doi:10.1038/nature13985

# Principles Of Regulatory Information Conservation Between Mouse And Human

*Yong Cheng*[*1], Zhihai Ma*[*1], Bong-Hyun Kim[2], Weisheng Wu[3], Philip Cayting[1], Alan P. Boyle[1], Vasavi Sundaram[4], Xiaoyun Xing[4], Nergiz Dogan[3], Jingjing Li[1], Ghia Euskirchen[1], Shin Lin[1, 5], Yiing Lin[1, 6], Axel Visel[7, 8,9], Trupti Kawli[1], Xinqiong Yang[1], Dorrelyn Patacsil[1], Cheryl A. Keller[3], Belinda Giardine[3], The mouse ENCODE Consortium, Anshul Kundaje[1], Ting Wang[4], Len A. Pennacchio[7, 8], Zhiping Weng[2], Ross C. Hardison[3#], Michael P. Snyder[1#]*

* These authors contributed equally to the work

# To whom correspondence should be addressed:

Michael P. Snyder (mpsnyder@stanford.edu)
Ross C. Hardison (rch8@psu.edu)

**Affiliations:**

1. Department of Genetics, Stanford University, Stanford, CA 94305, USA
2. Program in Bioinformatics and Integrative Biology, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01605, USA
3. Center for Comparative Genomics and Bioinformatics, Huck Institutes of the Life Sciences, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802, USA
4. Department of Genetics, Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO 63108, USA
5. Division of Cardiovascular Medicine, Stanford University, Stanford, CA 94304,USA
6. Department of Surgery, Washington University School of Medicine, St. Louis, MO 63110, USA
7. Lawrence Berkeley National Laboratory, Genomics Division, Berkeley, CA 94701,USA
8. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA
9. School of Natural Sciences, University of California, Merced, CA 95343,USA

The mouse ENCODE Consortium:

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,], Bing Ren[1]

**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.

# Supplementary Information

## Supplemental Methods and Analysis

### SMA1. Generating the human-mouse orthologous occupancy profiles

- SMA1.1. Rationale for TF selection
- SMA1.2. Peak call threshold (IDR vs FDR)
- SMA1.3. Rationale for using of cell lines
- SMA1.4. Function similarity between human and mouse model cell lines
- SMA1.5. Consistency between cell lines results and embryonic stem cells
- SMA1.6. TF protein sequences conservation
- SMA1.7. TFs with markedly different peaks numbers

### SMA2. Conserved and non-conserved features of TF OSs between mouse and human

- SMA2.1. Defining different genomic locations
- SMA2.2. Motif finding
- SMA2.3. ChromHMM
- SMA2.4. MeDIP-seq and MRE-seq

### SMA3. TF specific and location specific occupancy conservation

- SMA3.1. TF OSs sequence constraint
- SMA3.2. Mapping reciprocal orthologous sequences between human and mouse
- SMA3.3. Defining occupancy conserved TF OSs
- SMA3.4. Occupancy conservation of TFs with discordant peaks numbers

### SMA4. Divergence of TF occupancy is associated with changes of epigenetic signals

- SMA4.1. Comparing chromHMM states between TF OSs and orthologous sequences
- SMA4.2. Comparing DNA methylation level between TF OSs and orthologous sequences

### SMA5. Conservation of occupancy is associated with activity in multiple tissues

- SMA5.1. Calculating diversity index
- SMA5.2. Mouse transgenic enhancer assay

**SMA6. Conservation and diversity of TFs co-association**

- SMA6.1. Number of TFs associated with each binding region
- SMA6.2. Significance of pair-wise TF co-association
- SMA6.3. TF OSs occupancy conservation and association

**SMA7. Functional SNVs are significantly enriched in occupancy conserved TF OSs**

- SMA7.1. RegulomeDB SNVs and occupancy conservation
- SMA7.2. GWAS SNPs and occupancy conservation

## Supplemental Tables

**Table S1. Known biological functions of examined TFs**

**Table S2. Extended information for ChIP-Seq**

**Table S3. TF peaks and occupancy conserved TF peaks**

**Table S4. Extended information for TFs**

**Table S5. In vivo enhancer assay**

**Table S6. GO enrichment result for occupancy conserved TF OSs**

*Supplementary Tables and materials can be downloaded from http://mouse.encodedcc.org/publications/mcp02/

## Supplemental Figures

**Supplemental Figure1. Genome browser screenshot of ChIP-seq tracks**

**Supplemental Figure2. Function similarity between model cell lines**

**Supplemental Figure3. TFs protein sequences conservation**

**Supplemental Figure4. GO enrichment for occupancy conserved TF OSs targeted genes**

# Supplemental Methods and Analysis

## SMA1. Generating the human-mouse orthologous occupancy profiles

### SMA1.1. Rationale for TF selection

In this study, a diverse panel of TFs were chosen including those that bind DNA through specific consensus sequences, comprise part of the general transcriptional machinery such as RNA polymerase 2 (POL2), and modify or remodel chromatin (Details in Supplementary Table 1). For simplicity we refer to the entire collection as TFs, even though some are general factors. The new datasets reported here cover 100Mb (3%) and 85Mb (3%) of the human and mouse genomes respectively, and represent the largest collection of orthologous TFs occupancy repertoires between human and mouse to date. The screenshot of the ChIP-seq signal tracks are included in Supplementary Fig. 1.

The 34 TFs in the final list were chosen based on both unbiased representative selection and technical feasibility. All together, we tested 99 antibodies in all the four cell lines. 43 of them passed our antibody validation standard by western blot in at least one pair of human and mouse cell lines. Among the 43 TFs, the ChIP-seq results in 9 TFs could not meet the signal to noise ratio threshold set by ENCODE project. Thus, the 34 TFs in the final list were chosen based on both unbiased representative selection and technical feasibility.

### SMA1.2. Peak call threshold (IDR vs FDR)

One challenge when conducting ChIP-Seq comparisons across different TFs and species is setting peak-calling thresholds. Some studies rely on the FDR threshold based on the control data. However, a single empirical FDR threshold cannot be used for all different types of TFs across different sequencing depths. For the exact same ChIP dataset, if we change the sequencing depth of the control, for the same FDR threshold, the numbers of peaks passing the FDR threshold change dramatically. This makes it extremely difficult to use for automated analyses across large collections of datasets without explicit manual tuning. In contrast, IDR does not rely on thresholds based on the absolute value of the enrichment scores (FDR or signal strength) of peaks. Rather it looks for rank consistency

of peaks. It is largely immune to changes in sequencing depth of the control as well as the relative strength of different antibodies. Thus, IDR is more suitable to be used as a consistent threshold for large TF data sets and was used in our study. For a basic introduction to IDR please see http://www.personal.psu.edu/users/q/u/qul12/IDR101.pdf. For a detailed exposition, please see http://projecteuclid.org/euclid.aoas/1318514284.

**SMA1.3. Rationale for using cell lines**

Cell lines were used in this study because very large numbers of cells can be grown and treated in an identical fashion, and the cell lines represent relatively pure cell populations, whereas obtaining such purity from tissues or primary cells usually requires microdissection or fractionation by FACS. Each ChIP-seq assay (two replicates) conducted in this project used 100 million cells. It would be extremely difficult to obtain this amount of cells from primary cells/tissues for all the 34 TFs covered in this project. We note that previous comparative studies on TF binding examined a small number of factors in liver, which is one of the larger organs but is comprised of multiple cell types. Also, despite their utility, caveats also apply to the interpretation of data from primary cells or *ex vivo* expanded cells. For example, some isolation protocols induce stress response genes, and the expression patterns can change with passage. Such limitations can make it difficult to obtain consistent and reproducible results.

Another important point in our choice of cell lines is that not only did we want to generate these large datasets on the mapping of 34 TFs, but we also wanted to integrate the TF occupancy profiles with other functional genomics information like DNase I hypersensitive sites, chromatin modification states and DNA methylation profiles. Again, having enough cells so that all the assays are generated from the same type of cells was critical for the integrated analysis.

**SMA1.4 Function similarity between human and mouse model cell lines**

To further ensure that the cell lines we chose are analogous between the two species and do not differ on a global scale from primary cells. We used two approaches to examine the function similarity between the model cell lines used in this project.

1) Gene expression: RNA-seq for human and mouse cell lines and tissues were processed by the same pipeline[1]. FPKMs for each experiment were further normalized by quantile normalization. Cell types or tissues were clustered based on the expression level of 1vs1 orthologous genes. In order to generated tissue based cluster, genes showed significantly difference between the two species were removed. Hierarchical cluster was generated using hclust function in R (supplementary Fig. 2a).
2) Chromatin State: Chromatin states clustering analysis of the 15 mouse tissues and three human cells were conducted as described[1], basically, species-normalized chromatin activeness for each 1 vs 1 orthologous gene neighborhood were computed and used to cluster cells/tissues by hierarchical clustering (supplementary Fig. 2b).

**SMA1.5. Consistency between cell lines results and embryonic stem cells**

To further examine how robust are the results based on cell lines study, we generated five pairs of ChIP-seq data using human and mouse embryonic stem cells. As show in supplementary Fig. 3a, TF OSs in embryonic stem cells have similar genomic distribution as those in cell lines. TF OSs near promoter regions show higher level of occupancy conservation in primary cells compared with distal regions (supplementary Fig.3b). More important, TF OSs utilized by multiple tissues also tend to be occupancy conserved in embryonic stem cells (supplementary Fig.3c). Thus, the main conclusions that we draw based on experiments using cell lines are also supported by results using embryonic stem cells.

**SMA1.6 TF protein sequences conservation**

We examined TF evolutionary rates at the protein sequence level. We used the ratio of the number of nonsynonymous substitutions per non-synonymous site (Ka) to the number of synonymous substitutions per synonymous site (Ks) to estimate the selective pressure. As shown in supplementary Fig. 3, TFs generally have lower Ka/Ks compared with other human-mouse orthologous genes (median 0.07 vs 0.14, p-value < 2.2e-16 two-tail t-test), indicating that they are under negative selection. The TFs chosen in this project have been well studied in both human and mouse because of their important regulatory function, and not surprisingly, they have an even lower Ka/Ks ratio (median 0.04). We

further compared the functional domains TFs between human and mouse and found that both the functional domain compositions and amino acid sequences in the domains are almost identical in the two species.

**SMA1.7 TFs with markedly different peak numbers between human and mouse**

In order to minimize the difference introduced by technique, we thoroughly tested and validate ChIP-Seq antibodies, processed data with uniform pipeline. As a result, most the ChIP-Seq assays have similar total number of peaks between the two species. However, we do observed a few exceptions. E2F4, GATA1, JUND, MAFK and RDBP show markedly different peak numbers in erythroid progenitor cells; PAX5 and KAT2A in lymphoblast cells also have significantly difference peaks numbers.

We first examined the protein conservation of those TFs. We find that all those TFs show high level of protein conservation (SMA1.5). In addition, there is no correlation between peaks numbers difference and protein sequence conservation. Thus, protein sequences difference is not the reason for the observed peak number difference. We next examined the gene expression level of those TFs. Overall, TF expression in the same cell type show higher similarity than that in the same species. We find TFs show markedly different peaks tend to have different expression level between the two species. For example, the expression level for GATA1 expression level is about twice higher in mouse compared with those in human (20 vs 12). MAFK expression level in MEL is only less than 1/10 of that in K562 (0.3 vs 3.7). Thus, the different peak numbers for those TFs are likely to be caused by the different TF expression level. Other TFs, such E2F4 (MEL vs K562), KAT2A (CH12 vs GM12878) and PAX5 (CH12 vs GM12878) do have similar expression level in two species. One explanation is that TF occupancy is usually determined not only by the TF itself but also by its co-factors. Thus, the number of TF OSs cannot always be predicted accurately just from the abundance of the TF. Another possible explanation is that the actual TF concentration in nuclei can be different from the total RNA level examined by RNA-seq using whole cells.

**SMA2. Many general features of TF occupied sequences are well conserved between human and mouse**

### SMA2.1. Defining different genomic locations

Transcription start sites (TSSs) were defined by ENCOCDE consortium[1]. Promoter regions were defined as 2kb upstream and downstream of the TSSs. Distal regions were defined as 10kb away from TSSs. The rest genomic regions were defined as middle regions. All the three genomic locations are exclusive to each other and the priority during the definition is promoter, distal and middle. Each TF OS was assigned to one and only one the genomic location. If TF OS overlapped with multiple regions, the center of the OSs was used to define which region to assign.

### SMA2.2. Motif finding

To compare human and mouse regulatory networks, we applied the *de novo* motif discovery approach that we developed previously[2], and obtained a list of high-confidence sequence motifs using the ChIP-Seq datasets in human K562 cells and mouse MEL cells. For each ChIP-Seq dataset, we asked our computational pipeline to report up to five significant motifs. Typically one of the motifs is the canonical motif of the TF, reflecting its DNA-binding specificity, and we call this the primary motif. If the TF does not have a DNA binding domain, we define the strongest motif as its primary motif. We call the remaining motifs secondary motifs. When the primary motifs of a pair of orthologous TFs are compared, they are either "conserved" or "not conserved" based on whether the similarity between them passes cutoff 1e-5. Because a TF may have multiple secondary motifs, the secondary motifs of two orthologous TFs are "partly conserved" if a subset, but not all, of the motifs are conserved. When neither the human TF nor the mouse TF has a secondary motif, we assign the situation as motif "not available".

In order to examine if those secondary motifs have real biology meaning instead of noise, we compared those sequences to the consensus motifs of other TFs. As shown in following table, many of them turned out to be the consensus motif of known TFs. The co-localization among TFs identified by primary motif and secondary motifs are also supported by previous publications. For example, the predicted interaction between human JUND and GATA1 (the secondary motif) was consistent with previous publication[3]. Lack of secondary motif found from mouse JUND OSs was also consisted with the fact that mouse JUND was known to interact much fewer proteins than JUND in

human[4]. The conserved secondary motifs NF-E2 in MAFK OSs also supported by the previous observation that MAFK can form a heterodimer with the P45 subunit of NFE2 [5].

| | Primary motifs | | Secondary motifs | |
|---|---|---|---|---|
| | Human | Mouse | Human | Mouse |
| BHLHE40 | BHLHE40 | BHLHE40 | | GATA1 |
| CHD2 | UA1 | UA1 | ZNF143 | NFY |
| CTCF | CTCF | CTCF | | |
| E2F4 | E2F4 | E2F4 | NFY; SP-1; NRF1 | ZNF-143; NFY; NRF1 |
| GATA1 | GATA1 | GATA1 | TAL1 | ETS1 |
| JUND | AP-1 | | GATA1 | |
| MAFK | v-MAF | | NFE2 | NFE2 |
| MAX | MAX | MAX | USF; ELF1; SP-1 | USF |
| MAZ | MAZ | MAZ | CTCF; CREB; ELF1 | |
| MXI1 | MAX | MAX | RFX5; SP1 | |
| MYC | MYC | MYC | YY1 | |
| P300 | GATA1 | GATA1 | PU.1; AP-1 | PU.1; RUNX1 |
| RAD21 | CTCF | CTCF | | |
| RCOR1 | GATA1 | GATA1 | GFX1; AP-1 | PU.1; RUNX1 |
| RDBP | TBP | NRF1 | | |
| SMC3 | CTCF | CTCF | | |
| TAL1 | GATA1 | GATA1 | | |
| TBP | TBP | TBP | SP1 | A-box |
| UBTF | | | | |
| USF2 | USF | USF | NFY; YY1 | SP-1; NFY |

**SMA2.3. ChromHMM**

ChromHMM was applied on the ChIP-Seq data of five histone modifications to learn a multivariate HMM model for segmentation of mapped genome in each cell type. Specifically, the ChIP-Seq mapped reads were first pooled from replicates for each of the five histone modifications (H3K4me3, H3K4me1, H3K36me3, H3K27ac and H3K27me3). These mapped reads were first processed by ChromHMM into binarized data in every 200 bp window over the entire mapped genome, with ChIP "input" reads as the background control. To learn the model jointly from mouse and human, a pseudo genome table was first constructed by concatenating mouse mm9 table and human hg19 table, then the model was learned from the binarized data in all 4 cell lines, giving a

single model with a common set of emission parameters and transition parameters, which was then used to produce segmentations in all cell types based on the most likely state assignment of the model. We tried models with up to 20 states and selected an eight-state-model as it appeared most parsimonious in the sense that all eight states had clearly distinct emission properties, while the interpretability of distinction between states in models with additional states was less clear.

### SMA2.4 MeDIP-seq and MRE-seq

MeDIP-Seq and MRE-Seq experiments were performed as previously described[6]. The reads were aligned to hg19 and mm9 using BWA[7]. MRE-seq reads were further normalized for difference in enzyme efficiency.

## SMA3. TF specific and location specific occupancy conservation

### SMA3.1. TF OSs sequence constraint

phyloP wiggle track were downloaded from UCSC browser. Specifically, hg19 phyloP46way track was used for human and mm9 phyloP30way track was used for mouse. This average phyloP score were calculated at 1 base resolution in 200bp regions centered on the summit of TF peaks.

### SMA3.2. Mapping reciprocal orthologous sequences between human and mouse

Orthologous DNA sequences between human and mouse were mapped by bnMapper (https://bitbucket.org/james_taylor/bx-python/wiki/bnMapper) using reciprocal chain with default setting (bnMapper.py -f BED12).

### SMA3.3. Defining occupancy conserved TF OS and occupancy conservation level

For a given TF OS, if 1) It has reciprocal orthologous sequence in other species 2) The orthologous sequence has at least one base overlap with the binding regions of the same TF in the other species, this TF OS is defined as occupancy conserved TF OS. For a given TF, its occupancy conservation level was defined as the:

$$\frac{\#\ of\ \text{occupancy conserved TF OSs}}{\#\ \text{of total TF OSs}}$$

### SMA3.4. Occupancy conservation of TFs with markedly different peak numbers

To account for possible biases, we performed two related analyses. Since DNA sequences at promoters tend to be more conserved, we adjusted the occupancy conservation by the local sequence conservation difference (see Extended Data Fig. 4a).

For most of the TFs examined, the occupancy conservation level for the same TF is highly correlated between the two species. However, as described in SMA1.7, there are a few TFs show markedly different total binding peaks between the two species. Since we used the ratio between occupancy conserved peaks and total peaks to calculate the occupancy conservation level, specie with fewer peaks tends to have relative higher occupancy conservation level compared with species with more peaks. In order to avoid extremely high or low occupancy conservation level introduced by this peak number difference. We excluded those TFs and redid the analysis. As showed in Extended data Fig. 5b, our conclusion that the level of occupancy conservation differed both among TFs and with genomic location relative to the TSS still hold true.

## SMA 4. Divergence of TF occupancy is associated with changes of epigenetic signals

### SMA4.1. Comparing histone modification between TF OSs and orthologous sequences.

For a given TF binding sequences, the orthologous sequences in the other species were mapped as describe in SMA3.2. Each TF OS and its orthologous sequences are assigned to one and only one chromatin state. If given TF OS overlap with multiple states, it is assigned to the states that covered the largest proportion of the sequence.

### SMA4.2. Comparing DNA methylation level between TF OSs and orthologous sequences.

For each TF OS, we first calculated the MeDIP-seq signal in the 100bp window centered on the summit of the peak. We then calculated the MeDIP-seq signal in TF OS orthologous sequences using MeDIP-seq in the orthologous cell line. Orthologous sequences with large insertions (>100bp) or deletions (>50bp) were removed. MeDIP-seq signals were further normalized by 1) total number of MeDIP-seq reads 2) the length of the sequences 3) the number of CpG per region.

$$\frac{number\ of\ reads\ in\ region}{Total\ number\ of\ reads * lengh\ of\ seqeunces * \#\ of\ CpG}$$

## SMA 5. Conservation of occupancy is associated with activity in multiple tissues

### SMA5.1. Calculating diversity index

All TF OSs were merged into non-overlap genomic regions. For chromatin accessibility calculation, DHSs narrowPeaks files were downloaded from http://www.genome.ucsc.edu/ENCODE/. Accessibility of a given binding regions was presented by the signal of DHS narrowPeak located in this region. If there are multiple DHS narrowpeaks located within one binding region, the highest signal was chosen. For enhancer activity calculation, normalized H3K27ac ChIP-Seq signals ($^{Reads\ located\ in\ the\ given\ region}/_{Total\ number\ of\ reads}$) within +/- 50bp of the peak summits were used. Shannon Diversity was calculated using Vegan package [8] in R.

### SMA5.2. Mouse transgenic enhancer assay

We randomly picked 10 binding regions from the GATA1 OSs that were occupancy conserved between human K562 cell and mouse MEL cells. The human sequences of those 10 binding regions were used in the mouse transgenic enhancer assay. The enhancer assay was conducted as described before[9]. Other *in vivo* enhancer assay results were downloaded from VISTA Enhancer Browser[10].

### SMA5.3. GO enrichment for genes regulated by occupancy conserved TF OSs

GO enrichment analysis were conducted by using GREAT[11] with default setting. The result showed that the functions of occupancy conserved TF OSs target genes are highly conserved between human and mouse (Supplementary Fig. 4). Those functions include both general biological processes, such as metabolic process (CTCF OSs in all the four cell lines) and cell cycle (E2F4 OSs in all four cell lines), and tissue-specific functions,

such as immune system process (EP300 and IRF4 OSs in GM12878 and CH12) and erythrocyte differentiation (GATA1 OSs in MEL and K562). All the enriched GO categories for each TF in all the four cell lines are listed in Supplementary Table 6.

**SMA 6: Conservation and diversity of TFs co-association**

**SMA6.1. Number of TFs associated with each binding region**

OSs for different TFs were first merged into non-overlap genomic regions. If the binding region and its orthologous sequences are occupied by at least one orthologous TF, the regions are defined as occupancy conserved regions. The number of TFs associated with each binding region is defined by the number ChIP-Seq peaks that overlapped with the binding region. Multiple ChIP-Seq peaks from the same TF are count only as one.

**SMA6.2. Significance of pair-wise TF co-association**

In order to examine the significance of pair-wise TF association, we first calculated the number of regions that overlap between peaks of two TFs (A and B). We next generated 500 sets of pseudo ChIP-Seq peaks that have similar peaks length and genomic distribution as TF B binding peaks. We then calculated the number of overlapped peaks between TF A binding peaks and each pseudo peak set. The overlapped peak numbers between TF A and the 500 pseudo sets were used as the background distribution. The actual overlap number was then compared with the background distribution and its Z-score was used to represent the significance of the overlap.

**SMA6.3. TF OSs occupancy conservation and association**

In order to examine the conditional occupancy conservation between TF A and B, we first divided TF A binding peaks into two categories: 1) Binding peaks overlapped with TF B peaks; 2) Binding peaks not overlapped with TF B peaks. In each category, we further divided the peaks into two subgroups according to the occupancy conservation status. 1) Occupancy conserved. 2) Occupancy not conserved. Two-side Fisher's exact test was used to test the significance of the association between occupancy conservation and TF co-association. This test was then applied to every pair-wise TF association. P-value was further adjusted by BH procedure.

## SMA 7: Functional SNVs are significantly enriched in occupancy conserved TF OSs

### SMA7.1. RegulomeDB SNV and occupancy conservation

SNPs assigned with pre-calculated regulatory potentials were downloaded from http://www.regulomedb.org/downloads. We then calculated the number of SNPs with high regulatory potentials located within occupancy conserved TF OSs and non-conversed TF OSs. For background calculation, we downloaded Common SNPs (138) from UCSC genome table browser track and also calculated the overall common SNPs distribution within occupancy conserved TF OSs and human specific TF OSs. Fisher's exact test was conducted to examine the significance of enrichment.

### SMA7.2. GWAS SNPs and occupancy conservation

GWAS catalog file was downloaded from (http://www.genome.gov/admin/gwascatalog.txt) in Feb 2013. CEU SNPs linkage data were downloaded from hapmap. GWAS SNPs were defined as lead SNPs. For each lead SNP, we further located SNPs in linkage disequilibrium (LD) ($r^2 > 0.9$). If either the lead SNP or linked SNPs in high LD are located within a TF OS, we assigned the lead SNP to this TF OS. Lead SNPs that overlapped with exons were further removed. For each lead SNPs, if either the SNP itself or the LD SNPs are located within a given TF OS, it was assigned to that TF OS. Lead SNPs that can be assigned to multiple TF OSs were removed. Altogether, 1916 and 2231 lead SNPs were assigned to TF binding regions in the GM12878 and K562 cell lines, respectively. 734 (38.3%) and 639 (28.6%) reside in occupancy conserved TF OSs. Compared with the background distribution of all dbSNPs, GWAS SNPs are significantly enriched in occupancy conserved TF OSs (p-value < 2.2e-16 Fisher's Exact Test). For each phenotype, we calculated the number of its associated SNPs in occupancy conserved TF OSs and non-conversed TF OSs. We used the distribution of all common SNPs as the background. The significance of enrichment was calculated by two-side Fisher's exact test.

## Supplemental Tables

### Table S1. Functions of examined TFs

This table lists all the TFs examined this study. In total, there are 34 TFs. 23 are sequence-specific TFs (highlighted in pink), 6 are chromatin-associated factors (highlighted in blue), and 5 are general factors associated with POL2 or POL3 (highlighted in orange). Data from the Luscombe lab census of human transcription factors[12] was used to classify TFs into families based on the presence of DNA binding domains. The functions and disease association features were collected from GeneCards (www.genecards.org).

### Table S2. Extended information for ChIP-Seq

This table lists all the ChIP-Seq experiments conducted in this study. In total, there are 55 datasets for mouse and 60 datasets for human. Each dataset contain at least two replicates. ChIP-Seq using the same TF and same cell lines from different institutes were merged in the analysis. Column A lists all the individual replicates; Column B lists all the dataset; Column C and D list the TF and cell line name. Column E is the ENCODE antibody ID and Column F is the actual antibody catalog. Column G and H are the ChIP-Seq quality information for each replicates. Column I, J and K are the access ID to different database. NSC is Normalized Strand Cross-correlation and RSC is Relative Cross Correlation Coefficient. Both were used to measure the signal to noise ratio in the ChIP-Seq assay[13].

### Table S3. TF peaks and occupancy conserved TF peaks

This table lists both total and occupancy conserved binding peaks in each genomic location. Column A is the TF name, Column B is the cell line name. Column C-E list the number of peaks in each of the three genomic locations. Column F-H list the number of occupancy conserved peaks in each of the three genomic location.

**Table S4. Extended information for TFs**

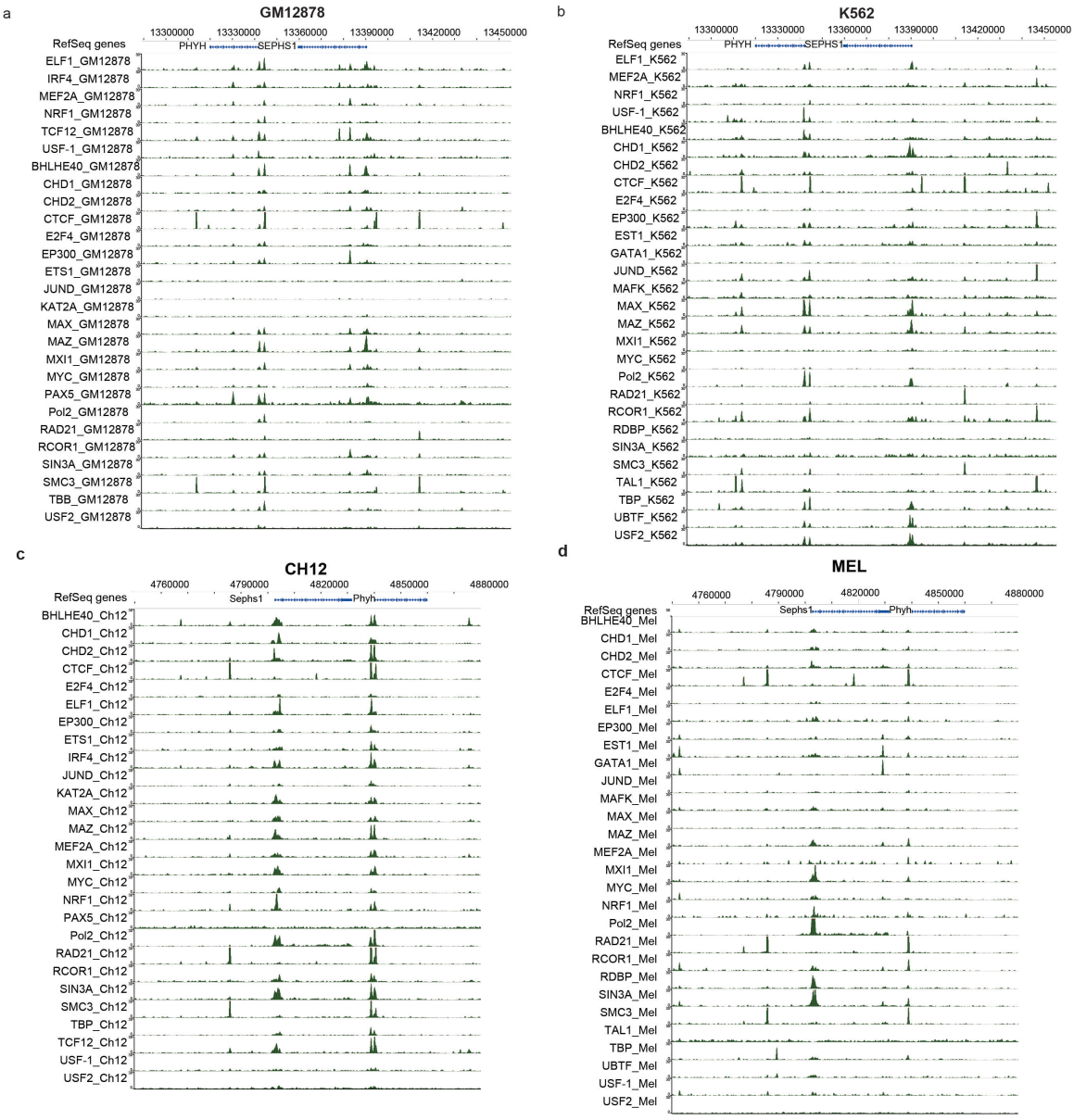
This table lists the expression level, total peaks numbers and protein conservation for the TFs examined in this study. Column B-E show the averaged expression level (FPKM) in each cell lines. Column F-I show the total number of binding peaks for each ChIP-Seq assay. TFs without ChIP-Seq information were labeled as "NA". The last column shows the Ka/Ks of each TF.

**Table S5. In vivo enhancer assay**

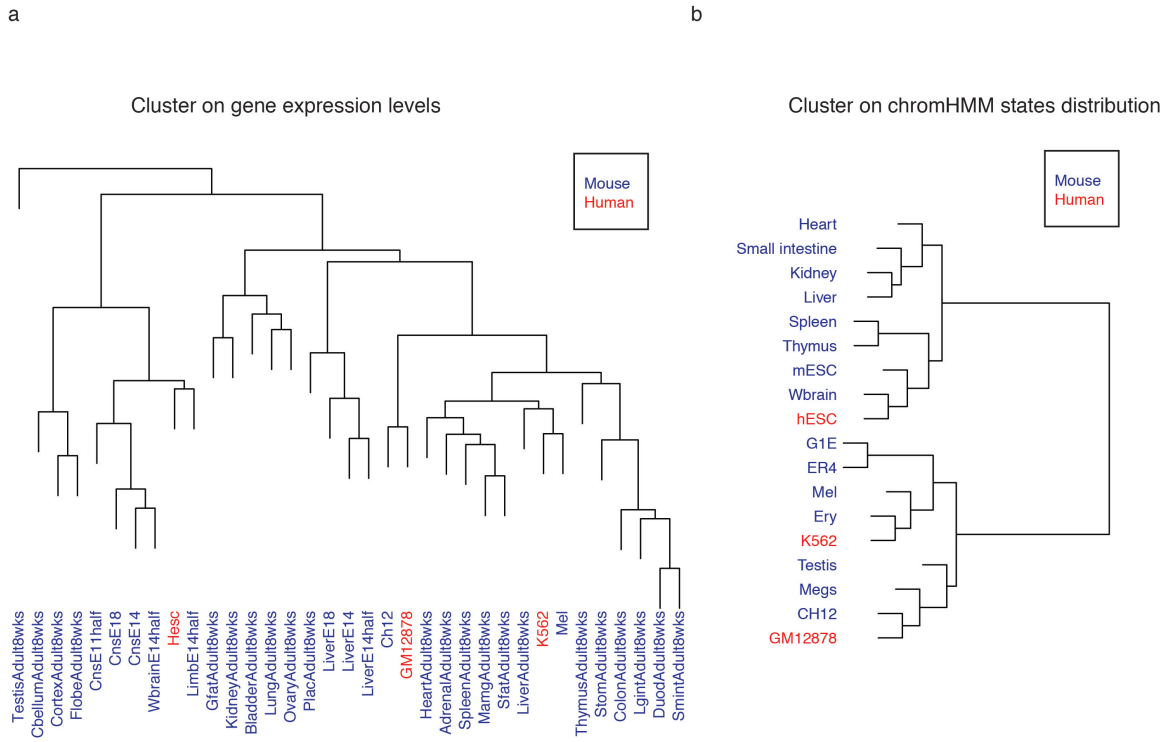
This table lists the in vivo enhancer assay results for different groups of GATA1 OSs. TF OSs highlighted by yellow are ten randomly picked occupancy conserved GATA1 OSs. The pink and blue highlighted regions are tested sequences deposited in Vista browser that overlapped with GATA1 OSs identified in this project. Pink ones overlapped with occupancy-conserved GATA1 OSs, blue ones overlapped with mouse specific GATA1 OSs.

**Table S6. GO enrichment result for occupancy conserved TF OSs**

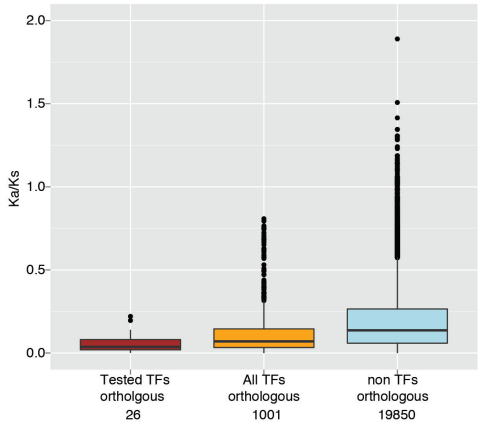
This table lists the functional enrichment of genes regulated by occupancy conserved TF OSs. The enrichment were calculated using GREAT (http://great.stanford.edu/)



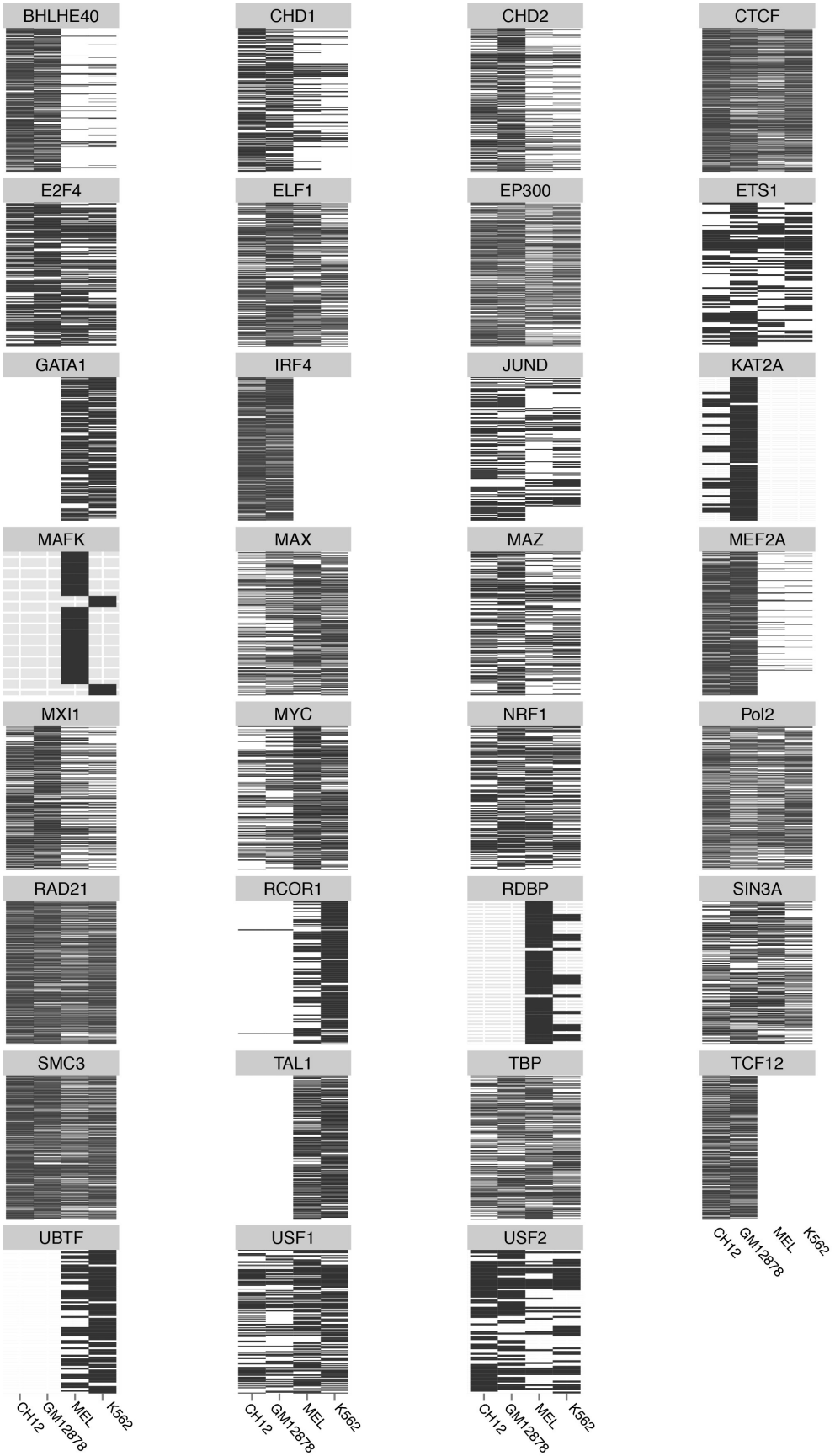

**Supplementary Figure 1 | Genome browser screenshot of ChIP-seq tracks.** Y-axis of the track is normalized ChIP-seq signal. Each panel represents one cell line.




**Supplementary Figure 2 | Function similarity between model cell lines. a.** Cell types or tissues were clustered based on the expression level of 1vs1 human-mouse orthologous genes. **b.** Chromatin states clustering analysis of the 15 mouse tissues and three human cells.




**Supplementary Figure 3 | TFs protein sequences conservation.** The boxplot represents the Ka/Ks distribution of human and mouse orthologous genes. "non TFs" represents all 1:1 human and mouse orthologous genes that are not TFs. "All TFs" are all 1:1 orthologous transcription factors. "Tested TFs" are the 32 TFs that were analyzed in this study.


BHLHE40
CHD1
CHD2
CTCF
E2F4
ELF1
EP300
ETS1
GATA1
IRF4
JUND
KAT2A
MAFK
MAX
MAZ
MEF2A
MXI1
MYC
NRF1
Pol2
RAD21
RCOR1
RDBP
SIN3A
SMC3
TAL1
TBP
TCF12
UBTF
USF1
USF2
CH12
GM12878
MEL
K562

**Supplementary Figure 4 | GO enrichment for occupancy conserved TF OSs targeted genes.** Each panel shows the GO categories that are significantly enriched in occupancy-conserved OSs of the given TF. Each row in the panel represent one GO category, each column represents one model cell line. If a given GO category show significantly enrichment, it is highlighted by black. Otherwise, it is highlighted by white.

## References


1. MouseENCODE Consortium. An Integrated and Comparative Encyclopedia of DNA Elements in the Mouse Genome. *Nature ,Accepted*
2. Wang, J. *et al. Genome research* (2012). doi:10.1101/gr.139105.112
3. Kawana, M., Lee, M. E., Quertermous, E. E. & Quertermous, T. Cooperative interaction of GATA-2 and AP1 regulates transcription of the endothelin-1 gene. *Mol. Cell. Biol.* **15,** 4225–4231 (1995).
4. Ravasi, T. *et al.* An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man. *Cell* **140,** 744–752 (2010).
5. Du, M.-J. *et al.* MafK/NF-E2 p18 is required for beta-globin genes activation by mediating the proximity of LCR and active beta-globin genes in MEL cell line. *Int. J. Biochem. Cell Biol.* **40,** 1481–1493 (2008).
6. Maunakea, A. K. *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466,** 253–257 (2010).
7. Li, H. & Durbin, R. *Bioinformatics* **25,** (2009).
8. Oksanen, J., Blanchet, F. G. & Kindt, R. vegan: Community Ecology Package. R package version 2.0-7.
9. Pennacchio, L. A. *et al. Nature* (2006). doi:10.1038/nature05295
10. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. *Nucleic acids research* (2007). doi:10.1093/nar/gkl822
11. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28,** 495–501 (2010).
12. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics* **10,** 252–263 (2009).
13. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research* **22,** 1813 (2012).