A Comparative Encyclopedia of DNA Elements in the Mouse Genome

Published in final edited form as:
Nature. 2014 November 20; 515(7527): 355–364. doi:10.1038/nature13992.

A Comparative Encyclopedia of DNA Elements in the Mouse Genome

A full list of authors and affiliations appears at the end of the article.

Summary

As the premier model organism in biomedical research, the laboratory mouse shares the majority of protein-coding genes with humans, yet the two mammals differ in significant ways. To gain greater insights into both shared and species-specific transcriptional and cellular regulatory programs in the mouse, the Mouse ENCODE Consortium has mapped transcription, DNase I hypersensitivity, transcription factor binding, chromatin modifications, and replication domains throughout the mouse genome in diverse cell and tissue types. By comparing with the human genome, we not only confirm substantial conservation in the newly annotated potential functional sequences, but also find a large degree of divergence of other sequences involved in transcriptional regulation, chromatin state and higher order chromatin organization. Our results illuminate the wide range of evolutionary forces acting on genes and their regulatory regions, and provide a general resource for research into mammalian biology and mechanisms of human diseases.

Introduction

Despite the widespread use of mouse models in biomedical research[1], the genetic and genomic differences between mice and humans remain to be fully characterized. At the sequence level, the two species have diverged substantially: approximately one half of

[#]To whom correspondence should be addressed: biren@ucsd.edu (B.R.); mbeer@jhu.edu (M.B.); ross@bx.psu.edu (R.C.H.); gilbert@bio.fsu.edu (D.M.G.); gingeras@cshl.edu (T.R.G.); roderic.guigo@crg.cat (R.G.); mpsnyder@stanford.edu (M.S.); jstam@u.washington.edu (J.S.).
[*]These authors contributed equally to the work
Current address for Feng Yue: Department of Biochemistry and Molecular Biology, School of Medicine, The Pennsylvania State University, Hershey, PA 17033, USA.
Current address for Rebecca Lowdon: Washington University in St. Louis, St. Louis, MO 63108
Current address for Leslie Adams: University of North Carolina Gillings School of Global Public Health, Chapel Hill, NC 27599, USA.
Current address for Weisheng Wu: Bioinformatics Core, Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA.

Note: The role of the NHGRI Project Management Group in the preparation of this paper was limited to coordination and scientific management of the mouse ENCODE consortium.

Author Contributions: F.Y., Y.C., A.B., J.V., W.W., T.R., M.B., R.C.H., J.S., M.P.S., R. G., T.R.G., D.M.G., and B.R. led the data analysis effort, R.S, Z.M., C.D., B.P., Y.S. R.C.H., J.S., M.P.S., R. G., T.R.G., D.M.G., and B.R.led the data production. F.Y., M.A.B., L.E., C.Y., P.C., A.B., A.K., S.L., Y.L., J.V., R.S., R.E.T., E.R., E.H., A.P.R., S.N., R.H., W.W., T.M., R.S.H., C.J., A.M., B.D.P., T.R., T.K., D.L., O.D., J.T., C.Z., A.D., D.P., S.D., P.P., J.L., G.B., A.T., K.B., M.P., P.F., and J.H. analyzed data. Y.S., D.M. L.P., Z.Y., S.K. Z.M., T.K., G.E., J.L., S.M.W. R.K., M.A.B., S.L., Y.L., M.Z., R.B., M.T.G., A.J., S.V., K.L., D.B., F.N., M.D., T.C., R.S.H., P.J. S., M.S.W., T.A.R., E.G., A.S., T.K., E.H., D.D., M.D.B., L.S., A.R., S.J., R.S., E.E.E., S.H.O., D.L., T.P., K.C., A.S., C.D., P.T., W.W., C.A.K., C.S.M., T.M., D.J., N.D., B.D.P., T.R., C.D., LH.S., M.F., J.D. produced data. F.Y., C.Y., W.W, T.R., B.D.P., S.L., Y.L., C.J., C.D., A.D., A.B., D.P., S.D., C.N., A.M., J.S., M.P.S., R. G., T.R.G., D.M.G., R.C.H., M.B., B.R. wrote the manuscript. The role of the NIH Project Management Group was limited to coordination and scientific management of the mouse ENCODE consortium.

human genomic DNA can be aligned to mouse genomic DNA, and only a small fraction (3–8%) is estimated to be under purifying selection across mammals [2]. At the cellular level, a systematic comparison is still lacking. Recent studies have revealed divergent DNA binding patterns for a limited number of transcription factors across multiple related mammals [3–6, 7,8], suggesting potentially wide-ranging differences in cellular functions and regulatory mechanisms[9,10]. To fully understand how DNA sequences contribute to the unique molecular and cellular traits in mouse, it is crucial to have a comprehensive catalog of the genes and non-coding functional sequences in the mouse genome.

Advances in DNA sequencing technologies have led to the development of RNA-seq, DNase-seq, ChIP-seq, and other methods that allow rapid and genome-wide analysis of transcription, replication, chromatin accessibility, chromatin modifications, and transcription factor binding in cells [11]. Using these large-scale approaches, the ENCODE consortium has produced a catalog of potential functional elements in the human genome [12]. Notably, 62% of the human genome is transcribed in one or more cell types [13], and 20% of human DNA is associated with biochemical signatures typical of functional elements including transcription factor binding, chromatin modification, and DNase hypersensitivity. The results support the notion that nucleotides outside the mammalian-conserved genomic regions could contribute to species-specific traits [6,12,14].

We have applied the same high throughput approaches to over 100 mouse cell types and tissues [15], producing a coordinated group of datasets for annotating the mouse genome. Integrative analyses of these datasets uncovered widespread transcriptional activities, dynamic gene expression and chromatin modification patterns, abundant *cis* regulatory elements, and remarkably stable chromosome domains in the mouse genome. The generation of these datasets also allowed an unprecedented level of comparison of genomic features of mouse and human. Described in the current manuscript and companion works, these comparisons revealed both conserved sequence features and widespread divergence in transcription and regulation. Some of the key findings are:

- Although much conservation exists, the expression profiles of many mouse genes involved in distinct biological pathways show considerable divergence from their human orthologs.
- A large portion of the *cis*-regulatory landscape has diverged between mouse and human, though the magnitude of regulatory DNA divergence varies widely between different classes of elements active in different tissue contexts.
- Mouse and human transcription factor networks are substantially more conserved than *cis*-regulatory DNA.
- Species-specific candidate regulatory sequences are significantly enriched for particular classes of repetitive DNA elements.
- Chromatin state landscape in a cell lineage is relatively stable in both human and mouse.
- Chromatin domains, interrogated through genome-wide analysis of DNA replication timing, are developmentally stable and evolutionarily conserved.

Results

Overview of data production and initial processing

To annotate potential functional sequences in the mouse genome, we used ChIP-seq, RNA-seq and DNase-seq to profile transcription factor binding, chromatin modification, transcriptome and chromatin accessibility in a collection of 123 mouse cell types and primary tissues (Fig. 1a, Supplementary Table 1–3). Additionally, to interrogate large-scale chromatin organization across different cell types, we also utilized a microarray-based technique to generate replication-timing profiles in 18 mouse tissues and cell types (Supplementary Table 3) [16]. Altogether, we produced over 1000 datasets. The list of the datasets and all the supporting material for this manuscript are also available at website http://mouseencode.org. Below we briefly outline the experimental approach and initial data processing for each class of sequence features.

RNA transcriptome—To comprehensively identify the genic regions that produce transcripts in the mouse genome, we performed RNA-seq experiments in 69 different mouse tissues and cell types with two biological replicates each (Supplementary Table 3, Supplemental Materials) and uncovered 436,410 contigs (Supplementary Table 4). Confirming previous reports [1317,18] and similar to the human genome, the mouse genome is pervasively transcribed (Fig. 1b), with 46% capable of producing polyadenylated messenger RNAs (mRNA). By comparison, 39% of the human genome is devoted to making mRNAs. In both species, the vast majority (87–93%) of exonic nucleotides were detected as transcribed, confirming the sensitivity of the approach. However, a higher percentage of intronic sequences were detected as transcribed in the mouse, and this might be due to a greater sequencing depth and broader spectrum of biological samples analyzed in mouse (Fig. 1b).

Candidate cis regulatory sequences—To identify potential *cis* regulatory regions in the mouse genome, we utilized three complementary approaches that involved mapping of chromatin accessibility, specific transcription factor (TF) occupancy sites and histone modification patterns. All of these approaches have previously been shown to uncover *cis* regulatory elements with high accuracy and sensitivity [19,20].

By mapping DNase I hypersensitive sites (DHSs) in 55 mouse cell and tissues types [21], we identified a combined total of ~1.5 million distinct DHSs at a false discovery rate (FDR) of 1% (Supplementary Table 5) (*Vierstra et al*. in press). Genomic footprinting analysis in a subset (25) of these cell types further delineated 8.9 million distinct TF footprints. De novo derivation of a *cis*-regulatory lexicon from mouse TF footprints revealed a recognition repertoire nearly identical with that of the human, including both known and novel recognition motifs.

We used ChIP-seq to determine the binding sites for a total of 37 TFs in various subsets of 33 cell/tissue types. Of these 37 TFs, 24 were also extensively mapped in the murine and human erythroid cell models (MEL and K562) and B-lymphoid cell lines (CH12 and GM12878) (Cite: *Cheng et al*. in press). In total we defined 2,107,950 discrete ChIP-seq

peaks, representing differential cell/tissue occupancy patterns of 280,396 distinct TF binding sites (supplemental methods) (Supplementary Table 6).

We also performed ChIP-seq for as many as nine histone H3 modifications (H3K4me1, H3K4me2, H3K4me3, H3K9me3, H3K27ac, H3K27me3, H3K36me3, H3K79me2, and H3K79me3) in up to 23 mouse tissues and cell types per mark. We applied a supervised machine learning technique, Random Forest based Enhancer prediction from Chromatin State (RFECS), to three histone modifications (H3K4me1, H3K4me3 and H3K27ac) [22], identifying a total of 291,200 candidate enhancers and 82,853 candidate promoters in the mouse genome (Supplementary Table 7 and 8). To functionally validate the predictions, we randomly selected 76 candidate promoter elements (avg. size 1000bp) and 183 candidate enhancer elements (avg. size 1000 bp), cloned these previously unannotated sequences into reporter constructs, and performed luciferase reporter assays via transient transfection in pertinent mouse cell lines (See supplemental methods, Supplementary Table 9 and 10). Overall, 66/76 (87%) candidate promoters and 129/183 (70.5%) candidate enhancers showed significant activity in these assays compared to 2/30 randomly selected negative controls (Supplementary Fig. 1c).

Collectively, our studies assigned potential regulatory function to 12.6% of the mouse genome (Fig. 1c).

Transcription factor networks—We explored the TF networks and combinatorial TF binding patterns in the mouse samples in two companion papers, and compared these networks to regulatory circuitry models generated for the human genome (*Cheng et al.* and *Stergachis et al*. in press). From genomic footprints, we constructed TF-to-TF cross-regulatory network in each of 25 cell/tissue types for a total of ~500 TFs with known recognition sequences. Analyses of these networks revealed regulatory relationships between transcription factor genes that are strongly preserved in human and mouse, in spite of the extensive plasticity of the *cis*-regulatory landscape (detailed below). Whereas only 22% of TF footprints are conserved, nearly 50% of cross-regulatory connections between mouse TFs are conserved in human through the innovation of novel binding sites. Moreover, analysis of network motifs shows that larger-scale architectural features of mouse and human TF networks are strikingly similar (*Stergachis et al*. in press).

Chromatin states—We produced integrative maps of chromatin states in 15 mouse tissue and cell types and six human cell lines (Supplementary Table 11) using a hidden Markov model (chromHMM) [23,24], which allowed us to segment the genome in each cell type into seven distinct combination of chromatin modification marks (or chromatin states). One state is characterized by the absence of any chromatin marks, while every other state features either predominantly one modification or a combination of two to three modifications (Extended Data Table 1, Supplementary materials). The portion of the genome in each chromatin state varied with cell type (Fig. 1d, Supplementary Fig. 2). Similar proportions of the genome are found in the active states in each cell type, for both mouse and human. Interestingly, the fraction of each genome that is in the H3K27me3-dominated, transcriptional repressed state is most variable, suggesting a profound role of transcriptional repression in shaping the *cis* regulatory landscape during mammalian development.

Replication domains (RDs)—Replication-timing, the temporal order in which megabase-sized genomic regions replicate during S-phase, is linked to the spatial organization of chromatin in the nucleus [25–28], serving as a useful proxy for tracking differences in genome architecture between cell types [29,30]. Since different types of chromatin are assembled at different times during the S phase [31], changes in replication timing during differentiation could elicit changes in chromatin structure across large domains. We obtained 36 mouse and 31 human replication-timing profiles covering 11 and 9 distinct stages of development, respectively (Supplementary Table 12). We defined "replication boundaries" as the sites where replication profiles change slope from synchronously replicating segments (discussed later). A total of 64,535 and 50,194 boundaries identified across all mouse and human datasets, respectively, were mapped to 4,322 and 4,675 positions, with each cell type displaying replication-timing transitions at 50–80% of these positions (Fig. 1e).

Annotation of orthologous protein coding and non-coding genes in the human and mouse genomes

To facilitate a systematic comparison of the transcriptome, *cis* regulatory elements and chromatin landscape between the human and mouse genomes, we built a high-quality set of human-mouse orthologs of protein coding and non-coding genes [32]. The list of protein-coding orthologs, based on phylogenetic reconstruction, contains a total of 15,736 1:1 and a smaller set of 1:many and many:many ortholog pairs (Supplementary Table 13, 14, 15). We also inferred orthologous relationships among short ncRNA genes using a similar phylogenetic approach. We established 1:1 human-mouse orthologues for 151,257 internal exon pairs (Supplementary Table 16) and 204,887 intron pairs (Supplementary Table 17), and predicted 2,717 (3,446) novel human (respectively, mouse) exons (Supplementary Table 18). Additionally, we mapped the 17,547 human long non-coding RNA (lncRNA) transcripts annotated in Gencode v10 onto the mouse genome. We found 2,327 (13.26%) human lncRNA transcripts (corresponding to 1,679, or 15.48% of the lncRNA genes) homologous to 5,067 putative mouse transcripts (corresponding to 3,887 putative genes) (Supplementary Fig. 3, Supplementary Table 19). Consistent with previous observations, only a small fraction of lncRNAs are constrained at the primary sequence level, with rapid evolutionary turnover[33]. Other comparisons of human and mouse transcriptomes covering areas including pre-mRNA splicing, antisense, and intergenic RNA transcription are detailed in a companion paper (Pervouchine et al., submitted).

Divergent and conserved gene expression patterns in human and mouse cells

Previous studies have revealed dramatic examples of species-specific gene expression patterns that underlie phenotypic changes during evolution [34–38]. In these cases changes in expression of a single gene between closely related species led to adaptive changes. However, it is not clear how extensive the changes in expression patterns are between more distantly related species, such as mouse and human, with some studies emphasizing similarities in transcriptome patterns of orthologous tissues [39–41] and others emphasizing substantial interspecies differences [42]. Our initial analyses revealed that gene expression patterns tended to cluster more by species rather than by tissue (Fig. 2a). To resolve the sets of genes contributing to different components in the clustering, we employed variance

decomposition (see Methods) to estimate, for each orthologous human-mouse gene pair, the proportion of the variance in expression that is contributed by tissue and by species (Fig. 2b). This analysis revealed the sets of genes whose expression varies more across tissues than between species, and those whose expression varies more between species than across tissues. As expected, the clustering of the RNA-seq samples is dominated either by species or tissues, depending on the gene set employed (Extended Data Fig. 1a, 1b). Furthermore, removal of the ~4800 genes that drive the species-specific clustering (see Shin et al., in press, Fig. S1D therein) or normalization methods that reduce the species effects reveals tissue specific patterns of expression in the same samples (Extended Data Fig. 1c). Categorizing orthologous gene pairs into these groups should enable more informative translation of research results between mouse and human. In particular, for gene pairs whose variance in expression is largest between tissues (and less between species), mouse should be a particularly informative model for human biology. In contrast, interpretation of studies involving genes whose variance in expression is larger between species needs to take into account the species variation. The relative contributions of species-specific and tissue-specific factors to each gene's expression are further explored in two companion papers (Lin et al., in press, Pervouchine et al. submitted).

To further identify genes with conserved expression patterns and those that have diverged between humans and mice, we developed a novel method, referred to as Neighborhood Analysis of Conserved Co-expression (NACC), to compare the transcriptional programs of orthologous genes in a way that did not require precisely matched cell lines, tissues, or developmental stages, as long as a sufficiently diverse panel of samples is used in each species (Supplemental Methods). Observing that the orthologs of most sets of co-expressed genes in one species remained significantly correlated across samples in the other species, we use the mean of these small correlated sets of orthologous genes as a reference expression pattern in the other species. We compute Euclidean distance to the reference pattern in the multi-dimensional tissue/gene expression space as a relative measure of conservation of expression of each gene. Specifically, for each human gene (the test gene), we defined the most similarly expressed set of genes (N=20) across all the human samples as that gene's co-expression neighborhood. We then quantify the average distance between the transcript levels of the mouse ortholog of the test gene and the transcript levels of each mouse ortholog of the neighborhood genes across the mouse samples. We then invert the analysis, and choose a mouse test gene and define a similar gene co-expression neighborhood in the mouse samples, and calculate the average distance between the expression of orthologs of the test gene and expression of neighborhood genes across the human samples. The average change in the human to mouse and mouse to human distances, referred herein as a NACC score, is a symmetric measure of the degree of conservation of co-expression for each gene. The distribution of this quantity for each gene is shown in Fig. 2c, showing that genes in one species show a strong tendency to be co-expressed with orthologs of similarly expressed genes in the other species compared to random genes (also see Supplemental Materials). We quantify the degree to which a specific biological process diverges between human and mouse as the average NACC scores of genes in each GO category by calculating a z-score using random sampling of equal size sets of genes. Fig. 2d shows that genes coding for proteins in the nuclear and intracellular organelle

compartments, and involved in RNA processing, nucleic acid metabolic processes, chromatin organization and other intracellular metabolic processes tend to exhibit more similar gene expression patterns between human and mouse, and genes involved in extracellular matrix, cellular adhesion, signaling receptors, immune responses and other cell membrane related processes are more diverged (for a complete list of all GO categories and conservation analysis, see Supplementary Table 21). As a control, when we applied the NACC analysis to two different replicates of RNA-seq datasets from the same species, no difference in biological processes can be detected (Supplementary Fig. 5).

Several lines of evidence indicate that NACC is a sensitive and robust method to detect conserved as well as diverged gene expression patterns from a panel of imperfectly matched tissue samples. First, when we applied NACC to a set of simulated datasets, we found that NACC is robust for the diversity and conservation of the mouse-human sample panel (in Supplementary Fig. 6). Second, we randomly sampled subsets of the full panel of samples and demonstrated that the categories of human-mouse divergence shown in Fig 2d are robust to the particular sets of samples we selected (Supplementary Fig. 7). Third, when we repeated NACC on a limited collection of more closely matched tissues and primary cell types (see supplemental methods), the biological processes detected as conserved and species specific in the larger panel of mis-matched human/mouse samples are largely recapitulated, although some pathways are detected with somewhat less significance, likely due to the smaller number of datasets used (Supplementary Fig. 8). In summary, the NACC results support and extend the principal component analysis, showing that while large differences between mouse and human transcriptome profiles can be observed (revealed in PC1), genes involved in distinct cellular pathways or functional groups exhibit different degrees of conservation of expression patterns between human and mouse, with some strongly preserved and others changing dramatically.

Prevalent species-specific regulatory sequences along with a core of conserved regulatory sequences

To better understand how divergence of *cis* regulatory sequences is linked to the range of conservation patterns detected in comparisons of gene expression programs between species, we examined evolutionary patterns in our predicted regulatory sequences. Previous studies have identified a wide range of evolutionary patterns and rates for *cis* regulatory regions in mammals[8,5], but there are still questions regarding the overall degree of similarity and divergence between the *cis* regulatory landscapes in the mouse and human. The variety of assays and breadth of tissue and cell type coverage in the mouse ENCODE data therefore provide an opportunity to address this problem more comprehensively.

We first determined sequence homology of the predicted *cis* elements in the mouse and human genomes. We established one-to-one and one-to-many mapping of human and mouse bases derived from reciprocal chained blastz alignments [43] and identified conserved *cis* regulatory sequences (*Denas et al., submitted*). This analysis showed that 79.3% of chromatin-based enhancer predictions, 79.6% of chromatin-based promoter predictions, 67.1% of the DHS, and 66.7% of the TF binding sites in the mouse genome have homologs in the human genome with at least 10% overlapping nucleotides, while by random chance

one expects 51.2%, 52.3%, 44.3%, and 39.3% respectively (Fig. 3a, supplementary materials for details). With a more stringent cutoff that requires 50% alignment of nucleotides, we found that 56.4% of the enhancer predictions, 62.4% of promoter predictions, 61.5% of DHS, and 53.3% of the TF binding sites have homologs, compared with an expected frequency of 34%, 33.8%, 33.6% and 33.7% by random chance (Supplementary Fig. 9). The candidate mouse regulatory regions with human homologs are listed in Supplementary Table 22–25. Thus, between half and two-thirds of candidate regulatory regions demonstrate a significant enrichment in sequence conservation between human and mouse. The remaining half to one-third have no identifiable orthologous sequence.

The candidate regulatory regions in mouse with no ortholog in human could arise either because they were generated by lineage-specific events, such as transposition, or because the ortholog in the other species was lost. Species-specific *cis* regulatory sequences have been reported before [3,14], but the fraction of regulatory sequences in this category remains debatable and may vary with different roles in regulation. We find that 15% (12,387 out of 82,853) of candidate mouse promoters and 16.6% (48,245 out of 291,200) of candidate enhancers (both predicted by patterns of histone modifications) have no sequence ortholog in humans (Supplementary Table 26, 28, for details please refer to supplementary method section). However, the question remains as to whether these species-specific elements are truly functional elements or simply correspond to false positive predictions due to measurement errors or biological noise. Supporting the function of mouse-specific *cis* elements, 18 out of 20 randomly selected candidate mouse-specific promoters tested positive using reporter assays in mouse ES cells, where they were initially identified (Fig. 3b, Supplementary Table 27). Further, when these 18 mouse-specific promoters were tested using reporter assays in the human ES cells, all of them also exhibited significant promoter activities (Extended Data Fig. 2a, Supplementary Table 27), indicating that the majority of candidate mouse-specific promoters are indeed functional sequences, which are either gained in the mouse lineage or lost in the human lineage. Similarly, a majority of the candidate mouse-specific enhancers discovered in ES cells are also likely bona fide *cis* elements, as 70.2% (26 out of 37) candidate enhancers randomly selected from this group were found to exhibit enhancer activities in reporter assays (Fig. 3b, Supplementary Table 29). Like the candidate mouse-specific promoters, 61.5% (16 out of 26) of the candidate mouse-specific enhancers also show enhancer activities in human ES cells (Extended Data Fig. 2a).

We next test whether the rapidly diverged *cis* regulatory elements would correspond to the same cellular pathways shown to be less conserved by the NACC analysis of gene expression programs. Indeed, GO analysis revealed that the mouse-specific regulatory elements are significantly enriched near genes involved in immune function (Fig. 3c), in agreement with the divergent transcription patterns for these genes reported above and an earlier report based on a smaller number of primate-specific candidate regulatory regions [44]. This suggests that regulation of genes involved in immune function tends to be species-specific [44], just as the protein-coding sequences coding for immunity, pheromones and other environmental genes are frequent targets for adaptive selection in each species [2,45]. The

target genes for mouse specific TF binding sites (Supplementary Table 30) are enriched in molecular functions such as histone acetyltransferase activity, high-density lipoprotein particle receptor activity, in addition to immune function (IgG binding).

We next investigated the mechanisms generating mouse-specific *cis* regulatory sequences: loss in human, gain in mouse, or both. 89% (42,947 out of 48,245) of mouse-specific enhancers and 85% (10535 out of 12387) of mouse-specific promoters overlap with at least one class of repeat elements (compared to 78% by random chance). Confirming earlier reports [46–48], we found that mouse-specific candidate promoters and enhancers are significantly enriched for repetitive DNA sequences, with several classes of repeat DNA highly represented (Fig. 3d and Extended Data Fig. 2b). Furthermore, mouse specific TF binding sites are highly enriched in mobile elements like SINE and LTR (*Sundaram et al.*, in press, Genome Research).

The 50% to 60% of candidate regulatory regions with sequences conserved between mouse and human are a mixture of: (i) sequences whose function has been preserved via strong constraint since these species diverged, (ii) sequences that have been co-opted (or exapted) to perform different functions in the other species, and (iii) sequences whose ortholog in the other species no longer has a discernable function, but divergence by evolutionary drift has not been sufficient to prevent sequence alignment between mouse and human. Several companion papers delve deeply into these issues (*Denas et al., Cheng et al., Vierstra et al.*). In particular, Cheng et al. show that the conservation of TF binding at orthologous positions (falling in category i) is associated with pleiotropic roles of enhancers, as evidenced by activity in multiple tissues. *Denas et al.* and *Vierstra et al.* describe the exaptation of conserved regulatory sequences for other functions.

We surveyed the conservation of function in the subset of mouse candidate *cis* elements that have sequence counterparts in the human genome. Of the 51,661 chromatin based promoter predictions that have human orthologs, 44% (22,655) of them are still predicted as a promoter in human based on the same analysis of histone modifications (Supplementary Table 31, see supplementary methods for details). Of the 164,428 chromatin based enhancer predictions that have human orthologs, 40% (64,962) of them are predicted as an enhancer in human (Supplementary Table 32). The remaining 56%–60% of candidate mouse regulatory regions with a human ortholog fall into category ii or iii (above), i.e. the orthologous sequence in human either performs a different function or does not maintain a detectable function.

One caveat of the above observation is that the tissues or cell samples used in the survey were not perfectly matched. To better examine the conservation of biochemical activities among these predicted *cis* regulatory elements with orthologs between mouse and human, we analyzed the chromatin modifications at the promoter or enhancer predictions in a broad set of 23 mouse tissue and cell types with the neighborhood co-expression association analysis (NACC) method described above. Instead of gene expression levels, we selected the histone modification H3K27ac as an indicator of promoter or enhancer activity as previously reported [49]. As shown in Fig. 4a, the promoter predictions (blue) show a significantly higher correlation in the level of H3K27ac in human and mouse than the

random controls (red). Similarly, most chromatin based enhancer predictions in the mouse genome exhibit conserved chromatin modification patterns in the human, albeit to a lesser degree than the promoters (Fig. 4b). NACC analysis on DNase-Seq signal resulted in very similar distributions of conserved chromatin accessibility patterns at promoters (Fig. 4c) and enhancers (Fig. 4d). Thus many sequence-conserved candidate *cis* regulatory elements appeared to have conserved patterns of activities in mice and humans.

Taken together, these analyses show that the mammalian *cis*-regulatory landscapes in the human and mouse genomes are substantially different, driven primarily by gain or loss of sequence elements during evolution. These species-specific candidate regulatory elements are enriched near genes involved in stress response, immunity and certain metabolic processes, and contain elevated levels of repeated DNA elements. On the other hand, a core set of candidate regulatory sequences are conserved and display similar activity profiles in humans and mice.

Chromatin state landscape reflects tissue and cell identities in both human and mouse

We examined gene-centered chromatin state maps in the mouse and human cell types (see Supplementary Methods) (Fig. 5a, Supplementary Fig. 10). In all cell types, the low-expressed genes were almost uniformly in chromatin states with the repressive H3K27me3 mark or in the state unmarked by these histone modifications. In contrast, expressed genes showed the canonical pattern of H3K4me3 at the TSS surrounded by H3K4me1, followed by H3K36me3-dominated states in the remainder of the transcription unit. A similar pattern was seen for all the active genes, regardless of the level of expression; the only exception was a tendency for the H3K4me3 to spread further into the transcription unit for the most highly expressed genes. The same binary relationship between chromatin state maps and expression levels of genes was observed in mouse and human cell types (Supplementary Fig. 10).

For both mouse and human cells, the majority of the genome was in the unmarked state in each cell type, consistent with previous observations in *Drosophila* [50] and human cell lines [12] (Supplementary Fig. 2). About 55% of the mouse genome was in an unmarked state in all the 15 cell types examined, while 65% is unmarked in all six human cell types. For genes that were in the unmarked state in mouse, their orthologs in human also tended to be in the unmarked state, and vice versa, leading to a positive correlation for the amount of gene neighborhoods in unmarked states (Supplementary Fig. 11). Strong correlations were also observed in profiles of other chromatin marks averaged over cell lines and tissues (Pervouchine et al., submitted). The genes in the unmarked zones were depleted of transcribed nucleotides relative to the number expected based on fraction of the genome included, and the levels of the transcripts mapped there were lower than those seen in the active chromatin states (Supplementary Fig. 12).

Previous studies revealed limited changes of the chromatin states in lineage-restricted cells as they undergo large-scale changes in gene expression during maturation [51–53]. The chromatin state maps recapitulated this result, showing very similar patterns of chromatin modification in a cell line model for proliferating erythroid progenitor cells (G1E) and in maturing erythroblasts (G1E-ER4 cells treated with estradiol) across genes whose

expression level changed significantly during maturation (Fig. 5b, Supplementary Fig. 10b). This limited change raised the possibility that the chromatin landscape, once established during lineage commitment, dictates a permissive (or restrictive) environment for the gene regulatory programs in each cell lineage [53], and the chromatin states may differ between cell lineages. We tested this by examining the chromatin state maps for genes that were differentially expressed between hematopoietic cell lineages (erythroblasts versus megakaryocytes), and we found dramatic differences between the two cell types (Fig. 5c and Supplementary Fig. 10b). Genes expressed at a higher level in megakaryocytes than in erythroblasts were all in active chromatin states in megakaryocytes, but many were in inactive chromatin states in erythroblasts (Fig. 5c). In the converse situation, genes expressed at a higher level in erythroblasts than in megakaryocytes showed more inactive states in the cells in which they were repressed (Supplementary Fig. 10b). These greater differences in chromatin states correlating with differential expression of genes between but not within cell lineages support the model that chromatin states are established during the process of lineage commitment. The clustering of cell types together by lineage based on chromatin state maps (Supplementary Fig. 10c) also supports the model that the landscape of active and repressed chromatin is established no later than lineage commitment, and that this landscape is a defining feature of each cell type. Greater differences in chromatin states correlating with differences in gene expression were also observed when comparing average chromatin profiles in human and mouse (Pervouchine et al., submitted).

Mouse chromatin states inform interpretation of human disease associated sequence variants

In order to investigate whether the mouse chromatin states were informative on sequence variants linked to human diseases by genome-wide association studies (GWAS), we combined the chromatin state segmentations of the fifteen mouse samples into a refined segmentation, which we used to train a self-organizing map (SOM) [54] on four histone modification ChIP-seq datasets (H3K4me3, H3K4me1, H3K36me3, and H3K27me3) for each mouse sample. We mapped 4,265 SNPs from the human GWAS studies uniquely onto the mouse genome and scored these Single Nucleotide Polymorphisms (SNPs) onto the trained SOM to determine whether SNP subsets were enriched in specific areas of the map. As shown in Fig. 6a, the highest enriched H3K4me1 unit in the kidney contains five GWAS hits (p-value<3.95e-14) on different chromosomes related to blood characteristics such as platelet counts (Fig. 6a, Extended Data Table 2a). Similarly, the second highest enriched unit in liver H3K36me3 contained six GWAS hits (p-value<7.54e-31) related to cholesterol and alcohol dependence out of twelve in that unit (Fig. 6b, Extended Data Table 2b). In contrast, one of the highest units in brain H3K27me3 has five GWAS hits (p-value<4.93e-33) on different chromosomes associated with brain disorders/response to addictive substances (Fig. 6c, Extended Data Table 2c). This unit is different from the other examples in that it is enriched for H3K27me3 signal in multiple tissues, with brain being the highest. 801 out of the 1350 units of the map showed statistical enrichment of SNPs of 0.05 after Holm-Bonferroni correction for multiple hypothesis testing, 55% of which (accounting for 1750 GWAS hits) had signal for at least one histone mark that ranked within the top 100 units on the map (Fig. 6d). The best histone marks for enriched GWAS units were primarily H3K4me1 (23%), H3K36me3 (18%), and H3K27me3 (12%), with H3K4me3 accounting for

a less than 2% of the remainder. Together these results suggest that the chromatin state maps can be used to identify potential sites for functional characterization in mouse for human GWAS hits. Indeed, Cheng et al. (in press) show that conserved DNA segments bound by orthologous TFs in human and mouse are enriched for trait-associated SNPs mapped by GWAS.

Large-scale chromatin domains are developmentally stable and evolutionarily conserved

We mapped the positions of early and late replication timing boundaries in each of 36 mouse and 31 human profiles (Fig. 7a). Significantly clustered boundary positions (above the 95th percentile of re-sampled positions) were identified and peaks in boundary density were aligned between cell types using a common heuristic (Extended Data Fig. 3a,b, Supplementary Fig. 13). After alignment, consensus boundaries were further classified by orientation and amount of replication timing separation, resulting in a more stringent filtering of boundaries (Supplementary Fig. 14, 15). Overall, we found that 88% of boundary positions (vs. 20% expected for random alignment; Fisher exact test p<2e-16) aligned position and orientation between two or more cell types in both mouse and human (i.e. 12% were cell-type specific, Fig 7b, Extended Data Fig. 3). Pair-wise comparisons of boundaries were consistent with developmental similarity between cell types (Supplementary Fig. 16). The earliest and latest replicating boundaries were most well preserved between cell types, while those of mid-S replicating boundaries were highly variable (Extended Data Fig. 3e, f).

Interestingly, the greatest number of boundaries was detected in embryonic stem cells (ESCs) in both species, with significant reduction in boundary numbers during differentiation (Supplementary Fig. 16) consistent with consolidation of domains and by proxy large-scale chromatin organization into larger “Constant Timing Regions” CTRs during differentiation [55]. Given that over half of the mouse and human genomes exhibit significant replication timing changes during development [16,56], these observations support the model that developmental plasticity in replication timing is derived from differential regulation of replication timing within CTRs whose boundaries are preserved during development.

Although conservation of replication timing between mouse and human has been reported [26,27], the conservation of replicating timing boundaries has not been examined. We converted boundary coordinates ±100kb across boundary positions between species, revealing significant overlap (Fig. 7c–d; p<2.2e-16 by Fisher’s exact test relative to a randomized boundary list). The level of conservation of the positions of boundaries improved from a median of 27% for cell type specific boundaries to 70% for boundaries preserved in 9 or more cell types (Fig. 7c), demonstrating that boundaries most highly preserved during development were the most conserved across species. This was consistent with results for transcription (Fig. 2), as well as the previous observation that suggest that an increased plasticity of replication timing during development is associated with increased plasticity of replication timing during evolution[57]. Together, these findings identify evolutionarily labile vs. constrained domains of the mammalian genome at the megabase scale.

Given the link between replication and chromatin assembly, we compared replication timing and levels of other chromatin properties in 200kb windows across the genome (Supplementary Fig. 17). Features associated with active enhancers (H3K4me1, H3K27ac, DNase I sensitivity) were more closely correlated to replication timing than features associated with active transcription (RNA pol II, H3K4me3, H3K36me3, H3K79me2). By contrast, the correlation of replication timing to repressive features, such as H3K9me3, was poor and cell-type-specific, consistent with prior results. A more stringent comparison of differences in chromatin to differences in replication timing between cell types (Extended Data Fig. 3c and 3g, Supplementary Fig. 17) again revealed that marks of enhancers, including p300, H3K4me1 and H3K27ac, and DNase I sensitivity were more strongly correlated to replication timing than marks of active transcription.

Conclusion

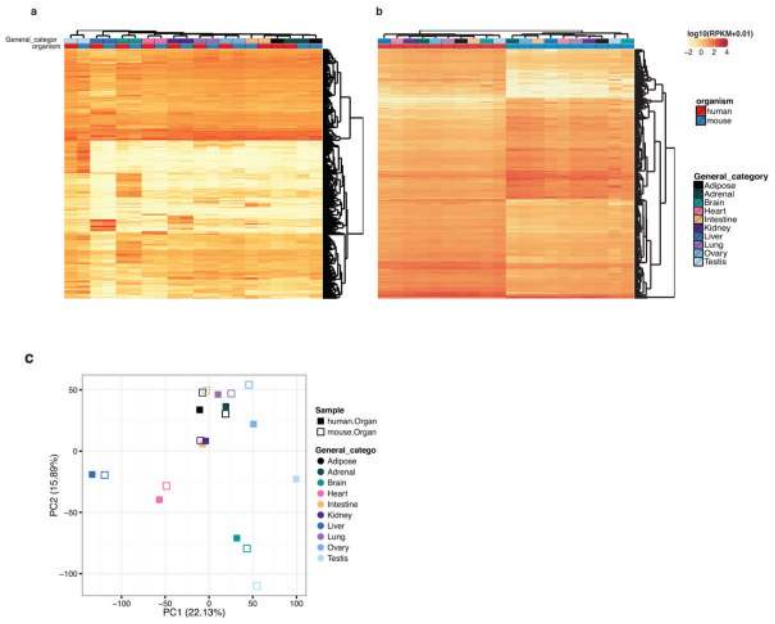
By comparing the transcriptional activities, chromatin accessibilities, transcription factor binding, chromatin landscapes and replication timing throughout the mouse genome in a wide spectrum of tissues and cell types, we have made significant progress toward a comprehensive catalog of potential functional elements in the mouse genome. The catalog described in the current study should provide a valuable reference to guide researchers to formulate new hypotheses and develop new mouse models, in the same way as the recent human ENCODE studies have impacted the research community [12].

We provide multiple lines of evidence that gene expression and their underlying regulatory programs have substantially diverged between the human and mouse lineages while a subset of core regulatory programs are largely conserved. The divergence of regulatory programs between mouse and human is manifested not only in the gain or loss of *cis* regulatory sequences in the mouse genome, but also in the lack of conservation in regulatory activities across different tissues and cell types. This finding is in line with previous observations of rapidly evolving transcription factor binding in mammals, flies and yeasts, and highlights the dynamic nature of gene regulatory programs in different species [3,4,7,58]. Furthermore, by comprehensively delineating the potential *cis* regulatory elements we demonstrated that specific groups of genes and regulatory elements have undergone more rapid evolution than others. Of particular interest is the finding that *cis* regulatory sequences next to immune system related genes are more divergent. The finding of species-specific *cis*-elements near genes involved in immune function suggests rapid evolution of regulatory mechanisms related to the immune system. Indeed, previous studies have uncovered extensive differences in the immune systems among different mouse strains and between humans and mice [59], ranging from relative makeup of the innate immune and adaptive immune cells [59], to gene expression patterns in various immune cell types [60], and transcriptional responses to acute inflammatory insults [61,62]. At least some of these differences may be attributed to distinct regulatory mechanisms [60], and our finding that many predicted mouse *cis* elements near genes with immune function lack sequence conservation supports the model that evolution of *cis* regulatory sequences contributes to differences in the immune systems between humans and mice. More generally, our findings are consistent with the view that changes in transcriptional regulatory sequences are a source for phenotypic differences in species evolution.

How can species-specific gains or loss of *cis* regulatory elements during evolution be compatible with their putative regulatory function? The finding of different rates of divergence associated with regulatory programs of distinct biological pathways suggests complex forces driving the evolution of the *cis* regulatory landscape in mammals. We discovered that specific classes of endogenous retroviral elements are enriched at the species-specific putative *cis* regulatory elements, implicating transposition of DNA as a potential mechanism leading to divergence of gene regulatory programs during evolution. Previous studies have shown that endogenous retroviral elements can be transcribed in tissue-specific manner [63,64], with a fraction of them derived from enhancers and necessary for transcription of genes involved in pluripotency [65,66]. Future studies will be necessary to determine whether retroviral elements at or near enhancers is generally involved in driving tissue-specific gene expression programs in different mammalian species.

Despite the divergence of the regulatory landscape between mouse and human, the pattern of chromatin states (defined by histone modifications) and the large-scale chromatin domains are highly similar between the two species. Half of the genome is well conserved in replication timing (and by proxy, chromatin interaction compartment) with the other half highly plastic both between cell types and between species. It will be interesting to investigate the significance of these conserved and divergent classes of DNA elements at different scales, both with regard to the forces driving evolution and for implications of the use of the laboratory mouse as a model for human disease.

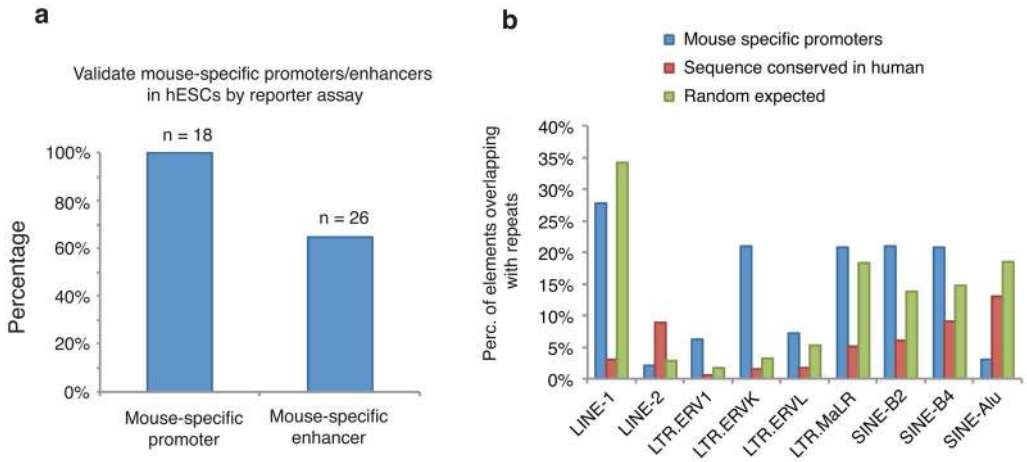
Extended Data



Extended Data Figure 1. Clustering analysis of human and mouse tissue samples

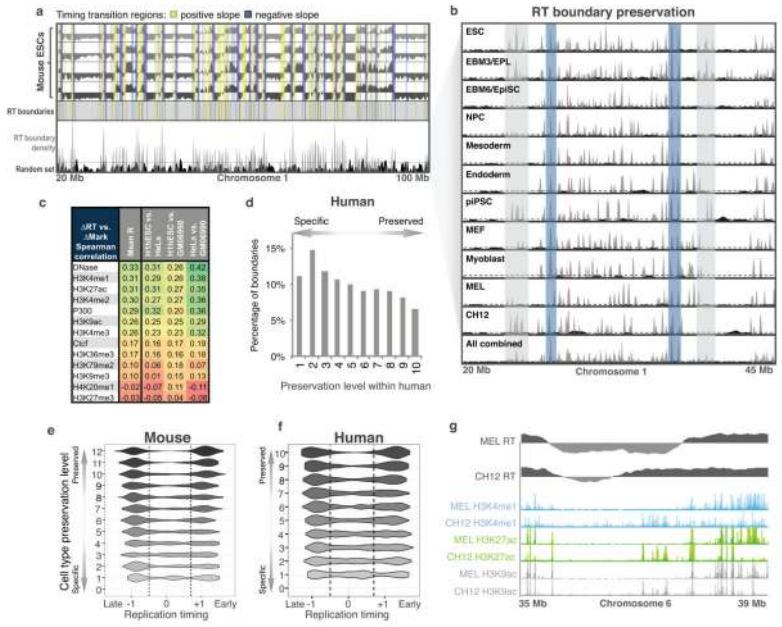
a, RNA-seq data from Ilumina Body Map (Adipose, Adrenal, Brain, Colon, Heart, Kidney, Liver, Lung, Ovary and Testis) were analyzed together with that from the matched mouse samples using clustering analysis. Genes with high variance across tissues were used,

resulting in cell samples clustering by tissues, not by species. **b,** Clustering employing genes with high variance between species shows clustering by species instead of tissues. **c,** Principal Component Analysis (PCA) was performed for RNA-seq data for 10 human and mouse matching tissues. The expression values are normalized within each species and we observed the clustering of samples by tissue types.



Extended Data Figure 2. Comparative analysis of sequence conservation in the cis elements predicted in the human and mouse genome

a, The predicted mouse-specific promoters and enhancers can function in human ES cells (hESCs). Percentages of predicted enhancers or promoters that test positive are shown in a bar chart. **b**, A bar chart shows the percentage of the predicted mouse-specific promoters containing various subclasses of LTR and SINE elements. As control, the predicted mouse cis elements with homologous sequences in the human genome or random genomic regions are included.



Extended Data Figure 3. Replication timing boundaries preserved among tissues are conserved during evolution

a, Heatmap of TTR overlap with positive (yellow) or negative (blue) slope. Replication timing (RT) boundaries were identified as clustered TTR endpoints (gray) above the 95th percentile (dashed line) of randomly resampled positions (black). **b**, Examples of constitutive boundaries (blue regions) and regulated boundaries (gray regions) highlighted. **c**, Spearman correlations between differences in chromatin feature enrichment and differences in RT in non-overlapping 200kb windows. **d,** Percentage of boundaries preserved between the indicated number of human cell types. (**e & f**). Distribution of boundary replication timing in mouse (e) and human (f) as a function of preservation level between cell types. **g**, Comparison of changes in replication timing vs. various histone marks across a segment of mouse Chr6.

Extended Data Table 1

A seven-state chromHMM model learned from four histone modifications in 15 mouse cell types or lines and six human cell lines is shown. The numbers represent the emission probabilities of each histone modification (column) in each chromatin state (row). The enriched histone modifications in each state are summarized in the first column. The fraction of genome assigned in each state was calculated (Supplementary Fig. 2). The average and variation of these fraction values across all included cell types/tissues are listed in the last two columns.

State	Feature	H3K27m3	H3K4m3	H3K4m1	H3K36m3	Average%	Variation
1	K4m3	0.07	0.92	0.05	0.03	0.75	0.07
2	K4m1/3	0.17	0.85	0.88	0.05	0.55	0.10
3	K4m1	0.01	0.01	0.47	0.02	3.35	0.57

State	Feature	H3K27m3	H3K4m3	H3K4m1	H3K36m3	Average%	Variation
4	K4m1+K36m3	0.01	0.05	0.59	0.71	0.58	0.23
5	K36m3	0.00	0.00	0.01	0.42	6.31	1.54
6	Unmarked	0.01	0.00	0.00	0.00	85.45	9.20
7	K27m3	0.29	0.00	0.02	0.00	3.01	3.87

Extended Data Table 2

Self-Organizing Map of histone modifications shows enrichment of human GWAS SNPs when mapped onto mouse

a, Kidney-specific H3K4me1 that shows enrichment of specific GWAS hits associated with urate levels and metabolites. **b**, Liver-specific H3K36me3 unit shows enrichment in GWAS hits related to cholesterol, alcohol dependence, and triglyceride levels. **c**, Brain-specific H3K27me3 signals show enrichment in GWAS SNPs associated with neurological disorders.

a	
rs6900341	Metabolite
rs1668871	Platelet counts
rs1063856	Coagulation factor levels
rs6798928	Immunoglobulin A
rs2079742	Urate levels
b	
rs1789891	Alcohol dependence
rs3811647	Hepcidin levels
rs10199768	Cardiovascular disease risk factors
rs17155315	QT interval
rs12686004	HDL cholesterol
rs3890182	HDL cholesterol
rs7758229	Colorectal cancer
rs6017342	Ulcerative colitis
rs603446	Triglycerides
rs2266788	HDL Cholesterol - Triglycerides
rs6056	Fibrinogen
rs641153	Age-related macular degeneration (CNV)
c	
rs6952808	Bipolar disorder and schizophrenia
rs2424635	Bipolar disorder and schizophrenia
rs2023454	Functional MRI
rs17115100	Parkinson’s disease
rs9312648	Response to amphetamines

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Feng Yue[1,*], Yong Cheng[2,*], Alessandra Breschi[3,*], Jeff Vierstra[4,*], Weisheng Wu[5,*], Tyrone Ryba[6,*], Richard Sandstrom[4,*], Zhihai Ma[2,*], Carrie Davis[7,*], Benjamin D. Pope[6,*], Yin Shen[1,*], Dmitri D. Pervouchine[3], Sarah Djebali[3], Bob Thurman[4], Rajinder Kaul[4], Eric Rynes[4], Anthony Kirilusha[8], Georgi K. Marinov[8], Brian A. Williams[8], Diane Trout[8], Henry Amrhein[8], Katherine Fisher-Aylor[8], Igor Antoshechkin[8], Gilberto DeSalvo[8], Lei-Hoon See[7], Meagan Fastuca[7], Jorg Drenkow[7], Chris Zaleski[7], Alex Dobin[7], Pablo Prieto[3], Julien Lagarde[3], Giovanni Bussotti[3], Andrea Tanzer[3], Olgert Denas[9], Kanwei Li[9], M. A. Bender[10,11], Miaohua Zhang[12], Rachel Byron[12], Mark T. Groudine[12,13], David McCleary[1], Long Pham[1], Zhen Ye[1], Samantha Kuan[1], Lee Edsall[1], Yi-Chieh Wu[14], Matthew D. Rasmussen[14], Mukul S. Bansal[14], Cheryl A. Keller[5], Christapher S. Morrissey[5], Tejaswini Mishra[5], Deepti Jain[5], Nergiz Dogan[5], Robert S. Harris[5], Philip Cayting[2], Trupti Kawli[2], Alan P. Boyle[2], Ghia Euskirchen[2], Anshul Kundaje[2], Shin Lin[2], Yiing Lin[2], Camden Jansen[16], Venkat S. Malladi[2], Melissa S. Cline[17], Drew T. Erickson[2], Vanessa M Kirkup[17], Katrina Learned[17], Cricket A. Sloan[2], Kate R. Rosenbloom[17], Beatriz Lacerda de Sousa[18], Kathryn Beal[19], Miguel Pignatelli[19], Paul Flicek[19], Jin Lian[20], Tamer Kahveci[21], Dongwon Lee[22], W. James Kent[17], Miguel Ramalho Santos[18], Javier Herrero[19,23], Cedric Notredame[3], Audra Johnson[4], Shinny Vong[4], Kristen Lee[4], Daniel Bates[4], Fidencio Neri[4], Morgan Diegel[4], Theresa Canfield[4], Peter J. Sabo[4], Matthew S. Wilken[24], Thomas A. Reh[24], Erika Giste[4], Anthony Shafer[4], Tanya Kutyavin[4], Eric Haugen[4], Douglas Dunn[4], Alex P. Reynolds[4], Shane Neph[4], Richard Humbert[4], R. Scott Hansen[4], Marella De Bruijn[25], Licia Selleri[26], Alexander Rudensky[27], Steven Josefowicz[27], Robert Samstein[27], Evan E. Eichler[4], Stuart H. Orkin[28], Dana Levasseur[29], Thalia Papayannopoulou[30], Kai-Hsin Chang[30], Arthur Skoultchi[31], Srikanta Gosh[31], Christine Disteche[32], Piper Treuting[33], Yanli Wang[34], Mitchell J. Weiss[35,36], Gerd A. Blobel[35,36], Peter J. Good[37], Rebecca F. Lowdon[37], Leslie B. Adams[37], Xiao-Qiao Zhou[37], Michael J. Pazin[37], Elise A. Feingold[37], Barbara Wold[8], James Taylor[9], Manolis Kellis[14,15], Ali Mortazavi[16], Sherman M. Weissman[20], John Stamatoyannopoulos[4,#], Michael P. Snyder[2,#], Roderic Guigo[3,#], Thomas R. Gingeras[7,#], David M. Gilbert[6,#], Ross C. Hardison[5,#], Michael A. Beer[22,#,*], and Bing Ren[1,#] **The mouse ENCODE Consortium**

Affiliations

[1]Ludwig Institute for Cancer Research and University of California, San Diego School of Medicine, 9500 Gilman Drive, La Jolla, CA 92093, USA

[2]Department of Genetics, Stanford University, 300 Pasteur Drive, MC-5477 Stanford, CA 94305, USA

[3]Bioinformatics and Genomics, Centre for Genomic Regulation (CRG) and UPF, Doctor Aiguader, 88, Barcelona 08003, Catalonia, Spain

[4]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA

[5]Center for Comparative Genomics and Bioinformatics, Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA

[6]Department of Biological Science, 319 Stadium Drive, Florida State University, Tallahassee, FL 32306-4295, USA

[7]Functional Genomics, Cold Spring Harbor Laboratory, Bungtown Road, Cold Spring Harbor, New York 11724, USA

[8]Division of Biology, California Institute of Technology, Pasadena, CA 91125

[9]Departments of Biology and Mathematics and Computer Science, Emory University, O. Wayne Rollins Research Center, 1510 Clifton Road NE, Atlanta, Georgia 30322, USA

[10]Departments of Pediatrics, University of Washington, Seattle, Washington 98195, USA

[11]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[12]Basic Science Division, Fred Hutchinson Cancer Research Center, Seattle, WA

[13]Departments of Radiation Oncology, University of Washington, Seattle, Washington 98195, USA

[14]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology (MIT), Cambridge, MA, 02139, USA

[15]Broad Institute of MIT and Harvard, Cambridge

[16]Department of Developmental and Cell Biology, University of California, Irvine, Irvine, CA 92697, USA

[17]Center for Biomolecular Science and Engineering, School of Engineering, University of California Santa Cruz (UCSC), Santa Cruz, CA 95064, USA

[18]Departments of Ob/Gyn and Pathology, and Center for Reproductive Sciences, University of California San Francisco, San Francisco, CA, USA

[19]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton Cambridge CB10 1SD, UK

[20]Yale University, Department of Genetics, PO Box 208005, 333 Cedar Street, New Haven, CT 06520-8005

[21]Computer & Information Sciences & Engineering, University of Florida, Gainesville, FL 32611, USA

[22]McKusick-Nathans Institute of Genetic Medicine and Department of Biomedical Engineering, Johns Hopkins University, 733 N. Broadway, BRB 573 Baltimore, Maryland 21205, USA

[23]Bill Lyons Informatics Centre, UCL Cancer Institute, University College London, London WC1E 6DD, UK

[24]Department of Biological Structure, University of Washington, HSB I-516, 1959 NE Pacific Street, Seattle, Washington 98195, USA

[25]MRC Molecular Haemotology Unit, University of Oxford, Oxford, UK

[26]Department of Cell and Developmental Biology, Weill Cornell Medical College, New York, NY 10065, USA

[27]HHMI and Ludwig Center at Memorial Sloan Kettering Cancer Center, Immunology Program, Memorial Sloan Kettering Cancer Canter, New York, NY 10065, USA

[28]Dana Farber Cancer Institute, Harvard Medical School, Cambridge MA 02138, USA

[29]University of Iowa Carver College of Medicine, Department of Internal Medicine, Iowa City, Iowa, IA 52242, USA

[30]Division of Hematology, Department of Medicine, University of Washington, Seattle, WA 98195, USA

[31]Department of Cell Biology, Albert Einstein College of Medicine, Bronx, NY 10461, USA

[32]Department of Pathology, University of Washington, Seattle, WA 98195, USA

[33]Department of Comparative Medicine, University of Washington, Seattle, WA 98195, USA

[34]Bioinformatics and Genomics program, The Pennsylvania State University, The Pennsylvania State University, University Park, PA 16802, USA

[35]Division of Hematology, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

[36]Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA

[37]NHGRI, National Institutes of Health, 5635 Fishers Lane, Bethesda, Maryland 20892-9307, USA

Acknowledgments

This work is funded by grants R01HG003991 (B.R.), 1U54HG007004 (T.R.G.), 3RC2HG005602 (M.P.S), GM083337 and GM085354 (D.M.G), F31CA165863(BDP), RC2HG005573 and R01DK065806 (R.C.H.) from the National Institutes of Health, and BIO2011-26205 from the Spanish Plan Nacional and ERC 294653 (to R.G.), J.V. is supported by a National Science Foundation Graduate Research Fellowship under grant no. DGE-071824.. KB, MP, JH and PF acknowledge the Wellcome Trust (grant number 095908), the NHGRI (grant number U01HG004695) and the European Molecular Biology Laboratory. We thank Gary Hon for helping the analysis of

high-throughput enhancer validation. L.S. is supported by R01HD043997-09. S.L. was supported by grants F32HL110473 and K99HL119617.

References

1. Paigen K. One hundred years of mouse genetics: an intellectual history. I. The classical period (1902–1980). Genetics. 2003; 163:1–7. [PubMed: 12586691]
2. Mouse Genome Sequencing C et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002; 420:520–562. [PubMed: 12466850]
3. Odom DT, et al. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nat Genet. 2007; 39:730–732. [PubMed: 17529977]
4. Schmidt D, et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science. 2010; 328:1036–1040. [PubMed: 20378774]
5. Stefflova K, et al. Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. Cell. 2013; 154:530–540. [PubMed: 23911320]
6. Wilson MD, Odom DT. Evolution of transcriptional control in mammals. Curr Opin Genet Dev. 2009; 19:579–585. [PubMed: 19913406]
7. Borneman AR, et al. Divergence of transcription factor binding sites across related yeast species. Science. 2007; 317:815–819. [PubMed: 17690298]
8. Zheng W, Gianoulis TA, Karczewski KJ, Zhao H, Snyder M. Regulatory variation within and between species. Annu Rev Genomics Hum Genet. 2011; 12:327–346. [PubMed: 21721942]
9. Wray GA. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet. 2007; 8:206–216. [PubMed: 17304246]
10. King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. Science. 1975; 188:107–116. [PubMed: 1090005]
11. Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. Nat Rev Genet. 2010; 11:476–486. [PubMed: 20531367]
12. Consortium EP, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]
13. Djebali S, et al. Landscape of transcription in human cells. Nature. 2012; 489:101–108. [PubMed: 22955620]
14. Consortium EP, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447:799–816. [PubMed: 17571346]
15. Stamatoyannopoulos JA, et al. An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol. 2012; 13:418. gb-2012-13-8-418 [pii]. 10.1186/gb-2012-13-8-418 [PubMed: 22889292]
16. Hiratani I, et al. Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. Genome Res. 2010; 20:155–169. [PubMed: 19952138]
17. Jacquier A. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nat Rev Genet. 2009; 10:833–844. [PubMed: 19920851]
18. Xu Z, et al. Bidirectional promoters generate pervasive transcription in yeast. Nature. 2009; 457:1033–1037. [PubMed: 19169243]
19. Maston GA, Landt SG, Snyder M, Green MR. Characterization of enhancer function from genome-wide analyses. Annu Rev Genomics Hum Genet. 2011; 13:29–57. [PubMed: 22703170]
20. Hardison RC, Taylor J. Genomic approaches towards finding cis-regulatory modules in animals. Nat Rev Genet. 2012; 13:469–483. [PubMed: 22705667]
21. Thurman RE, et al. The accessible chromatin landscape of the human genome. Nature. 2012; 489:75–82. [PubMed: 22955617]
22. Rajagopal N, et al. RFECS: a random-forest based algorithm for enhancer identification from chromatin state. PLoS Comput Biol. 2013; 9:e1002968. [PubMed: 23526891]
23. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012; 9:215–216. [PubMed: 22373907]

24. Hoffman MM, et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 2013; 41:827–841. [PubMed: 23221638]
25. Dixon JR, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012; 485:376–380. [PubMed: 22495300]
26. Ryba T, et al. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Res. 2010; 20:761–770. [PubMed: 20430782]
27. Yaffe E, et al. Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture. PLoS Genet. 2010; 6:e1001011. [PubMed: 20617169]
28. Baker A, et al. Replication fork polarity gradients revealed by megabase-sized U-shaped replication timing domains in human cell lines. PLoS Comput Biol. 2012; 8:e1002443. [PubMed: 22496629]
29. Moindrot B, et al. 3D chromatin conformation correlates with replication timing and is conserved in resting cells. Nucleic Acids Res. 2012; 40:9470–9481. [PubMed: 22879376]
30. Takebayashi, S-i; Dileep, V.; Ryba, T.; Dennis, JH.; Gilbert, DM. Chromatin-interaction compartment switch at developmentally regulated chromosomal domains reveals an unusual principle of chromatin folding. Proc Natl Acad Sci U S A. 2012; 109:12574–12579. [PubMed: 22807480]
31. Lande-Diner L, Zhang J, Cedar H. Shifts in replication timing actively affect histone acetylation during nucleosome reassembly. Mol Cell. 2009; 34:767–774. [PubMed: 19560427]
32. Wu Y-C, Bansa lS, Rasmussen MD, Herrero J, Kellis M. Phylogenetic identification and functional validation of orthologous genes across human, mouse, fly, worm, yeast. in preperation.
33. Derrien T, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 2012; 22:1775–1789. [PubMed: 22955988]
34. McLean CY, et al. Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature. 2011; 471:216–219. [PubMed: 21390129]
35. Shubin N, Tabin C, Carroll S. Deep homology and the origins of evolutionary novelty. Nature. 2009; 457:818–823. [PubMed: 19212399]
36. Jones FC, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012; 484:55–61. [PubMed: 22481358]
37. Grossman SR, et al. Identifying recent adaptations in large-scale genomic data. Cell. 2013; 152:703–713. [PubMed: 23415221]
38. Fraser HB. Gene expression drives local adaptation in humans. Genome Res. 2013; 23:1089–1096. [PubMed: 23539138]
39. Brawand D, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011; 478:343–348. [PubMed: 22012392]
40. Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science. 2012; 338:1593–1599. [PubMed: 23258891]
41. Barbosa-Morais NL, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science. 2012; 338:1587–1593. [PubMed: 23258890]
42. Sabeti PC, et al. Positive natural selection in the human lineage. Science. 2006; 312:1614–1620. [PubMed: 16778047]
43. Schwartz S, et al. Human-mouse alignments with BLASTZ. Genome Res. 2003; 13:103–107. [PubMed: 12529312]
44. King DC, et al. Finding cis-regulatory elements using comparative genomics: some lessons from ENCODE data. Genome Res. 2007; 17:775–786. [PubMed: 17567996]
45. Ponting CP. The functional repertoires of metazoan genomes. Nat Rev Genet. 2008; 9:689–698. [PubMed: 18663365]
46. Bourque G, et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res. 2008; 18:1752–1762. [PubMed: 18682548]
47. Kunarso G, et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet. 2010; 42:631–634. [PubMed: 20526341]

48. Jacques PE, Jeyakani J, Bourque G. The majority of primate-specific regulatory sequences are derived from transposable elements. PLoS Genet. 2013; 9:e1003504. [PubMed: 23675311]
49. Calo E, Wysocka J. Modification of enhancer chromatin: what, how, and why? Mol Cell. 2013; 49:825–837. [PubMed: 23473601]
50. Filion GJ, et al. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell. 2010; 143:212–224. [PubMed: 20888037]
51. John S, et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat Genet. 2011; 43:264–268. [PubMed: 21258342]
52. Jin F, Li Y, Ren B, Natarajan R. PU.1 and C/EBP(alpha) synergistically program distinct response to NF-kappaB activation through establishing monocyte specific enhancers. Proc Natl Acad Sci U S A. 2011; 108:5290–5295. [PubMed: 21402921]
53. Wu W, et al. Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration. Genome Res. 2011; 21:1659–1671. [PubMed: 21795386]
54. Mortazavi A, et al. Integrating and mining the chromatin landscape of cell-type specificity using self-organizing maps. Genome Res. 2013; 23:2136–2148. [PubMed: 24170599]
55. Hiratani I, et al. Global reorganization of replication domains during embryonic stem cell differentiation. PLoS Biol. 2008; 6:e245. [PubMed: 18842067]
56. Hansen RS, et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc Natl Acad Sci U S A. 2010; 107:139–144. [PubMed: 19966280]
57. Ryba T, et al. Replication timing: a fingerprint for cell identity and pluripotency. PLoS computational biology. 2011; 7:e1002225.10.1371/journal.pcbi.1002225 [PubMed: 22028635]
58. Moses AM, et al. Large-scale turnover of functional transcription factor binding sites in Drosophila. PLoS Comput Biol. 2006; 2:e130. [PubMed: 17040121]
59. Mestas J, Hughes CCW. Of mice and not men: differences between mouse and human immunology. J Immunol. 2004; 172:2731–2738. [PubMed: 14978070]
60. Shay T, et al. Conservation and divergence in the transcriptional programs of the human and mouse immune systems. Proc Natl Acad Sci U S A. 2013; 110:2946–2951. [PubMed: 23382184]
61. Seok J, et al. Genomic responses in mouse models poorly mimic human inflammatory diseases. Proc Natl Acad Sci U S A. 2013; 110:3507–3512. [PubMed: 23401516]
62. Wells CA, et al. Genetic control of the innate immune response. BMC immunology. 2003; 4:5.10.1186/1471-2172-4-5 [PubMed: 12826024]
63. Faulkner GJ, et al. The regulated retrotransposon transcriptome of mammalian cells. Nat Genet. 2009; 41:563–571.10.1038/ng.368 [PubMed: 19377475]
64. Xie W, et al. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell. 2013; 153:1134–1148. [PubMed: 23664764]
65. Lu X, et al. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. Nature structural & molecular biology. 2014; 21:423–425.10.1038/nsmb.2799
66. Fort A, et al. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. Nat Genet. 2014; 46:558–566.10.1038/ng.2965 [PubMed: 24777452]

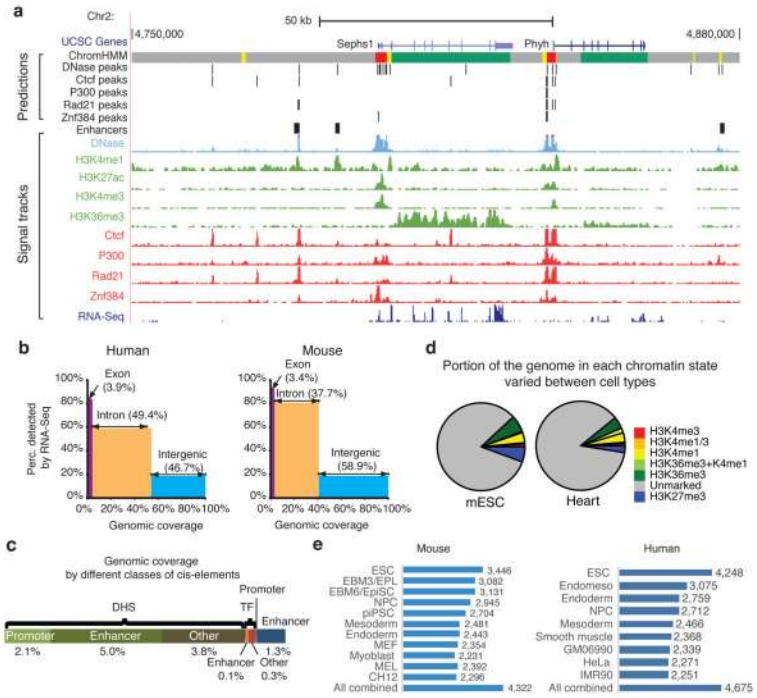


Figure 1. Overview of the mouse ENCODE data sets

a, A genome browser snapshot shows the primary data and annotated sequence features in the mouse CH12 cells (methods). **b,** Chart shows that much of the human and mouse genomes is transcribed in one or more cell and tissue samples. **c**, A bar chart shows the percentages of the mouse genome annotated as various types of cis-regulatory elements (Method). **d**, Pie charts show the fraction of the entire genome that is covered by each of the seven states in the mouse embryonic stem cells (mESC) and adult heart. **e**, Charts showing the number of replication timing (RT) boundaries in specific mouse and human cell types, and the total number of boundaries from all cell types combined.

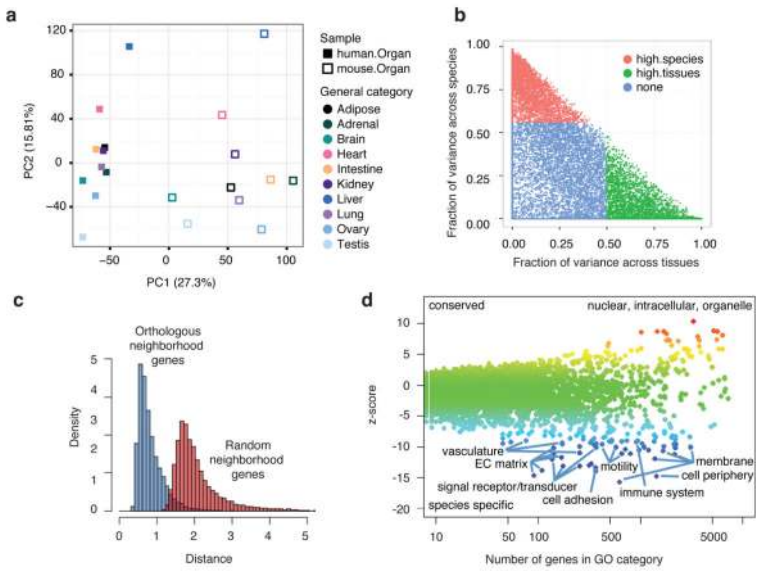


Figure 2. Comparative analysis of the gene expression programs in human and mouse samples

a. Principal Component Analysis (PCA) was performed for RNA-seq data for 10 human and mouse matching tissues. The expression values are normalized across the entire dataset. Solid squares denote human tissues. Open squares denote mouse tissues. Each category of tissue is represented by a different color. **b**. Gene expression variance decomposition (see Method) estimates the relative contribution of tissue and species to the observed variance in gene expression for each orthologous human-mouse gene pair. Green dots indicate genes with higher between-tissue contribution and red dots genes with higher between-species contributions. **c**, Neighborhood analysis of conserved co-expression (NACC) in human and mouse samples. The distribution of NACC scores for each gene is shown. **d**, A scatter plot shows the average of NACC score over the set of genes in each functional GO category. Highlighted are those biological processes that tend to be more conserved between human and mouse and those processes that have been less conserved (See Supplementary Table 21 for list of genes).

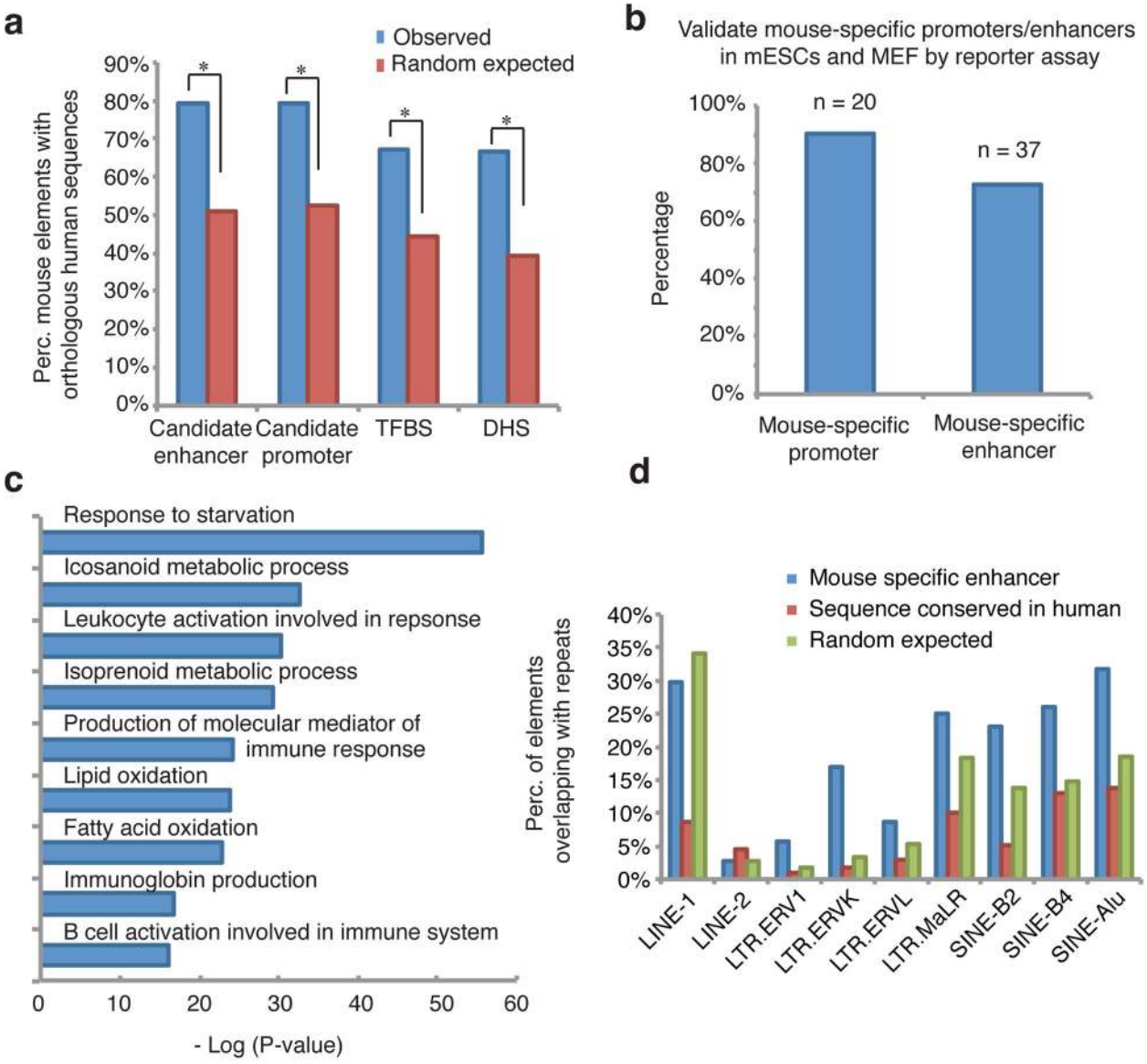


Figure 3. Comparative analysis of the cis elements predicted in the human and mouse genome

a, Charts show the fractions of the predicted mouse cis regulatory elements with homologous sequences in the human genome (Methods). **b,** A bar chart shows the fraction of the DNA fragments tested positive in the reporter assays performed either using the mESCs or Mouse embryonic fibroblasts (MEF). **c,** A chart shows the Gene ontology (GO) categories enriched near the predicted mouse-specific enhancers. **d**, A bar chart shows the percentage of the predicted mouse-specific enhancers containing various subclasses of LTR and SINE elements. As control, the predicted mouse cis elements with homologous sequences in the human genome or random genomic regions are included.

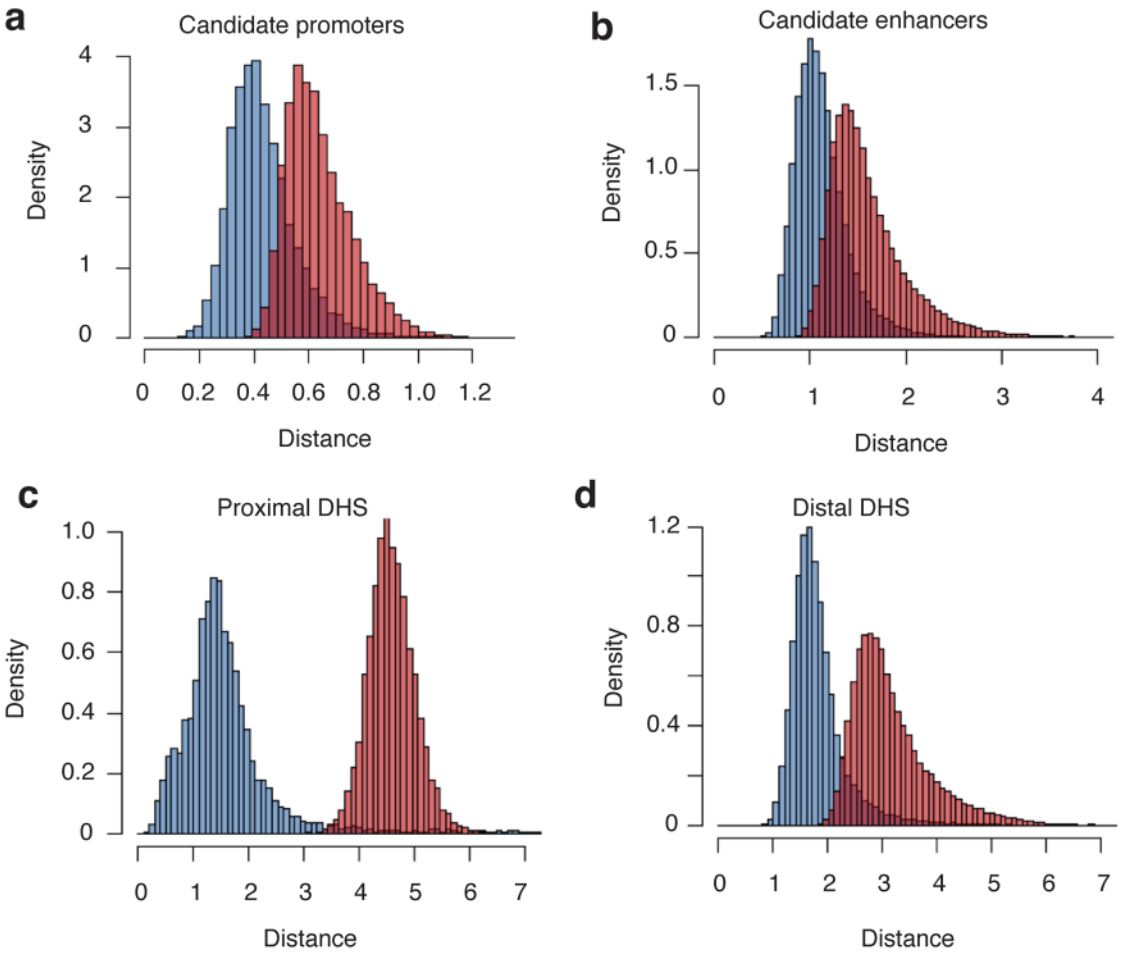


Figure 4. Analysis of conservation in biochemical activities at the predicted mouse *cis* regulatory sequences with human orthologs

(a & b) Histograms show the distribution of the NACC score for the chromatin modification H3K27ac signal at the predicted mouse promoters (**a**) or enhancers (**b**). (**c & d**) Histograms show the distributions of NACC scores for DNase I signal at the promoter proximal (**c**) and distal (**d**) DNase hypersensitive sites (DHS).

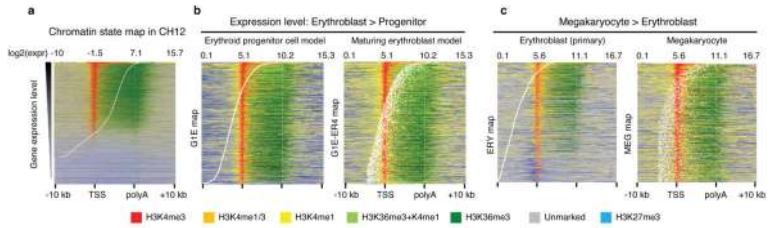


Figure 5. Chromatin landscape is stable within individual cell lineages

a, Map displaying the distribution of chromatin states over the neighborhoods of human-mouse one-to-one ortholog genes in CH12 cells. The gene neighborhood intervals were sorted by the transcription level of each gene, shown by white dots. **b** & **c**, Distribution of chromatin states in human-mouse one-toone orthologs that are differentially expressed genes between (**b**) erythroid progenitor and erythroblasts models and (**c**) erythroblast and megakaryocyte.

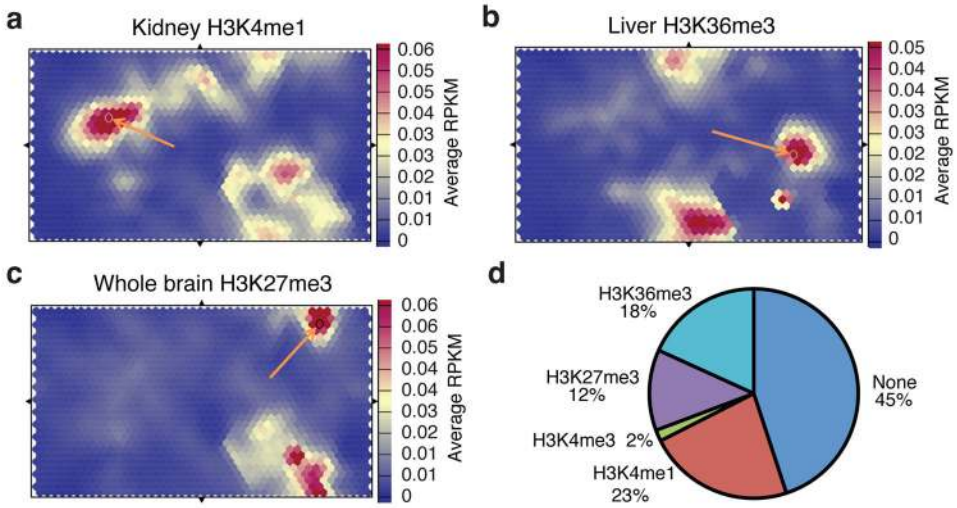


Figure 6. Human GWAS hits when mapped onto mouse genome are associated with specific chromatin states

a, A self-organization map of histone modification H3K4me1 shows association between kidney H3K4me1 state and specific GWAS hits associated with urate levels (Methods). **b**, Liver-specific H3K36me3 unit shows enrichment in GWAS hits related to cholesterol, alcohol dependence, and triglyceride levels. **c**, Brain-specific H3K27me3 high unit shows enrichment in GWAS SNPs associated with neurological disorders. **d**, Characterization of every unit with statistically significant GWAS enrichments in terms of highest histone modification signal in at least one sample. Units with no signal in top 100 map units for every histone modification are listed as none.

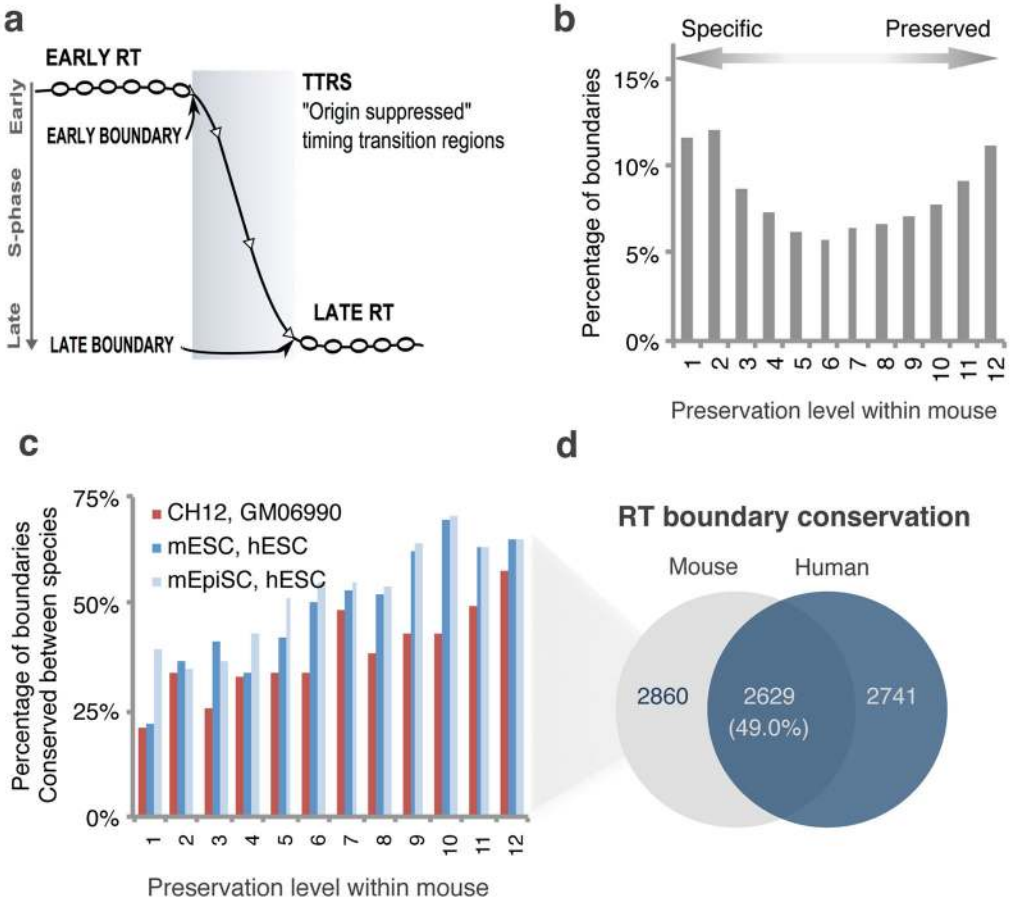


Figure 7. RT boundaries preserved among tissues are conserved in mice and humans

a, Depiction of a timing transition region (TTR) between the early and late replication domains. Early and late boundaries are defined as slope changes at either end of TTRs. **b.** Boundaries conserved between species for matched mouse and human cell types as a function of preservation among mouse cell types. **c**. Percentage of boundaries conserved between species (bar graph) and overall conservation of boundaries between comparable mouse and human cell types (CH12 vs. GM06990, mESC vs. hESC, mEpiSC vs. hESC) as a function of preservation among mouse cell types. (**d**) A Venn diagram compares the replication timing boundaries identified in the mouse and human genome.