Mapping Dynamic Histone Acetylation Patterns to Gene Expression in Nanog-Depleted Murine Embryonic Stem Cells

PLoS COMPUTATIONAL BIOLOGY

Mapping Dynamic Histone Acetylation Patterns to Gene Expression in Nanog-Depleted Murine Embryonic Stem Cells

Florian Markowetz[1,2]*, Klaas W. Mulder[1], Edoardo M. Airoldi[3], Ihor R. Lemischka[4], Olga G. Troyanskaya[5]*

[1] Cancer Research UK Cambridge Research Institute, Cambridge, United Kingdom, [2] Department of Oncology, University of Cambridge, Cambridge, United Kingdom, [3] Department of Statistics and FAS Center for Systems Biology, Harvard University, Cambridge, Massachusetts, United States of America, [4] Department of Gene and Cell Medicine and The Black Family Stem Cell Institute, Mount Sinai School of Medicine, New York, New York, United States of America, [5] Lewis-Sigler Institute for Integrative Genomics and Department of Computer Science, Princeton University, Princeton, New Jersey, United States of America

Abstract

Embryonic stem cells (ESC) have the potential to self-renew indefinitely and to differentiate into any of the three germ layers. The molecular mechanisms for self-renewal, maintenance of pluripotency and lineage specification are poorly understood, but recent results point to a key role for epigenetic mechanisms. In this study, we focus on quantifying the impact of histone 3 acetylation (H3K9,14ac) on gene expression in murine embryonic stem cells. We analyze genome-wide histone acetylation patterns and gene expression profiles measured over the first five days of cell differentiation triggered by silencing Nanog, a key transcription factor in ESC regulation. We explore the temporal and spatial dynamics of histone acetylation data and its correlation with gene expression using supervised and unsupervised statistical models. On a genome-wide scale, changes in acetylation are significantly correlated to changes in mRNA expression and, surprisingly, this coherence increases over time. We quantify the predictive power of histone acetylation for gene expression changes in a balanced cross-validation procedure. In an in-depth study we focus on genes central to the regulatory network of Mouse ESC, including those identified in a recent genome-wide RNAi screen and in the PluriNet, a computationally derived stem cell signature. We find that compared to the rest of the genome, ESC-specific genes show significantly more acetylation signal and a much stronger decrease in acetylation over time, which is often not reflected in a concordant expression change. These results shed light on the complexity of the relationship between histone acetylation and gene expression and are a step toward dissecting the multilayer regulatory mechanisms that determine stem cell fate.

Citation: Markowetz F, Mulder KW, Airoldi EM, Lemischka IR, Troyanskaya OG (2010) Mapping Dynamic Histone Acetylation Patterns to Gene Expression in Nanog-Depleted Murine Embryonic Stem Cells. PLoS Comput Biol 6(12): e1001034. doi:10.1371/journal.pcbi.1001034

Editor: Ilya Ioshikhes, Ottawa University, Canada

Received May 12, 2010; **Accepted** November 12, 2010; **Published** December 16, 2010

Funding: FM is fully supported by Cancer Research UK. KWM was supported by a Marie Curie Fellowship (PIEF-GA-2008-220642). EMA was partially supported by the National Science Foundation under grants no. DMS-0907009 and no. IIS-1017967, and by the National Institute of Health under grant no. R01 GM096193. This research was partly supported by NIH grant R01 GM071966, NSF grant IIS-0513552 and NSF CAREER award DBI-0546275 to OGT and by NIGMS Center of Excellence grant P50 GM071508. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: florian.markowetz@cancer.org.uk (FM); ogt@cs.princeton.edu (OGT)

Introduction

Embryonic stem cells (ESC) are pluripotent cells that have the potential to self-renew indefinitely and to differentiate into any of the three germ layers. Molecular regulation of embryonic stem cell fate is implemented by a coordinated interaction between epigenetic [1–5], transcriptional [6–11] and translational [12,13] mechanisms.

The molecular mechanisms for self-renewal, maintenance of pluripotency and lineage specification are poorly understood [14], but recent results point to key roles for a network of transcription factors [9,15,16] and a wide range of epigenetic mechanisms [2,17–19]. For example, recent work showed the importance of chromatin remodeling factors like polycomb proteins [20,21] and the SWI/SNF complex [22] for ES cell regulation. ES cells are richer in less compact euchromatin and, as differentiation progresses, accumulate highly condensed, transcriptionaly inactive heterochromatin regions [23]. Major architectural chromatin proteins are hyper-dynamic and bind loosely to chromatin in ES cells. Upon differentiation, the hyperdynamic proteins become immobilized on chromatin [24]. Bivalent domains – consisting of large regions of H3 lysine 27 methylation harboring smaller regions of H3 lysine 4 methylation– silence developmental genes in ES cells while keeping them poised for action [1,3].

Multi-layered time-course data in Nanog-depleted mouse ESC

The number of data sets in ESC linking epigenetic mechanisms to other molecular regulatory mechanisms and following that relationship over time is very limited. Recently, however, Lu and coworkers [25] presented a dynamic systems-level study to assess how different molecular regulatory mechanisms interact in stem cell fate decisions in mouse ESC. Lu *et al* initiated cell differentiation by experimentally down-regulating Nanog, a key pluripotency regulator. Over the following five days they measured changes on four different molecular levels: histone acetylation (H3K9,14ac), chromatin-bound RNA polymerase II, messenger RNA (mRNA) expression and nuclear protein abundance. This

Author Summary

Stem cell differentiation and the maintenance of self-renewal are intrinsically complex processes that require coordinated regulation on many different cellular levels. Here we focus on the relationship between two important layers and follow it over the first five days of differentiation. The first layer – measured by acetylation of one of the histone proteins – describes which parts of the DNA are tightly wrapped up and which lie open. The second layer describes the activity of genes measured by their mRNA expression. Using a wide array of statistical approaches we show that changes in histone acetylation are very predictive for gene expression and that the concordance between the two levels increases over time. Concentrating on genes central to the regulatory networks in embryonic stem cells we find that key genes show very high acetylation signal in the beginning that decreases quickly over time, indicating that they lie in initially open regions that are rapidly closing down. These results are a step forward to a better understanding of the complexities of the relationship between histone acetylation and gene expression, which will help to dissect the multilayer regulatory mechanisms that determine stem cell fate.

data set provides a rich resource to untangle the complexity of the multi-layer regulatory mechanism responsible for stem cell fate. Lu *et al* anchored their analyses on changes in nuclear protein expression and found that many lacked concordant changes in mRNA expression, pointing to important roles for translational and post-translational regulation of ESC fate. Here, we complement theses analyses with an in-depth study of the relation between histone acetylation and gene expression in the same data set.

Histone acetylation and gene expression

The acetylation of lysine residues is among the best characterized histone modifications. It has long been correlated with transcriptional activation [26,27]. This observation has been verified in many recent high-throughput studies [28–30]. For example, histone acetylation was found to be positively correlated with expression in yeast [31,32] and human T cells [33,34]. The last study also suggests that acetylation sites often cluster together in so called 'acetylation islands' [34].

Several models have been suggested to explain how histone acetylation and other modifications regulate gene expression [35], including charge neutralization [36] and a signalling pathway model [37]. However, the detailed mechanism is still poorly understood. This problem is highlighted by two recent studies, one experimental and one statistical. Günther *et al.* [38] stress the importance of additional regulatory events by showing that acetylated and methylated nucleosomes, as well as RNA polymerase II, occupy the promoters of most protein-coding genes in human ES cells, even those that are not expressed. Yuan *et al.* [39] assessed the global regulatory role of histone acetylation in *Saccharomyces cerevisiae* by controlling for confounding effects like transcription factor binding sites and nucleosome occupancy. They find a clear effect of histone 3 acetylation, but no significant direct impact of histone 4 acetylation or combinatorial effects, even though they correlate with expression.

These results indicate that further experimental results and statistical analyses are required to untangle the regulatory role of histone acetylation and the mechanism by which it acts. The need for a better understanding of histone acetylation is especially urgent in ES cells, where many key regulatory mechanisms are epigenetic and act by chromatin modifications and remodeling. For example, embryonic stem cells in which histone de-acetylation is inhibited, undergo morphological and gene expression changes indicative of differentiation [40].

Overview of results

In the following, we first start by analyzing the internal structure of the histone acetylation profiles and their change during differentiation. We investigate the dynamics of acetylation over time and find that the location of acetylation islands remain stable. We find that differentially down-regulated genes are accompanied by a much stronger loss of acetylation than up-regulated genes are by a gain of acetylation. In a next step we assess the dynamics of the correlation between mean acetylation levels and expression and find that coordination increases over time. Using statistical classification methods we then quantify the predictive power of acetylation profiles for gene expression changes. Finally, we focus on genes playing key roles in the regulatory networks governing fate decisions in embryonic stem cells. We show that these genes show highly increased acetylation profiles. Over time the high levels of acetylation get reduced more strongly than in other, not ESC specific genes. This behaviour is far less pronounced in the gene expression data, pointing to a key role in non-transcriptional regulation of pluripotency for important ESC genes.

Results

Our central questions are how changes in gene expression are reflected in histone acetylation, how predictive histone acetylation is for gene expression changes, and how this relationship changes over time. To answer them, in the following we employ different statistical approaches to describe the internal structure of histone acetylation profiles and to map them to changes in gene expression.

Histone acetylation changes in differentially expressed genes

Location of acetylation islands is stable over time. As examples of the data we work with, Figure 1A shows acetylation profiles of Pou5f1/Oct4 and Klf4. The plots show acetylation levels at four time points: before silencing Nanog (day 0) and at days 1, 3, and 5 afterwards. The plots show large internal variation of acetylation signal for each gene. As a first preparatory step in our analysis we investigated if there is evidence that the location of acetylation signal changes over time. If the signal location does not change, then only the quantitative level of acetylation are important when mapping it to gene expression in the next steps of our study.

We identified acetylation islands [34] by comparing probe signal to background distribution of control probes on the array (see *Materials and Methods*). Figure 1A depicts the background distribution as a grey area, all probes above it are counted as 'acetylated', all probes inside as 'unacetylated'. (This is a slight abuse of terminology since technically it is not the probe that is acetylated but the histone protein bound to a piece of DNA complementary to the probe.) With these results, we investigated dynamical changes on the probe level and asked for each gene: Are the same probes acetylated over time, or does the position of acetylation signal change over time? To answer this question, we represented each gene by two numbers: the percentage of probes staying un-acetylated and the percentage of probes staying acetylated between time-points. Figure 1B shows that the distribution of these values is concentrated in the upper right

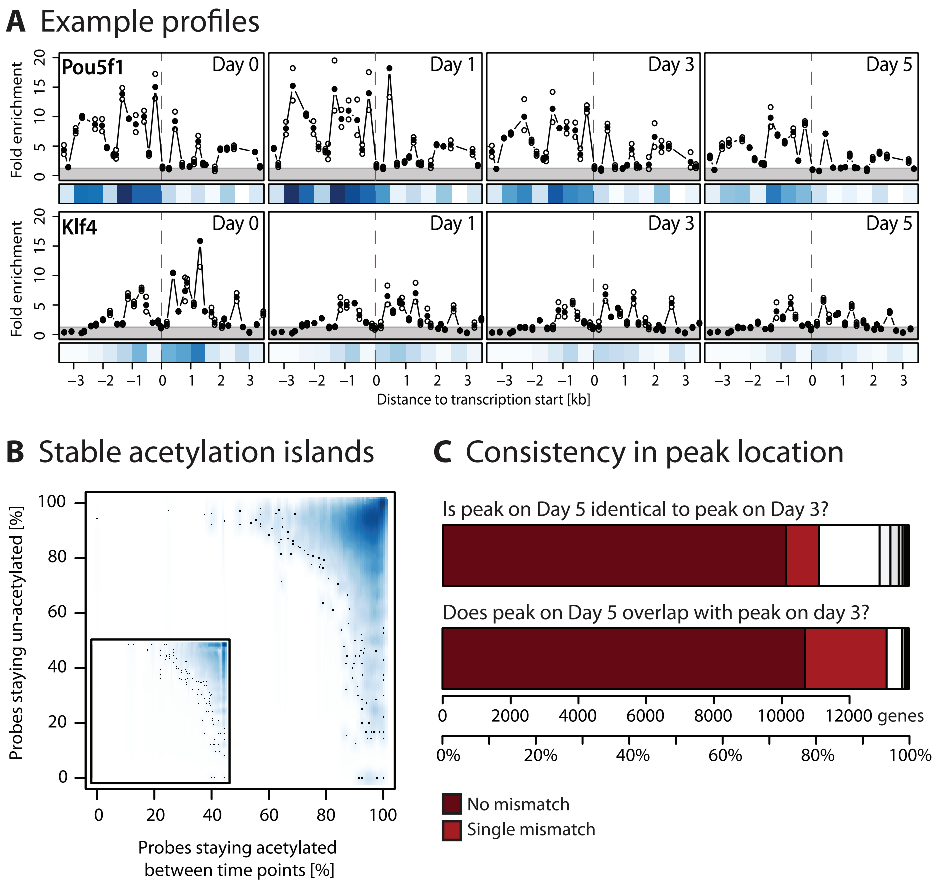

Figure 1. Acetylation profiles over time. A Histone acetylation profiles of Pou5f1/Oct4 and Klf4 before Nanog-knockdown (day 0) and on days 1, 3, and 5 afterwards. All plots are centered at the transcription start site (TSS; red dashed line). The gray area shows the background signal, circles indicate replicate measurements, dots averages. The blue heatmap underneath each plot shows quantitative data averaged over 5 kb intervals to make it comparable between genes with different numbers and positions of probes. **B** To test for evidence of location changes, we counted probes as 'acetylated' if they were above the noise (gray area in panel A). The smoothed scatterplot shows for each gene the percentages of probes staying acetylated (x-axis) or un-acetylated (y-axis) over time. The mass of the distribution lies in the upper right corner indicating high stability of acetylation islands. This is independent of particular gene sets of days as the inlay exemplifies by plotting only the changes between day 3 and 5 for genes differential on day 5. **C** We defined a peak in the acetylation profile as the smallest region covering 30% of the total signal. Peaks stay very stable over time. The plot shows that for example between days 3 and 5 ca. 70% of peaks are at exactly the same position and for almost 80% of peaks the location on day 5 overlapped the location on day 3 completely. If we allow one mismatch between peak locations the numbers go up to 80% and 95% respectively.

doi:10.1371/journal.pcbi.1001034.g001

corner of the plot which corresponds to perfect conservation of acetylation location over time. Probes that are acetylated at any time-point stay acetylated and un-acetylated probes stay un-acetylated. In a second step we investigated if regions of peak signal in the binned profiles change over time. We defined a peak as those bins that include the maximum of the profile and together carry $\geq 30\%$ of the signal. Figure 1C shows that not only acetylated probes but also peaks stay stable over time.

Thus, in summary, in our data we find no evidence that the location of acetylation signal changes over time. This simple analysis plays only a preparatory role in our study: it allows us to focus on quantitative changes in signal intensity in the next steps of our analysis. The data we work with from now is exemplified by the blue heatmaps underneath the profiles in Figure 1A. For each gene, our data captures the quantitative acetylation signal in a region of ± 3.5kb around the transcription start site (TSS).

Loss of acetylation is more pronounced than gain of acetylation. We find clear correlations between histone acetylation and gene expression. For example, Figure 2A shows all genes differentially expressed on day 5. In this plot, genes transcriptionally up-regulated also show increased levels of acetylation, while down-regulated genes show a decrease.

However, these plots also indicate that the loss of acetylation for down-regulated genes is much stronger pronounced than the gain of acetylation for up-regulated genes. This is particularly visible in Figure 2B, which plots the acetylation distributions separately for up-regulated, down-regulated and stable genes. The down-regulated genes show a very strong loss over the whole width of the profile, while the up-regulated genes show a much weaker signal and only close to the TSS. Genes without significant expression changes show a strong bias towards loss of acetylation, but the size of the effect is much smaller than in the down-regulated genes.

Partial correlation analysis resolves spatial and temporal dependencies in acetylation profiles. We were interested in the internal correlation structure of the histone acetylation profiles and used partial correlation analysis (see *Materials and Methods*) to measure the direct relations between regions around TSS (i.e. the bins in the profile). Figure 3A shows partial correlation matrices combining data from day 1, 3 and 5. We computed one matrix for

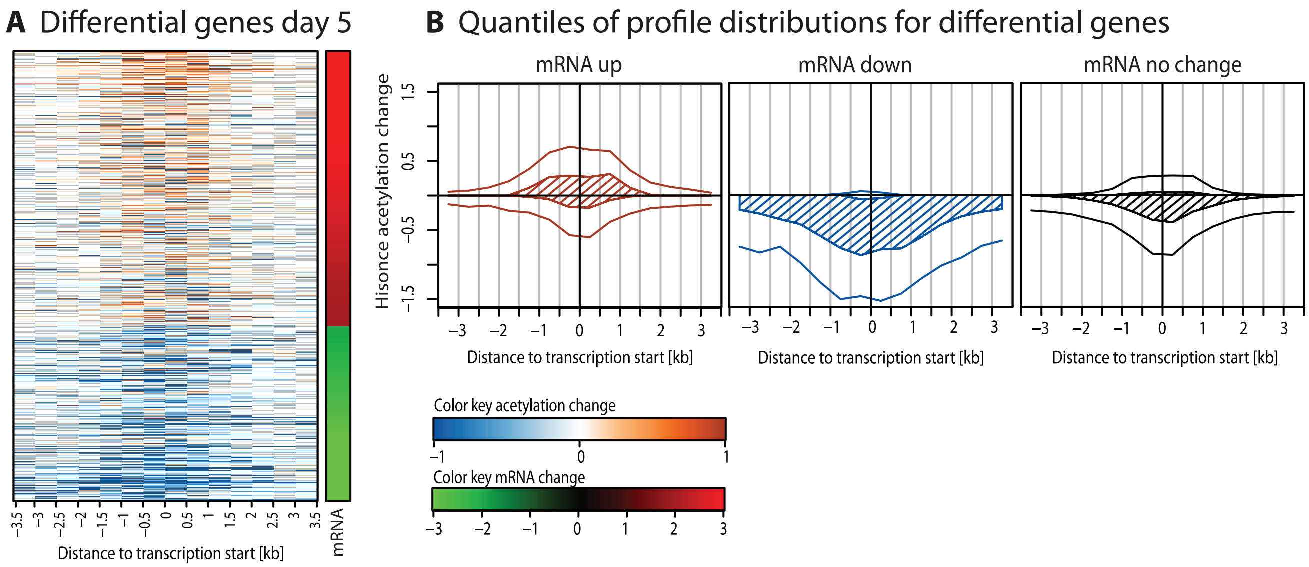

Figure 2. Acetylation profiles for differential genes. A A heatmap of acetylation changes between days 0 and 5 for all genes with significantly differential mRNA levels on day 5. Transcriptionally up-regulated genes show an increase in acetylation signal, while down-regulated genes show a decrease. **B** Visualization of the distributions of changes in acetylation signal from day 0 to day 5 for genes transcriptionally up-regulated, down-regulated or non-changing on day 5. Each plot shows four lines corresponding to the 10%, 25%, 75%, and 90% quantiles of the distributions in each bin. The hatched area emphasizes the inter-quartile range between the 25% and 75% quantile. Up-regulated genes show elevated acetylation levels close to TSS, while down-regulated genes show a broad decrease in acetylation across several kb around TSS.
doi:10.1371/journal.pcbi.1001034.g002

genes differentially expressed on day 5 and a second one for genes with stable expression. Both matrices show a strong stripe-pattern indicating high correlation for neighboring bins and for the same bin at different days. The differences between the two matrices are minimal, as can be seen in the right-most matrix of Figure 3A. Only between day 3 and 5 do the acetylation profiles in differential genes show a little bit more correlation than those in the non-differential genes.

The significant entries of the partial correlation matrix can be depicted by the graph structure shown in Figure 3B. The three layers of the graph correspond to the three days and each edge indicates a significant partial correlation coefficient. We show the graph for the non-differential genes since their larger number results in higher power. The graph shows that the spatial and temporal dependencies between variables very clearly show in the correlation structure of the data, for example almost all neighboring bins at the same time-point are connected. However, close to the TSS the graph is much less connected than in more distant regions. This possible reflects the presence of nucleosome free regions around the TSS in many active genes [41].

Coordination of histone acetylation and gene expression increases over time

We assessed the correlation between histone acetylation changes and gene expression changes versus day 0 for all pairings of days. This analysis is anchored on the ESC state (day 0) and assesses the coordination of *cumulative* changes away from it. In a first step, we summarized each acetylation profile by the mean and computed the standard Pearson correlation between the resulting acetylation vector and expression vector (Figure 4A, left matrix). The results show that changes in acetylation demonstrate significant correlation to changes in mRNA expression. Correlations with acetylation changes on day 1 are generally small (in general <0.1), but correlations between changes on days 3 and 5 show very significant values, e.g. on day 5 Pearson correlation is 0.344. Even though this value is small, the level of coherence is very surprising given the large number of genes ($>17\,000$). The correlation table shows coherence between histone acetylation and gene expression increases over time and is biggest on day 5.

Correlation results are statistically significant. We assessed the significance of observed correlations in two ways. First, we used the analytic Null-distributions known for the correlation measures we used [42]. Significance is a function of sample size and with $>17\,000$ genes we find all correlations between days 3 and 5 to be significant with p-values smaller than 10^{-100}. Correlations with day 1 (first row or column in Figure 4A left matrix) are much weaker, but still almost always significant on a level of 10^{-4}. One reason for these extremely small p-values is that the analytic Null-distributions assume independence between genes, which is an unreasonable assumption for genomic data. To correct for this bias, we used a permutation approach that keeps the correlation structure of genes intact for a second assessment of significance. We compared the correlations measured in the actual data with the distribution of 10^4 correlation values computed on permuted versions of the data. However, qualitatively the results were identical to the first approach: correlations between days 3 and 5 are very significant (no permutation yielded a correlation exceeding the value on the actual data) and correlations to day 1 are much weaker.

Correlation results are robust to gene selection and correlation measures. In the next step we assessed the robustness of the observed correlation pattern by using different types of correlation measures, different ways to average the acetylation changes and different subsets of the data (right matrix of Figure 4A). In particular, we used the Spearman rank correlation between the *median* (instead of *mean*) acetylation change and expression, as well as the correlations computed by Canonical Correlation Analysis, a statistical method to find directions of maximal correlation between datasets (see *Materials and Methods*). For each of these different ways to compute

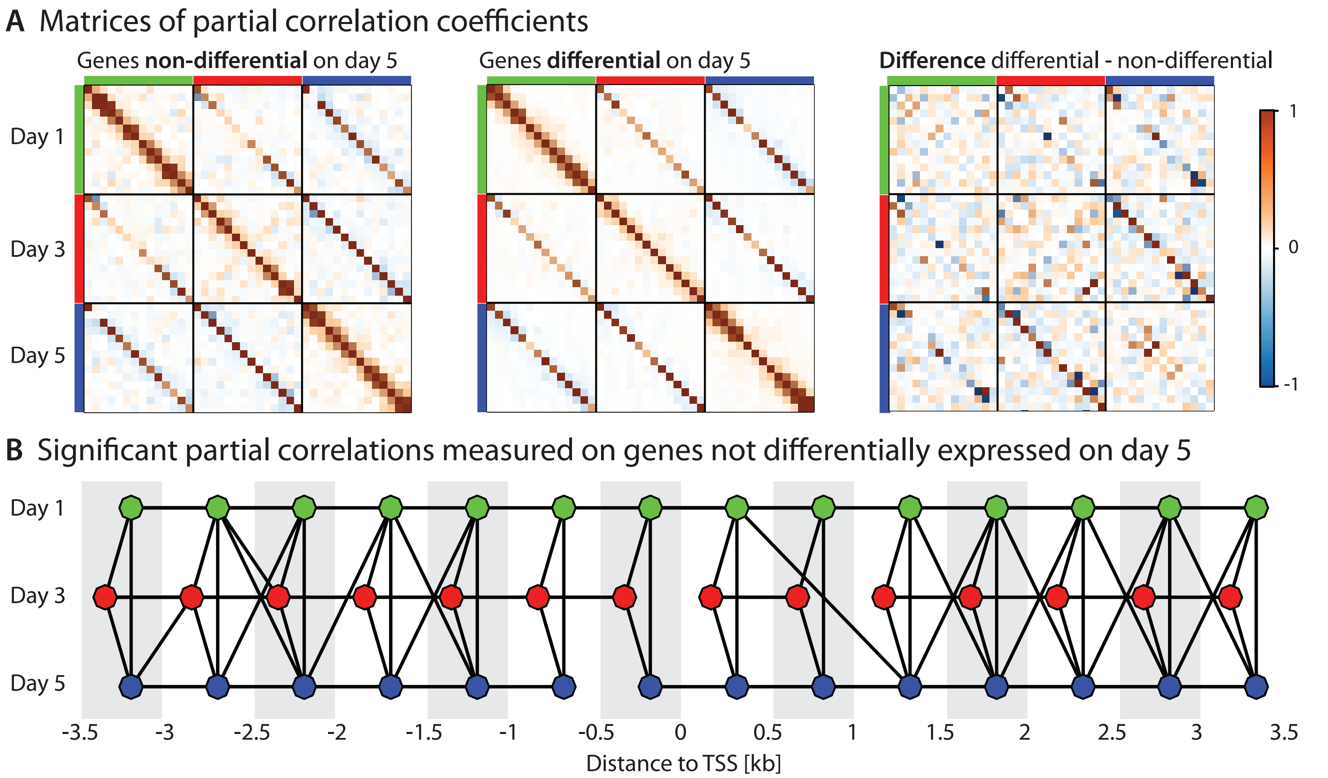

Figure 3. Partial correlation analysis of acetylation profiles. We analyzed spatial and temporal dependencies between regions around TSS by partial correlation coefficients. **A** Matrices of partial correlation coefficients for histone acetylation profiles on days 1 (green), 3 (red) and 5 (blue) computed on non-differential genes only (left) and differential genes only (middle). The right matrix shows the difference of the other two. **B** A graph representation of significant partial correlations (multiple testing corrected p-value <0.05). We show the graph computed on non-differential genes only. Partial correlations on differential genes are very similar, as panel A shows, but since there are many more non-differential than differential genes we gain in power to detect significant correlations. We find that spatial and temporal relationships are largely preserved in the partial correlation structure. However, regions closer to TSS [$\leq$1.5 kb] are less densely connected than the regions further away and in particular show gaps at positions right next to TSS on days 3 and 5.

doi:10.1371/journal.pcbi.1001034.g003

correlations, we asked whether the results are global or driven by a small subset of genes, e.g. the differentially expressed genes. Figure 4A summarizes our finding that the pattern of increased correlation over time was preserved for all subsets of genes and definitions of correlation. This indicates that our results are reproducible and describe a global event not limited to a specific subset of genes or a particular correlation estimate.

Acetylation changes are highly predictive for gene expression changes

Correlation analysis showed global coherence between *averaged* acetylation profiles and gene expression. Next, we analyzed the predictive power of the *complete* profile using a wide array of statistical classification methods. We investigate the predictive power of histone acetylation for gene expression by asking: Can changes in histone acetylation patterns project changes in gene expression? If acetylation is a marker for open chromatin, does it predict expression change in general, and how well can it distinguish up- from down-regulation?

Setup of classification analysis. To address these questions we applied a comprehensive collection of classification methods in an unbiased repeated 10-fold cross-validation study (see *Materials and Methods*) to four different classification problems: (i) Distinguishing transcriptionally up- from down-regulated genes, (ii) distinguishing down-regulated genes from un-responsive genes, (iii) distinguishing up-regulated genes from un-responsive genes, or (iv) distinguishing differential genes (up or down) from un-responsive genes. On each of these four problems we used (a.) Support Vector Machines with different kernel functions; (b.) versions of Gaussian discriminant analysis; (c.) several classification tree methods; (d.) k-nearest neighbor classification with varying numbers of neighbors; as well as (e.) naive Bayes classification, neural networks and logistic regression (see *Materials and Methods*).

Different classifiers may respond to different signal in the data. For example, naive Bayes classifiers assume independence of features (here: the bins in the acetylation profiles), while SVM and other non-linear classifiers can make use of interactions between features. Our selection of classification methods offers a comprehensive overview of current state-of-the-art methodology and makes our results independent of an arbitrary choice of some particular classification method.

Results of classification analysis. Figure 4B shows the results of the cross validation study. In all problems all classifiers clearly beat the baseline of 50% accuracy, but there are obvious differences in performance: Distinguishing up- from down-regulated genes is the easiest problem with performances reaching 80% and above. This margin of improvement over baseline is quite large given that predicting expression from sequence information is a notoriously hard problem (see the discussion of [43] in [44]) and that the acetylation marks we are using ranked far behind others in predictive power for expression in a recent comparison [30].

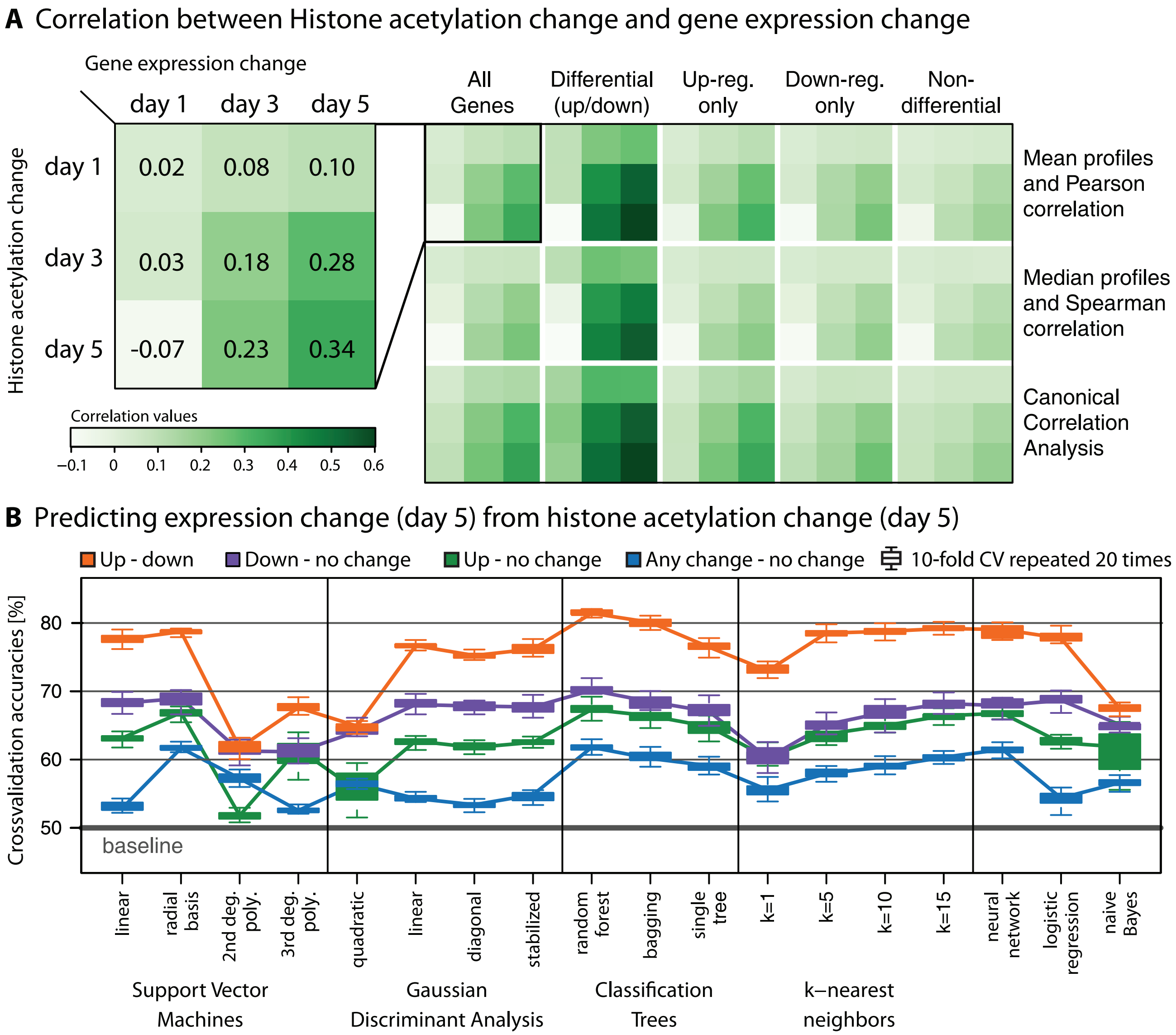

Figure 4. Predictive power of acetylation changes for gene expression changes. A The left matrix shows the correlation between genome-wide mean acetylation changes and gene expression changes using Pearson correlation. Correlation values are small, but highly significant (see discussion in the main text). The right matrix shows correlation results when using other gene sets defined by differential expression on day 5 (columns of the matrix) or other measures of correlation (rows of the matrix). **B** Cross-validation results for a wide array of statistical classifiers predicting gene expression change from histone acetylation change. For each classifier four boxplots show the results of 10-fold cross-validation repeated 20 times sampling balanced data sets. The color of the boxplots corresponds to one of four classification problems: Up- versus down-regulation (Orange), Down-regulation versus no-change (Purple), Up-regulation versus no-change (Green) and any change (up or down) versus no change (Blue).

doi:10.1371/journal.pcbi.1001034.g004

The other three classification problems are harder, especially for distinguishing differential from unresponsive genes classifier performances only reach a level of around 60% accuracy. This can be explained by the set of differential genes containing two opposing signals, which makes it hard to clearly separate it.

The other two curves in Figure 4B show that down-regulated genes can be better distinguished from un-responsive genes than up-regulated genes can. This might be surprising since we saw in Figure 2B that the acetylation profile distributions for down-regulated genes overlapped more with the un-responsive genes than the profiles for up-regulated genes did. However, it can be explained by the fact that loss of acetylation affects wider regions than gain of acetylation signal as can be seen in Figure 2B.

For all classification problems, more highly regularized and constrained methods beat less regularized ones; for example, a larger number of neighbors improves k-nearest neighbor classification, quadratic Gaussian discriminant analysis performs worse than the three linear versions, and the higher degree polynomial SVMs are in most cases out-performed by the linear SVM.

ESC genes show very strong acetylation changes, which are not all reflected in gene expression

Our results so far investigated the general relationship between histone acetylation and gene expression. Now we focus on sets of genes central to the regulatory network governing ES cell state. We will call them *ESC genes* for short. We used several freely available data sources, which complement each other in describing ESC from different perspectives including transcriptional, proteomic and functional. In particular, we used five different descriptions of key ESC genes given by (1) the PluriNet [45], a computationally derived stem cell signature; (2) hits of a recent RNAi screen for self-renewal [46]; (3) gene ontology [47] term GO:0019827 'stem cell maintenance'; (4) members of an ESC-specific protein-interaction network [15]; (5) key transcriptional regulators of ESC [8].

Average histone acetylation signal is very high in ESC genes. The left panel in Figure 5A plots the sorted mean acetylation signal on day 0, before Nanog-silencing triggers differentiation, and underneath the ranks where the five ESC gene lists fall in this ordering. In all five gene lists we observe a strong trend for ESC genes to have a very high average acetylation signal, i.e. the bars representing the gene sets all cluster on the right-hand side of the plot. The trends are strong and easily visible by eye; we quantify their significance by Gene Set Enrichment Analysis (GSEA [48], see *Materials and Methods*) and observe p-values $\leq 10^{-4}$ for four gene sets and $p < 10^{-3}$ for the fifth one.

Decrease in histone acetylation signal is not accompanied by similarly strong decrease in gene expression. Over time the acetylation signal generally diminishes, but this trend is especially pronounced in ESC genes (Figure 5A, middle panel). All five gene sets have p-values $\leq 10^{-2}$ and three of them even $\leq 10^{-4}$. This shows that compared to all other genes, ESC genes are predominantly affected by de-acetylation during the first days of differentiation. If we take histone acetylation as a marker of open chromatin, this result could indicate that the chromatin regions, at which the ESC genes are located, are closing down over time.

We were then interested in seeing how this strong de-acetylation is reflected in gene expression (right-most panel in Figure 5A). Qualitatively, the correlation results of Figure. 4A also hold for the sets of ESC genes. However, when comparing expression changes in ESC genes to other genes, we only found a strong trend to negative expression changes in the set of transcriptional regulators [8] ($p \leq 10^{-4}$) but only much less in the other gene sets. Members of the protein interaction network [15] show moderate down-regulation, but in particular the PluriNet genes [45] and the RNAi hits [46] are uniformly spread out over the spectrum. One way to interpret this observation are other major regulatory influences on key ESC genes that can not be explained by accumulation of condensed and transcriptionally inactive heterochromatin regions (as far as these are indicted by histone de-acetylation).

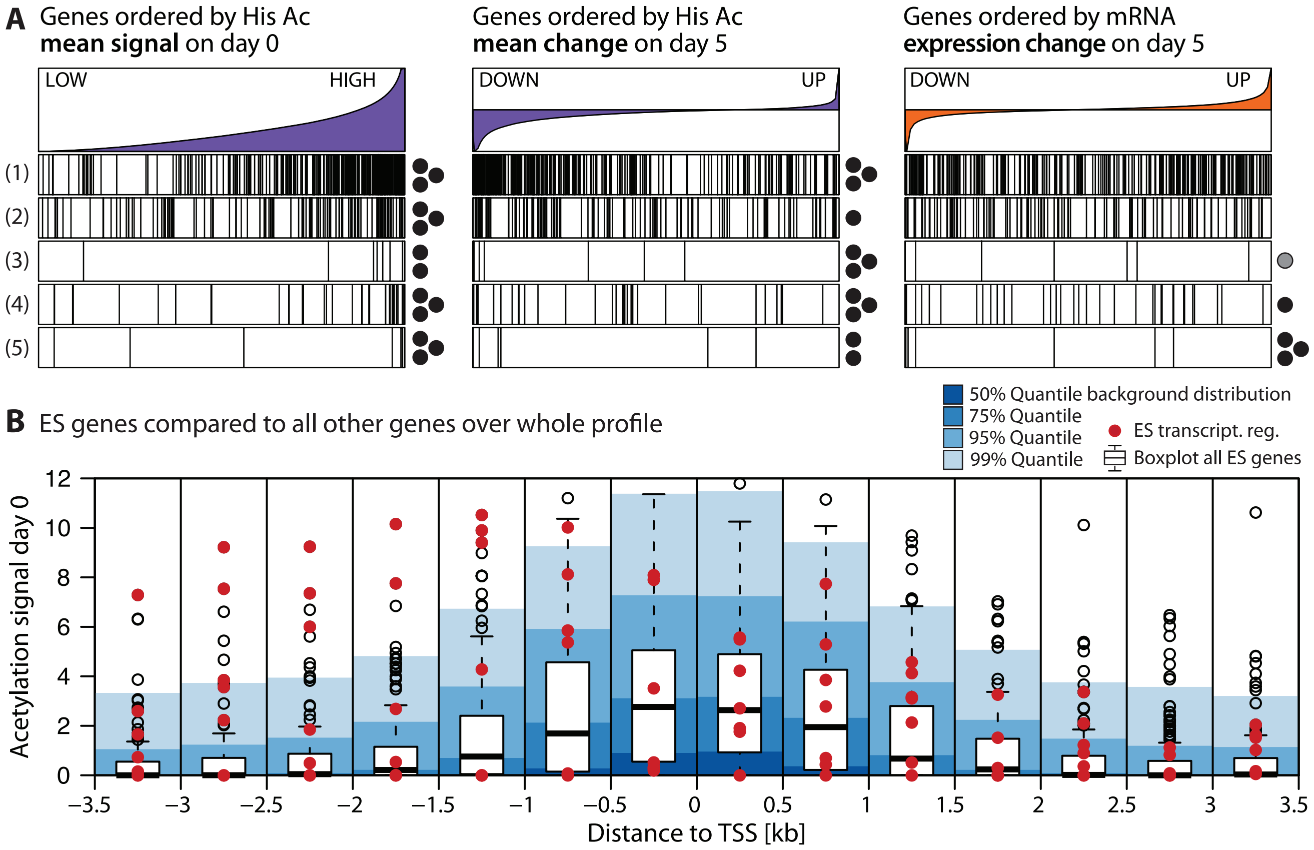

Figure 5. ESC genes show distinct histone acetylation patterns. We compare five sets of ESC specific genes to all other genes in terms of their histone acetylation and gene expression changes: (1) members of the PluriNet [45]; (2) hits of a recent RNAi screen [46]; (3) gene ontology term GO:0019827 'stem cell maintenance'; (4) members of an ESC-specific protein-interaction network [15]; (5) key transcriptional regulators [8]. **A** All genes are ordered by their mean acetylation signal on day 0, their acetylation change on day 5 and their expression change on day 5. The positions of the five ES specific gene sets in this ordering are then indicated by bars. The dots and circles indicate statistical significance of observed trends evaluted by GSEA: three dots for $p \leq 10^{-4}$, two for $p \leq 10^{-3}$ and one for $p \leq 10^{-2}$, while a circle represents $p \leq 0.05$. **B** Here we compare ES genes to all others over the whole acetylation profile. The blue areas indicate quantiles of the genome-wide distribution of acetylation signal. The ES specific gene sets (white boxplots) show overall very high acetylation levels, in particular the transcriptional regulators (red dots) show surprisingly high histone acetylation levels before TSS.

doi:10.1371/journal.pcbi.1001034.g005

ESC genes show overall very strong histone acetylation profiles. The high acetylation signal of ESC genes is not only found in the mean value, but over the whole profile. Figure 5B compares the distribution of acetylation signal between ESC genes and all other genes for each bin individually. Quantiles for the global acetylation distribution across all genes are shown in blue and white boxplots represent the distributions for the union of ESC gene sets described above. Because of their important regulatory function, the set of transcriptional regulators [8] are additionally highlighted in red. We see a significant upwards shift for ESC genes in over the whole range of the profile. This shift is especially pronounced for stem cell transcriptional regulators directly before the TSS.

Discussion

In this paper we have addressed several questions central to an understanding of the relationship between histone acetylation and gene expression. Using a wide array of methods we have investigated how changes in gene expression are reflected in histone acetylation, how predictive histone acetylation is for gene expression changes, and how this relationship changes over time. In the following we will give a short discussion of our main results.

Gain and loss of acetylation over time

While there are less genes transcriptionally down-regulated than up-regulated (Figure. 2A) we find that the accompanying de-acetylation events are much more pronounced than the acetylation events. The wider impact of de-acetylation could be seen in the heatmap (Figure 2A) and the distribution plots (Figure 2B). Its effects could be seen in the results of correlation analysis (Figure 3) and classification analysis (Figure 4).

While Roh *et al* [34] observe main changes in a region of ±1kb around TSS, we observe wider changes especially for down-regulated genes. In particular for ESC genes we find strong acetylation changes over time (Figure 5A) and for several transcriptional regulators we see that acetylation is extremely high before TSS (Figure 5B). These differences in acetylation signal could point to mechanistic differences in how acetyatlion acts and which transcriptional co-factors it recruits in activated and repressed genes.

Predictive power of acetylation changes for gene expression changes

We have seen from the classification results (Figure 4B) that histone acetylation changes are highly predictive of gene expression changes. We have also found that the coordination between histone acetylation measurements and gene expression increases over time. This pattern is stable to varying correlation measures and selecting subsets of genes (Figure 4A).

One way to interpret this trend is a time-lag before changes in chromatin structure (as far as these are indicated by histone acetylation) result in coordinated changes in gene expression. In this scenario, chromatin changes induce gene expression changes, which only become visible at a later time-point and thus increase correlation over time. However, the time-delay in our case would span several days and it is not clear which mechanism causes it, since (de-)acetylation dynamics –at least in yeast– are known to work in the order of minutes [29]. Another question we can not answer from predictive models alone is whether chromatin structure changes are *causative* for gene expression changes or whether it is the other way round: chromatin changes could be induced by expression changes and activation of chromatin modelling proteins.

Distinct acetylation patterns in key ESC genes

It is known that ES cells in general are rich in less compact euchromatin [23] and high histone acetylation levels are one of the indicators for these open chromatin regions. Thus, the strong acetylation signal of ESC genes we observed could indicate that they are located at open chromatin and thus easily accessible to transcription factors. Our results show that ESC genes are enriched for strong de-aceylation (Figure 5A; middle panel). This observation could point to the fact that in early development, as soon as the cell commits to a certain lineage, ESC are located in genomic regions that are de-acetylated and compacted much faster than other regions of the genome. Our interpretation depends on how close the link between histone acetylation and chromatin structure actually is. Not all chromatin changes will be reflected in histone acetylation and in future work it will be important to also probe other markers of chromatin organization, like *e.g.* histone methylation, in ESC over time. Integrated analyses of different markers will give a much richer picture of epi-genetic gene regulation than any individual marker can [30].

The stability of acetylation islands we observe and the strong de-acetylation over time agree with a *global accessibility model* of lineage commitment [17] in which ES cells are subject to global active histone modifications that get lost in a lineage-specific way during differentiation. In contrast, our observations do not agree with a *localised marking model* [17] in which short regions of accessible chromatin are expanded during development. This expansion would be visible as location changes in acetylation islands which we did not observe. However, the situation could change if the time-course was repeated using ChIP-seq instead of ChIP-chip technology which offers a higher resolution of acetylation changes.

Our results have two important implications: First, the pattern in Figure 5A shows that the expressions of some of the key ESC genes, especially PluriNet and the RNAi hits, are not regulated completely by chromatin accessibility (as far as it is visible in histone acetylation patterns). Second, the uniform distribution of gene expression changes in many ESC genes shows that they do not regulate pluripotency on a transcriptional level.

The differences in behaviour we see between transcriptional regulators on the one hand and the PluriNet genes and RNAi hits on the other hand could possibly be attributed to differences in how specific these genes actually are for ES cells. The transcriptional regulators are all well-known and very specific, while the computational and functional predictions from PluriNet or RNAi screens can also capture many non-specific genes. For example, the MATISSE algorithm [49] used to derive the PluriNet signature uses protein-interactions and gene expression to find genes connected to key ESC markers. The genes 'pulled in' by the algorithm can help to better understand the mechanisms behind the known marker genes, without being specific regulators themselves. Similar considerations hold for RNAi screens. Many genes contributing to basic cellular functions can potentially be found to be essential for self-renewal, without being stem-cell specific.

In summary

Our results are a step forward to a better understanding of the complexities of the relationship between histone acetylation and gene expression, which will help to dissect the multilayer regulatory mechanisms that determine stem cell fate. The data of Lu *et al* [25] is an example of a very rich and complex dynamic phenotype of a single-gene perturbation. Future work will need to integrate this data with similar phenotypes of other genes and then use statistical methods [50] to uncover the cellular networks underlying the observed phenotypes.

Materials and Methods

Software

The complete analysis was performed in the statistical computing language R [51] using packages available from the Bioconductor website at http://www.bioconductor.org [52]. In addition to the basic distribution we mainly used the packages limma [53,54], GeneNet [55], CCA [56], MLInterfaces [57], and all packages implied by these. All code is available from the first author upon request.

Data preprocessing

Data generation, pre-processing and mapping of genes between datasets is done in exactly the same way as in [25]. Per day we use for each gene the average of three replicates of gene expression measurements and the average of two replicates of H3K9,14ac ChIP-chip. We apply simple quality filters to the histone acetylation data: 19, 413 genes are represented by probes on the chip. For each gene, the probes are concentrated in a ± 3.5 kb region around transcription start. Out of the 19, 413 genes, we select the 17,268 that have more than 10 probes within 3.5 kb of transcription start. On average, we find ~ 30 probes per gene, which typically have a distance of ~ 248 bases pairs. For all of these genes the data set also contains gene expression measurements.

Identification of acetylation islands

To find acetylation islands [34], we compared the measurement for each probe against the distribution of measurements of the control probes on the array. The control probes are designed to be un-acetylated and thus constitute a negative control. Comparing the probe values against the Null distribution yields a *p*-value for every probe. Using a hierarchical model and an empirical Bayesian estimation strategy [58] we computed day-to-day variability of the acetylation profiles. We used the day-specific variability estimates to compare the probe values against the Null distribution, which yields a (model-based) p-value for every probe. We use an FDR cutoff of $\alpha = 0.1$ on the *p*-value distribution to decide which probes to call acetylated and which not.

Frequency of acetylation changes

We computed for each gene the conditional distribution of probe acetylation states given the previous time-point. The distribution table can be represented by two numbers: the percentage α of probes staying un-acetylated and the percentage β of probes staying acetylated. In this way, each gene can be mapped to a point in $[0,100] \times [0,100]$. Genes with too few (≤ 3) acetylated or un-acetylated probes ($= 3510$ genes) were discarded because their estimates would be unstable. Results for the remaining 13758 genes are shown in Figure 2A. Plotted are the frequencies computed by assuming that the change distribution is the same for all time points; results don't change qualitatively if we compute individual changes between days (see inlay in Figure 2A for genes differential on day 5).

Step-wise linear approximation of acetylation profiles

Genes are represented by different number of probes with varying distances between each other and to the transcription start site. To make acetylation profiles comparable between genes we map them onto vectors of equal length by averaging all probes in equi-distant bins around transcription start. We chose a binning of 0.5kb, thus covering the ± 3.5kb region with 14 bins and mapping each acetylation profile into R^{14}. We only considered the signal above background, bins with no probes above background were set to zero. Examples of raw and binned profiles can be seen in Figure 1A. This binning and averaging makes the data comparable between genes, while preserving most of the quantitative variation in the data.

Partial-correlation analysis

To delineate the correlation structure of the data we used partial correlation analysis, also called a Gaussian graphical model [42,59]. In contrast to regular correlation, *partial* correlation corrects for the influence of all other variables in the model: Vanishing partial correlation (under a Gaussian assumption) means that two variables are independent given all other variables (genomic regions in our case). Thus, partial correlation coefficients measure the direct relationship between two variables, while regular correlation coefficients also measure indirect effects. We used a shrinkage approach [60] for robust estimation of partial correlations. The results can be depicted in a graph, where each node corresponds to a variable (a genomic region) and each edge a partial correlation that is different from zero. Missing edges indicate vanishing partial correlation and thus conditional independence. We select the network containing only edges with probability > 0.9 corresponding to a local FDR cutoff of 0.1 [55].

Canonical correlation analysis

Canonical correlation analysis (CCA, see [42]) is a way of measuring the linear relationship between two multidimensional variables. In general, CCA finds vectors a and b such that the random variables $a'X$ and $b'Y$ maximize the correlation $\rho = \mathrm{cor}(a'X, b'Y)$. Vectors a and b are unique up to scalar multiplication. The random variables $U = a'X$ and $V = b'Y$ are the first pair of canonical variables and ρ is called the canonical correlation. In our application X corresponds to the histone acetylation data (a 14 dimensional random variable) and Y to the RNA data per day (a one dimensional random variable). Thus, we only need to find vector a to maximize the correlation between the two data sets. Computing the correlation between mean acetylation profiles and expression is closely related to CCA, since it corresponds to the choice of $a_{\mathrm{mean}} = \frac{1}{14}(1, \ldots, 1)$, but it is not guaranteed to find the maximal correlation.

Classification methods

(a.) Support Vector Machines (SVM, [61]) construct the hyperplane with maximal margin of separation between the positive and negative training examples. Using non-linear distance measures, so-called kernel functions, this approach can be extended to non-linear classification. We use a linear kernel, a radial basis function kernel and polynomial kernels of degrees 2 and 3. (b.) Gaussian Discriminant Analysis [62] assumes that the positive and negative examples follow a multivariate normal distribution. Versions of Discriminant Analysis differ by the constraints they put on the covariance matrices: no constraints (Quadratic DA); or the same covariance matrix for both classes (Linear DA); or the same *diagonal* covariance matrix (Diagonal Linear DA). Stabilized Linear DA is linear discriminant analysis based on left-spherically distributed linear scores. (c.) Classification trees [63] recursively partition the dataset by splitting along most-informative single features. Bagging [64] (short for 'bootstrap aggregating') aggregates many classification trees built on resampled versions of the training data. Similar to bagging, a Random Forest [65] is an aggregation of many classification trees built on resampled versions of the data and on a randomly chosen subset of features. (d.) k-nearest neighbors predicts a gene into the class represented by the majority of the k genes closest to it. We use $k = 1$, 5, 10, and 15. (e.) Naive Bayes classification assumes independence of features (hence *naive*) and classifies according to the class posterior probability. The neural network [66] is a single-hidden-layer network. Logistic Regression [62] combines a linear model of the data together with a logistic function to model class probabilities. All classifiers were used via the R-package MLInterfaces [57] and with the default parameters defined there.

Balanced evaluation of prediction accuracy

The datasets we use for classification can be very unbalanced, for example only $\sim 5\%$ of all genes show a significant expression change. Thus, the baseline for classification is already at 95% accuracy (when we predict all genes as 'unchanged'). To be able to compare between methods and different classification scenarios, we resorted to a random sampling strategy: We sampled from the larger part of the training set 20 times sets of the size of the smaller part. This created 20 instances of balanced training sets with a baseline of 50%. On each training set we computed the 10-fold cross-validation (CV) accuracy. The variance we see in the CV results is thus a sum of the variance introduced by sampling the training set and the variance from randomly splitting the data into 10 subsets inside CV procedure. It is reassuring that Figure 4B overall shows very consistent results, only individual boxplots are spread out widely.

Gene Set Enrichment Analysis (GSEA)

The goal of GSEA [48] is to determine whether members of a gene set (for example ES genes) tend to occur toward the top (or bottom) of a list of phenotypes (in our case: mean acetylation or expression). GSEA is especially suited to find coherent changes in a group of genes, even if the individual changes are small. GSEA calculates an enrichment score for a given gene set using rank of genes and infers statistical significance of each ES against ES background distribution calculated by permutation of the original data set. We report the empirical *p*-value after $2{\cdot}10^4$ permutations, *i.e.* in how many permutations did we observe a result more extreme than the one on real data. We did no multiple-testing correction, since with only 15 tests altogether even the most conservative correction ($p' = 15{\cdot}p$) would not qualitatively change our results.

Author Contributions

Conceived and designed the experiments: FM IRL OGT. Performed the experiments: FM. Analyzed the data: FM KWM. Contributed reagents/materials/analysis tools: FM EMA IRL. Wrote the paper: FM KWM OGT.

References

1. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125: 315–326.
2. Boyer LA, Mathur D, Jaenisch R (2006) Molecular control of pluripotency. Curr Opin Genet Dev 16: 455–462.
3. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553–560.
4. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, et al. (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454: 766–770.
5. Vastenhouw NL, Zhang Y, Woods IG, Imam F, Regev A, et al. (2010) Chromatin signature of embryonic pluripotency is established during genome activation. Nature 464, 922-926.
6. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, et al. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122: 947–956.
7. Loh YH, Wu Q, Chew JL, Vega VB, Zhang W, et al. (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat Genet 38: 431–440.
8. Ivanova N, Dobrin R, Lu R, Kotenko I, Levorse J, et al. (2006) Dissecting self-renewal in stem cells with RNA interference. Nature 442: 533–538.
9. Kim J, Chu J, Shen X, Wang J, Orkin SH (2008) An extended transcriptional network for pluripotency of embryonic stem cells. Cell 132: 1049–1061.
10. Chickarmane V, Peterson C (2008) A computational model for understanding stem cell, trophectoderm and endoderm lineage determination. PLoS One 3: e3478.
11. Ying QL, Wray J, Nichols J, Batlle-Morera L, Doble B, et al. (2008) The ground state of embryonic stem cell self-renewal. Nature 453: 519–523.
12. Sampath P, Pritchard DK, Pabon L, Reinecke H, Schwartz SM, et al. (2008) A hierarchical network controls protein translation during murine embryonic stem cell self-renewal and differentiation. Cell Stem Cell 2: 448–460.
13. Chang WY, Stanford WL (2008) Translational control: a new dimension in embryonic stem cell network analysis. Cell Stem Cell 2: 410–412.
14. Macarthur BD, Ma'ayan A, Lemischka IR (2009) Systems biology of stem cell fate and cellular reprogramming. Nat Rev Mol Cell Biol 10: 672–681.
15. Wang J, Rao S, Chu J, Shen X, Levasseur DN, et al. (2006) A protein interaction network for pluripotency of embryonic stem cells. Nature 444: 364–368.
16. Nishiyama A, Xin L, Sharov AA, Thomas M, Mowrer G, et al. (2009) Uncovering early response of gene regulatory networks in ESCs by systematic induction of transcription factors. Cell Stem Cell 5: 420–433.
17. Szutorisz H, Dillon N (2005) The epigenetic basis for embryonic stem cell pluripotency. Bioessays 27: 1286–1293.
18. Reik W (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. Nature 447: 425–432.
19. Hemberger M, Dean W, Reik W (2009) Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal. Nat Rev Mol Cell Biol 10: 526–537.
20. Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA, et al. (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature 441: 349–353.
21. Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, et al. (2006) Control of developmental regulators by polycomb in human embryonic stem cells. Cell 125: 301–313.
22. Schaniel C, Ang Y, Ratnakumar K, Cormier C, James T, et al. (2009) Smarcc1/Baf155 couples self-renewal gene repression with changes in chromatin structure in mouse embryonic stem cells. Stem Cells 27: 2979–91.
23. Meshorer E, Misteli T (2006) Chromatin in pluripotent embryonic stem cells and differentiation. Nat Rev Mol Cell Biol 7: 540–546.
24. Meshorer E, Yellajoshula D, George E, Scambler PJ, Brown DT, et al. (2006) Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. Dev Cell 10: 105–116.
25. Lu R, Markowetz F, Unwin RD, Leek JT, Airoldi EM, et al. (2009) Systems-level dynamic analyses of fate change in murine embryonic stem cells. Nature 462: 358–362.
26. Allfrey VG, Faulkner R, Mirsky AE (1964) Acetylation and methylation of histones and their possible role in the regulation of rna synthesis. Proc Natl Acad Sci U S A 51: 786–794.
27. Pogo BG, Allfrey VG, Mirsky AE (1966) RNA synthesis and histone acetylation during the course of gene activation in lymphocytes. Proc Natl Acad Sci U S A 55: 805–812.
28. Sterner DE, Berger SL (2000) Acetylation of histones and transcription-related factors. Microbiol Mol Biol Rev 64: 435–459.
29. Kurdistani SK, Grunstein M (2003) Histone acetylation and deacetylation in yeast. Nat Rev Mol Cell Biol 4: 276–284.
30. Karlić R, Chung HR, Lasserre J, Vlahovicek K, Vingron M (2010) Histone modification levels are predictive for gene expression. Proc Natl Acad Sci U S A 107: 2926–2931.
31. Kurdistani SK, Tavazoie S, Grunstein M (2004) Mapping global histone acetylation patterns to gene expression. Cell 117: 721–733.
32. Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, et al. (2005) Genome-wide map of nucleosome acetylation and methylation in yeast. Cell 122: 517–527.
33. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, et al. (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet 40: 897–903.
34. Roh TY, Cuddapah S, Zhao K (2005) Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. Genes Dev 19: 542–552.
35. Schones DE, Zhao K (2008) Genome-wide approaches to studying chromatin modifications. Nat Rev Genet 9: 179–191.
36. Shogren-Knaak M, Ishii H, Sun JM, Pazin MJ, Davie JR, et al. (2006) Histone H4-K16 acetylation controls chromatin structure and protein interactions. Science 311: 844–847.
37. Schreiber SL, Bernstein BE (2002) Signaling network model of chromatin. Cell 111: 771–778.

38. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA (2007) A chromatin landmark and transcription initiation at most promoters in human cells. Cell 130: 77–88.
39. Yuan GC, Ma P, Zhong W, Liu JS (2006) Statistical assessment of the global regulatory role of histone acetylation in Saccharomyces cerevisiae. Genome Biol 7: R70.
40. Karantzali E, Schulz H, Hummel O, Huebner N, Hatzopoulos A, et al. (2008) Histone deacetylase inhibition accelerates the early events of stem cell differentiation: transcriptomic and epigenetic analysis. Genome Biol 9: R65.
41. Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, et al. (2005) Genome-scale identification of nucleosome positions in S. cerevisiae. Science 309: 626–630.
42. Anderson TW (2004) An Introduction to Multivariate Statistical Analysis. 2nd edition. New York: Wiley.
43. Beer MA, Tavazoie S (2004) Predicting gene expression from sequence. Cell 117: 185–198.
44. Yuan Y, Guo L, Shen L, Liu JS (2007) Predicting gene expression from sequence: a reexamination. PLoS Comput Biol 3: e243.
45. Müller FJ, Laurent LC, Kostka D, Ulitsky I, Williams R, et al. (2008) Regulatory networks define phenotypic classes of human stem cell lines. Nature 455: 401–405.
46. Hu G, Kim J, Xu Q, Leng Y, Orkin SH, et al. (2009) A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. Genes Dev 23: 837–848.
47. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. the gene ontology consortium. Nat Genet 25: 25–29.
48. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545–15550.
49. Ulitsky I, Shamir R (2007) Identification of functional modules using network topology and high-throughput data. BMC Syst Biol 1: 8.
50. Markowetz F (2010) How to understand the cell by breaking it: network analysis of gene perturbation screens. PLoS Comput Biol 6: e1000655.
51. R Development Core Team (2007) R: A language and environment for statistical computing.
52. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.
53. Smyth, K G (2005) Limma: linear models for microarray data. In: Gentleman, Carey V, Dudoit S, Irizarry R, Huber W, eds. Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. New York: Springer. pp 397–420.
54. Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, et al. (2007) A comparison of background correction methods for two-colour microarrays. Bioinformatics 23: 2700–2707.
55. Opgen-Rhein R, Schäfer J, Strimmer K (2007) GeneNet: Modeling and Inferring Gene Networks. R package version 1.2.1. Available: http://cran.r-project.org/.
56. González I, Déjean S, Martin PGP, Baccini A (2007) CCA: An R package to extend canonical correlation analysis. J Stat Softw 23: 1–14.
57. Mar J, Gentleman R, Carey V (2008) MLInterfaces: Uniform interfaces to R machine learning procedures for data in Bioconductor containers. R package version 1.24.0. Available: http://www.bioconductor.org.
58. Airoldi EM (2007) Getting started in probabilistic graphical models. PLoS Comp Biol 3: e252.
59. Lauritzen SL (1996) Graphical Models. Oxford: Clarendon Press.
60. Schäfer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol 4: Art32.
61. Schölkopf B, Smola A (2002) Learning with Kernels. Cambridge: MIT Press.
62. Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning. New York: Springer.
63. Breiman, Friedman, Olshen, Stone (1984) Classification and Regression Trees. Wadsworth.
64. Breiman L (1996) Bagging predictors. Mach Learn 24: 123140.
65. Breiman L (2001) Random forests. Mach Learn 45: 5–32.
66. Ripley BD (1996) Pattern Recognition and Neural Networks. Cambridge: Cambridge University Press.