Resource

Cell

Comparative Epigenomic Annotation of Regulatory DNA

Shu Xiao,[1,2,6] Dan Xie,[1,2,6] Xiaoyi Cao,[2,3,6] Pengfei Yu,[2,3,6] Xiaoyun Xing,[5] Chieh-Chun Chen,[1,2] Meagan Musselman,[1] Mingchao Xie,[5] Franklin D. West,[4] Harris A. Lewin,[2] Ting Wang,[5] and Sheng Zhong[1,2,3,*]

[1]Department of Bioengineering

[2]Institute for Genomic Biology

[3]Center for Biophysics and Computational Biology

University of Illinois at Urbana-Champaign, 1304 West Springfield Avenue, Urbana, IL 61801, USA

[4]Department of Animal and Dairy Science, University of Georgia, 425 River Road, Athens, GA 30602, USA

[5]Department of Genetics, Center for Genome Sciences and Systems Biology, Washington University School of Medicine, 4444 Forest Park Parkway, St. Louis, MO 63108, USA

[6]These authors contributed equally to this work

*Correspondence: szhong@illinois.edu

DOI 10.1016/j.cell.2012.04.029

SUMMARY

Despite the explosive growth of genomic data, functional annotation of regulatory sequences remains difficult. Here, we introduce "comparative epigenomics"—interspecies comparison of DNA and histone modifications—as an approach for annotation of the regulatory genome. We measured in human, mouse, and pig pluripotent stem cells the genomic distributions of cytosine methylation, H2A.Z, H3K4me1/2/3, H3K9me3, H3K27me3, H3K27ac, H3K36me3, transcribed RNAs, and P300, TAF1, OCT4, and NANOG binding. We observed that epigenomic conservation was strong in both rapidly evolving and slowly evolving DNA sequences, but not in neutrally evolving sequences. In contrast, evolutionary changes of the epigenome and the transcriptome exhibited a linear correlation. We suggest that the conserved colocalization of different epigenomic marks can be used to discover regulatory sequences. Indeed, seven pairs of epigenomic marks identified exhibited regulatory functions during differentiation of embryonic stem cells into mesendoderm cells. Thus, comparative epigenomics reveals regulatory features of the genome that cannot be discerned from sequence comparisons alone.

INTRODUCTION

In eukaryotic cells, the genomic DNA is etched with a number of chemical modifications called epigenomic modifications (epi-modifications). These epi-modifications add an extra layer of information to the genomic sequence and enable it to encode a more complex program of gene regulation (Karlić et al., 2010; Maunakea et al., 2010). Different epi-modifications affect how the DNA interacts with transcription factors, although many mechanisms remain unknown (Campos and Reinberg, 2009). Adding to the complexity, the genomes are far from being completely annotated on the functional level, making it necessary to first find regulatory genomic sequences before we can understand their complex regulatory roles.

Evolutionary comparisons provide a powerful tool to study genome functions. This became obvious when it was recognized that the majority of DNA can mutate freely without deleterious effects, whereas certain sequence elements are more constrained (Kimura, 1968). Leveraging this theory, researchers have inferred functional genomic segments by examining genomic sequence conservation (Hardison, 2003) and have identified human-specific regulatory DNA by looking for sequences with accelerated rates of evolutionary change (Pollard et al., 2006).

The successes in genomic comparisons beg the question: can we also use evolution to study the functions of the epigenome? To do so, the basic evolutionary properties of the epigenome must be established first, preferably in the contexts of both genomic and transcriptomic evolution. To explore relationships among evolutionary changes to the genome, the epigenome, and the transcriptome, several specific questions were of critical interest. First, evolutionary selection has left clear traces on the human genome (Ren, 2010); what are the traces of evolutionary selection on the human epigenome? Second, are evolutionary changes to the epigenome merely a consequence of genomic sequence changes or, rather, has the epigenome made the genome more or less susceptible to evolutionary selection? Third, the degree of gene expression conservation correlates poorly with the extent to which nonexonic sequences are conserved among vertebrates (Chan et al., 2009; Wilson et al., 2008); might this discrepancy be explained by the epigenome? Fourth, mammalian orthologous transcription factors (TF) often do not bind to orthologous DNA sequences (Jegga et al., 2008), as only ~5% of the OCT4- and NANOG-binding sites occupy homologous sequences in human and mouse embryonic stem (ES) cells (Kunarso et al., 2010); do epigenetic modification enzymes apply the same types of modifications to orthologous sequences in mammals?

Among many types of epi-modifications (Tan et al., 2011), a subset is known to correlate with gene transcription. For

example, DNA cytosine methylation (C^m) (Maunakea et al., 2010), histone 3 lysine 27 trimethylation (H3K27me3), and histone 3 lysine 9 trimethylation (H3K9me3) may repress gene transcription, whereas histone 3 lysine 4 mono-, di-, and trimethylation (H3K4me1/2/3), lysine 27 acetylation (H3K27ac), and lysine 36 trimethylation (H3K36me3) are positively associated with transcription (Karlić et al., 2010). The roles of some epigenomic marks (epi-marks) remain controversial. For example, histone variant H2A.Z is generally assumed to be associated with active promoters because it anticorrelates with C^m in plants, insects, and fish (Zemach et al., 2010). Consistent with this, H2A.Z is associated with active promoters in flies (Weber et al., 2010). However, H2A.Z is associated with inactive promoters in yeasts (Guillemette et al., 2005; Raisner et al., 2005). The role of H2A.Z has yet to be tested in mammals. Even for the epi-modifications whose roles are better established, they may have undiscovered functions.

The functions of many epi-modifications have so far only been evaluated individually, primarily due to the difficulty of assessing the functional significances of colocalized epi-marks. Any two epi-marks can colocalize in some genomic regions, but such colocalizations do not necessarily serve any regulatory functions. The best documented epi-marks colocalization is probably the bivalent domain (H3K27me3+H3K4me3), which is hypothesized to be poised for activation during differentiation of embryonic stem cells (ESCs) (Mikkelsen et al., 2007). We wish to develop a method to systematically examine the functions of epi-modifications and, more importantly, the functions of combinations of epi-marks. We propose to leverage the connection between evolutionary conservation and functional importance to achieve this goal.

Here, we introduce ''comparative epigenomics''—interspecies comparison of epigenomes—as an approach for annotation of the regulatory sequences of the genome. We created a multispecies epigenomic data set from pluripotent stem cells of humans, mice, and pigs, which is comprised of genomic distributions of DNA methylation and eight histone modifications, the binding intensities of four transcription regulators (NANOG, OCT4, P300, and TAF1), and transcribed RNA sequences. We first examined the coevolution properties among the epigenome, the genome, and the transcriptome. Comparing epigenomic changes to genomic changes, we observed strong epigenomic conservation for both rapidly evolving and slowly evolving DNA sequences, but not on neutrally evolving DNA sequences. These data suggest that epigenomic conservation is not completely dictated by genomic sequences. On the other hand, interspecies epigenomic changes are linearly correlated with evolutionary changes of transcription factor binding and gene expression, suggesting that comparative epigenomics can directly reveal critical information on gene regulation. Based on these initial analyses, we set out to discover regulatory sequences by conserved colocalization of different epi-marks. To test the functions of these putative regulatory sequences, we developed a differentiation assay in which mouse embryonic stem cells were differentiated into mesendoderm cells. Our time course chromatin immunoprecipitation sequencing (ChIP-seq) and RNA sequencing (RNA-seq) data in this differentiation process confirmed the regulatory functions of all seven pairs of epi-marks identified by conserved colocalization. Thus, conserved colocalization is an efficient approach to identify functional epi-mark combinations from a large (combinatorial) number of random combinations of epi-marks. More importantly, comparative epigenomics reveals regulatory features of the genome that cannot be discerned from sequence comparison alone.

RESULTS

Epigenomes of Human, Mouse, and Pig Pluripotent Stem Cells

To answer the above questions, we conducted a comparative epigenomics study (Mikkelsen et al., 2010) with a focus on evolution. We generated and compiled from published work the genomic distributions of nine epigenetic modifications, including C^m, H2A.Z, H3K4me1/2/3, H3K9me3, H3K27me3, H3K27ac, and H3K36me3; and also the binding of four transcription regulators, P300, TAF1, OCT4, and NANOG, in pluripotent stem cells of humans, mice, and pigs (*Sus scrofa*) (West et al., 2010). C^m was assayed by both methylated DNA immunoprecipitation followed by sequencing (MeDIP-seq) and by DNA digestion by methyl-sensitive restriction enzymes followed by sequencing (MRE-seq) (Maunakea et al., 2010). Histone modifications and binding of transcription regulators were assayed with ChIP-seq. Gene expression was measured by RNA-seq technology. Taken together, a total of 48 sequencing data sets were compiled, among which 29 data sets (73 billion bases) were generated from this study (GEO accession number GSE36114), and the 19 other data sets (27 billion bases) were compiled from three published works (Chen et al., 2008; Goren et al., 2010; Lister et al., 2009) (Table S1, top, available online).

Comparative Epigenome Browser

To manage, visualize, and compare these data, we built an interactive data analysis system, the Comparative Epigenome Browser (http://sysbio.igb.uiuc.edu/cpbrowser). A useful analytical feature of this system is a side-by-side comparison of epi-modifications on orthologous genomic regions.

Interspecies Epigenomic Variation Is Greater Than Within-Species Variation

We compared interspecies and within-species epigenomic variations. Based on histone modifications, human and mouse cell lines cluster separately. Hierarchical clustering showed that two human embryonic stem cell (hESC) lines and three human induced pluripotent stem cell (hiPS) lines cluster tightly together, and it also showed that mouse embryonic stem cell (mESC) and two mouse epiblast stem cell (mEpiSC) lines cluster together separately from the human cells (Figures S1A–S1F). The tight clustering of different sources of pluripotent cells within a species was based on comparison of all human-mouse orthologous genes, and the results were consistent for both the activation mark H3K4me3 and the repressive mark H3K27me3, regardless of the distance metric used (Figures S1C and S1D). These data indicate that interspecies epigenomic differences are greater than within-species differences. H3K4me3 and H3K27me3 are the only two marks that have been assayed in both naive and

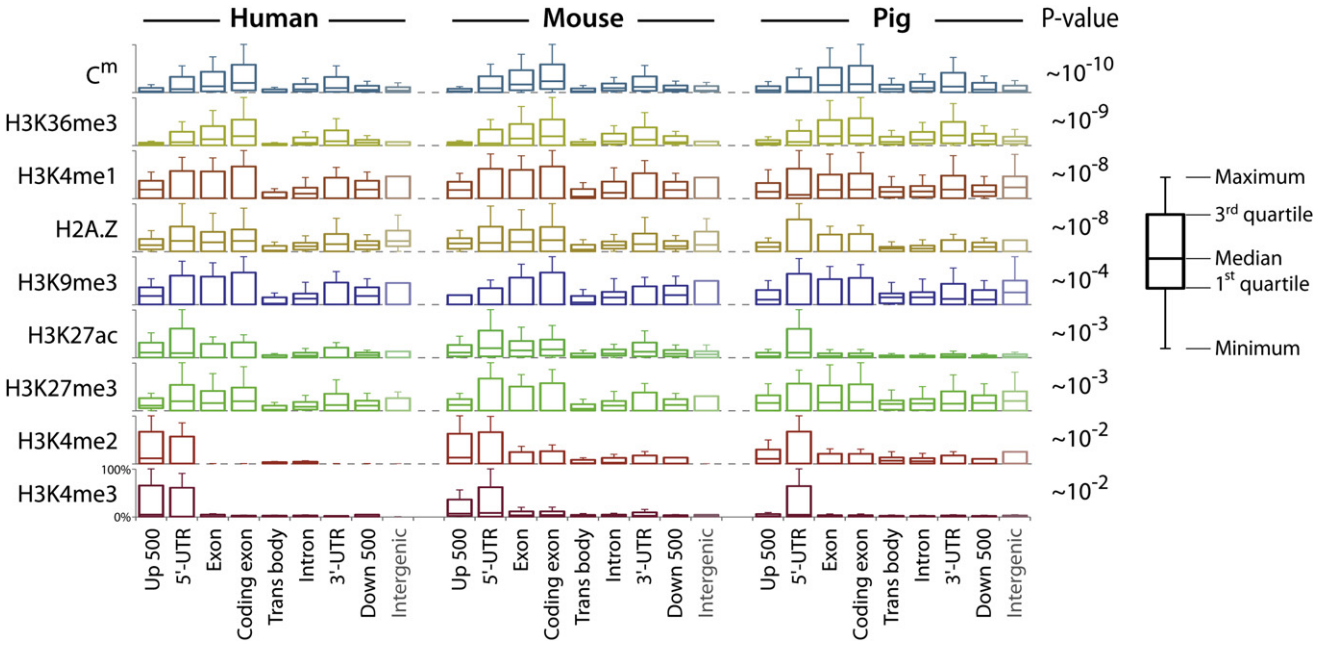

Figure 1. Interspecies Conservation of Epigenomic Modifications

Each box plot represents the distribution of the normalized intensities of the indicated epi-modifications (e.g., C^m, uppermost row) in various genomic regions (e.g., 500 bp upstream of genes, leftmost column). Median, quartiles, maximum, and minimum intensity values are shown in each box plot (see insert). The assembly of nine box plots shows the distribution of relative intensities of an epi-modification on different genomic regions in a species (e.g., C^m in human, uppermost and leftmost panels). p value indicates the support to conservation of each epi-modification, calculated from a nonparametric test comparing the data in the left, middle, and right. See also Figure S1.

primed pluripotent stem cells (Nichols and Smith, 2009) in a genome-wide manner. Lab-to-lab variation in ChIP-seq data is relatively small (Figure S1G) and thus is unlikely to affect this result. Consistently, hierarchical clustering of gene expression data showed strong within-species similarities (Pearson correlation = 0.86 ± 0.19; rank correlation = 0.91 ± 0.09) and a clear interspecies difference (Pearson correlation = 0.36 ± 0.03; rank correlation = 0.60 ± 0.04) (Figure S1E).

Traces of Evolutionary Conservation of the Epigenome

The pronounced interspecies epigenomic differences provoked the question of whether there are any traces of conservation of the epigenome. We started by analyzing the epi-modification intensities for various features of the genome, including intergenic regions, promoters, exons, introns, and 5′ and 3′ untranslated regions (UTRs). The distribution of the intensities of each epi-modification on each type of genomic feature was summarized (Figure 1, box plots), and then these distributions were combined into an "epi-mark intensity" distribution for all genomic regions for a species (Figure 1, any panel of nine box plots). A nonparametric test with the null hypothesis that epi-mark intensity distributions are different across species generated a p value for every epi-mark. These p values ranged from 10^{-10} (C^m) to 10^{-2} (H3K4me3), indicating that the relative difference in epi-modification intensities on different genomic features is in general consistent across species, although the consistency levels vary from modification to modification. This pattern of conservation provides evolutionary support for the idea of using epigenomic data to predict functional noncoding genomic features (Ernst and Kellis, 2010).

We then asked whether the co-occupancy of two epi-marks is conserved across species. In the human genome, nonrandomly coappearing epi-marks include H3K4me1/2/3 (null hypothesis: epi-marks appear independently in the genome; minimum odds ratio = 4.55; p value < 10^{-20}), H3K27ac, and H3K4me1/2/3 (minimum odds ratio = 10.3; p value < 10^{-20}), as well as H3K27me3 and H3K4me1/2/3 (minimum odds ratio = 4.14; p value < 10^{-20}). Nonrandom, mutual-avoiding epi-modifications include C^m versus H3K4me2/3 (odds ratio = 0.70; p value < 10^{-20}). These nonrandom co-occupancy patterns are conserved in mouse and swine genomes (Figures 2A and S2A–S2C). Even weak co-occupancy patterns between any two of H3K27me3, H3K36me3, and H3K4me1/2/3 are conserved in all three species (p value < 10^{-9}). In contrast to being antagonistic chromatin marks in plants, insects, and fish (Zemach et al., 2010; Zilberman et al., 2008), H2A.Z and C^m are not clearly anticorrelated in any of the three mammals (observed total length of comarked regions ≥ expected total length of comarked regions). Instead, H3K4me2/3 and H3K9me3 exhibit stronger anticorrelation to C^m (Figures 2A and S2A–S2C).

Next, we tested whether the genomic regions with one or a pair of epi-modifications are correlated with conserved genomic sequences. Genomic regions with all assayed epi-modifications, except for H3K9me3, are correlated with sequence-conserved regions (odds ratio = 1.43; p value < 10^{-20}) (Figures 2B and S2D–S2F). With the exception of

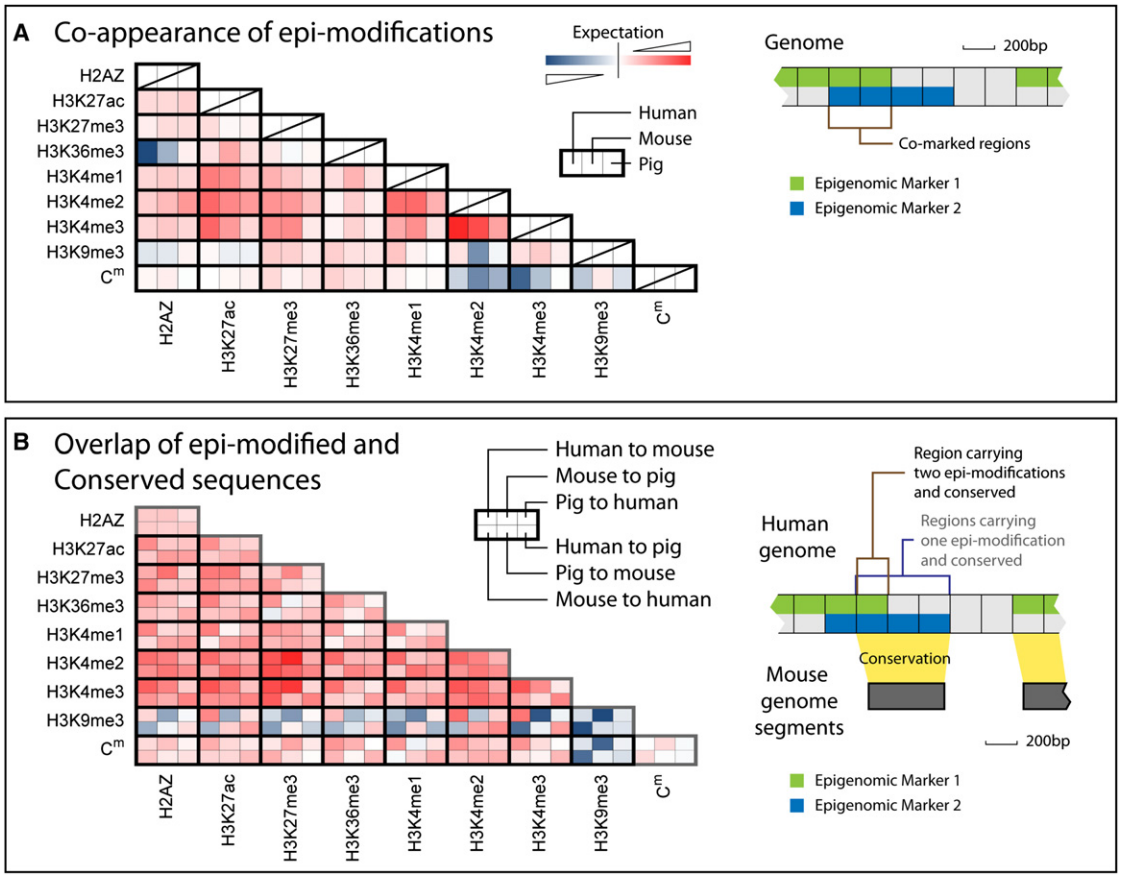

Figure 2. Interspecies Conservation of Co-occupancy of Different Epi-Modifications

(A) Log ratio between the number of genomic regions carrying two epi-modifications (shown as row and column names) and the expected number, calculated from a null model that the epi-modifications appear independently of each other (each small box, red: log ratio > 0, co-occupancy; blue: log ratio < 0, antico-occupancy). With a few exceptions, both positive and negative co-occupancies of any epi-marks are conserved across species, as seen in similar colors of the three consecutive boxes in a row.

(B) Log ratio between the number of conserved regions carrying one (diagonal boxes) or two (nondiagonal boxes) epi-modifications and the expected number, calculated from a null model in which conserved regions and epi-modified regions appear independently. Conserved genomic regions are determined by six pairwise comparisons shown in six small boxes outlined with a darker edge. For example, the leftmost upper box refers to the human genomic regions conserved in a human versus mouse comparison. All genomic regions with epi-modifications, except H3K9me3, were positively associated with conserved regions (red). H3K9me3 selectively marks nonconserved regions (blue). Bivalent domains (comarked by repression mark H3K27me3 and activation mark H3K4me2/3) exhibited the strongest association with conserved regions.

See also Figures S1 and S2.

H3K9me3-marked regions, the genomic regions with two epi-modifications are also correlated with conserved sequences (p value < 10^{-9}), often with stronger correlations than single epi-modification regions.

If the epigenome is evolutionarily conserved, one would expect to see not only that epi-modification marked and conserved sequences are positively correlated, but also that orthologous sequences in two genomes share the same epi-modifications. To test this hypothesis, we categorized conserved regions of the human genome into three distinct sets, which are strongly, moderately, and weakly conserved in 46 vertebrate species, respectively (Siepel et al., 2005). In each set, we quantified epigenomic conservation by the ratio between the observed number of sequences whose orthologous sequences share the same epi-modification and the expected number of epi-sharing orthologous sequences (Figure S2K). H3K27me3, H3K36me3, H3K4me2, and H3K4me3 showed 3- to 20-fold increases in co-occupancy of human-mouse or human-pig orthologous sequences than would be expected by random chance (maximum p value < 10^{-7}). For H3K27me3, H3K36me3, H3K4me1/2/3, and C^m, as the level of sequence conservation increases, the chance of orthologous sequences sharing the same epi-modification also increases. Taken together, these data suggested that examinations of the co-occupancy of epi-marks, the correlation between epi-modification and conserved regions, and the co-occupancy of orthologous sequences may be viable approaches to assess epigenomic conservation.

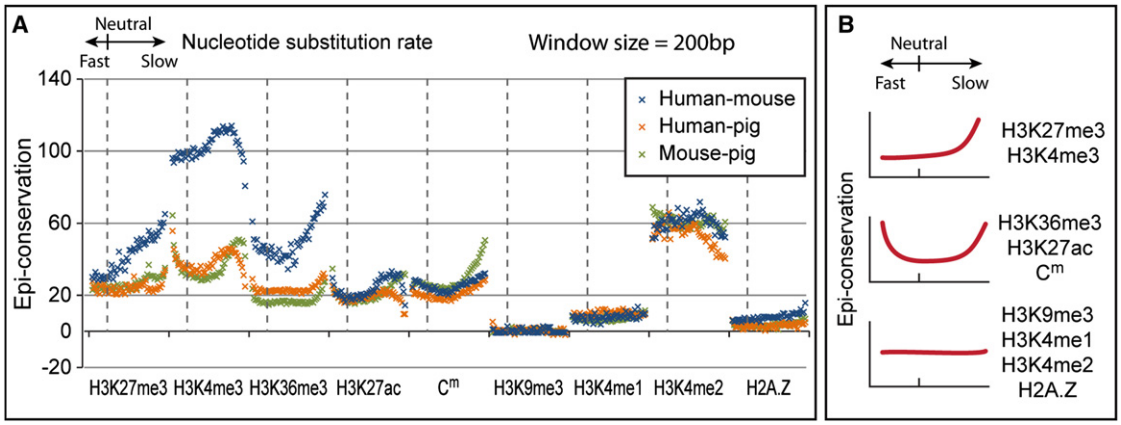

Figure 3. Global Comparison of Genomic and Epigenomic Conservations

(A) The human genome was categorized into 50 distinct sets by nucleotide substitution rates (x axis). These sets were ordered from the fastest changing (1st), to neutral (17th), and to slowest changing (50th). Epi-conservation levels by human-mouse (green) and human-pig (orange) comparisons are plotted on the y axis. Similarly, the mouse genome was categorized into 50 sets, and the epi-conservation levels in a mouse-pig comparison were plotted (blue).

(B) Schematic representations of the correlations between sequence selection and epi-conservation. Some epi-marks exhibit a U-shaped correlation, whereas others can be represented by the right half or the flat bottom of the U curve.

See also Figure S2.

The "U-Shaped" Correlation between Epigenomic and Genomic Conservations

To globally examine the relationship between genomic evolution and epigenomic changes, we categorized the human genome into 50 distinct sets of sequence segments and ordered these sets by nucleotide substitution rate (see Experimental Procedures). We then identified the epigenomic conservation levels in every set (Figures 3A and S2G–S2J). The enhancer mark H3K27ac, gene-body mark H3K36me3, and C^m exhibited increased conservation in both the accelerated substitution rate (rapidly changing) and reduced substitution rate (conserved) sets. H3K4me3 and H3K27me3 exhibited increased conservation in reduced substitution rate sets. In summary, a U-shaped correlation was observed between genomic selection and epigenomic conservation. The large portion of evolutionary neutral or near-neutral sequences exhibit a baseline epigenomic conservation level that forms the bottom of the U shape, and the sequences with accelerated or reduced substitution rates, respectively, exhibit enhanced epigenomic conservation, making the two ends of the U shape tilt upward (Figure 3B).

The increased conservation levels for H3K27ac, H3K36me3, and C^m in rapidly changing sequences indicate that epigenomic conservation is not completely determined by interspecies sequence similarity. To further test this hypothesis, we directly correlated epigenomic conservation with interspecies sequence similarity. This is a different test from the vertebrate conservation analysis (Figure 3A) because the nucleotide substitution rate estimated from 46 vertebrates does not necessarily correlate with pairwise sequence similarity between two species (Prabhakar et al., 2008). Except for H3K9me3, pairwise comparisons among human, mouse, and pig genomes consistently rule out a direct correlation between sequence similarity and epigenomic conservation (p value < 10^{-20}) (Figure S2L). These data indicate that either pairwise alignment is not sufficient to detect epigenomic conservation or epigenomic conservation is not a simple consequence of sequence similarity. Therefore, we hypothesized that either: (1) some epi-modifications may directly facilitate nucleotide substitution or (2) some conserved epi-modifications may buffer negative selective pressure, providing the genome greater freedom to change. Consistent with the former hypothesis, C^m is more mutagenic than C (Coulondre et al., 1978); the C→T change occurs more frequently than other changes between human and chimpanzee genomes, and such a change depends on local GC content (Jiang and Zhao, 2006). Because mechanisms that associate histone modifications with DNA mutations remain unidentified, we explored the plausibility of the latter hypothesis.

To buffer sequence changes from negative selective pressure, the epigenome must buffer genomic changes from generating phenotypic outcomes through, for example, concomitant transcriptome changes. To explore this possibility, we started by asking whether the same combination of epi-modifications is predictive of gene expression in every species. In each species, we used a linear regression model to fit the expression value of every gene to the nine measured epi-mark intensities in its promoter, and we used a model selection procedure to choose the epi-modifications that are predictive of gene expression. With only four epi-mark intensity values, the expression of every gene can be predicted in each species (largest p value < 10^{-16}) (Figure S3A). The models did not overfit (Figure S3D), and the epi-marks to expression predictive power matches 52.7%–81.3% of using one RNA-seq data set to predict another (as measured by R^2; Figure S3C). The epi-modifications predictive of gene expression levels were almost identical among humans, mice, and pigs, including H3K4me3, H3K36me3, H3K27me3, and H3K27ac (Figure S3B). The only exception was that, in pigs, H3K9me3 replaced H3K27ac in the final

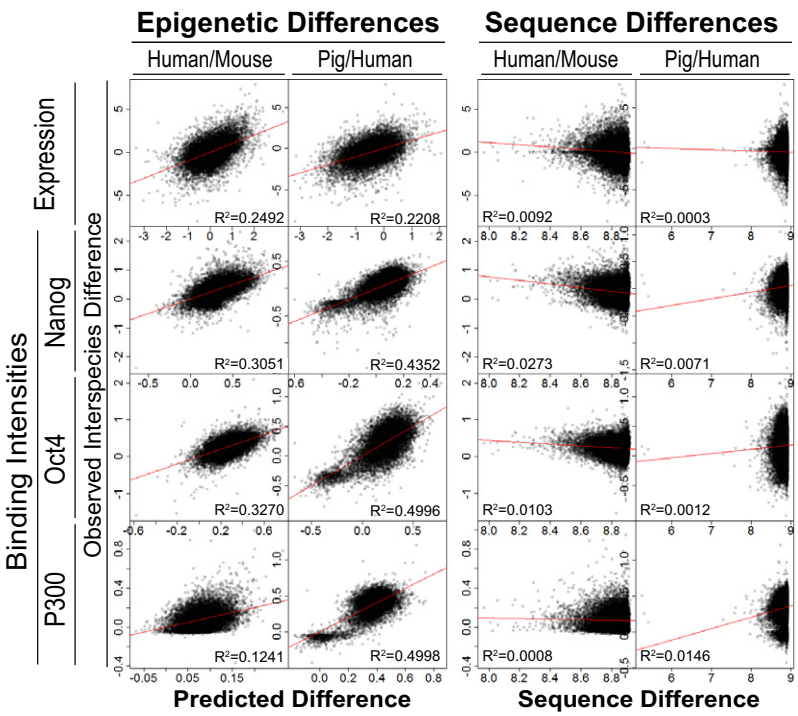

Figure 4. Correlations among Evolutionary Changes of Epi-Modification Intensities, Gene Expression Levels, TF Binding Intensities, and Genomic Sequences

(Left) Evolutionary changes of epi-modification intensities are predictive of gene expression changes and TF binding intensity changes. x axis, predicted gene expression or TF binding intensity changes with a linear model of interspecies epi-intensity changes; y axis, observed interspecies changes. (Right) Scatter plots between interspecies gene expression difference (y axis) and promoter sequence difference (x axis). For every orthologous gene pair, sequence difference was measured by log(m)-log(n), where m is the maximum log blastn score of all orthologous promoters (4,000 bp centered at TSS), and n is the blastn score of the orthologous promoter pair under consideration.

R^2, square of the sample correlation coefficient. See also Figures S3 and S4 and Table S2.

model due to a large correlation (0.91) between H3K4me3 and H3K27ac data. These data show that gene expression can be predicted by a conserved set of epi-marks, reiterating the idea that epigenomic conservation can be used to study gene regulation.

Interspecies Epigenomic Changes Are Predictive of Interspecies Changes of Gene Expression and Transcription Factor Binding

We then asked whether interspecies epigenomic changes are correlated to transcriptomic changes. The interspecies differences of epi-modification intensities are predictive of interspecies gene expression differences (p value < 10^{-16}) (Figure 4). In a control experiment, when interspecies epi-mark intensity differences were considered, the original epi-modification intensities in each species did not contribute to further explain gene expression difference (Table S2). This implies that the epigenomic information associated with changes of gene expression between species is distinct from the epigenomic information associated with gene expression variation within a single species. For example, the human RNA-processing gene *DDX17* (FPKM = 70.89) expressed higher than the human *AVPI1* (FPKM = 2.99), a gene related to mitogen-activated protein kinase (MAPK) activity. Although this within-species expression difference can be predicted by human epi-mark intensities, the interspecies differences of expression for these genes cannot be predicted (Figure S3E). However, the moderate decrease of *DDX17* expression (from humans to mice) and the 438% increase of *AVPI1* expression were associated with interspecies changes of epi-modifications. In contrast, published cross-species analysis of tissue expression data found no identifiable sequence-to-expression correlation in vertebrates (Chan et al., 2009). Similarly, we could not find any apparent correlation between interspecies sequence difference and expression difference by using a simple model (Figure 4). Take the calcium channel, voltage-dependent, gamma subunit 7 (*CACNG7*) gene as an example: its 12,000 bp upstream sequences are conserved in humans, mice, and pigs, whereas its expression levels are high in humans and mice but low in pigs. However, marks of active promoters, H3K4me2/3, are conserved in humans and mice, but not in pigs, providing a correlation for the interspecies expression differences with epi-marks (Figure 5).

One possible mechanism to explain how transcriptomic evolutionary changes might relate to epigenomic changes is that epigenomic changes may influence TF binding intensities. In each species, epi-mark intensities are strongly correlated with the total binding intensities of all assayed TFs (Figures S4A–S4C). As a representative TF, NANOG's in vivo binding sites are surrounded by cell-type-specific epigenomic patterns (Figure S4D). The evolutionary changes of epi-mark intensities are predictive of binding intensity changes of OCT4 (p value < 10^{-22}), NANOG (p value < 10^{-22}), and P300 (p value < 10^{-22}) (Figure 4, left). The weak association between interspecies sequence changes and TF binding changes (Figure 4, right) is related to the finding that very few OCT4 and NANOG proteins bind to orthologous sequences (Kunarso et al., 2010). These data are in line with the hypothesis that the epigenome may buffer the evolutionary changes of genome sequences from generating evolutionary changes in TF binding, gene expression, and phenotypic outcomes. Such a process would alleviate evolutionary selection pressure on sequence changes by masking genetic changes from immediate phenotypic changes. These data do not rule out the idea that genetic changes can introduce coding changes and therefore can be evolutionarily selected.

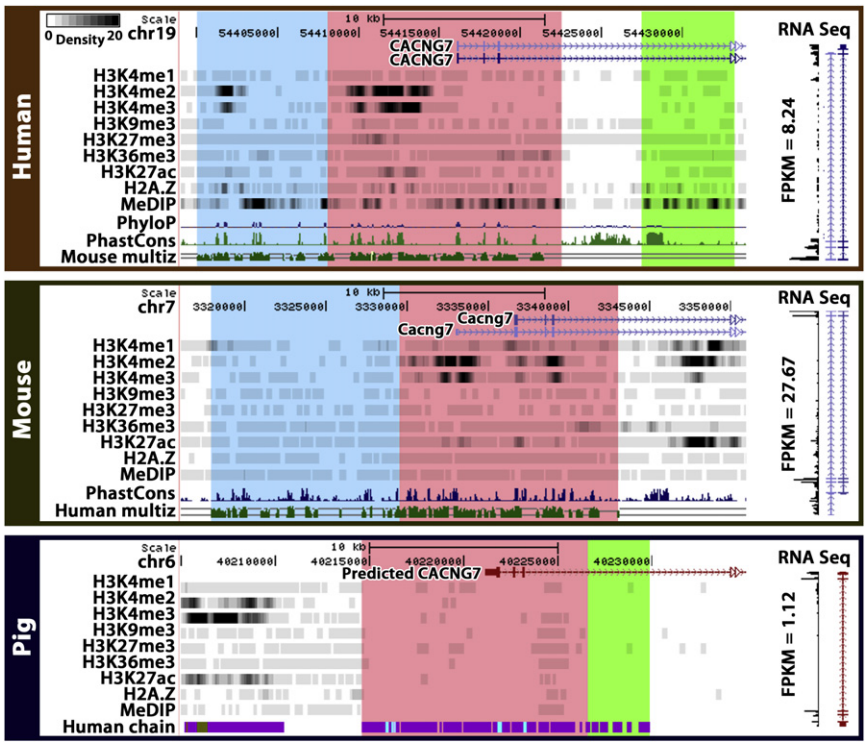

Figure 5. An Example of Correlated Interspecies Epi- and Gene Expression Changes

The genomic and epigenomic neighborhoods of CACNG7 in three species are displayed by the Comparative Epigenome Browser. The orthologous regions determined by the LiftOver program are shaded in the same color. Densities of ChIP-seq counts and MeDIP-seq counts are plotted in grayscale. Approximately 12 kb upstream sequences of the gene are conserved (pink). H3K4me2 and H3K4me3 are present and conserved in upstream regions in humans and mice, coinciding with conserved expression of the gene (RNA-seq data drawn vertically on the right). The conserved pig upstream sequence (in pink) is devoid of H3K4me2/3 marks, coinciding with a much lower expression level of the pig gene. See also Figure S3.

Evolutionarily Conserved Colocalization of Different Epigenomic Marks Defines Several Classes of *cis*-Regulatory Sequences

We set out to test the functions of evolutionarily conserved epi-mark combinations (Figure 2 and Table 1) by using a cell differentiation assay. During ESC differentiation, the directions of epigenomic changes and expression changes of nearby genes were expected to reflect the function of an epi-mark combination (Rada-Iglesias et al., 2011; Zentner et al., 2011). We differentiated mESCs into mesendoderm cells (Tada et al., 2005; Yasunaga et al., 2005), a lineage in which the dynamic changes of the epigenome has not been examined. On day 6 of differentiation, almost all cells expressed the mesendoderm protein Goosecoid (GSC) and endoderm protein SOX17 (Figure S5A) and exhibited typical mesendoderm morphology (Figure S5B). Pluripotency genes *Oct4*, *Sox2*, and *Nanog* were downregulated, and mesendoderm markers *Gsc* and *Chordin* (*Chrd*) and endoderm markers *Foxa2*, *Sox17*, *Lim1*, and *Hnf4* were upregulated (Figure S5C). We mapped the epigenomes and the transcriptomes of day 4 and day 6 differentiated cells by using ChIP-seq and RNA-seq, adding a total of 18 data sets (38 billion bases) to our overall data set (Table S1, bottom and Figure S6).

Seven pairs of epi-marks were identified as conserved comodifications in pluripotent stem cells, namely H2A.Z+H3K4me2/3, H3K27ac+H3K4me1/2, H3K27ac+H3K4me2/3, H3K27me3+H3K4me1/2, H3K27me3+H3K4me2/3, H3K36me3+H3K27ac, and H3K36me3+H3K4me1 (Table 1). The bivalent domain (H3K27me3+H3K4me2/3) was the most conserved epi-mark combination among all 36 pairs of modifications (Figure 2B), lending credence to the approach of using epigenomic comparison for identifying gene regulatory regions. To illustrate how our time course experiment can reveal the functions of an epi-mark combination, we examined how bivalent domains regulate the early stages of mESC differentiation in a lineage-specific manner. It resulted that not all bivalent domains behave the same. Four subclasses of bivalent domains with different dynamic behaviors were discovered. On day 6, the majority of sequences either retained both marks (40.7% of the sequences) or lost H3K27me3 while retaining H3K4me2/3 (36.7%). These sequences were preferentially located near transcription start sites (TSS) (Figure S7A, I and III). As expected, the genes whose promoters lost H3K27me3 but retained H3K4me2/3 exhibited higher expression (Figure 6A, red line) than those that kept both marks (purple line), and those that kept both marks were higher than those that kept H3K27me3 but lost H3K4me2/3 (green line). These data indicate that there are subclasses of bivalent promoters, which may be activated, still-poised, repressed, or suffer a loss of both marks during mesendoderm formation. We further examined the functions of genes regulated by each subclass of bivalent promoters. The genes with activated promoters are enriched for Gene Ontology (GO) terms of "transforming growth factor β (TGF-β) and receptor binding" (p value < 10^{-17}), "mesoderm formation" (p value < 10^{-13}), and "positive regulation of BMP pathway" (p value < 10^{-9}), consistent with the mesendoderm differentiation process. For example, the bivalent promoters of mesendoderm marker gene *Gsc* and mesoderm regulatory gene *Bmp7* were activated (Figure S6). On the other hand, "neuron fate commitment" was enriched in both the still-poised (p value < 10^{-118}) and the repressed (p value < 10^{-13}) subclasses. These data reveal an intricate coordination among distinct subclasses of bivalent promoters that facilitates lineage-specific differentiation.

We expected the conserved comodifications H3K27me3+H3K4me1/2 to mark poised enhancers (Bernstein et al., 2006; Rada-Iglesias et al., 2011). On day 6 of differentiation, 31.8%

Table 1. Most Conserved Comarks for Each Epi-Modification

Epi-Mark	Most Conserved Comarks	Putative Function	Biological Inference
H2A.Z	H3K4me2/3	poised promoter	negatively associated with gene activation in ES cells and during ESC differentiation
H3K27ac	H3K4me1/2/3	active enhancer (me1/me2) or promoter (me2/me3)	H3K27ac marks promoters as well
H3K27me3	H3K4me1/2/3	poised enhancer (me1/me2) or bivalent promoter (me2/me3)	poised enhancers regulate as many genes as bivalent promoters do
H3K36me3	H3K27ac, H3K4me1	active enhancer	not correlated with H3K27me3 may be a neglected mark of active enhancers
H3K4me1	H3K27me3, H3K4me2	poised enhancer	
H3K4me2	H3K4me3	active or poised regulatory regions	
H3K4me3	H3K4me2	active or poised regulatory regions	
H3K9me3		repressed region	negatively correlated with sequence conservation
C^m			C^m either only mildly influences gene regulation or influences it in a way that is independent from histone modifications

Comments in the Biological Inference column only relate to mammalian species.

of H3K27me3+H3K4me1/2-marked regions removed repression mark H3K27me3 and kept activation mark H3K4me1/2. The genes next to this subset of hypothetically activated enhancers exhibited increased expression on days 4 and 6 (Figure 6B, red line), which was even greater than the expression of genes associated with other dynamic epigenomic patterns (purple, green, and blue lines). Assuming that an enhancer may regulate a target gene within 50,000k bp, we estimated that about 4,618 genes could be regulated by poised enhancers. The number is on the same order as that of bivalent-promoter-regulated genes (~5,194). In addition, conserved H3K27ac+H3K4me1/2 and H3K27ac+H3K4me2/3 marked active enhancers and promoters as expected (Figures 6C and 6D).

The most conserved comark of H2A.Z was H3K4me2/3. H2A.Z is a variant of H2A and is required for early mammalian development (Faast et al., 2001). Despite the usual assumption that H2A.Z is associated with active gene expression in multicellular organisms (Weber et al., 2010; Zemach et al., 2010), we found H2A.Z not to be positively associated with gene expression levels in mESCs (Pearson correlation = −0.0066; Figure S7D). This is consistent with the lack of global anticorrelation of H2A.Z and C^m in all three mammals (Figure 2A). Thus, H2A.Z could be a repressor mark in mammals, and H2A.Z+H3K4me2/3 could mark poised promoters rather than active promoters as is generally assumed. Indeed, H2A.Z+ and H3K4me2/3- promoters were less active than H2A.Z- and H3K4me2/3- promoters, and H2A.Z+ and H3K4me2/3+ promoters were less active than H2A.Z- and H3K4me2/3+ (Figure S7C). More importantly, during differentiation, the H2A.Z+ and H3K4me2/3+ promoters in ESCs that lost the H2A.Z mark became more active (Figure 6E, red line), and those that lost H3K4me2/3 but kept H2A.Z were downregulated (Figure 6E, green line). Thus, we propose that H2A.Z is a repressor mark in mammalian pluripotent stem cells and that H2A.Z+H3K4me3 marks a class of poised promoters.

H3K36me3, H3K27ac, and H3K4me1/2 exhibited pairwise conservation. Whereas H3K4me1 and H3K27ac were previously associated with enhancers, H3K36me3 has been typically regarded as a mark for actively transcribed regions. The conserved colocalization of H3K36me3 with H3K27ac and H3K4me1/2 tempted us to explore H3K36me3 as an enhancer mark as well. Consistent with this thought, active enhancers make transcripts (eRNA) (Ren, 2010); H3K36me3 could be associated with any transcribed regions, including active enhancers. If this hypothesis holds, we would predict that H3K36me3 should avoid overlapping with bivalent (poised) enhancers. Indeed, the epi-mark that has the least colocalization with H3K27me3 is H3K36me3 (Figure 2A). During differentiation, the genes near (not overlapping with) sequences that lost H3K36me3 and H3K27ac (or H3K4me1) exhibited lower expression than those with one mark lost, which in turn exhibited lower expression than those that retained two marks (Figures 6F and 6G). Thus, we propose that H3K36me3, when coappearing with H3K27ac or H3K4me1/2, is a mark of active enhancers. In summary, it is powerful to use epi-mark combinations to annotate regulatory sequences and to form hypotheses about their functions. The difficulties of having too many epi-mark combinations and not knowing how to distinguish random versus functional colocalizations can be overcome by using evolutionary conservation.

DISCUSSION

Many of the functional regions in the human genome have been identified by comparative genomic approaches based on evolutionary principles. Here, we provide a view of the evolutionary properties of the mammalian epigenome and illustrate coevolutionary relationships among genomes, transcriptomes, and epigenomes. These results show how comparative epigenomics, an emerging field that studies evolutionary patterns of

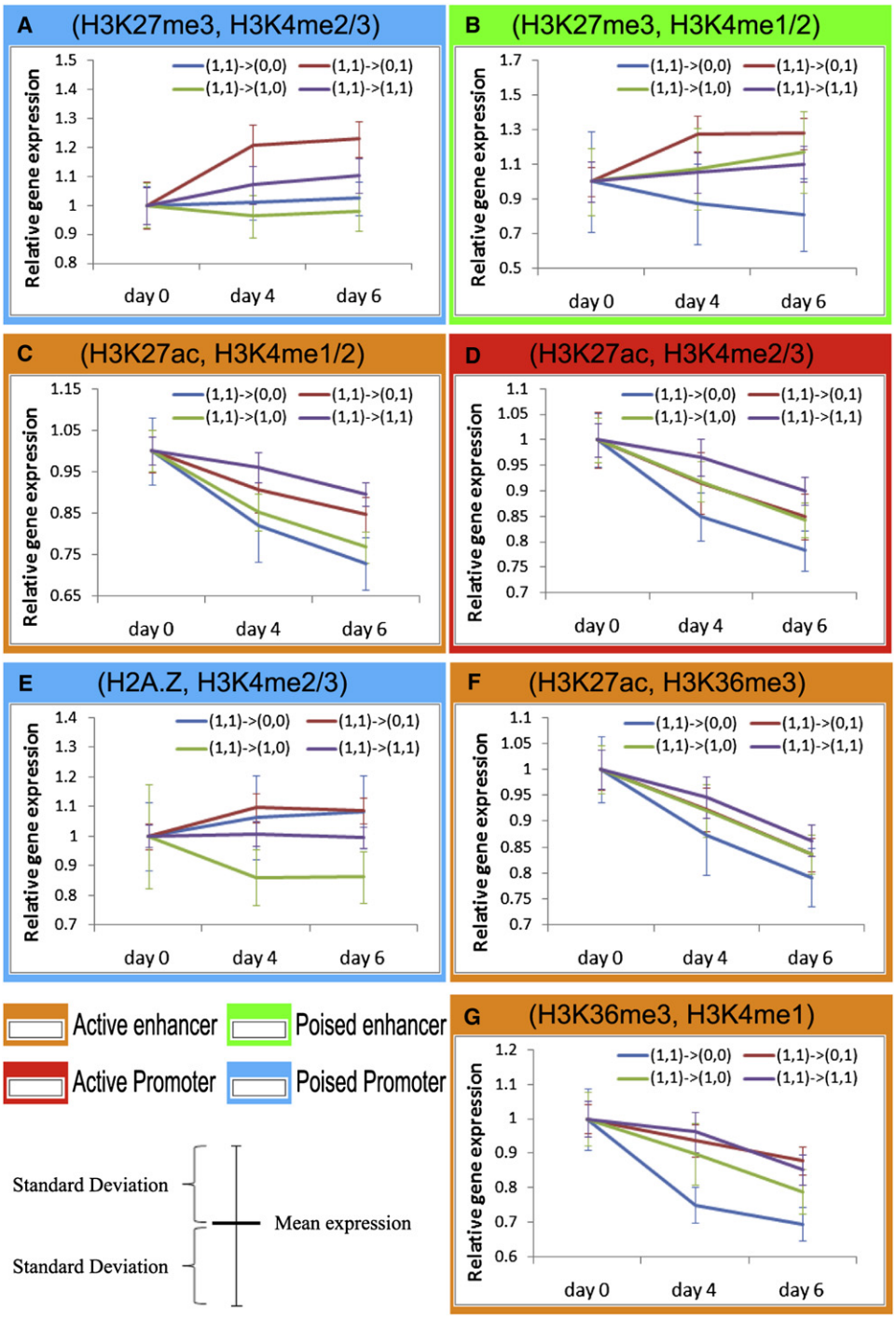

Figure 6. Epi-Changes and Gene Expression Changes during Differentiation

(A–G) Each panel represents a set of genomic regions associated with a pair of epi-marks. Each set of regions is categorized into four subclasses, i.e., kept both marks during differentiation (1,1→1,1), lost the first mark (1,1→0,1), lost the second mark (1,1→1,0), and lost both marks (1,1→0,0). For example, the red line (1,1→0,1) in (A) (H3K27me3, H3K4me2/3) represents sequences with loss of H3K27me3 (the first sign changes from 1 to 0) and retention of H3K4me2/3 (the second sign stays at 1). Relative gene expression values of the nearest genes to the comarked regions are plotted on the y axis.
Error bars show SD of the mean. See also Figures S5–S7 and Table S1.

epigenomes, can use epigenomic information to functionally annotate genomes.

We compared interspecies epigenomic changes to both genomic and transcriptomic changes. Our data show that the degree of epigenomic conservation is not always correlated with the degree of genomic conservation but that epigenomic conservation can yield additional information to genomic conservation. More importantly, the conservation levels of epigenomes are indicative of the conservation levels of gene expression, further illustrating that epigenomic comparison can shed light on regulatory functions of the genome.

Evolution appears to have left traces on mammalian epigenomes, and one identifiable trace is in the combination of epi-marks. Some combinations coappear (colocalize) in a conserved manner. The conservation of colocalized epi-marks is much stronger than the conservation level of each epi-mark, thus making the combinations computationally identifiable. We used a stem cell differentiation assay to test the regulatory functions of the conserved epi-mark combinations. These tests confirmed the regulatory functions of all (seven out of seven) conserved epi-mark combinations, suggesting that interspecies comparison can efficiently distinguish functional colocalization of epi-marks from nonfunctional combinations. This highlights an efficient approach to identify functional epi-mark combinations from a large (combinatorial) number of candidate combinations.

Our studies also reveal the phenomenon that the conservation levels of three epi-marks increase for genomic regions with accelerated sequence changes (related to positive selection [Sabeti et al., 2006]). This is a surprising finding because we expected the epigenomic conservation to be weak at locations where the genomic conservation is also weak. This suggests that epigenomic conservation may be used in conjunction with sequence comparison to identify positively selected regions to reveal functional sequences that make humans unique. Finally, the correlated evolutionary changes of the epigenome, the transcriptome, and TF binding suggest the functional importance of the epigenome in mammalian transcription networks (TNs). This may explain the limited successes in human TN reconstruction using only the information of DNA sequence motifs and gene expression, which were sufficient for reconstruction of yeast TNs (Segal et al., 2003).

EXPERIMENTAL PROCEDURES

Processing Sequencing Data

ChIP-seq and RNA-seq data were mapped to genome assemblies hg19, mm9, and susScr2 by using Bowtie software. MeDIP-seq and MRE-seq were analyzed by using methods described in Maunakea et al. (2010). Epi-modification intensities of a 200 bp region were estimated based on the overlapping ChIP-seq reads in this region. A 200-bp-long sliding window was used to scan every genome to compute TF binding and epi-modification intensities. All analyses were repeated with 100 and 300 bp sliding windows, as well as a 200 bp window with a different conservation threshold (Figures S2A–S2F).

Global Analysis of Genomic and Epigenomic Conservations

Categorizing the Human Genome by Rate of Sequence Change

The human genome was divided into 15 million 200 bp segments. A PhyloP score was computed for every base from 46 vertebrate genomes (Cooper et al., 2005; Pollard et al., 2010), and an average PhyloP score was computed for each 200 bp segment. These genomic segments were put into 50 equal-sized sets with increasing average PhyloP scores. The first set with the smallest PhyloP scores are the fastest-changing sequences. The last set with the largest PhyloP scores are the most conserved.

Quantifying Epigenomic Conservation

In the human-mouse comparison, a 200 bp human genomic segment was determined to be epigenomically conserved if the mouse orthologous sequence was marked by the same epi-modification. For each set of genomic segments, the average number of epigenomically conserved segments was calculated. This average number was then divided by its expectation to obtain a ratio. The expectation was derived from an independence model in which epigenomic conservation was assumed to be independent of sequence conservation. The ratio was used as a quantitative measure of epigenomic conservation for each sequence set. Human-pig and mouse-pig comparisons were done similarly. All analyses were repeated with 100 and 300 bp segment sizes, as well as 200 bp segments with a different conservation threshold (Figures S2G–S2J). This conservation threshold was only used to determine the orthologous sequences on which epigenomic data should be retrieved for estimating epigenomic conservation levels. To account for gene conversion, transposon, and multigene families, all "one-to-many" alignable sequences between any two species were removed from this analysis. Whether any sequence has a one-to-one or one-to-many alignment to another species was determined by processing UCSC pairwise alignment chain files.

Conservation Level and Sequence Identity

Sequence conservation level was estimated by Phastcons (Siepel et al., 2005) scores computed by the UCSC Genome Browser group using 46 vertebrate genomes. The human genome was put into strongly, moderately, and weakly conserved groups with Phastcons scores in [1, 1], [0.5, 1] and [0.1, 0.5], respectively. Human-mouse and human-pig orthologous regions were identified by the LiftOver results stored in the UCSC Genome Browser. Sequence identities of these LiftOver regions were estimated based on the proportion of matched nucleotides in local alignments to the total length of the two sequences.

Linear Regression

In each species, the expression level of every gene is regressed to the eight epi-modification intensities on the upstream sequence of the gene. A forward stepwise model selection procedure was used to select the epi-modifications that were predictive. In evolutionary analysis, the interspecies gene expression differences were regressed to epi-modification differences. The same forward selection procedure was applied.

Control for Hot Spots of Aberrant Epigenomic Reprogramming

All analyses were repeated with and without the genomic regions with epigenetic differences between iPS cells and ESCs (Table S3). The differences in computed statistics were all below 0.1% of the SDs of these statistics, making no changes to the order of computed p values.

In Vitro Differentiation of mESCs

Undifferentiated E14 mESCs were cultured under feeder-free conditions. Guided differentiation was carried out as previously described (Tada et al., 2005). Briefly, 2×10^5 cells were seeded on Collagen-IV-coated 10 cm dishes (BD, 08-774-33) in serum-free medium ESF-B (Itochu Corporation) supplemented with 0.1% bovine serum albumin (BSA), 50 mM 2Me, and 10 ng/ml Activin A. Medium was changed every day.

Immunofluorescence Staining

10,000 to 20,000 cells were seeded on Collagen-IV-coated 35 mm dishes (Ibidi 45074) with the same medium as in guided differentiation. Cells were fixed with 4% paraformaldehyde. Primary antibodies for GSC (Origene, TA500087) and Sox17 (Millipore 09-038) were incubated at the same time for 2 hr at 37°C. Secondary antibodies, goat anti-mouse conjugate Alexa 568 (Invitrogen A-11031) and goat anti-rabbit conjugate Alexa 488 (Invitrogen A-11034), were each subsequently incubated for 2 hr at 37°C. DNA was stained by Hoechst 33342 (Invitrogen, H3570) for 15 min at room temperature. Images were taken by a Zeiss LSM 700 microscope.

ACCESSION NUMBERS

The GEO accession number for sequencing data reported in this paper is GSE36114.

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and three tables and can be found with this article online at doi:10.1016/j.cell.2012.04.029.

ACKNOWLEDGMENTS

This work was supported by NIH DP2-OD007417, NSF DBI 09-60583, and Sloan Research Fellowship to S.Z.; and by NIH 5U01ES017154, March of Dimes Foundation, Edward Jr. Mallinckrodt Foundation, DDRCC P30DK52574, and Siteman Cancer Center P50CA134254 to T.W. The authors thank Wei Huang and Darina McDee and Drs. Lisa Stubbs, Tetsuya Tanaka, Phillip Newmark, Rex Gaskins, Gene Robinson, David Clayton, Nigel Goldenfeld, Fei Wang, and Taekjip Ha for useful discussions.

Received: May 9, 2011
Revised: September 10, 2011
Accepted: April 4, 2012
Published: June 7, 2012

REFERENCES

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

Campos, E.I., and Reinberg, D. (2009). Histones: annotating chromatin. Annu. Rev. Genet. *43*, 559–599.

Chan, E.T., Quon, G.T., Chua, G., Babak, T., Trochesset, M., Zirngibl, R.A., Aubin, J., Ratcliffe, M.J., Wilde, A., Brudno, M., et al. (2009). Conservation of core gene expression in vertebrate tissues. J. Biol. *8*, 33.

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

Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., and Sidow, A.; NISC Comparative Sequencing Program. (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. *15*, 901–913.

Coulondre, C., Miller, J.H., Farabaugh, P.J., and Gilbert, W. (1978). Molecular basis of base substitution hotspots in Escherichia coli. Nature *274*, 775–780.

Ernst, J., and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat. Biotechnol. *28*, 817–825.

Faast, R., Thonglairoam, V., Schulz, T.C., Beall, J., Wells, J.R., Taylor, H., Matthaei, K., Rathjen, P.D., Tremethick, D.J., and Lyons, I. (2001). Histone variant H2A.Z is required for early mammalian development. Curr. Biol. *11*, 1183–1187.

Goren, A., Ozsolak, F., Shoresh, N., Ku, M., Adli, M., Hart, C., Gymrek, M., Zuk, O., Regev, A., Milos, P.M., and Bernstein, B.E. (2010). Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA. Nat. Methods *7*, 47–49.

Guillemette, B., Bataille, A.R., Gévry, N., Adam, M., Blanchette, M., Robert, F., and Gaudreau, L. (2005). Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. PLoS Biol. *3*, e384.

Hardison, R.C. (2003). Comparative genomics. PLoS Biol. *1*, E58.

Jegga, A.G., Inga, A., Menendez, D., Aronow, B.J., and Resnick, M.A. (2008). Functional evolution of the p53 regulatory network through its target response elements. Proc. Natl. Acad. Sci. USA *105*, 944–949.

Jiang, C., and Zhao, Z. (2006). Directionality of point mutation and 5-methylcytosine deamination rates in the chimpanzee genome. BMC Genomics *7*, 316.

Karlić, R., Chung, H.-R., Lasserre, J., Vlahoviček, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. Proc. Natl. Acad. Sci. USA *107*, 2926–2931.

Kimura, M. (1968). Evolutionary rate at the molecular level. Nature *217*, 624–626.

Kunarso, G., Chia, N.Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.S., Ng, H.H., and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat. Genet. *42*, 631–634.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.

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

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

Mikkelsen, T.S., Xu, Z., Zhang, X., Wang, L., Gimble, J.M., Lander, E.S., and Rosen, E.D. (2010). Comparative epigenomic analysis of murine and human adipogenesis. Cell *143*, 156–169.

Nichols, J., and Smith, A. (2009). Naive and primed pluripotent states. Cell Stem Cell *4*, 487–492.

Pollard, K.S., Salama, S.R., King, B., Kern, A.D., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J.S., Bejerano, G., Baertsch, R., et al. (2006). Forces shaping the fastest evolving regions in the human genome. PLoS Genet. *2*, e168.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. *20*, 110–121.

Prabhakar, S., Visel, A., Akiyama, J.A., Shoukry, M., Lewis, K.D., Holt, A., Plajzer-Frick, I., Morrison, H., Fitzpatrick, D.R., Afzal, V., et al. (2008). Human-specific gain of function in a developmental enhancer. Science *321*, 1346–1350.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.

Raisner, R.M., Hartley, P.D., Meneghini, M.D., Bao, M.Z., Liu, C.L., Schreiber, S.L., Rando, O.J., and Madhani, H.D. (2005). Histone variant H2A.Z marks the 5′ ends of both active and inactive genes in euchromatin. Cell *123*, 233–248.

Ren, B. (2010). Transcription: enhancers make non-coding RNA. Nature *465*, 173–174.

Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D., and Lander, E.S. (2006). Positive natural selection in the human lineage. Science *312*, 1614–1620.

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

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

Tada, S., Era, T., Furusawa, C., Sakurai, H., Nishikawa, S., Kinoshita, M., Nakao, K., Chiba, T., and Nishikawa, S. (2005). Characterization of mesendoderm: a diverging point of the definitive endoderm and mesoderm in embryonic stem cell differentiation culture. Development *132*, 4363–4374.

Tan, M., Luo, H., Lee, S., Jin, F., Yang, J.S., Montellier, E., Buchou, T., Cheng, Z., Rousseaux, S., Rajagopal, N., et al. (2011). Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. Cell *146*, 1016–1028.

Weber, C.M., Henikoff, J.G., and Henikoff, S. (2010). H2A.Z nucleosomes enriched over active genes are homotypic. Nat. Struct. Mol. Biol. *17*, 1500–1507.

West, F.D., Terlouw, S.L., Kwon, D.J., Mumaw, J.L., Dhara, S.K., Hasneen, K., Dobrinsky, J.R., and Stice, S.L. (2010). Porcine induced pluripotent stem cells produce chimeric offspring. Stem Cells Dev. *19*, 1211–1220.

Wilson, M.D., Barbosa-Morais, N.L., Schmidt, D., Conboy, C.M., Vanes, L., Tybulewicz, V.L., Fisher, E.M., Tavaré, S., and Odom, D.T. (2008). Species-specific transcription in mice carrying human chromosome 21. Science *322*, 434–438.

Yasunaga, M., Tada, S., Torikai-Nishikawa, S., Nakano, Y., Okada, M., Jakt, L.M., Nishikawa, S., Chiba, T., Era, T., and Nishikawa, S. (2005). Induction and monitoring of definitive and visceral endoderm differentiation of mouse ES cells. Nat. Biotechnol. *23*, 1542–1550.

Zemach, A., McDaniel, I.E., Silva, P., and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science *328*, 916–919.

Zentner, G.E., Tesar, P.J., and Scacheri, P.C. (2011). Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome Res. *21*, 1273–1283.

Zilberman, D., Coleman-Derr, D., Ballinger, T., and Henikoff, S. (2008). Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. Nature *456*, 125–129.