when knocked down by Ivanova et al are in this

module (Oct4, Nanog, Sox2, Esrrb, Dppa4, and Tcf1, Fisher's exact test, $p\text{-value} = 1.01 \times 10^{-5}$). The blue module consists of genes that become expressed when Oct4 is down-regulated upon differentiation and contains many highly connected genes related to ES cell differentiation (Gata4, Gata6, Cited2, Bmp2, Tcf7l2, and Foxa12). These results confirm our earlier finding that the signed black module is a pluripotency module while the signed blue is a differentiation module.

A Comparison of Transcription Factor Binding Enrichment in Unsigned and Signed Networks

This conclusion is further supported by our analysis of TF binding patterns in the modules of the Zhou *et al* network. Table 3 shows enrichment for binding by the Oct4 and cMyc groups of TFs and Suz12 from ChIP-seq data from Chen *et al* [16]. In the signed network the black, brown, and red modules are enriched for binding by transcription factors in the Oct4 and cMyc groups and are under-enriched for Suz12 binding. This is not surprising since the black, red, and brown modules are co-located on the same large branch in the signed network (Figure 4b), which shows that they have similar expression profiles and share many of the same gene connections. The blue module, on the other hand, is under-enriched for genes bound by the Oct4 group TFs. These results were also found when using Oct4, Nanog, and PcG binding data from two other studies (Loh *et al* (2006) and Boyer *et al* (2006)) (Additional File 6). The identification of the brown module and the increased enrichment of the black

module in the signed network show that signed WGCNA is better at grouping genes into modules based on transcriptional regulation. Furthermore, the black and blue modules are well separated in the signed network (Figure 4b), while in the unsigned network branch cutting could erroneously group them into one large module (Figure 4a).

Functional Enrichment Analysis of the Pluripotency and Differentiation Modules

Given that there was significant overlap between the pluripotency modules and differentiation modules of the Ivanova and Zhou networks, we focused further analysis on the network constructed from the Zhou *et al* data set as this network was based solely on differentiation induced expression changes. We determined functional enrichment using the Database for Annotation Visualization and Integrated Discovery (DAVID) [53]. Table 4 shows the significantly enriched GO terms for genes with the 5% highest k_{ME}^{blue} . In agreement with the finding that the blue module is related to differentiation, genes within this module are significantly enriched for a functional group containing organ development, system development, and cell differentiation ($p\text{-value} = 0.002$). Other highly enriched groups are involved in regulating protein localization ($p\text{-value} = 9.7 \times 10^{-17}$) and membrane composition ($p\text{-value} = 5.6 \times 10^{-8}$).

Table 3: Transcription Factor Binding in Zhou et al Networks

Module	No. of genes	Unsigned Network		
		Oct4Complex	cMycComplex	Suz12
black	2659	1.55, 2.08e-09	1.4, 4.18e-46	0.666, 6.21e-10
blue	1484	0.872, 0.172	1.09, 0.0125	1.01, 0.447
brown	992	1.2, 0.093	1.02, 0.315	1.11, 0.127
green	1493	0.758, 0.028	0.863, 0.000186	1.36, 4.79e-06
grey	7318	0.766, 1.87e-08	0.803, 1.69e-51	1.2, 7.19e-14
magenta	1583	0.919, 0.28	1.04, 0.147	1.1, 0.0982
red	2593	1.48, 2.3e-07	1.64, 4.02e-107	0.437, 2.49e-26
turquoise	838	0.997, 0.459	1.3, 1.74e-08	0.829, 0.0631
yellow	2002	1.08, 0.213	1.1, 0.00171	0.844, 0.0106
Signed Network				
black	1859	1.94, 2.94e-15	1.4, 4.22e-30	0.523, 2.19e-13
blue	1972	0.656, 0.000465	1.03, 0.18	1.13, 0.03
brown	1267	1.45, 0.000755	1.72, 1.22e-59	0.411, 5.78e-14
green	1548	0.818, 0.0763	0.933, 0.0414	1.27, 0.000329
grey	7175	0.729, 1.64e-10	0.784, 2.34e-59	1.24, 2.7e-18
pink	659	0.941, 0.432	1.27, 5e-06	0.474, 2.48e-06
red	2184	1.48, 2.18e-06	1.63, 1.7e-85	0.493, 9.37e-18
turquoise	2317	0.896, 0.156	1.05, 0.0498	1.09, 0.0716
yellow	1870	1.17, 0.0587	1.05, 0.0545	0.885, 0.054

Module enrichments for binding by the Oct4 and cMyc TF Groups and Suz12. Each enrichment score (before the comma) is followed by its corresponding p-value calculated from the hyper-geometric distribution (uncorrected for multiple comparisons).

Table 4: Functional Pathways in Highly Connected Pluripotency and Differentiation Related Genes in the Zhou et al Network

p-value	Functional Groups	Blue Module highly connected genes (kME)
9.72E-17	er-golgi transport; protein localization; protein transport; vesicle-mediated transport; secretion by cell; cellular localization; secretory pathway; intracellular transport;	Myl6 (0.994), Sh3glb1 (0.993), Tm9sf3 (0.993), Tram1 (0.992), Der11 (0.991), Serinc1 (0.991), Lman1 (0.991), Lrp10 (0.991), Mcfd2 (0.99), Mcfd2 (0.99), Tmed10 (0.99), Tpcn1 (0.989), Ar11 (0.989), Tinagl (0.987), Rab2 (0.987), Txndc1 (0.987), Col4a1 (0.987)
5.65E-08	Glycan structures – biosynthesis I; signal-anchor; transferase activity, glycosyltransferase	Glt8d1 (0.993), Creb3 (0.991), Fut8 (0.99), Fkrp (0.99), Extl2 (0.989), Glt8d3 (0.987), Itm2c (0.986), Hs3st1 (0.986), Pofut2 (0.986), Dpagt1 (0.985), Mgat2 (0.983), Abhd6 (0.982), Ddost (0.982), Ndst2 (0.981), B4galnt1 (0.981), St3gal6 (0.98)
2.98E-06	membrane; transmembrane; transmembrane region; topological domain: Cytoplasmic	H13 (0.996), Pdgfra (0.994), Cd59a (0.994), Glt8d1 (0.993), Sh3glb1 (0.993), Tm9sf3 (0.993), Tram1 (0.992), Gdpd5 (0.991)
0.00185	organ development; system development; anatomical structure morphogenesis; cell differentiation; organ morphogenesis	Pdgfra (0.994), Myl6 (0.994), Sh3glb1 (0.993), Lmo4 (0.992), Rgnef (0.989), Syvn1 (0.988), Kit (0.988), Fndc3b (0.988), Txndc1 (0.987), Lama1 (0.987), Barx1 (0.986), Col4a2 (0.986), Ctgf (0.985), Fgf3 (0.985), Crim1 (0.983), Pthr1 (0.983)
p-value	Functional Groups	Black Module highly connected genes (kME)
1.98E-08	response to DNA damage stimulus; DNA damage; DNA repair	Msh6 (0.993), Rfl1 (0.983), Mrel1a (0.982), Setx (0.974), Xrcc5 (0.971), Chek1 (0.968), Xab2 (0.967), Xrn2 (0.967), Trp53 (0.959), Npm1 (0.958), Tdp1 (0.955), Bccip (0.954)
3.75E-08	Mitochondrion; transit peptide; Mitochondrion	Mrpl15 (0.992), Ppif (0.991), Mrps5 (0.987), Hspa9 (0.984), Coq3 (0.984), Tst (0.981), Mrpl45 (0.98), Akap1 (0.979), L2hgdh (0.978), Mrps31 (0.978), Chchd4 (0.976), Abcel (0.975), Dci (0.975), Fpgs (0.974), Mrpl39 (0.973), Bdh1 (0.971)
5.83E-08	nucleus; biopolymer metabolic process; DNA binding; cellular metabolic process; Transcription regulation;	Msh6 (0.993), Pes1 (0.991), Zic3 (0.991), Uchl1 (0.99), Rnfl38 (0.99), Rnfl38 (0.99), Wdr36 (0.989), Pou5f1 (0.989), Rbpj (0.987), Glo1 (0.987), Tdgf1 (0.987), OTTMUSG00000010173 (0.986), Aarsd1 (0.986), Nup133 (0.985), Xpo1 (0.985), Xpo1 (0.985), Dnajc6 (0.985), Klhl13 (0.984), Dppa4 (0.984)
5.26E-04	cell cycle phase; cell cycle process; cell cycle; mitotic cell cycle; mitosis; cell division	Pes1 (0.991), Rfl1 (0.983), Mrel1a (0.982), Gtbbp4 (0.972), Chek1 (0.968), Mnat1 (0.966), Rcc2 (0.964), Gadd45gip1 (0.963), Rpa1 (0.961), Hells (0.96), Trp53 (0.959), Terf1 (0.959)

Enriched GO terms of genes with the highest 5% black and blue k_{ME} 's. Highly connected genes in each functional group are given. All k_{ME} values are highly significant (Bonferroni corrected p-value < 0.00015).

Table 4 also shows significant GO terms for genes with the 5% highest k_{ME}^{black} . Given that many pluripotency TFs are in the black module, it is not surprising that the functional classifications, DNA binding and transcriptional regulation, are significantly enriched (p-value = 5.4×10^{-8}). However, two functional classifications, DNA damage/repair and mitochondrial function, are more significantly enriched than the transcriptional regulation group (p-values = 2.0×10^{-8} and 3.8×10^{-8} , respectively) suggesting that these pathways play important roles in maintaining pluripotency and self-renewal.

We also used Ingenuity Pathway Analysis, IPA, to compare functional enrichment in the pluripotency and differentiation modules (Ingenuity Systems, <http://www.ingenuity.com>). Additional File 7 shows that functional groups similar to those found using DAVID are enriched in the black and blue module respectively. Cell cycle and DNA replication, recombination, and repair are enriched in the black module compared to the blue module and skeletal, muscular, and cardiovascular system development are enriched in the blue module.

Comparison to a Standard Differential Expression Analysis

Here we compare some of our WGCNA results with those of a standard differential expression analysis. In Figure 2c and Figure 5 we showed that for some modules a strong relationship between module membership (k_{ME}) and differential expression (gene significance/fold change) can be observed. In [33], we provide a geometric description of modules for which such a relationship can be observed. While a close relationship may exist between k_{ME} and a Student t-test statistic, it does not imply that corresponding gene ranking procedures are equivalent.

Here we compare signed WGCNA to standard differential expression methods using three different approaches. First, we show that a gene ranking based on k_{ME} is more consistent (reproducible) than that based on the Student t-test in our data. Specifically, we computed two gene rankings for the Ivanova *et al* data set, one ranked by t-statistic and the other by connectivity to a module of interest. We similarly computed two such rankings for the Zhou *et al* data set and studied the overlap between the two data sets (Additional File 8). Of the 1000 genes most signifi-

cantly down regulated upon differentiation in each data set 139 overlap (hyper-geometric p-value = 1.0×10^{-20}). However, when ranking genes by connectivity to each data set's pluripotency model there is an increase in overlap to 230 (p-value = 1.7×10^{-75} , Additional File 8). This increased consistency is also seen in genes up regulated upon differentiation where 77 genes overlap between the two data sets (p-value = 0.02) when ranking by t-statistic and 161 genes overlap when ranking by connectivity to the differentiation modules (p-value = 2.8×10^{-31} , Additional File 8).

A second approach for comparing gene rankings is to use the functional enrichment with regard to known gene ontologies. To compare the abilities of k_{ME} versus the conventional t-statistic in identifying functionally interesting groups of genes we consider functional enrichment of genes found by one ranking method but not the other (Figure 6). Of the 1000 genes most strongly connected to the Ivanova *et al* pluripotency module (black), 463 overlap with the 1000 genes most significantly down regulated upon Oct4 RNAi in the same data set. Figure 6 shows the functional enrichment from Ingenuity Pathway Analysis, IPA, of the 537 genes in each group that do not overlap. Genes found only by signed WGCNA are significantly enriched for functions important to ES cells like DNA Replication, Recombination, and Repair, Cell Cycle, Cancer, Protein Synthesis, and RNA Post-Transcriptional Modification, which have previously been found using network methods [54]. Generally, genes identified by signed WGCNA exhibit more significant enrichment of functional classifications than those found by standard differential analysis.

We similarly compare the 1000 most highly connected genes in the differentiation (blue) module and the 1000 genes that are most significantly up regulated upon Oct4 RNAi in the Ivanova network. Interestingly, only five genes overlap (Figure 6). By examining those genes that do not overlap we see that ranking by connectivity yields greater significant enrichment for many functional groups important in ES cell differentiation including Organ Development, Tissue Development, Cell Morphology etc.

Similar analysis of pluripotency genes in Zhou *et al* yields consistent results with DNA Replication, Recombination, and Repair being more enriched when ranked by connectivity (Additional File 9) while analysis of highly connected genes in the differentiation module shows that differential analysis moderately out performs ranking by connectivity. The differences in functional enrichment in the Zhou *et al* data set are subtle given that there is more overlap between the two rankings (Additional File 8). This large overlap is likely due to the simplicity of the expression array samples which are filtered into only two

groups, those that exhibit Oct4 expression and those that do not. Meanwhile, signed WGCNA is especially useful in Ivanova *et al* where smaller overlap is caused by the complexity of the expression samples which are made of many different RNAi treatments.

A third approach for comparing gene rankings is to use the enrichment with regard to epigenetic and transcriptional regulators. In Additional File 4 we relate different gene rankings to enrichment significance with regard to the following variables (a) histone H3K4 alone versus all others, (b) bivalent H3K4 & H3K27 versus all others [50], (c) High CPG class versus all others (i.e. HCG versus ICG and LCG), (d) promoter CPG methylation status [51], (e) Oct 4 complex binding status, (f) cMyc complex binding status. We report results for 3 different gene rankings using the Ivanova data: the black and blue curve represent gene rankings according to k_{ME}^{black} and k_{ME}^{blue} , respectively. The grey curve represents ranking according to a Student T-test of differential expression. Additional File 4 shows that black and blue module genes can have very different enrichment results that tend to be very different from those of a standard analysis. This analysis illustrates how module membership provides important complementary variables along the Student t-test for understanding differences between genes.

The increased functional enrichment and improved consistency between data sets suggest that signed WGCNA is a complementary method to standard differential analysis. In practice, we recommend to use both k_{ME} and the Student t-test to find highly differentially expressed intramodular hub genes.

Pluripotency Module Genes involved in Transcriptional Regulation and Chromatin Structure

To gain a better understanding of the regulatory network involved in maintaining ES cell pluripotency and self-renewal, we first looked at genes with high black k_{ME}^{black} and GO terms related to transcriptional regulation or chromatin structure. Figure 7 lists such genes and contains information about how these genes are bound by TFs in the Oct4 and cMyc groups, and Suz12 plus Klf4, Esrrb, and Tcfcp2l1. The sign of black k_{ME}^{black} allows us to distinguish genes that promote pluripotency from those that repress it. As expected, many of the genes with high positive module membership measure (k_{ME}^{blue}) are known to participate in ES cell regulation (Zic3, Mkrn1, Phc1, Esrrb, Jarid2, Nodal, Jarid1b, Tgif1, Utf1, Hells, and Rest) [20,55,56]. Importantly, four TFs capable of reprogramming differentiated cells into a ES cell like state are in this list (Sox2, Klf2, Nanog, and Pou5f1(Oct4)) [10,11], con-



Figure 6
Comparison of Genes Ranked by Network Connectivity and Differential Expression in the Ivanova et al data set. Ingenuity Pathway Analysis of functional enrichments in the set of genes ranked within the top 1000 by Student's t-test and k_{ME} and yet do not overlap with each other. Venn diagrams show the amount of gene overlap between the top 1000 black (pluripotency) module genes and the top 1000 genes most significantly down-regulated upon Oct4 RNAi (left); gene overlap between the top 1000 blue (differentiation) module genes and the 1000 genes most significantly up-regulated with Oct4 RNAi (right). Significance of differential expression was determined using Student's t-statistic. p-values have been corrected for multiple hypothesis tests (Benjamini-Hochberg). Only significantly enriched functional groups are shown.

firming that the black module captures the known core transcriptional regulatory network responsible for maintaining a cell's stemness. Genes with high positive values of black module membership (k_{ME}^{black}) but have not been implicated in maintaining pluripotency and self-renewal

should therefore be strong candidates for further functional study. Such genes include Msh6 (involved in DNA damage and repair [57]), Rbpj (involved in Notch signaling [58]), Zfp39 (spermatogenesis [59]), and Nrf1 (mitochondrial organization and biogenesis [60]). Similarly, many of the highly negatively connected genes in the

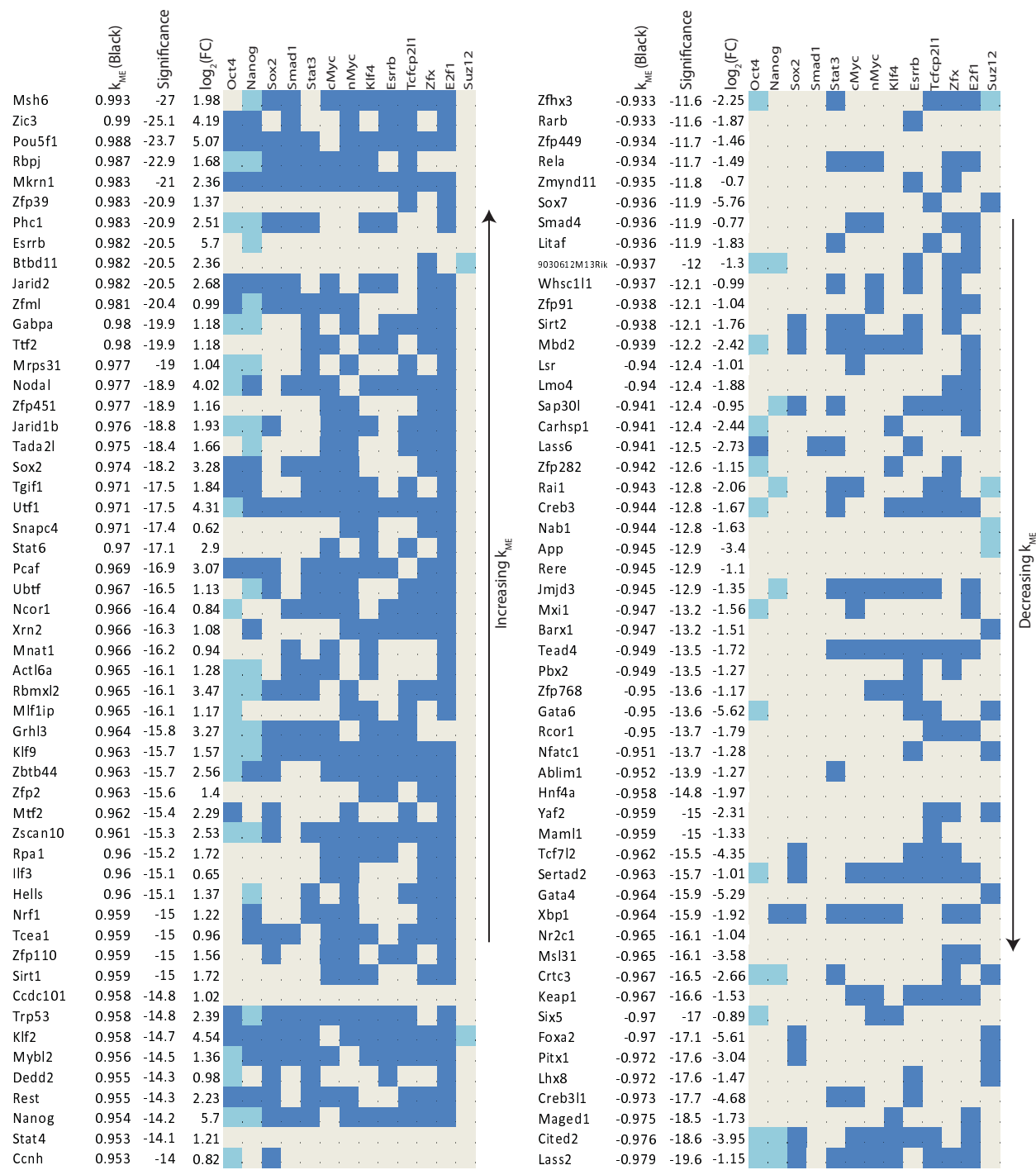


Figure 7
Transcriptional Regulators related to Pluripotency and Differentiation in the Zhou Network. TF and Suz12 binding in the promoter regions of highly connected genes related to ES cell pluripotency and self-renewal with GO terms of transcriptional regulation or chromatin structure. Genes are listed by black k_{ME} (positive, left and negative, right) along with their corresponding significance level ($\log_{10}$ of the Bonferroni corrected p-value generated by a correlation test). Binding data from Chen *et al* (2008), Boyer *et al* (2006), and Loh *et al* (2006), are marked in blue (bound) and beige (unbound). For Oct4, Sox2, and Suz12, where binding is given by two studies, binding will be blue if it is found in both studies and light blue if found in only one.

black module are known to play a role in ES cell differentiation (Cited2, Gata4, Gata6, Tead4, Foxa2, and Sox7). Because k_{ME}^{black} and k_{ME}^{blue} are highly negatively correlated ($r = -0.99$), genes with high positive k_{ME}^{blue} have negative k_{ME}^{black} and vice versa. Highly negatively connected genes that are not known to be involved in ES cell differentiation are also candidates for functional investigation. Some genes like Maged1 are known to play a role in cell differentiation but have not been shown to be important in ES cell differentiation specifically [61]. Other genes like Lass2 have little known about their role in cell differentiation. Figure 7 also shows that genes positively connected to the black module tend to be bound by more ES cell related TFs compared to the negatively connected genes, supporting the idea that these TFs bind and activate pluripotency and self-renewal genes [13].

Pluripotency Module Genes not Involved in Transcriptional Regulation or Chromatin Structure

Genes that are not involved in transcriptional regulation or chromatin modification (as defined by GO analysis) but have high average k_{ME}^{black} are also of interest. Figure 8 lists such genes along with TF and Suz12 binding information, connectivity, and fold change. Once again the importance of some genes has been validated, while others should be candidates for further research. Genes like Dppa5, Dppa4, and Tcl1 are markers of pluripotency and Nup133, a nuclear pore complex subunit, has recently been shown to be necessary in the maintenance of pluripotency [62]. Nup133 highlights the usefulness of signed WGCNA. Using the t-statistic from standard differential analysis Nup133 is ranked 222th most significantly down regulated upon differentiation while using connectivity its rank moves to 28th. Other candidate genes include Sh3gl2, which binds lipids and proteins [63], Mrpl15, a mitochondrial ribosomal protein, and Ppif, involved in mitochondrial function and oxidative stress-induced cell death [64]. Of the negatively connected genes in Figure 8, Ctsl (cathepsin L) has recently been shown to cleave the histone H3 N-terminus during ES cell differentiation [65], while little is known about Ctsz, also a cathepsin, while Gnas and Ctgf are differentiation genes [66,67]. The high negative connectivity of Uqcrh, a mitochondrial inner membrane protein [68], along with the high positive connectivity of Mrpl15 and Ppif, confirms that mitochondrial regulation may be distinct in ES cells [69-71] and suggests that they may be important regulators of mitochondrial function in ES cells.

Pluripotency Module Genes that Lack Binding by Known Pluripotency TFs

Figure 9 shows genes with relatively high k_{ME}^{black} that lack binding by the TFs Nanog, Oct4, Sox2, and Klf4 [10,11] and have low binding (≤ 2) by other TFs that maintain pluripotency (Smad1, Stat3, cMyc, nMyc, Esrrb, Tcfcp2l1, Zfx, and E2f1). Transcriptional regulators that have high positive module membership in the black module but lack binding by the pluripotency TFs are of interest since their strong module membership cannot be explained via regulation by these TFs. Their high average k_{ME}^{black} suggests that they may be upstream regulators of Oct4, Sox2, Nanog, and other genes important to pluripotency.

To further investigate the role of these genes, we used motif scanning methods described in Zhou *et al* (2007) [8] to determine if the binding sites of these genes are contained in regions co-bound by TFs in the Oct4 group or cMyc group in ChIP-seq data from Chen *et al* (2008). We concentrated solely on Lrh1 (Nr5a2) and Elk1 since their motifs have known position specific weight matrices while the other genes lack known motifs. Table 5 shows the enrichment and significance of motifs scanned. Sox2 and Oct4 bind to the composite SoxOct motif besides their own, Stat3 binds the Stat1 motif, and cMyc binds the Ebox motif. Both the Oct4 and cMyc groups' expected motifs are enriched. For example, the SoxOct motif has over three fold enrichment in regions bound by TFs in the Oct4 group. Interestingly, the Lrh1 motif is more enriched than the Nanog motif in sequences bound by the Oct4 group, which contains Nanog binding by definition. This reinforces the hypothesis that Lrh1 co-binds regions bound by TFs in the Oct4 group [8]. Furthermore, Lrh1 sites are found in the promoter regions of Pou5f1 (Oct4), Klf4, Dppa5, and Suz12 with Pou5f1 having three separate sites. These motif sites and Lrh1's known importance in ES cells, suggest that it may be an upstream regulator of these pluripotency factors and as such is a candidate for experimental validation [72]. The Elk1 motif is also significantly enriched in sequences bound by the cMyc group, thus Elk1 may co-regulate genes bound by TFs in this group.

A Geometric Interpretation of Signed WGCNA Modules

To understand how signed WGCNA is better able to separate genes into functional modules in the Ivanova data set, we plotted genes in the signed black or turquoise module relative to the unsigned turquoise module eigengene (Additional File 10). Note that genes located in the black and turquoise modules in the signed network are clearly separated into two clusters. Because a module eigengene is defined as the first principle component of its module,

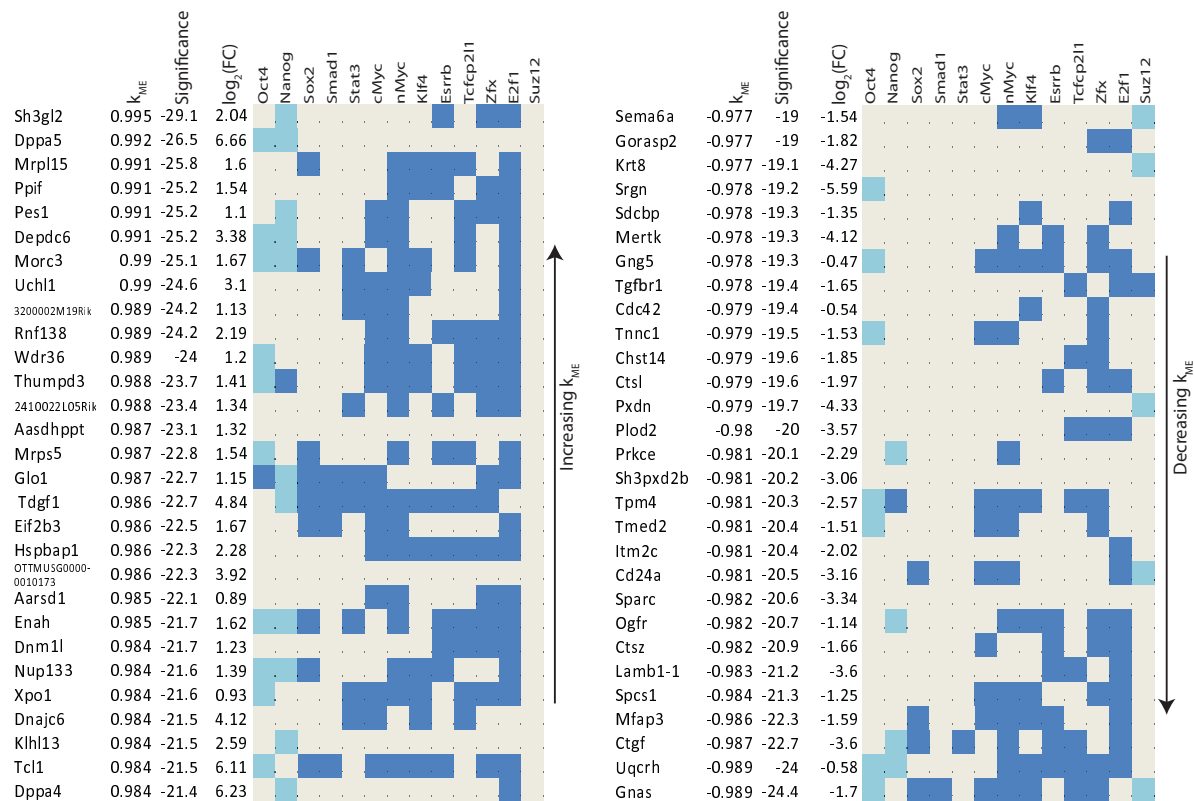


Figure 8
Non-Transcriptional Regulators Related to Pluripotency and Differentiation. TFs and Suz12 binding of highly connected genes related to ES cell pluripotency and self-renewal lacking GO terms for transcriptional regulation or chromatin structure. Genes are tabulated in the same format as Figure 7.

it describes the main direction in which the module's gene expressions vary. Note that the signed module eigengenes are oriented in the direction of their clusters. The direction of the unsigned turquoise module eigengene is more difficult to interpret. Because the turquoise module in the unsigned network contains two distinct signed modules (black and turquoise), its module eigengene describes the variance between these two sub-modules and the variance within the larger sub-module, the signed turquoise. As such, the unsigned turquoise module eigengene fails to quantify the true importance of highly connected genes in the signed black module. For example, Oct4's $k_{ME}^{turquoise}$ in the unsigned turquoise module is -0.74 while it is 0.94 in the signed black module. Thus, Oct4 is not identified as a hub gene in the unsigned network while it is clearly a hub gene in the signed network.

Conclusion

We show that a systems biology approach, which utilizes gene expression, transcription factor binding, genomic,

epigenetic and gene ontology data, can be improved by accounting for the sign of co-expression relationships. We also show that signed WGCNA has advantages over standard differential expression methods. Specifically, signed WGCNA has more consistent gene rankings between data sets (see Additional File 8), is better able to identify functionally enriched groups of genes (Figure 6), and its focus on module eigengenes circumvents the multiple testing problems that plague standard gene-based expression analysis. Below, we highlight several novel stem cell related genes that would not have been found using a standard differential expression analysis.

Signed WGCNA provides novel insight into murine ES cell biology, which unsigned WGCNA is unable to provide. Applying these signed methods to previously published data, we identified pluripotency and differentiation gene modules not found in unsigned networks or differential analysis. The results of signed WGCNA are robust as it identifies similar modules in independently published data sets. We show that module eigengene based connectivity k_{ME} is valuable for annotating genes with regard to

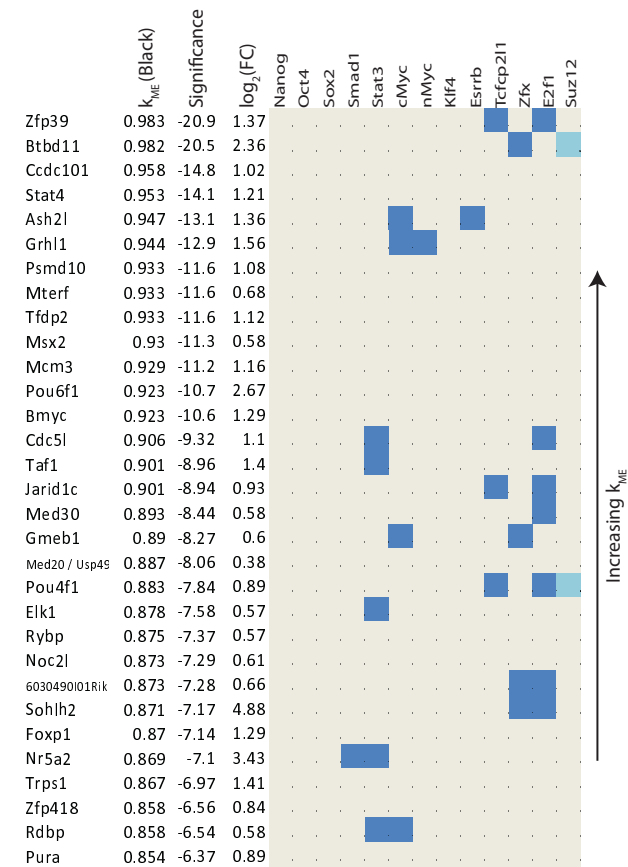


Figure 9
Pluripotency Transcriptional Regulators that are not Bound by the Core TF Machinery in ES Cells. Genes related to ES cell pluripotency and self-renewal with GO terms of transcriptional regulation or chromatin structure and little pluripotency TF binding. Genes are listed by black k_{ME} along with their corresponding significance level ($\log_{10}$ of the Bonferroni corrected p-value generated by a correlation test). Binding data from Chen *et al*, Boyer *et al* and Loh *et al*, are marked in blue (bound) and beige (unbound). For Suz12, light blue indicates that a gene is called bound in Chen *et al* or Boyer *et al* but not in both.

module membership and for identifying genes related to pluripotency and differentiation. As a resource, we provide a module membership annotation for each gene with regard to the signed modules (Additional Files 11 and 12).

Many current studies focus on the role transcriptional regulators play in ES cell maintenance. As expected, the pluripotency module is enriched with genes active in transcriptional regulation, e.g. Oct4, Sox2, Klf2, Nanog, Jarid1b, Jarid2, Nodal, Tgif1, and Esrrb, and contains other genes expected to play a role in ES cell function, such as Dppa4 and Dppa5. The module also contains genes that have recently been shown to be necessary for

maintaining the pluripotent state, Nup133 and Utf1 [45,62].

Interestingly, the pluripotency module contains genes with roles in two other pathways, DNA repair and mitochondrial function, which are not found by standard differential analysis. The enrichment for genes that respond to DNA damage is not surprising given that ES cells spend a larger portion of their cell cycle in S phase and have a shorter G1 phase than differentiated cells [73]. An emphasis on accurate DNA replication is expected since it helps ES cells maintain a stable genome and prevents errors from being inherited by differentiated cells. Mitochondria in ES cells may assist in the prevention of DNA damage [71]. During aerobic production of adenosine triphosphate (ATP), mitochondria leak superoxides leading to the creation of reactive oxygen species (ROS), which damage DNA. ES cells, however, produce ATP anaerobically and thus minimize the amount of DNA damaging ROS [69,71]. ES cells also have fewer mitochondria than differentiated cells and their mitochondria are smaller, have fewer cristae, lack dense matrices, and are perinuclearly located [69-71]. Our use of signed WGCNA reveals that in addition to genes involved in transcriptional regulation, genes that prevent or repair DNA damage are key to maintaining pluripotency and self-renewal.

Figure 3 reports significant relationships between module membership, chromatin structure and epigenetic modifications (histone modifications and DNA methylation), which are known to play a role in controlling gene expression during ES cell self-renewal and differentiation. While the relationships are highly significant, we find that epigenetic variables and binding data explain only 8.3% of the variation in module membership k_{ME}^{black} and 4.3% of the variation of k_{ME}^{blue} (Table 2). In Additional File 5, we provide gene annotations with regard to module membership, transcription factor bindings, histone trimethylation status, CpG DNA methylation etc.

Using module eigengene based connectivity k_{ME}^{blue} we find that many known differentiation related genes are highly connected in the differentiation (blue) module, Cited2, Gata4, and Gata6, along with Ctstl, which has recently been shown to be active in differentiation [65]. We also find that Uqcrh, a gene involved in the electron transport chain, is highly connected in this module, lending support to the argument that ES cell mitochondria differ from those in differentiated cells. Module eigengene based connectivity enabled us to identify novel candidate genes in the differentiation module, like Uqcrh, that warrant

Table 5: Motif Enrichment in Genes bound by Oct4 or cMyc TF Groups

Motif	Binding TF	Oct4 Group			cMyc Group		
		No. of genes	ratio	p-value	No. of genes	ratio	p-value
SoxOct	(Oct4, Sox2)	66	3.41	2.0E-21	103	0.45	6.4E-24
Oct4	(Oct4)	37	2.33	9.9E-08	151	0.81	2.5E-03
Sox2	(Sox2)	51	3.06	3.8E-15	178	0.86	1.4E-02
Nanog	(Nanog)	18	2.20	5.0E-04	122	1.35	3.9E-04
Stat1	(Stat3)	34	2.83	7.3E-09	145	1.32	1.0E-04
Ebox	(cMyc)	6	0.59	9.5E-01	373	4.02	2.6E-207
E2f1	(E2f1)	1	1.44	9.4E-02	284	2.25	1.2E-40
Klf4	(Klf4)	38	1.02	4.0E-01	235	0.61	1.4E-19
Lrh1	(Lrh1)	28	2.22	1.4E-05	158	1.12	6.9E-02
Elk1	(Elk1)	10	0.19	3.1E-02	170	1.64	8.6E-11

Sequences co-bound by all TFs of the Oct4 group (Oct4, Sox2, Nanog, Smad1, and Stat3) or all TFs of the cMyc group (cMyc, nMyc, E2f1, and Zfx) in the Chen *et al* data set were scanned for relevant motifs, plus the Lrh1 (Nr5a2) and Elk1 motifs. There were 122 genes bound by all TFs in the Oct4 group and 1173 genes bound by all TFs in the cMyc group. Enrichments, computed by comparing against motif scans in control sequences, and p-values are shown below.

experimental validation (Figure 8). For the pluripotency module interesting candidate genes are Msh6, Ppif, Sh3gl2, Rbpj, Elk1, Nrf1, Nup133, Mrpl15, and Zfp39 (Figures 7 and 8). These genes lack significant fold change but are highly connected and thus would not be found using standard differential analysis. Using sequence data with motif analysis we confirm the importance of two genes, Nr5a2 and Elk1, computationally.

We use gene ontology information and literature results to provide strong statistical evidence that these candidate genes are very promising and justify further biological study. Our article provides a resource in form of module based gene annotation tables that could form the starting point of future biological validation studies. Depending on their function, these candidate genes can be tested by RNAi knock down, viral infection in order to increase the efficiency of reprogramming, or, if they bind DNA, analyzing their binding sites. Our article demonstrates that signed WGCNA not only identifies many well known ES cell regulators; it also yields novel insights regarding ES cell function.

Methods

Our statistical methods are implemented in the WGCNA R software package [34]. For example, a signed network using the power $\beta = 12$ is constructed with the R command `ADJ = adjacency(datExpr, power = 12, type = "signed")`.

The Topological Overlap Matrix

The topological overlap, reflects the relative interconnectivity between genes i and j . It takes into account the relationship between the two genes and their shared connection pattern to other genes [19,35-37]. The topological overlap between two genes is defined as follows:

$$t_{ij} = \frac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}} \quad (6)$$

where a_{ij} is the above defined adjacency, $l_{ij} = \sum_{u \neq i, j} a_{iu} a_{uj}$ and $k_i = \sum_{u \neq i} a_{iu}$.

The Effect of the Co-expression Similarity Measure on the Topological Overlap Measure

The choice of co-expression similarity measure (i.e.

$s_{ij}^{\text{unsigned}} = |\text{cor}(x_i, x_j)|$ versus $s_{ij}^{\text{signed}} = \frac{\text{cor}(x_i, x_j) + 1}{2}$) has strong implications for the resulting topological overlap measure. Note that $\text{cor}(x_1, x_u) = -1$ implies $s_{1u}^{\text{unsigned}} = 1$ and $s_{1,u}^{\text{signed}} = 0$. In case of a soft threshold $\beta = 1$, we find the following corresponding weighted adjacency measures: $a_{1u}^{\text{unsigned}} = 1$ and $a_{1,u}^{\text{signed}} = 0$. In the following, consider the simple network in Additional File 1 (part a) where adjacencies $a_{1u} = a$ and $a_{u2} = a$ for $u = 3, \dots, n$. Using equation (6), t_{12} simplifies as follows:

$$t_{1,2} = \frac{\sum_{u \neq 1,2} a_{1u} a_{u2} + a_{12}}{\min\{k_1, k_2\} + 1 - a_{12}} = \frac{(n-2)a^2}{(n-2)a + 1} \approx a,$$

where the latter approximation assumes that the number of genes n is large. For the above situation (with $\beta = 1$) this implies $t_{1,2}^{\text{unsigned}} \approx a^{\text{unsigned}} = 1$ and $t_{1,2}^{\text{signed}} \approx a^{\text{signed}} = 0$. Thus, the two genes have high interconnectedness in an unsigned network but zero interconnectedness in a signed network. Additional File 1 (part b) shows a simple network where genes 1 and 2 are oppositely correlated with their neighbors. Here the choice of gene co-expression

measure results in a very different topological overlap measures, which in turn, leads to different modules.

The Module Eigengene and Module Membership

For the q -th module, we summarize its expression data, by a module eigengene, $E^{(q)}$, found by singular value decomposition of the expression data [39]:

$$X^{(q)} = U^{(q)} D^{(q)} (V^{(q)})^T, \quad (7)$$

where $X^{(q)}$ is the $n^{(q)} \times m$ matrix of standardized expression profiles of the $n^{(q)}$ genes in the module across m samples, $U^{(q)}$ is an $n^{(q)} \times m$ matrix with orthogonal columns, $D^{(q)}$ is an $m \times m$ diagonal matrix of singular values, and $V^{(q)}$ is an $m \times m$ orthogonal matrix of singular vectors. Then $E^{(q)}$ is defined by:

$$E^{(q)} = v_1^{(q)}, \quad (8)$$

where $v_1^{(q)}$ is the singular vector in $V^{(q)}$ corresponding to $|d_1^{(q)}|$, the largest absolute singular value in $D^{(q)}$. The module eigengene, $E^{(q)}$, can be used to define the module eigengene based connectivity, k_{ME} , or fuzzy module membership [25,27,33,39] via

$$k_{ME,i}^{(q)} = \text{cor}(x_i, E^{(q)}). \quad (9)$$

The Relationship between k_i and k_{ME} , when $\beta = 1$

To study the relationship between module eigengene based connectivity, k_{ME} , and intramodular connectivity, k_i , we consider a special case of a signed weighted network with $\beta = 1$ (i.e. $a_{ij} = \frac{1 + \text{cor}(x_i, x_j)}{2}$). Then the intramodular connectivity is given by

$$k_i^{(q)} = \sum_{j \neq i} a_{ij} = \sum_{j \neq i} \frac{1 + \text{cor}(x_i, x_j)}{2} = \frac{n^{(q)} - 1}{2} + \sum_{j \neq i} \frac{\text{cor}(x_i, x_j)}{2},$$

where $n^{(q)}$ is the number of genes in the q^{th} module. It has been shown that network modules are approximately factorizable (i.e. $\text{cor}(x_i, x_j) \approx \text{cor}(x_i, E^{(q)}) \times \text{cor}(x_j, E^{(q)})$) [33,39]. This approximation implies

$$k_i^{(q)} \approx \frac{n^{(q)} - 1}{2} + \text{cor}(x_i, E^{(q)}) \sum_{j \neq i} \frac{\text{cor}(x_j, E^{(q)})}{2} \approx a + b k_{KE}^{(q)} \quad (10)$$

where $E^{(q)}$ is the module eigengene of module q , $a = \frac{n^{(q)} - 1}{2}$, and $b = \frac{1}{2} \sum_j \text{cor}(x_j, E^{(q)})$. Equation (11) implies an

approximate linear relationship between $k_i^{(q)}$ and $k_{ME,i}^{(q)}$ if $\beta = 1$. Using real data, we illustrate this relationship in Additional File 2.

Motif Enrichment

Methods developed in Zhou *et al* [8] were used to scan sequences for motifs with pre-defined position specific weight matrices. Sequences were determined by extending bound ChIP-seq sites 150 bp up and downstream resulting in regions approximately 330 bp long. Any overlapping regions were then joined into larger meta-regions. A set of control sequences was scanned to determine a motif enrichment ratio. The control group was created by randomly sampling 5,000 probes from the Agilent Mouse Promoter Whole Genome ChIP-on-chip Microarray Set. These probes are distributed -5.5 kb upstream to +2.5 kb downstream of approximately 17,000 known gene transcription start sites from UCSC's version mm8 genome.

Overlapping control probes were merged into meta-regions as described above. Enrichment of a motif is defined as $\frac{M_T}{\lambda}$ where M_T is the number of observed sites

and λ is the number of expected sites, $\lambda = \frac{M_C N_T}{N_C}$, where M_C is the number of sites in the control, N_C the length of all control sequences, and N_T the length of all bound sequences. Statistical significance is determined by the Poisson distribution with λ as the mean.

Authors' contributions

MJM carried out the study, analyzed the data, and drafted the article. KP and GF provided embryonic stem cell expertise in interpreting data. QZ provided guidance in statistical analysis. SH conceived of the study and assisted in its design and data analysis. All authors helped write the article.

Additional material

Additional file 1

A Simple Illustration of How the Choice of a Similarity Measure Affects TOM. The TOM measure of interconnectedness is often used to define clusters of highly interconnected genes. Here we use very simple networks to highlight properties of the TOM measure. (a) Computing the topological overlap between genes 1 and 2 when all connection strengths between intermediate genes equal the constant a . (b) The numbers on the edges of the left network are correlations while the numbers on the edges of the networks on the right hand side equal corresponding unsigned adjacencies (upper network) and signed adjacencies (lower network). In a signed network, the topological overlap between genes 1 and 2 is very low because intermediate genes 3 and 4 have negative correlations with gene 1. In contrast, the topological overlap between genes 1 and 2 is high in an unsigned network.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-10-327-S1.pdf]

Additional file 2

Intramodular Connectivity is Highly Correlated with Module Eigengene Based Connectivity k_{ME} . For each module from the Zhou et al data, we plot intramodular connectivity (defined using a weighted network with power $\beta = 1$) versus module eigengene based connectivity k_{ME} . We find that the two connectivity measures are highly correlated. A theoretical derivation between network concepts and eigengene based analogs is presented in [32].

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-327-S2.pdf>]

Additional file 3

Binding Enrichments for Unsigned and Signed Ivanova et al (2006) Networks. This file contains enrichments and corresponding p-values for binding from Loh et al (2006), Boyer et al (2007), and Chen et al (2008).

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-327-S3.xls>]

Additional file 4

Comparing Gene Rankings to Regulators of Gene Expression. Here we relate different gene rankings to enrichment significance with regard to the following variables (a) histone H3K4 alone versus all others, (b) bivalent H3K4&H3K27 versus all others [49], (c) high CPG class versus all others (i.e. HCG versus ICG and LCG), (d) promoter CPG methylation status [50], (e) Oct 4 complex binding status, (f) cMyc complex binding status. We report results for 3 different gene rankings using the Ivanova data: the black and blue curve represent gene rankings according to

k_{ME}^{black} and k_{ME}^{blue} , respectively. The grey curve represents ranking according to a Student T-test of differential expression. Additional File 4 shows that black and blue module genes can have very different enrichment results that tend to be quite different from those of a standard analysis.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-327-S4.pdf>]

Additional file 5

Data for Cross-Referencing Module Membership to Epigenetic Regulators. In this Additional File, we merged Additional File 11 (Module Membership k_{ME} etc in the Ivanova data) with a Table (S4) from [50] that contained promoter CPG methylation and lysine trimethylation data from [49]. This Additional file reports module membership values, histone modifications, promoter CpG Status, and polycomb, Oct4 complex, cMyc complex binding etc. The table reports data regarding genes whose promoters have H3K4me3, H3K27me3, both, or neither histone mark. Further, column Class reports CPG promoter classifications (high HCP, intermediate ICP, or low LCP). The column Methylated reports which genes are known to be methylated (value 1) or unmethylated (0) [50]. For completeness, it also includes information regarding genes bound by Nanog or Oct4 in the proximal promoter (within 10 kb) or bound by Nanog and Oct4 long range (within 500 kb), or bound by Polycomb [5,6].

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-327-S5.csv>]

Additional file 6

Binding Enrichments for Unsigned and Signed Zhou et al (2007) Networks. This file contains enrichments and corresponding p-values for binding from Loh et al (2006), Boyer et al (2007), and Chen et al (2008).

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-327-S6.xls>]

Additional file 7

Ingenuity Pathway Analysis of the Pluripotency and Differentiation Modules from Zhou et al (2007).

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-327-S7.pdf>]

Additional file 8

Comparison of Overlap in Ivanova et al and Zhou et al (2007) when Ranking by t-statistic and Connectivity.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-327-S8.pdf>]

Additional file 9

Ingenuity Pathway Analysis of Genes Ranked by Connectivity and Differential Expression in Zhou et al (2007).

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-327-S9.pdf>]

Additional file 10

Understanding Signed Module Membership. Here we visualize the relative position of unsigned and signed similarity modules. We used module eigengene based connectivity, $k_{ME} = \text{cor}(x_i, E)$, to visualize a gene's module membership, focusing on genes located in the signed turquoise or black modules. For any vectors a and b with angle θ between them, their correlation can be interpreted as $\text{cor}(a, b) = \cos(\theta)$. Using this relationship we plotted genes in polar coordinates (radially) relative to the unsigned turquoise module. The figure shows the angle, θ , between the gene's expression profile and the turquoise module eigengene from the unsigned network, indicated by the solid turquoise line. Each gene's radius is defined as the absolute value of its $\log_2$ expression fold change (FC). FC is the ratio between the average expression in the control RNAi samples and the average expression in the Oct4 RNAi knock down samples. For reference the signed turquoise and black module eigengenes are indicated by dashed turquoise and black lines, respectively, and genes are colored by signed module membership. Known ES cell regulators and differentiation markers are labeled. Note that the signed module eigengenes (dashed lines) reflect the relationship within their corresponding modules while the unsigned module eigengene (solid line) reflects the relationship between the two signed modules.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-327-S10.pdf>]

Additional file 11

Module Membership and Binding Information in the Signed Ivanova et al (2006) Network. This file contains module membership, k_{ME} , and binding data from Loh et al (2006), Boyer et al (2007), and Chen et al (2008) for each gene on the microarray.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-327-S11.xls>]

Additional file 12

Module Membership and Binding Information in the Signed Zhou et al (2007) Network. This file contains module membership, k_{ME} , and binding data from Loh et al (2006), Boyer et al (2007), and Chen et al (2008) for each gene on the microarray.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-327-S12.xls>]

Acknowledgements

SH is supported IU19AI063603-01, 5P30CA016042-28, P50CA092131, and DK072206. KP is supported by DP2 OD001686-01, RNI-00564-1, RLI-00681-1, and P01 GM081621-01A1. QZ is supported by NSF DMS-0805491. We thank Mark Chin, Noah Dowell, Tom Drake, Tova Fuller, Dan Geschwind, Peter Langfelder, Jake Lusis, Paul Mischel, Mike Oldham, Bernadett Papp, Rupa Sridharan, and Jason Tchiew for useful discussions.

References

- Chien K: **Regenerative medicine and human models of human disease.** *Nature* 2008, **453**:302-305.
- Passier R, van Laake L, Mummery C: **Stem-cell-based therapy and lessons from the heart.** *Nature* 2008, **453**:322-329.
- Takahashi K, Yamanaka S: **Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors.** *Cell* 2006, **126**:663-676.
- Ivanova N, Dobrin R, Lu R, Kotenko L, Levorse J, DeCoste C, Schafer X, Lun Y, Lemischka I: **Dissecting self-renewal in stem cells with RNA interference.** *Nature* 2006, **442**:533-538.
- Loh Y, Wu Q, Chew J, Vega V, Zhang W, Chen X, Bourque G, George J, Leong B, Liu J: **The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells.** *Nature Genetics* 2006, **38**:431-440.
- Boyer L, Plath K, Zeitlinger J, Brambrink T, Medeiros L, Lee T, Levine S, Wernig M, Tajonar A, Ray M, Bell G, Otte A, Vidal M, Gifford D, Young R, Jaenisch R: **Polycomb complexes repress developmental regulators in murine embryonic stem cells.** *Nature* 2006, **441**:349-353.
- Maherali N, Sridharan R, Xie W, Utikal J, Eminli S, Arnold K, Stadtfeld M, Yachechko R, Tchiew J, Jaenisch R, Plath K, Hochedlinger K: **Directly Reprogrammed Fibroblasts Show Global Epigenetic Remodeling and Widespread Tissue Contribution.** *Cell Stem Cell* 2007, **1**:55-70.
- Zhou Q, Chipperfield H, Melton DA, Wong WH: **A gene regulatory network in mouse embryonic stem cells.** *Proc Natl Acad Sci* 2007, **104**(42):16438-16443.
- Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S: **Induction of pluripotent stem cells from adult human fibroblasts by defined factors.** *Cell* 2007, **131**(5):861-872.
- Yu J, Vodyanik M, Smuga-Otto K, Antosiewicz-Bourget J, Frane J, Tian S, Nie J, Jonsdottir G, Ruotti V, Stewart R, Slukvin I, Thomson J: **Induced pluripotent stem cell lines derived from human somatic cells.** *Science* 2007, **318**(5858):1917-1920.
- Nakagawa M, Koyanagi M, Tanabe K, Takahashi K, Ichisaka T, Aoi T, Okita K, Mochizuki Y, Takizawa N, Yamanaka S: **Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts.** *Nature Biotechnology* 2008, **26**:101-106.
- Viswanathan S, Daley G, Gregory R: **Selective blockade of micro-RNA processing by Lin28.** *Science* 2008, **320**(5872):58-59.
- Kim J, Chu J, Shen X, Wang J, Orkin S: **An extended transcriptional network for pluripotency of embryonic stem cells.** *Cell* 2008, **132**(6):1049-1061.
- Park I, Zhao R, West J, Yabuuchi A, Huo H, Ince T, Lerou P, Lensch M, Daley G: **Reprogramming of human somatic cells to pluripotency with defined factors.** *Nature* 2008, **451**(7175):141-146.
- Lowry W, Richter L, Yachechko R, Pyle A, Tchiew J, Sridharan R, Clark A, Plath K: **Generation of human induced pluripotent stem cells from dermal fibroblasts.** *Proc Natl Acad Sci USA* 2008, **105**(8):2883-2888.
- Chen X, Xu H, Yuan P, Fang F, Huss M, Vega V, Wong E, Orlov Y, Zhang W, Jiang J, Loh Y, Yeo H, Yeo Z, Narang V, Govindarajan K, Leong B, Shahab A, Ruan Y, Bourque G, Sung W, Clarke N, Wei C, Ng H: **Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells.** *Cell* 2008, **133**:1106-1117.
- Mitsui K, Tokuzawa Y, Itoh H, Segawa K, Murakami M, Takahashi K, Maruyama M, Maeda M, Yamanaka S: **The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells.** *Cell* 2003, **113**(5):631-642.
- Stuart JM, Segal E, Koller D, Kim SK: **A gene-coexpression network for global discovery of conserved genetic modules.** *Science* 2003, **302**(5643):249-255.
- Zhang B, Horvath S: **A General Framework for Weighted Gene Co-Expression Network Analysis.** *Stat Appl Genet Mol Biol* 2005, **4**:Article17.
- Walker E, Ohishi M, Davey R, Zhang W, Cassar P, Tanaka T, Der S, Morris Q, Hughes T, Zandstra P, Stanford W: **Prediction and Testing of Novel Transcriptional Networks Regulating Embryonic Stem Cell Self-Renewal and Commitment.** *Cell Stem Cell* 2007, **1**:71-86.
- Huang Y, Li H, Hu H, Yan X, Waterman M, Huang H, Zhou X: **Systematic discovery of functional modules and context-specific functional annotation of human genome.** *Bioinformatics* 2007, **23**(13):222-229.
- Chen C, Weirauch M, Powell C, Zamboni A, Stuart J: **A search engine to identify pathway genes from expression data on multiple organisms.** *BMC Systems Biology* 2007, **1**:20.
- Horvath S, Zhang B, Carlson M, Lu K, Zhu S, Felciano R, Laurance M, Zhao W, Shu Q, Lee Y, Scheek A, Liao L, Wu H, Geschwind D, Febo P, Kornblum H, TF C, Nelson S, Mischel P: **Analysis of Oncogenic Signaling Networks in Glioblastoma Identifies ASPM as a Novel Molecular Target.** *PNAS* 2006, **103**(46):17402-17407.
- Wei H, Persson S, Mehta T, Srinivasasainagendra V, Chen L, Page G, Somerville C, Loraine A: **Transcriptional Coordination of the Metabolic Network in Arabidopsis.** *Plant Physiol* 2006, **142**(2):762-774.
- Ghazalpour A, Doss S, Zhang B, Plaisier C, Wang S, Schadt E, Thomas A, Drake T, Lusis A, Horvath S: **Integrating Genetics and Network Analysis to Characterize Genes Related to Mouse Weight.** *PLoS Genet* 2006, **2**(8):e130.
- Carlson M, Zhang B, Fang Z, Mischel P, Horvath S, Nelson SF: **Gene Connectivity, Function, and Sequence Conservation: Predictions from Modular Yeast Co-expression Networks.** *BMC Genomics* 2006, **7**(40):.
- Oldham M, Horvath S, Geschwind D: **Conservation and evolution of gene co-expression networks in human and chimpanzee brain.** *Proc Natl Acad Sci U S A* 2006, **103**(47):17973-17978 [<http://www.pnas.org/content/103/47/17973.full>].
- Keller MP, Choi Y, Wang P, Belt Davis D, Rabaglia ME, Oler AT, Stapleton DS, Argmann C, Schueler KL, Edwards S, Steinberg HA, Chai-bub Neto E, Kleinhanz R, Turner S, Hellerstein MK, Schadt EE, Yandell BS, Kendzierski C, Attie AD: **A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility.** *Genome Res* 2008, **18**(5):706-716.
- Weston D, Gunter L, Rogers A, Wulfschleger S: **Connecting genes, coexpression modules, and molecular signatures to environmental stress phenotypes in plants.** *BMC Systems Biology* 2008, **2**:16 [<http://www.biomedcentral.com/1752-0509/2/16>].
- Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, Geschwind DH: **Functional organization of the transcriptome in human brain.** *Nature Neuroscience* 2008, **11**(11):1271-1282.

31. Shieh G, Chen CM, Yu CY, Huang J, Wang WF, Lo YC: **Inferring transcriptional compensation interactions in yeast via stepwise structure equation modeling.** *BMC Bioinformatics* 2008, **9**:134 [<http://www.biomedcentral.com/1471-2105/9/134>].
32. Presson A, Sobel E, Papp J, Suarez C, Whistler T, Rajeevan M, Vernon S, Horvath S: **Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome.** *BMC Systems Biology* 2008, **2**(95):.
33. Horvath S, Dong J: **Geometric Interpretation of Gene Coexpression Network Analysis.** *PLoS Computational Biology* 2008, **4**(8):.
34. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis.** *BMC Bioinformatics* 2008, **9**:559.
35. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297**:1551-1555.
36. Li A, Horvath S: **Network Neighborhood Analysis with the multi-node topological overlap measure.** *Bioinformatics* 2006, **23**(2):222-231.
37. Yip A, Horvath S: **Gene network interconnectedness and the generalized topological overlap measure.** *BMC Bioinformatics* 2007, **8**(22):.
38. Langfelder P, Zhang B, Horvath S: **Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut library for R.** *BMC Syst Biol* 2007:1-24 [<http://bioinformatics.oxfordjournals.org/cgi/reprint/btm563v1>].
39. Dong J, Horvath S: **Understanding network concepts in modules.** *BMC Systems Biology* 2007, **1**:24.
40. Fuller T, Ghazalpour A, Aten J, Drake T, Lusis A, Horvath S: **Weighted gene coexpression network analysis applied to mouse weight.** *Mammalian Genome* 2007, **18**(6):463-472.
41. Langfelder P, Horvath S: **Eigengene networks for studying the relationships between co-expression modules.** *BMC Systems Biology* 2007, **1**:54.
42. Velkey J, O'Shea K: **Oct4 RNA interference induces trophectoderm differentiation in mouse embryonic stem cells.** *Genesis* 2003, **37**(1):18-24.
43. Nishimoto M, Miyagi S, Yamagishi T, Sakaguchi T, Niwa H, Muramatsu M, Okuda A: **Oct-3/4 maintains the proliferative embryonic stem cell state via specific binding to a variant octamer sequence in the regulatory region of the UTF1 locus.** *Molecular Cell Biology* 2005, **25**(12):5084-5094.
44. Boom V van den, Kooistra SM, Boesjes M, Geverts B, Houtsmuller A, Monzen K, Komuro I, Essers J, Drenth-Diephuis L, Eggen BJ: **UTF1 is a chromatin-associated protein involved in ES cell differentiation.** *Journal of Cell Biology* 2007, **10**(178):913-924.
45. Zhao Y, Yin X, Qin H, Zhu F, Liu H, Yang W, Zhang Q, Xiang C, Hou P, Song Z, Liu Y, Yong J, Zhang P, Cai J, Liu M, Li H, Li Y, Qu X, Cui K, Zhang W, Xiang T, Wu Y, Zhao Y, Liu C, Yu C, Yuan K, Lou J, Ding M, Deng H: **Two Supporting Factors Greatly Improve the Efficiency of Human iPSC Generation.** *Cell Stem Cell* 2008, **35**:475-479.
46. Keller G: **Embryonic stem cell differentiation: emergence of a new era in biology and medicine.** *Genes Development* 2005, **19**(10):1129-1155.
47. Bernstein B, Mikkelsen T, Xie X, Kamal M, Huebert D, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber S, Lander E: **A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells.** *Cell* 2006, **125**(2):315-326.
48. Bernstein B, Meissner A, Lander E: **The Mammalian Epigenome.** *Cell* 2007, **128**(4):669-681.
49. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA: **A chromatin landmark and transcription initiation at most promoters in human cells.** *Cell* 2007, **130**:77-88.
50. Mikkelsen T, Ku M, Jaffe D, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim T, Koche R, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander E, Bernstein B: **Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.** *Nature* 2007, **448**(7153):553-60.
51. Fouse S, Shen Y, Pellegrini M, Cole S, Meissner A, VanNeste L, Jaenisch R, Fan G: **Promoter CpG Methylation Contributes to ES Cell Gene Regulation in Parallel with Oct4/Nanog, PcG Complex, and Histone H3 K4/K27 Trimethylation.** *Cell Stem Cell* 2008, **2**(2):160-169.
52. Lee T, Jenner R, Boyer L, Guenther M, Levine S, Kumar R, Chevalier B, Johnstone S, Cole M, Isono K, Koseki H, Fuchikami T, Abe K, Murray H, Zucker J, Yuan B, Bell G, Herbolzheimer E, Hannett N, Sun K, Odom D, Otte A, Volkert T, Bartel D, Melton D, Gifford D, Jaenisch R, Young R: **Control of developmental regulators by Polycomb in human embryonic stem cells.** *Cell* 2006, **125**(2):301-13.
53. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** *Genome Biol* 2003, **4**(9):R60.
54. Müller F, Laurent L, Kostka D, Ulitsky I, Williams R, Lu C, Park I, Rao M, Shamir R, Schwartz P, Schmidt N, Loring J: **Regulatory networks define phenotypic classes of human cell lines.** *Nature* 2008, **455**(18):401-406.
55. Sharov A, Piao Y, Matoba R, Dudekula D, Qian Y, VanBuren V, Falco G, Martin P, Stagg C, Bassey U, Wang Y, Carter M, Hamatani T, Aiba K, Akutsu H, Sharova L, Tanaka T, Kimber W, Yoshikawa T, Jaradat S, Pantano S, Nagaraja R, Boheler K, Taub D, Hodes R, Longo D, Schlessinger D, Keller J, Klotz E, Kelsoe G, Umezawa A, Vescovi A, Ros-sant J, Kunath T, Hogan B, Curci A, D'Urso M, Kelso J, Hide W, Ko M: **Transcriptome analysis of mouse stem cells and early embryos.** *PLoS Biology* 2003, **1**(3):.
56. Jiang J, Chan Y, Loh Y, Cai J, Tong G, Lim C, Robson P, Zhong S, Ng H: **A core Klf circuitry regulates self-renewal of embryonic stem cells.** *Nature Cell Biology* 2008, **10**(3):353-360.
57. Young L, Keuling A, Lai R, Nation P, Tron V, Andrew S: **The associated contributions of p53 and the DNA mismatch repair protein Msh6 to spontaneous tumorigenesis.** *Carcinogenesis* 2007, **28**:2131-2138.
58. Hori K, Cholewa-Waclaw J, Nakada Y, Glasgow S, Masui T, Henke R, Wildner H, Martarelli B, Beres T, Epstein J, Magnuson M, MacDonald R, Birchmeier C, Johnson J: **A nonclassical bHLH-Rbpj transcription factor complex is required for specification of GABAergic neurons independent of Notch signaling.** *Genes and Development* 2008, **22**:166-178.
59. Nocea T, Fujiwara Y, Sezakia M, Fujimoto H, Higashinakagawa T: **Expression of a mouse zinc finger protein gene in both spermatocytes and oocytes during meiosis.** *Developmental Biology* 1992, **153**(2):356-367.
60. O'Hara M, Nibbio B, Craig R, Nemeth K, Charlap J, Knudsen T: **Mitochondrial benzodiazepine receptors regulate oxygen homeostasis in the early mouse embryo.** *Reproductive Toxicology* 2003, **17**(4):365-375.
61. Kendall S, Battelli C, Irwin S, Mitchell J, Glackin C, Verdi J: **NRAGE mediates p38 activation and neural progenitor apoptosis via the bone morphogenetic protein signaling cascade.** *Molecular and Cellular Biology* 2005, **25**(17):7711-7724.
62. Lupu F, Alves A, Anderson K, Doye V, Lacy E: **Nuclear pore composition regulates neural stem/progenitor cell differentiation in the mouse embryo.** *Developmental Cell* 2008, **14**:831-842.
63. Howard L, Nelson K, Maciewicz R, Blobel C: **Interaction of the metalloprotease disintegrins MDC9 and MDC15 with two SH3 domain-containing proteins, endophilin I and SH3PX1.** *Journal of Biological Chemistry* 1999, **274**(44):31693-31699.
64. Baines C, Kaiser R, Purcell N, Blair N, Osinska H, Hambleton M, Brunskill E, Sayen M, Gottlieb R, Dorn G, Robbins J, Molkentin J: **Loss of cyclophilin D reveals a critical role for mitochondrial permeability transition in cell death.** *Nature* 2005, **434**(7033):658-662.
65. Duncan E, Muratore-Schroeder T, Cook R, Garcia B, Shabanowitz J, Hunt DF, Allis C: **Cathepsin L Proteolytically Processes Histone H3 During Mouse Embryonic Stem Cell Differentiation.** *Cell* 2008, **135**(2):284-294.
66. Sakamoto A, Chen M, Nakamura T, Xie T, Karsenty G, Weinstein L: **Deficiency of the G-protein alpha-subunit G(s)alpha in osteoblasts leads to differential effects on trabecular and cortical bone.** *Journal of Biological Chemistry* 2005, **280**(22):21369-21375.
67. Shimo T, Kanyama M, Wu C, Sugito H, Billings P, Abrams W, Rosenbloom J, Iwamoto M, Pacifici M, Koyama E: **Expression and roles of connective tissue growth factor in Meckel's cartilage development.** *Developmental Dynamics* 2004, **231**:136-147.
68. Da Cruz S, Xenarios I, Langridge J, Vilbois F, Parone P, Martinou J: **Proteomic analysis of the mouse liver mitochondrial inner membrane.** *Journal of Biological Chemistry* 2003, **278**(42):41566-41471.

69. Cho Y, Kwon S, Pak Y, Seol H, Choi Y, Park D, Park K, Lee HK: **Dynamic changes in mitochondrial biogenesis and antioxidant enzymes during the spontaneous differentiation of human embryonic stem cells.** *Biochemical and Biophysical Research Communications* 2006, **348**:1472-1478.
70. Lonergan T, Bavister B, Brenner C: **Mitochondria in stem cells.** *Mitochondrion* 2007, **7**:289-296.
71. Saretzki G, Walter T, Atkinson S, Passos JF, Bareth B, Keith W, Stewart R, Hoare S, Stojkovic M, Armstrong L, von Zglinicki T, Lako M: **Downregulation of Multiple Stress Defense Mechanisms During Differentiation of Human Embryonic Stem Cells.** *Stem Cells* 2008, **26**:455-464.
72. Gu P, Goodwin B, Chung A, Xu X, Wheeler D, Price R, Galardi C, Peng L, Latour A, Koller B, Gossen J, Kliewer S, Cooney A: **Orphan Nuclear Receptor LRH-1 Is Required To Maintain Oct4 Expression at the Epiblast Stage of Embryonic Development.** *Molecular Cell Biology* 2005, **25**:3492-3505.
73. White J, Dalton S: **Cell Cycle Control of Embryonic Stem Cells.** *Stem Cell Reviews* 2005, **1**:131-138.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

