Supplementary Information

Integrative analysis of haplotype-resolved epigenomes across human tissues

Inventory of Supplementary Information

Methods

References

Supplementary Tables

Methods:

Obtaining human tissue samples

Adrenal, stomach, lung, heart, muscle, ovary, small bowel, colon, spleen, adipose, bladder, thymus and liver tissues were obtained from deceased donors at the time of organ procurement at the Barnes-Jewish Hospital (St. Louis, USA). Samples were flash frozen with liquid nitrogen. Research consent from family was obtained, and this study was approved by Mid-American Transplant Services. Tissues were derived from donors with identification numbers STL001 (donor 1), STL002 (donor 2), STL003 (donor 3) and STL011 (donor 4).

ChIP-seq

ChIP-seq was conducted as previously described[2] with 20ug of chromatin and 5ug of antibodies. The following antibodies were used: H3K27ac (Active Motif: 39133), H3K4me3 (Millipore:04-745), H3K4me1 (Abcam: ab8895), H3K36me3 (Active Motif: 61021), H3K27me3 (Active Motif: 61017) and H3K9me3 (Abcam, ab8898).. Chip and input library preparation and sequencing procedures were carried out as described previously according to Illumina protocols with minor modifications (Illumina, San Diego, CA)

Hi-C on human tissue samples

Human tissue samples were flash frozen and pulverized prior to formaldehyde cross-linking. Hi-C was then conducted on the samples as previously described[3].

Cis-regulatory elements prediction

We used a random-forest based algorithm, RFECS (Random Forest for Enhancer Identification using Chromatin States), for the purpose of enhancer and promoter prediction[4]. Briefly, the enhancer identification procedure was as follows. We used histone modification profiles at distal p300 binding sites in H1 to train a random-forest for enhancer prediction. We constructed the forest using a selected set of histone modifications that provide largely non-redundant information, including H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, and H3K9me3. The enrichment of these modifications was determined and used in 100 bp bins from -1 to +1kb along the p300-binding sites or selected non-p300 background sites to train the RFECS classifier. Using this classifier, we predicted enhancers genome wide in all 22 human tissues and combined them with predictions in H1, H1-derived cell-line, and IMR-90, for which predictions had been previously made[5]. In case of promoter prediction, our training set comprised of known UCSC TSS overlapping DNase I hypersensitive sites in H1 as representative of promoters, and a set of randomly selected genomic regions and distal p300 sites as representative of non-TSS background. The features were the same 6 core histone modifications mentioned above. Using the overlap of known UCSC TSS with predicted promoters at various voting percentage cutoffs[4], we selected a cutoff in each cell-type that gave us at least 50% overlap with UCSC TSS. We considered any enhancer that lay within 2.5kb of a predicted promoter within the same cell-type as a false positive and filtered this out from our final list of predicted enhancers.

To filter for strong promoters, we combined all promoter predictions across 28 cell-types merging predictions within 1kb of each other. We applied a Z-score normalization followed by logit transformation to the input-normalized H3K4me3

RPKM values at this combined list within each cell-type. Then we clustered the Z-score normalized H3K4me3 levels within each cell-type using fast k-means ++[6] and selected the optimal number of clusters using a Davies-Bouldin measure[7]. Based on the clustering we assigned present/strong enrichment or absent/weak (1 or 0) values of H3K4me3 in each cell-type to a particular promoter.

Identification of tissue-restricted and non-restricted regulatory elements

In order to identify tissue-restricted enhancers or promoters, we combined all enhancer or promoter predictions across 28 cell-types merging predictions within 1kb of each other. We applied a Z-score normalization followed by logit transformation to the input-normalized H3K27ac RPKM values at this combined list of enhancers within each cell-type. We then clustered this normalized H3K27ac level within each cell-type using fast k-means ++[6] and selected the optimal number of clusters using a Davies-Bouldin measure[7]. Based on the clustering we assigned present/strong enrichment or absent/weak (1 or 0) values of H3K27ac in each cell-type to a particular enhancer or promoter. If the enhancer or promoter had presence of H3K27ac in two or less cell-types among the 28 cell-types considered, we called it tissue-restricted. If the enhancer or promoter had enrichment of H3K27ac in 10 (or 15) or more cell-types we declared it to be non-restricted.

TF Motif enrichment analysis in each tissue

We used HOMER[8] to find motif enrichment at -0.5 to +0.5kb around the tissue-specific enhancers within each cell-type. We selected any motif instance that was

enriched at p-value < 10e-10 in at least one tissue or cell-type. We performed hierarchical clustering on all these motif instances using the following distance metric:

$$D(a,b) = \frac{1}{w}\sum_{i=1}^{w}\frac{1}{\sqrt{2}}\sum_{L\in\{ACGT\}}(a_{i,L} - b_{i,L})^2$$

If the motifs were of different length, we considered all possible alignments of the two motifs and retained the minimum distance measure.

We generated clusters from the hierarchical clustering by testing several different distance cutoffs. For a particular clustering, we calculated the enrichment of cell-type or tissue within a cluster of motifs using the hyper-geometric distribution. We select the minimum number of clusters, at which each of the 28 cell-types was enriched in at least one cluster at a p-value <0.05. This reduced the overall 824 motif instances to a set of 29 clusters. We found that many of the enriched putative binding motifs were of particular transcription factors known to be important in maintaining the cell/tissue-type's identity and function[9-20].

Hierarchical clustering of tissues based strong enhancers and promoters

We selected all enhancers and promoters that showed strong enrichment of H3K27ac in at least one along the 28 cell-types using procedure described above. We used the overlap of these "strong enhancer" or "strong promoters" assignments between cell-types as a distance measure and performed hierarchical clustering using MATLAB. Further, the cell-types were ordered using the optimal leaf ordering algorithm[21], also in MATLAB.

Repetitive element annotation and Shannon-entropy analyses

Repetitive element annotations were downloaded from Repeatmasker track in UCSC genome browser in hg18 and filtered for length of over 1kb to remove fragments and poorly annotated elements. Shannon-entropy-based analysis was conducted as previously described[22]. A threshold of H3K27ac RPKM >0.15 was used to reduce bias resulting from poorly covered regions. Tissue specificity was plotted as 2 to the power of the entropy score.

Mappable HERV-H annotations in Extended Data Fig. 3c were obtained from a previously published study[23]. For matrix of H3K27ac enrichment in Extended Data Fig. 3f, average H3K27ac enrichment (RPKM) was calculated for each subfamily of class I ERVs, as annotated in Repeatmasker. The average is then normalized in terms of each cell- or tissue-type (normalized by rows) and displayed as a heatmap.

***c*REDS filtering and RNA-seq analyses**

All predicted enhancers that reside within 500bp of a strong H3K4me3 promoter (defined previously) were selected as *c*REDS. For analysis of RNA-seq signal surrounding *c*REDS as described in Fig 1d, we calculated RPKM values of RNA-seq signal surrounding defined *c*REDS for the 16 tissues. As enhancer controls, we selected enhancers that were at least 2.5kb away from any predicted promoter. As promoter controls, we selected any strong H3K4me3 enriched promoter that was at least 2.5kb away from a predicted enhancers.

Transfection and luciferase reporter assay

Luciferase reporter assays were carried out as previously described[24] with several modifications. Briefly, *c*REDS harboring either enhancer or promoter marks in H1 hESCs and the converse signature in IMR-90, were selected. These regions along with 2 negative control sites, which have no detectable enrichment of any tested histone modifications, were amplified by PCR (primers sequences provided in table below) and cloned into pGL3-enhancer or pGL3-promoter vectors after restriction digest with appropriate enzymes to generate cohesive ends. After validation of sequence by Sanger sequencing, constructs were transfected in H1 hESCs with Fugene HD (Roche) at a 4:1 reagent to DNA ratio. Transfected cells were cultured for an additional 2 days prior to harvest for screening. The Dual-Luciferase Reporter Assay kit (Promega Cat#:E1960) was used according to manufacturer's protocol. The adjusted firefly luciferase activity of each sample was normalized to the average of active of 2 negative control regions.

Primer sequences used for PCR amplification of cREDS

Genomic location (hg18)	5' primer	3' primer
chr11:67,533,737-67,535,686	CAGCAGACTTGCGTCAAGAG	GGAGATTCCAGTCCACCTGA
chr22:43,986,093-43,988,247	GTCCGACCTTTGCTCTACCA	CAAAGCGACTCTGCAGACAG
chrX:53,326,364-53,328,474	AAAGGGCGAGACAGAAGACA	CAGGAGCCCTACATCCTTCA
chr3:113,841,497-113,843,650	TGCACGGAGGTCATAAAACA	CTGGCAAGGAGGTTTCTGAG
chr17:72,826,255-72,828,421	CTATCCCTGGGGCCATTATT	TGGCACACTGGAAAATGAAA
chr9:5,439,805-5,441,366	GGGCTTTCTTAACCCTCACC	AATCGAATGCAGCAATGAAA
chr4:154,397,418-154,399,374	GGCTTGTGGAACTGGACTGT	TGCATGAACAAATGGCTCTC

chr19:60,752,339-60,754,449	GCCAAGATGGACCACTGAAC	TTGAACCCAGGAAGTCAAGG
chr3:126,321,501-126,323,500	ACTGGGGGAAGAAGAAGAGC	CCAGCATTCAGGGTTCTCTC
chr2:74,460,022-74,461,828	CCCAGCCTATCACTGCCTAA	GGCCAGTGAATGAAAGCAGT
chr3:113,841,057-113,842,982	ATTCTTGACCGAGCTGAGGA	AAAAGCACAAAGCAGGGAGA
chr2:74,459,856-74,462,020	GCAACAGCAACTCCATGAACC	AGGGTAGAGCGGGGTAAAGT
chr3:150,574,945-150,577,215	ACATCCGTTTCTATCAGCTGTGC	ACAGCATGGAAAGAATGTGAGCA
chr6:42,801,658-42,803,569	TGAAGCCTGTGAGCTCTTGG	AACAGACAACCGGCACTAGG
chrX:19,597,483-19,599,676	GACATCACAGGACCAAGGCA	ATCTCAAACACTCCGGCTCC
chr11:67,563,195-67,565,272	AGAATTTCCTGCCCACCCTCT	CCACAGACAGTTCCCAACCAC
chr3:113,837,192-113,839,320	GTTGGGCGCACATAGGATCAA	CCTACACATGAGCCCAGGAGA
chr11:65,534,889-65,536,864	AGTTTGTGTCCTGTGGGCTC	ATAGGGTAGGGGCAGGTCAG
chr15:93,188,885-93,190,727	GGCCTTGACTCTCCCAGAAC	GTGCATCCTCTGTCCCCAAA
chr16:48,857,088-48,858,837	GTGTCTGACCCTGGATGTGG	TCCCTTCTCCTCTCCAGCAA

Zebrafish reporter assay

Zebrafish reporter assays were conducted as previously described[5]. Selected *c*REDS regions were amplified by PCR and cloned into the pT2MX vector 3' of the GFP gene by Infusion cloning according to manufacturer's protocol (Clontech). For each construct and control vector, 100 embryos were injected. Approximately 50% of fertilized embryos survived to day 3 for imaging. Images were generated with a Nikon C2 confocal microscope.

Variant calling and haplotyping

In order to call variants of each of the four tissue donors, whole genome sequencing data for each individual were mapped to hg18 reference genome using Novoalign. We excluded unmapped and non-uniquely mapped reads and also removed PCR

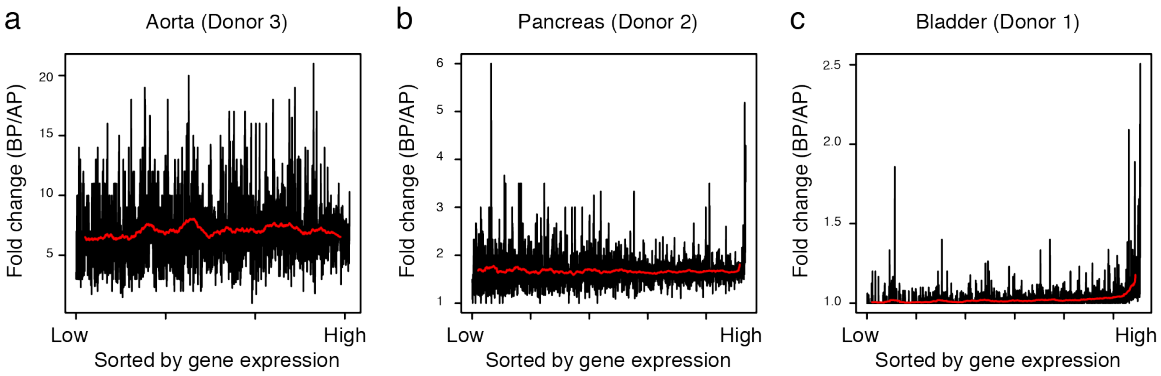
duplicate reads using Picard tools. To call variant, the mapped reads were processed according to Genome Analysis Toolkit (GATK) best practices guidelines, including indel recalibration, variant realignment, variant calling using the Unified Genotyper, and variant recalibration. Based on the variant information, new reference genomes (hg18) for each individual were constructed after masking variants.

Haplotypes were constructed using the previously described HaploSeq method[25]. Firstly, Hi-C reads from each donor were used as input sequence into the HapCUT software[26] in order to generate haplotype predictions. Next, Hi-C data was combined with WGS mate-pair data for the donor genomes to generate haplotype blocks. For each chromosome, several haplotype blocks were generated and the "Most Variants Phased" block is the one containing the majority of variants on each chromosome. This block was used as a seed haplotype for local conditional phasing by utilizing population sequencing data obtained from 1000 genome project. After performing local conditional phasing we can generate two haplotypes for each chromosome, one for the maternal allele and one for the paternal allele. Since we do not have information regarding the parent of origin in each donor genome, one allele is designated as P1 (parent1) and the other as P2 (parent2). Unfortunately, we cannot phase across the two chromosomal arms of chromosome 9, thus we independently phased both arms. The performance of haplotyping was evaluated based on resolution and completeness as shown in Supplementary Table 6. The resolution was defined as the fraction of phased SNPs among identified SNPs. The completeness was defined as the size, in base pairs, of the span of haplotype block in each chromosome.

Sequence read alignment in haplotype-resolved context

We realigned ChIP-seq datasets for six core histone modification marks in haplotype-resolved context. Single end sequencing data and paired-end H3K27ac sequencing data were aligned to the variants masked individual reference genome (hg18) using Stampy[27]. Unmapped and non-uniquely mapped reads were removed. The reads spanning variant loci were split into the P1 and P2 alleles according to the sequence match in each variant between two alleles. After splitting reads into the P1 and P2 alleles we removed PCR duplicate reads with Picard.

For 37 mRNA-seq data, we mapped the paired-end reads to a variant masked transcriptome genome using Novoalign. We constructed transcriptome genome using Useq software based on Gencode annotation (hg18). The mapped reads were split into the P1 and P2 allele informative reads according to the sequence match in each variant between two alleles. The duplicate reads were considered as PCR duplicates and removed with Picard. To determine whether duplicate reads in mRNA-seq datasets originated from PCR duplication, we investigate the distribution of duplicate reads in terms of gene expression levels. If the duplicate reads are biased to the highly expressed genes the duplicate reads reflect gene expression levels. If not, the duplicate reads can be considered as PCR duplicate reads. We observed that the samples containing high duplicate reads showed uniformly distributed duplicate reads regardless of gene expression levels (Supplementary Fig. 1), indicating that the duplicate reads contain a lot of PCR duplicate reads.



Supplementary Fig 1. The distribution of duplicate reads in mRNA-seq data according to gene expression. The fold changes of mRNA-seq read numbers are shown before and after removing duplicates for a) Aorta from donor 3, b) pancreas from donor 2 and c) bladder from donor 1, from which 84%, 44%, and 10% were duplicate reads respectively. The x-axis indicates ranked genes in terms of gene expression levels. In bladder, highly expressed genes show more duplicate reads compared to lower expressed genes. In aorta there is no relationship between duplicates reads and gene expression levels, suggesting such reads potentially originated from PCR duplicates. When we identify allele biased genes, the sample containing high duplicate reads shows dramatic change of number of allele biased genes (aorta and pancreas) while the sample containing low duplicate reads shows subtle change of number of allele biased genes (bladder). The number of allelically biased genes were 1229, 479, and 165 before removing duplicate reads in aorta, pancreas, and bladder, respectively, which drop to 15, 117, and 160 upon duplicate reads removal.

The high duplicate reads can cause identification of more allelically biased genes compared to after removing duplicate reads. To avoid any statistical bias

during downstream analysis we decided to remove duplicate reads across all samples. We also exclude Esophagus (donor 3) sample because this sample contains an excessive number of PCR duplicate reads.

For Hi-C datasets, read pairs were aligned independently to the variant masked individual genome (hg18) using Stampy and merged using in house script by excluding non-uniquely mapped, non-mapped and PCR duplicate read pairs. Hi-C read pairs spanning variant loci were also split into the P1 or P2 alleles based on the variants information.

Removing alignment biased variants

Although we aligned sequencing reads to variants masked genome there are still local biases favoring either allele. To correctly identify allelically biased patterns we removed those biased alleles. We removed local biases through following three steps. Firstly, we removed alignment biases by aligning simulated reads spanning variants location. If there is more than 5% difference between alleles those variants were considered to underlie an inherent mapping bias. Secondly, we removed alleles located in copy number variable regions and allelic biased copy number variable regions by comparing the coverage between two alleles based on WGS data. Any variants located in the region with higher coverage than three times standard deviation above the mean of each haplotype were excluded. Any variants showing biased WGS coverage between two alleles were also excluded (binomial test p-value 0.05 after Benjamini correction). Lastly, we remove erroneously called as heterozygous variant during genotyping. We calculated the probability of each

heterozygous variants were actually homozygous from the likelihood of observing the coverage on each allele from whole genome sequencing. Only heterozygous SNPs that had a FDR of less than 0.5% were included in downstream analysis.

Annotation of genes, imprinted loci, enhancers and TF motif calling used for allelically delineated analyses

To identify allelically biased genes we used gene annotations defined by GENCODE database (hg18) by taking only level 1 and 2 genes. During allelic analyses across multiple individuals we re-defined active enhancers based on H3K27ac ChIP-seq signals. First we identified H3K27ac peaked regions with MACS with default parameters and removed any peaks less than 2.5kb from transcription start site defined in GENCODE (hg18) annotation. In total across 31 samples, we defined 240,238 peaks. After that, we defined 217,029 H3K27ac peaked regions as active enhancers for the allelic analyses after excluding peaks overlapping with known TSS. The rationale behind using this method for allelic analyses was that it was necessary to focus only on the strong enhancers, as weak elements (ones covered by few reads) are not testable for allelic biases. RFECS, which was used in the non-allelic analyses, utilized all 6 of the histone modifications for accurate predictions of both strong and weak *cis*-regulatory elements, which is important in prediction of novel elements. To allow for greater statistical power, we generated additional H3K27ac ChIP-seq datasets with longer reads and deeper coverage. As the ChIP-seq datasets with equivalent coverage were not generated for the other 5 marks, we instead opted to use a second prediction algorithm for the allelic analyses.

The predictions generated by the 2 methods show high degree of overlap when comparing the strong enhancers (p-value < $2.2e10^{-16}$)

For imprinted genes, we obtained 59 known imprinting genes downloaded from publicly available imprinted gene database (http://www.geneimprint.com/).

To identify enriched TF motifs in active enhancer regions of each sample, we performed HOMER motif search analysis with default parameters. The motifs that show a p-value of less than 10e-6 were used for downstream analysis.

Identification of allelically biased chromatin activity, enhancer activity, and gene expression

To identify allele-biased genes we performed binomial test for the number of aligned reads between two alleles. We only counted aligned reads spanning exonic regions. The genes containing at least more than 10 aligned reads were considered as informative genes. Allele biased genes were defined based on 5% FDR after Benjamini & Hochberg correction.

Allele biased enhancers were identified as similar to identifying allele biased genes. P values between two alleles were calculated based on binomial test after counting number of aligned reads at enhancer regions. We defined allele-biased enhancers in terms of 5% or 1% FDR during downstream analyses.

Allele biased chromatin activities at promoter regions were defined by performing binomial test for the number of aligned reads between two alleles. The promoter regions were defined as upstream and downstream of 1.5kb surrounding transcription start site. The allele biased chromatin activities at transcribed regions

were also calculated as similar as promoter regions. The transcribed regions were defined based on GENCODE annotation (hg18). Allele biased chromatin activity was defined if the binomial test p value is less than 0.05. Due to the very limited number of allelically informative reads in single-end ChIP-seq data we did not apply FDR correction.

Experimental validation of allelically biased enhancer activity

ChIP with thymus (donor 1) and pancreas (donor 2 and 3) tissues were conducted with the same protocol and antibodies as done for ChIP-seq. Rabbit IgG antibody was included as background for each set of chromatin. The immunoprecipitated DNA was analyzed by qPCR with primers targeting the two alleles of selected enhancers, where the differentiating SNP was the 3' most base of the oligo. Two negative control regions, with no allelic differences, were also selected. The primers used are included below:

Genomic location (hg18)	**Allele**	**5' primer**	**3' primer**
chr1:58630667-58631339	P1	GGGGAAACATGAGCTATATGC	TACCTTAGCCAAGAGCCAGT
	P2	GGGGAAACATGAGCTATATGT	TACCTTAGCCAAGAGCCAGT
chr15:22750615-22753669	P1	GGGTAGAAAAATCGCACCAAAT	GTCTTCCTATGTGCGGTACA
	P2	GGGTAGAAAAATCGCACCAAAT	GTCTTCCTATGTGCGGTACG
chr12:3706715-3708947	P1	AACTGACTCCCTCCCCAACC	CATGCAACAGCATCTGTCATC
	P2	AACTGACTCCCTCCCCAACA	CATGCAACAGCATCTGTCATC

chr9:72281386-72287497	P1	TGTGGGTCCCCACCTTCG	GCTGGGCTGCTCTGTGTAAAAC
	P2	TGTGGGTCCCCACCTTCT	GCTGGGCTGCTCTGTGTAAAAC

Analyses of allelically biased gene expression in terms of tissue-specificity and individual-specificity

As shown in Figure 2d, to test whether allelically biased gene expression is common between two or more individuals or individual-restricted manner, we considered duplicates or triplicates tissue-types. For each tissue-type, we only selected commonly informative genes across duplicates or triplicates and calculated how many genes are individual-restricted (AD(n=133), GA(n=98), LV(n=167), LG(n=233), PA(n=91), PO(n=276), RV(n=151), SG(n=108), SB(n=102), and SX(n=260)). If the genes are allelically expressed in only one individual the genes were defined as individual-restricted genes, otherwise as commonly biased genes. To avoid the random variance effect, when detecting individual-restricted allelically biased genes, we only considered those that are identified at least two samples. We also simulated sample-restricted allele biased gene expression by using randomly selected trios of different samples. Sample-restricted allele biased genes were defined if those that were allelically expressed in only one sample. We iterated 10,000 times to estimate sample-restricted allele biased gene expression.

The direction of allelically biased enhancer activity in the same genotype

In order to investigate whether the direction of enhancer activity biases was dependent on genotype, we first compared allele biased enhancer activities

identified by two tissue-types derived from the same individual and hence had the same genome. We considered all pairs of allele biased enhancers that were allelically biased in both tissue-types. Secondly, we compared allele biased enhancer activities that were identified in different donors but in the same tissue-type. Importantly, we focused on loci where the two donors had identical genotypes. We only considered commonly allele biased SNPs located within allele biased enhancers between donors. Only five tissue types are available for comparison. For each SNP within allele biased enhancers, we performed binomial test for allelically biased activity and we only considered allele biased SNPs if the p values is less than 0.05. Lastly, we compared allelic enhancers defined in this study to allele biased H3K27ac regions defined by McVicker and colleagues[1] (downloaded from the author's website (GW_H3K27ac.signif_merged.txt)). We defined reference allele biased activity at each SNP as A.EST/(A.EST+B.EST) where 2*A.EST, 2*B.EST, and A.EST+B.EST can be roughly interpreted as expected read depths from individuals who are homozygous for the reference allele, homozygous for the non-reference allele and heterozygous respectively. During analysis we only considered allele biased enhancers defined by 1% FDR.

The overlap between allelic enhancer and QTL regions

In order to validate the functionality of the allelically biased enhancers, we compared the allelic enhancers identified in our study to QTL regions from multiple studies. We compared allele biased enhancers to DHS-QTLs[28], H3K27ac-QTLs[1], and e-QTLs[29]. If the distance between allelic enhancer regions and QTL regions are less

than 5kb, we considered the regions as overlapping. For eQTL data sets, we only considered *cis*-eQTLs in EUR population. We also calculated the overlapping regions by using randomly selected testable enhancers as the same number of allele biased enhancers. We iterated this test for 10,000 times. For the random data sets, we defined random genomic regions with the same number and same length as allele biased enhancers.

Distance between allelic enhancer and allelic genes

In Figure 4a, we compared the distribution of shortest distance of allelic enhancer-gene pairs to non-allelic biased enhancers and allelic expressed genes pairs. Since the number of non-allelic biased enhancers is much higher than allelic biased enhancers, we randomly selected non-allelic biased enhancers as the same number of allelic biased enhancers. We used the set of allele biased enhancers defined by 5% FDR. We generated the shortest distance between randomly selected non-allelic enhancers and allelic expressed genes ten times.

We also calculated the shortest distance between concordant allele biased enhancer-gene pairs for whole chromosome-spanning haplotype blocks and simulated 300kb-haplotype blocks (Extended Data Fig. 10a). Simulated 300kb-haplotype blocks were generated by equally binned the genome with 300kb windowed regions.

Fraction of concordant allelic enhancer-gene pairs

In Figure 4c, all possible enhancers-gene pairs within the indicated distance window, encompassing both functional pairs and those simply residing in close proximity, are defined with either concordant or discordant allelic bias. The fractions of concordant pairs are shown. Regardless of FDR cutoff, allelic enhancer-gene pairs are significantly higher in concordance as compared to permutated controls. The random permutation data was generated by randomly assigned allele bias direction to enhancers and genes.

Correlation coefficient between tissue-restricted, allelically biased gene-enhancer pairs

In order to test whether tissue-restricted, allelically biased gene expression is associated to tissue-restricted allelic enhancers, we calculated the Pearson correlation coefficients between gene expression levels and enhancer activities at an allele resolution. We only considered allele biased gene-enhancer pairs less than 500kb.

TF motif disruption investigation in allele biased enhancers

In order to identify motif disrupted allelic enhancers, we performed STORM[30] motif search with ‑f ‑t 0.8 options for all enriched TF motifs identified by HOMER in each sample. If the motif sequence matches to only one allele, those allele biased enhancers defined by 1% FDR were considered as motif disrupted enhancer candidates. We assume that motif disrupted alleles tend to show less enhancer activities. For each TF motif, we calculated number of allelically biased enhancers

that are concordant and discordant with motif disruption, respectively. The significantly associated motif disrupted TF candidates were selected by performing binomial test between the numbers of concordant and discordant enhancers with motif disruption. We defined potential allelic enhancer associated motifs with 10% FDR after Benjamini & Hochberg correction.

When we compared motif disrupted scores and allele biased enhancer activities, the P1 allele cooperative motif disrupted scores were calculated by subtracting the P2 allele motif score from the P1 allele motif score based on motif position weight matrix (PWM). The positive score indicates P2 allele disrupted motif, otherwise P1 allele disrupted motif.

Linking motif disrupted allele biased enhancers to their target genes

In order to link motif disrupted allelic enhancers to potential target genes, enhancer-promoter physical interactions were predicted based on Hi-C interaction frequencies. Normalized Hi-C interaction frequencies of any enhancer-promoter pairs lying within 1M were converted to virtual 4C-seq scores as described in another companion paper (Dixon, Jung, and Selvaraj et al. submitted to Nature as a companion paper). Briefly, Hi-C interaction frequencies were calculated in terms of 5kb window and normalized using HiCNorm[31]. After that, interaction frequencies between promoter-enhancer pairs were calculated where promoter regions were fixed as +/- 7.5kb surrounding TSS and enhancer regions were defined by using different windowed regions as 5kb, 10kb, 20kb, 30kb, 40kb, 50kb, 75kb, and 100kb. For each windowed enhancer regions the interaction frequencies were defined as

(Interaction frequency / window size)*5kb. The summation of these interaction frequencies was defined as virtual 4C-seq score for enhancer-promoter pairs. The virtual 4C-seq scores were normalized again by considering the distance between enhancer and promoter. Based on the skewed normal distribution, we can calculate p values for the given virtual 4C-seq scores. We calculated virtual 4C-seq scores for thymus, aorta, and left ventricle. The enhancer-promoter long-range strong interactions are defined if the virtual 4C-seq score p value is less than 0.01 in each tissue. Based on those predicted enhancer-promoter strong interactions we can link motif disrupted allele biased enhancers to genes with allelically biased expression.

Due to the availability of Hi-C data in thymus, aorta, and left ventricle, we only considered those three tissues with corresponding donors. The short-range enhancer-gene pairs were defined as any enhancers localized less than 20kb from target genes. The motif disrupted enhancers were defined if there is more than 0.1 motif score difference between two alleles according to STORM motif hit score.

Supplementary Tables

Supplementary Table 1. All predicted enhancers across single sample of all 28 tissues/cell-types. Chromosomal locations and the prediction statuses of each enhancer element are provided.

Supplementary Table 2. All predicted promoters across single sample of all 28 tissues/cell-types.

Supplementary Table 3. List of *c*REDS elements defined across all tissues.

Supplementary Table 4. Number of Hi-C reads

Donor	Tissue	Total Hi-C reads	Hi-C reads cis	Hi-C reads (>20kb)
DONOR 1	Thymus	354,180,782	147,536,727	42,169,368
DONOR 2	Aorta	251,406,978	175,923,355	73,459,928
DONOR 3	Left ventricle	194,568,331	105,538,995	44,719,823
DONOR 4	Liver	366,637,018	276,717,743	98,882,631

Supplementary Table 5. Completeness and resolution of haplotypes

Donor	Number of SNPs	Resolution (%)	Completeness (%)
DONOR 1	2,512,926	78	99.82
DONOR 2	1,940,733	78	99.60
DONOR 3	1,952,614	85	99.74
DONOR 4	1,942,753	89	99.91

Supplementary Table 6. Haplotype resolved samples

Type	Number of samples
H3K4me1 single-end	29
H3K4me3 single-end	28
H3K9me3 single-end	23
H3K27ac single-end	31
H3K27me3 single-end	28
H3K36me3 single-end	28
H3K27ac paired-end	20 : AD_DONOR 3, AO_DONOR 2, AO_DONOR 3, EG_DONOR 2, EG_DONOR 3, GA_DONOR 1, GA_DONOR 2, GA_DONOR 3, LG_DONOR 2, LV_DONOR 1, LV_DONOR 3, OV_DONOR 2, PA_DONOR 2, PA_DONOR 3, PO_DONOR 3, RA_DONOR 3, RV_DONOR 3, SG_DONOR 3, SX_DONOR 3, TH_DONOR 1
mRNA-seq	36

Supplementary Table 7. The number of informative and allelic genes defined in each of the 18 tissues with 36 samples.

Sample	Allelically expressed genes	Informative genes	Fraction of allelic genes
AD_DONOR 2	307	2,356	0.13
AD_DONOR 3	174	2,162	0.08
AO_DONOR 2	58	1,386	0.04
AO_DONOR 3	22	448	0.05
BL_DONOR 1	161	3,004	0.05
EG_DONOR 2	305	2,872	0.11
FT_DONOR 1	225	3,022	0.07
FT_DONOR 2	179	2,588	0.07
FT_DONOR 3	45	1,137	0.04
GA_DONOR 1	176	2,663	0.07
GA_DONOR 2	22	450	0.05
GA_DONOR 3	96	1,840	0.05
LG_DONOR 1	473	3,946	0.12
LG_DONOR 2	296	2,805	0.11
LI_DONOR 4	137	2,219	0.06
LV_DONOR 1	375	3,297	0.11
LV_DONOR 3	196	2,330	0.08
OV_DONOR 2	298	2,988	0.10
PA_DONOR 2	127	2,234	0.06
PA_DONOR 3	196	2,490	0.08
PO_DONOR 1	305	2,922	0.10
PO_DONOR 2	104	1,566	0.07
PO_DONOR 3	314	2,670	0.12
RA_DONOR 3	297	2,963	0.10
RV_DONOR 1	396	3,409	0.12
RV_DONOR 3	176	2,068	0.09
SB_DONOR 1	405	4,094	0.10
SB_DONOR 2	32	669	0.05
SB_DONOR 3	49	756	0.06
SG_DONOR 1	410	4,053	0.10
SG_DONOR 2	25	465	0.05
SG_DONOR 3	64	1,087	0.06
SX_DONOR 1	230	2,860	0.08
SX_DONOR 2	160	1,872	0.09
SX_DONOR 3	270	2,913	0.09
TH_DONOR 1	394	3,507	0.11

Supplementary Table 8. The number of informative and allelic enhancers defined in each of the 14 tissues with 20 samples.

Sample	Informative enhancer	Allelic enhancer (FDR=1%)	Fraction	Allelic enhancer (FDR=5%)	Fraction
AD_DONOR 3	27999	186	0.007	825	0.030
GA_DONOR 2	28988	836	0.029	2,032	0.070
PA_DONOR 2	6960	187	0.027	419	0.060
SG_DONOR 3	40698	1,548	0.038	3,425	0.084
AO_DONOR 2	41855	700	0.017	2,228	0.053
GA_DONOR 3	50881	810	0.016	2,669	0.052
PA_DONOR 3	21906	539	0.025	1,257	0.057
SX_DONOR 3	45788	2,065	0.045	4,288	0.094
AO_DONOR 3	29879	1,035	0.035	2,306	0.077
LG_DONOR 2	28139	402	0.014	1,460	0.052
PO_DONOR 3	8266	244	0.030	536	0.065
TH_DONOR 1	31171	507	0.016	1,639	0.053
EG_DONOR 2	25392	856	0.034	1,846	0.073
LV_DONOR 1	26550	994	0.037	2,213	0.083
RA_DONOR 3	25675	901	0.035	2,142	0.083
EG_DONOR 3	44473	1,508	0.034	3,337	0.075
LV_DONOR 3	17776	617	0.035	1,322	0.074
GA_DONOR 1	12302	159	0.013	579	0.047
OV_DONOR 2	18468	283	0.015	1,032	0.056
RV_DONOR 3	29836	1,036	0.035	2,434	0.082

Supplementary Table 9. List of potential motif disruption TF candidates in allele-biased enhancer activity

Sample	Motif	Binomial test p value	FDR
AD_DONOR 3	ZNF711(Zf)/SH-SY5Y-ZNF711-ChIP-Seq/Homer	0.028	0.074
AD_DONOR 3	Gata4(Zf)/Heart-Gata4-ChIP-Seq(GSE35151)/Homer	0.044	0.039
PA_DONOR 2	TEAD4(TEA)/Tropoblast-Tead4-ChIP-Seq(GSE37350)/Homer	0.010	0.018
PA_DONOR 2	E2F4(E2F)/K562-E2F4-ChIP-Seq(GSE31477)/Homer	0.053	0.057
PA_DONOR 2	Gata4(Zf)/Heart-Gata4-ChIP-Seq(GSE35151)/Homer	0.033	0.052
PA_DONOR 2	AP-2alpha(AP2)/Hela-AP2alpha-ChIP-Seq/Homer	0.025	0.008
SG_DONOR 3	Jun-AP1(bZIP)/K562-cJun-ChIP-Seq/Homer	0.024	0.063
SG_DONOR 3	Fli1(ETS)/CD8-FLI-ChIP-Seq(GSE20898)/Homer	0.017	0.034
SG_DONOR 3	Elk4(ETS)/Hela-Elk4-ChIP-Seq(GSE31477)/Homer	0.003	0.006
SG_DONOR 3	EWS:ERG-fusion(ETS)/CADO_ES1-EWS:ERG-ChIP-Seq/Homer	0.002	0.003
SG_DONOR 3	Elk1(ETS)/Hela-Elk1-ChIP-Seq(GSE31477)/Homer	0.007	0.008
SG_DONOR 3	SPDEF(ETS)/VCaP-SPDEF-ChIP-Seq/Homer	0.000	0.001
SG_DONOR 3	Foxo1(Forkhead)/RAW-Foxo1-ChIP-Seq/Homer	0.001	0.004
SG_DONOR 3	FOXP1(Forkhead)/H9-FOXP1-ChIP-Seq(GSE31006)/Homer	0.025	0.045
SG_DONOR 3	Tlx?/NPC-H3K4me1-ChIP-Seq/Homer	0.021	0.025
SG_DONOR 3	Znf263(Zf)/K562-Znf263-ChIP-Seq/Homer	0.049	0.057
SG_DONOR 3	Smad3(MAD)/NPC-Smad3-ChIP-Seq(GSE36673)/Homer	0.051	0.083
SG_DONOR 3	Gata1(Zf)/K562-GATA1-ChIP-Seq/Homer	0.019	0.036
SG_DONOR 3	Maz(Zf)/HepG2-Maz-ChIP-Seq(GSE31477)/Homer	0.007	0.011
SG_DONOR 3	PU.1-IRF(ETS:IRF)/Bcell-PU.1-ChIP-Seq(GSE21512)/Homer	0.068	0.094
SG_DONOR 3	TEAD4(TEA)/Tropoblast-Tead4-ChIP-Seq(GSE37350)/Homer	0.029	0.054
SG_DONOR 3	PQM-1(?)/cElegans-L3-ChIP-Seq(modEncode)/Homer	0.024	0.025
SG_DONOR 3	Klf4(Zf)/mES-Klf4-ChIP-Seq/Homer	0.041	0.083
AO_DONOR 2	ETV1(ETS)/GIST48-ETV1-ChIP-Seq/Homer	0.051	0.059
AO_DONOR 2	ELF5(ETS)/T47D-ELF5-ChIP-Seq(GSE30407)/Homer	0.005	0.005
GA_DONOR 3	Fli1(ETS)/CD8-FLI-ChIP-Seq(GSE20898)/Homer	0.049	0.095
GA_DONOR 3	SUT1?/SacCer-Promoters/Homer	0.032	0.068
GA_DONOR 3	SPDEF(ETS)/VCaP-SPDEF-ChIP-Seq/Homer	0.036	0.079
GA_DONOR 3	Maz(Zf)/HepG2-Maz-ChIP-Seq(GSE31477)/Homer	0.009	0.014
GA_DONOR 3	Gata1(Zf)/K562-GATA1-ChIP-Seq/Homer	0.042	0.077
GA_DONOR 3	Nr5a2(NR)/mES-Nr5a2-ChIP-Seq/Homer	0.063	0.082
PA_DONOR 3	TEAD(TEA)/Fibroblast-PU.1-ChIP-Seq/Homer	0.031	0.053
PA_DONOR 3	NF1(CTF)/LNCAP-NF1-ChIP-Seq/Homer	0.095	0.098
PA_DONOR 3	Stat3+il23(Stat)/CD4-Stat3-ChIP-Seq/Homer	0.018	0.032
PA_DONOR 3	Maz(Zf)/HepG2-Maz-ChIP-Seq(GSE31477)/Homer	0.012	0.013
PA_DONOR 3	CRE(bZIP)/Promoter/Homer	0.072	0.051
PA_DONOR 3	STAT4(Stat)/CD4-Stat4-ChIP-Seq/Homer	0.062	0.078
PA_DONOR 3	MyoD(HLH)/Myotube-MyoD-ChIP-Seq/Homer	0.040	0.062

PA_DONOR 3	STAT1(Stat)/HelaS3-STAT1-ChIP-Seq/Homer	0.036	0.06
PA_DONOR 3	Unknown3/Arabidopsis-Promoters/Homer	0.036	0.053
SX_DONOR 3	ETV1(ETS)/GIST48-ETV1-ChIP-Seq/Homer	0.000	0
SX_DONOR 3	ERG(ETS)/VCaP-ERG-ChIP-Seq/Homer	0.000	0
SX_DONOR 3	Fli1(ETS)/CD8-FLI-ChIP-Seq(GSE20898)/Homer	0.000	0
SX_DONOR 3	GABPA(ETS)/Jurkat-GABPa-ChIP-Seq/Homer	0.000	0
SX_DONOR 3	Ets1-distal(ETS)/CD4+-PolII-ChIP-Seq/Homer	0.042	0.088
SX_DONOR 3	ELF1(ETS)/Jurkat-ELF1-ChIP-Seq/Homer	0.000	0
SX_DONOR 3	PU.1(ETS)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	0.000	0
SX_DONOR 3	SPDEF(ETS)/VCaP-SPDEF-ChIP-Seq/Homer	0.010	0.013
SX_DONOR 3	PU.1-IRF(ETS:IRF)/Bcell-PU.1-ChIP-Seq(GSE21512)/Homer	0.048	0.071
SX_DONOR 3	Esrrb(NR)/mES-Esrrb-ChIP-Seq/Homer	0.007	0.006
SX_DONOR 3	RUNX-AML(Runt)/CD4+-PolII-ChIP-Seq/Homer	0.010	0.025
SX_DONOR 3	Gata1(Zf)/K562-GATA1-ChIP-Seq/Homer	0.023	0.035
AO_DONOR 3	NF1-halfsite(CTF)/LNCaP-NF1-ChIP-Seq/Homer	0.033	0.058
AO_DONOR 3	AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	0.020	0.026
AO_DONOR 3	TEAD(TEA)/Fibroblast-PU.1-ChIP-Seq/Homer	0.000	0
AO_DONOR 3	PU.1(ETS)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	0.016	0.03
AO_DONOR 3	JunD(bZIP)/K562-JunD-ChIP-Seq/Homer	0.059	0.057
AO_DONOR 3	CRE(bZIP)/Promoter/Homer	0.010	0.028
LG_DONOR 2	MafA(bZIP)/Islet-MafA-ChIP-Seq(GSE30298)/Homer	0.095	0.081
LG_DONOR 2	RUNX2(Runt)/PCa-RUNX2-ChIP-Seq(GSE33889)/Homer	0.043	0.054
LG_DONOR 2	STAT6(Stat)/CD4-Stat6-ChIP-Seq/Homer	0.002	0.005
LG_DONOR 2	STAT6/Macrophage-Stat6-ChIP-Seq/Homer	0.059	0.041
LG_DONOR 2	Nur77(NR)/K562-NR4A1-ChIP-Seq(GSE31363)/Homer	0.046	0.032
PO_DONOR 3	ETV1(ETS)/GIST48-ETV1-ChIP-Seq/Homer	0.048	0.043
PO_DONOR 3	PHA-4(Forkhead)/cElegans-Embryos-PHA4-ChIP-Seq(modEncode)/Homer	0.035	0.05
PO_DONOR 3	Six1(Homeobox)/Myoblast-Six1-ChIP-Chip(GSE20150)/Homer	0.022	0.035
TH_DONOR 1	ETV1(ETS)/GIST48-ETV1-ChIP-Seq/Homer	0.004	0.004
TH_DONOR 1	EWS:ERG-fusion(ETS)/CADO_ES1-EWS:ERG-ChIP-Seq/Homer	0.026	0.049
TH_DONOR 1	RUNX1(Runt)/Jurkat-RUNX1-ChIP-Seq/Homer	0.041	0.094
TH_DONOR 1	Maz(Zf)/HepG2-Maz-ChIP-Seq(GSE31477)/Homer	0.034	0.057
TH_DONOR 1	Unknown5/Drosophila-Promoters/Homer	0.057	0.099
EG_DONOR 2	AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	0.055	0.099
EG_DONOR 2	Jun-AP1(bZIP)/K562-cJun-ChIP-Seq/Homer	0.008	0.014
EG_DONOR 2	EWS:ERG-fusion(ETS)/CADO_ES1-EWS:ERG-ChIP-Seq/Homer	0.009	0.014
EG_DONOR 2	PU.1(ETS)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	0.024	0.026
EG_DONOR 2	ZNF711(Zf)/SH-SY5Y-ZNF711-ChIP-Seq/Homer	0.005	0.011
EG_DONOR 2	Smad2(MAD)/ES-SMAD2-ChIP-Seq(GSE29422)/Homer	0.037	0.077
LV_DONOR 1	SPDEF(ETS)/VCaP-SPDEF-ChIP-Seq/Homer	0.031	0.083
LV_DONOR 1	Esrrb(NR)/mES-Esrrb-ChIP-Seq/Homer	0.022	0.025
LV_DONOR 1	TEAD4(TEA)/Tropoblast-Tead4-ChIP-Seq(GSE37350)/Homer	0.020	0.029

LV_DONOR 1	GRE/RAW264.7-GRE-ChIP-Seq/Homer	0.011	0.006
LV_DONOR 1	MafA(bZIP)/Islet-MafA-ChIP-Seq(GSE30298)/Homer	0.019	0.032
LV_DONOR 1	Maz(Zf)/HepG2-Maz-ChIP-Seq(GSE31477)/Homer	0.028	0.046
LV_DONOR 1	CRE(bZIP)/Promoter/Homer	0.054	0.057
RA_DONOR 3	ERG(ETS)/VCaP-ERG-ChIP-Seq/Homer	0.035	0.07
RA_DONOR 3	ETV1(ETS)/GIST48-ETV1-ChIP-Seq/Homer	0.028	0.06
RA_DONOR 3	AR-halfsite(NR)/LNCaP-AR-ChIP-Seq(GSE27824)/Homer	0.001	0.003
RA_DONOR 3	Elk4(ETS)/Hela-Elk4-ChIP-Seq(GSE31477)/Homer	0.046	0.066
RA_DONOR 3	MYB(HTH)/ERMYB-Myb-ChIPSeq(GSE22095)/Homer	0.030	0.08
RA_DONOR 3	PR(NR)/T47D-PR-ChIP-Seq(GSE31130)/Homer	0.018	0.017
RA_DONOR 3	PQM-1(?)/cElegans-L3-ChIP-Seq(modEncode)/Homer	0.035	0.038
RA_DONOR 3	PU.1(ETS)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	0.000	0
RA_DONOR 3	EWS:FLI1-fusion(ETS)/SK_N_MC-EWS:FLI1-ChIP-Seq/Homer	0.004	0.005
RA_DONOR 3	Smad3(MAD)/NPC-Smad3-ChIP-Seq(GSE36673)/Homer	0.040	0.084
RA_DONOR 3	JunD(bZIP)/K562-JunD-ChIP-Seq/Homer	0.008	0.019
RA_DONOR 3	Maz(Zf)/HepG2-Maz-ChIP-Seq(GSE31477)/Homer	0.015	0.028
EG_DONOR 3	AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	0.000	0
EG_DONOR 3	Jun-AP1(bZIP)/K562-cJun-ChIP-Seq/Homer	0.000	0.001
EG_DONOR 3	Bach2(bZIP)/OCILy7-Bach2-ChIP-Seq(GSE44420)/Homer	0.001	0.001
EG_DONOR 3	Fli1(ETS)/CD8-FLI-ChIP-Seq(GSE20898)/Homer	0.043	0.081
EG_DONOR 3	PU.1(ETS)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	0.028	0.04
EG_DONOR 3	Nrf2(bZIP)/Lymphoblast-Nrf2-ChIP-Seq(GSE37589)/Homer	0.073	0.031
EG_DONOR 3	EKLF(Zf)/Erythrocyte-Klf1-ChIP-Seq(GSE20478)/Homer	0.033	0.046
EG_DONOR 3	IRF2(IRF)/Erythroblas-IRF2-ChIP-Seq(GSE36985)/Homer	0.007	0.006
EG_DONOR 3	CArG(MADS)/PUER-Srf-ChIP-Seq/Homer	0.027	0.05
LV_DONOR 3	ERG(ETS)/VCaP-ERG-ChIP-Seq/Homer	0.004	0.001
LV_DONOR 3	ETV1(ETS)/GIST48-ETV1-ChIP-Seq/Homer	0.000	0
LV_DONOR 3	GABPA(ETS)/Jurkat-GABPa-ChIP-Seq/Homer	0.001	0
LV_DONOR 3	Unknown5/Drosophila-Promoters/Homer	0.013	0.016
LV_DONOR 3	Elk1(ETS)/Hela-Elk1-ChIP-Seq(GSE31477)/Homer	0.055	0.063
LV_DONOR 3	Gata2(Zf)/K562-GATA2-ChIP-Seq/Homer	0.058	0.072
LV_DONOR 3	Bach2(bZIP)/OCILy7-Bach2-ChIP-Seq(GSE44420)/Homer	0.040	0.068
LV_DONOR 3	EWS:FLI1-fusion(ETS)/SK_N_MC-EWS:FLI1-ChIP-Seq/Homer	0.009	0.007
LV_DONOR 3	PU.1(ETS)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	0.010	0.014
LV_DONOR 3	Smad3(MAD)/NPC-Smad3-ChIP-Seq(GSE36673)/Homer	0.021	0.034
LV_DONOR 3	PQM-1(?)/cElegans-L3-ChIP-Seq(modEncode)/Homer	0.024	0.025
LV_DONOR 3	ELF1(ETS)/Jurkat-ELF1-ChIP-Seq/Homer	0.001	0.002
LV_DONOR 3	ELF5(ETS)/T47D-ELF5-ChIP-Seq(GSE30407)/Homer	0.002	0
GA_DONOR 1	Rbpj1(?)/Panc1-Rbpj1-ChIP-Seq(GSE47459)/Homer	0.049	0.095
GA_DONOR 1	RUNX1(Runt)/Jurkat-RUNX1-ChIP-Seq/Homer	0.068	0.07
OV_DONOR 2	PHA-4(Forkhead)/cElegans-Embryos-PHA4-ChIP-Seq(modEncode)/Homer	0.009	0.009
OV_DONOR 2	TEAD4(TEA)/Tropoblast-Tead4-ChIP-Seq(GSE37350)/Homer	0.005	0.01

OV_DONOR 2	Stat3(Stat)/mES-Stat3-ChIP-Seq/Homer	0.011	0.003
OV_DONOR 2	E2F4(E2F)/K562-E2F4-ChIP-Seq(GSE31477)/Homer	0.064	0.071
OV_DONOR 2	Unknown1(NR/Ini-like)/Drosophila-Promoters/Homer	0.072	0.056
RV_DONOR 3	ETV1(ETS)/GIST48-ETV1-ChIP-Seq/Homer	0.008	0.015
RV_DONOR 3	EWS:ERG-fusion(ETS)/CADO_ES1-EWS:ERG-ChIP-Seq/Homer	0.002	0.004
RV_DONOR 3	GABPA(ETS)/Jurkat-GABPa-ChIP-Seq/Homer	0.003	0.001
RV_DONOR 3	EWS:FLI1-fusion(ETS)/SK_N_MC-EWS:FLI1-ChIP-Seq/Homer	0.016	0.038
RV_DONOR 3	SPDEF(ETS)/VCaP-SPDEF-ChIP-Seq/Homer	0.036	0.082
RV_DONOR 3	ELF1(ETS)/Jurkat-ELF1-ChIP-Seq/Homer	0.066	0.095
RV_DONOR 3	Foxo1(Forkhead)/RAW-Foxo1-ChIP-Seq/Homer	0.002	0
RV_DONOR 3	GRE/RAW264.7-GRE-ChIP-Seq/Homer	0.058	0.036

Supplemental References:

1 McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science* **342**, 747-749, doi:10.1126/science.1242429 (2013).

2 Hawkins, R. D. *et al.* Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6**, 479-491, doi:10.1016/j.stem.2010.03.018 (2010).

3 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).

4 Rajagopal, N. *et al.* RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS computational biology* **9**, e1002968, doi:10.1371/journal.pcbi.1002968 (2013).

5 Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134-1148, doi:10.1016/j.cell.2013.04.022 (2013).

6 Arthur, D. V., S. k-means++: The Advantages of Careful Seeding. *Technical Report. Stanford* (2006).

7 Davies, D. L. & Bouldin, D. W. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence* **1**, 224-227 (1979).

8 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).

9 Ashizawa, S., Brunicardi, F. C. & Wang, X. P. PDX-1 and the pancreas. *Pancreas* **28**, 109-120 (2004).

10 Daubas, P., Tajbakhsh, S., Hadchouel, J., Primig, M. & Buckingham, M. Myf5 is a novel early axonal marker in the mouse brain and is subjected to post-transcriptional regulation in neurons. *Development* **127**, 319-331 (2000).

11 Fayard, E., Auwerx, J. & Schoonjans, K. LRH-1: an orphan nuclear receptor involved in development, metabolism and steroidogenesis. *Trends in cell biology* **14**, 250-260, doi:10.1016/j.tcb.2004.03.008 (2004).
12 Flandez, M. *et al.* Nr5a2 heterozygosity sensitises to, and cooperates with, inflammation in KRas(G12V)-driven pancreatic tumourigenesis. *Gut* **63**, 647-655, doi:10.1136/gutjnl-2012-304381 (2014).
13 Hirai, H. *et al.* Involvement of Runx1 in the down-regulation of fetal liver kinase-1 expression during transition of endothelial cells to hematopoietic cells. *Blood* **106**, 1948-1955, doi:10.1182/blood-2004-12-4872 (2005).
14 Hwang, D. H. *et al.* Transplantation of human neural stem cells transduced with Olig2 transcription factor improves locomotor recovery and enhances myelination in the white matter of rat spinal cord following contusive injury. *BMC neuroscience* **10**, 117, doi:10.1186/1471-2202-10-117 (2009).
15 Inoue, Y., Inoue, J., Lambert, G., Yim, S. H. & Gonzalez, F. J. Disruption of hepatic C/EBPalpha results in impaired glucose tolerance and age-dependent hepatosteatosis. *The Journal of biological chemistry* **279**, 44740-44748, doi:10.1074/jbc.M405177200 (2004).
16 Jahan, I., Kersigo, J., Pan, N. & Fritzsch, B. Neurod1 regulates survival and formation of connections in mouse ear and brain. *Cell and tissue research* **341**, 95-110, doi:10.1007/s00441-010-0984-6 (2010).
17 Lee, C. S. *et al.* Loss of nuclear factor E2-related factor 1 in the brain leads to dysregulation of proteasome gene expression and neurodegeneration. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 8408-8413, doi:10.1073/pnas.1019209108 (2011).
18 Massari, M. E. & Murre, C. Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. *Molecular and cellular biology* **20**, 429-440 (2000).
19 Moya, M. *et al.* Foxa1 reduces lipid accumulation in human hepatocytes and is down-regulated in nonalcoholic fatty liver. *PloS one* **7**, e30014, doi:10.1371/journal.pone.0030014 (2012).
20 Nagy, P., Bisgaard, H. C. & Thorgeirsson, S. S. Expression of hepatic transcription factors during liver development and oval cell differentiation. *The Journal of cell biology* **126**, 223-233 (1994).
21 Bar-Joseph, Z., Gifford, D. K. & Jaakkola, T. S. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics* **17 Suppl 1**, S22-29 (2001).
22 Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**, 116-120, doi:10.1038/nature11243 (2012).
23 Lu, X. *et al.* The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nature structural & molecular biology* **21**, 423-425, doi:10.1038/nsmb.2799 (2014).
24 Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics* **39**, 311-318, doi:10.1038/ng1966 (2007).
25 Selvaraj, S., J, R. D., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat Biotechnol*, doi:10.1038/nbt.2728 (2013).

26 Bansal, V. & Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* **24**, i153-159, doi:10.1093/bioinformatics/btn298 (2008).
27 Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome research* **21**, 936-939, doi:10.1101/gr.111120.110 (2011).
28 Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390-394, doi:10.1038/nature10808 (2012).
29 Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-511, doi:10.1038/nature12531 (2013).
30 Schones, D. E., Smith, A. D. & Zhang, M. Q. Statistical significance of cis-regulatory modules. *BMC bioinformatics* **8**, 19, doi:10.1186/1471-2105-8-19 (2007).
31 Hu, M. *et al.* HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* **28**, 3131-3133, doi:10.1093/bioinformatics/bts570 (2012).