Supplementary Materials for

Evolutionary Changes in Promoter and Enhancer Activity During Human Corticogenesis

Steven K. Reilly[1,5] Jun Yin[1,5], Albert E. Ayoub[2,3], Deena Emera[1], Jing Leng[1,4,+], Justin Cotney[1], Richard Sarro[1], Pasko Rakic[2,3], James P. Noonan[1,2,4*]

correspondence to: james.noonan@yale.edu

This PDF file includes:

Materials and Methods
Legends for Tables S1-S5
References (*27-44*)
Figs. S1 to S12

Materials and Methods

Tissue collection & ChIP-Seq

Human cortical tissues were collected, staged and flash frozen at -80°C by the Joint MRC/Wellcome Trust Human Developmental Biology Resource (www.hdbr.org). Human tissues were staged using the Carnegie system. The use of human embryonic tissue in this study was reviewed and approved by the Yale Human Investigation Committee. For embryonic human stages 7 p.c.w. Carnegie Stage 16 (CS16) & 8.5 p.c.w. (CS23), and the comparable stages in rhesus and mouse (Table S1), the entirety of the cortex was microdissected, homogenized, cross-linked, and stored at -80°C. Rhesus and mouse cortical tissues were harvested according to approved Yale IACUC protocols. In human 12 p.c.w. (Fetal Stage 2) cortex and the corresponding 12 p.c.w. rhesus and E17.5 mouse cortical tissues, the primitive frontal and occipital lobes were microdissected and treated separately.

Chromatin preparation, including extraction, sonication and immunoprecipitation, were all carried out as previously described (*8,9*). Briefly, between 10 and 25 µg of total chromatin were incubated with either 2 µg of H3K27ac antibody (Abcam ab4729) or 10 µg of H3K4me2 (Millipore 07-030) bound to protein G Dynabeads for 15-20 hours at 4°C with rotation. The Dynabeads were washed and the chromatin eluted, purified, and quantified as previously described (*8*). Purified chromatin was prepared for sequencing using the standard Illumina multiplexing protocol. 75 base pair single end sequencing was performed on Illumina HiSeq 2500 instruments at the Yale Center for Genome Analysis (YCGA).

Identification of H3K27ac and H3K4me2 enriched regions

ChIP-seq reads were aligned to their respective genome references (human: hg19, rhesus: rheMac2, mouse: mm9) using Bowtie (v0.12.7) *(27)*. Summary information for alignments is in Table S1. Only uniquely mapping reads were retained for further analysis. Duplicate reads (i.e., reads with the same alignment start site) were further filtered out, with only one copy of the read retained for peak detection. ChIP-seq signals for H3K27ac and H3K4me2 were generated by extending each aligned read to 300 bp (based on sonication size). The ChIP-seq signal value at each nucleotide in the genome is equal to the number of extended fragments that overlap it. Fragment counts were then normalized per 1 million aligned reads.

Peaks were identified in each species using a previously described sliding window approach tuned specifically for histone modifications based on a Poisson model (*9*). 500 bp windows with a 25 bp step size were tiled across the genome. In each window the number of ChIP and input control reads were counted and then the input scaled to match the sequencing depth of the ChIP experiment. Significance was assessed using a Poisson model where the null model has more input than ChIP reads, assuming a uniform distribution of total mapped reads in a given chromosome. Significantly enriched windows (P value $\leq 1x10^{-5}$) within 1 kb of each another were merged. Merged regions were assigned to promoters if they overlapped a 1 kb upstream segment of any transcription start site in Gencode (v10). The remaining regions that did not overlap an annotated exon in Gencode (v10) were labeled as putative enhancers.

Reproducibly enriched regions were defined as those that had 1 bp minimum overlap between two ChIP-seq replicates from the same time point and species. The coordinates for these

reproducible regions were defined by merging the coordinates for both replicates. The ChIP-seq signal tracks and peak calls are available for visualization in the UCSC Genome Browser:

H3K27ac: http://130.132.212.207/~sreilly/Track_Hub_me2/hub.txt
H3K27me2: http://130.132.212.207/~sreilly/Track_Hub_ac/hub.txt

K-means clustering

H3K27ac data from human cortex was compared with publicly available H3K27ac datasets as follows. A single composite multi-sample enhancer annotation from developing cortex, limb, embryonic stem cells, and select adult tissues profiled by the Roadmap Epigenomics Project (http://www.roadmapepigenomics.org) (*28*) was generated by merging replicate peaks across all samples using a 1 bp overlap. The level of H3K27ac signal in each region for each sample was quantified by averaging read counts per kilobase per million mapped reads (RPKM) in each region from each replicate. Each region was represented by a vector of a length equal to the total number of tissues considered, with each point representing the RPKM value of marking in that region for a single tissue. Each vector was normalized by subtracting the mean of all tissue quantifications from each individual tissue quantification, divided by the standard deviation of values for that vector. The matrix of these normalized tissue quantification values was then subjected to k-means clustering using R to identify sets of sites exhibiting the strongest marking in each tissue compared to all other samples in the comparison. We used GREAT (*18*) to identify biological functions and processes showing significant enrichment for each set of enhancers.

Principal component analysis and Spearman correlation analysis

To compare human, rhesus and mouse H3K27ac at three way orthologous sites by PCA, we constructed a single composite annotation for all human H3K27ac regions in multiple tissues. This was generated by using BEDtools to merge, using a 1 bp overlap, all human cortex and limb H3K27ac regions(*9*, *29*). Corresponding orthologous H3K27ac annotations in rhesus and mouse were generated from the human annotation via liftOver. We required three-way orthology (described in detail below) for regions to be included in the annotation. The level of H3K27ac marking was quantified as described above, and the resulting RPKM values for all three species were assembled into a quantification matrix. Each row in this matrix represented a given H3K27ac region, and each column represented the normalized level of H3K27ac signal at that region in a particular species and tissue. Principal component analysis was performed on this matrix using R. Spearman correlation analysis on the multi-species matrix was also carried out using R.

To compare human and mouse H3K27ac signal in multiple homologous tissues, reproducible H3K27ac enhancer regions identified in developing cortex, developing limb (*9*) and six other tissues profiled by the Roadmap Epigenomics Project (embryonic stem cells, spleen, intestine, heart, liver, kidney) were merged using BEDtools as described above (*28*). An orthologous mouse enhancer annotation was generated using liftOver. Although only mouse and human samples were used, to be consistent with other analyses in this study, we required three-way orthology (described in detail below) for regions to be included in the annotation. H3K27ac data from mouse cortex was generated in this study, while data from other mouse tissues was obtained from previous studies (*9*, *30*). H3K27ac signal was quantified and PCA carried out as described above.

Identification of epigenetic gains on the human lineage

Human-lineage epigenetic gains were identified using a previously described approach (*9*). To identify three-way orthologous sites in human, rhesus and mouse, H3K27ac and H3K4me2 regions were mapped between genome assemblies (hg19, rheMac2, mm9) using the liftOver tool and chain files from the UCSC Genome Browser. A region was considered as two-way orthologous if all following criteria were satisfied: mapping from the query to the target assembly was unique, the reciprocal mapping from the target assembly back to the query was unique, and the coordinates derived after reciprocal mapping had at least 50% overlap with the original peak coordinates. Alignments and reciprocal mapping to random and unassembled chromosomes were excluded. Three-way orthologous regions were identified by intersecting two-way orthologous regions identified in human-rhesus and human-mouse comparisons. Summary information for reciprocal orthology of human peaks is provided in Table S1. To ensure proper mapping of short reads and allow quantitative comparisons of epigenetic marking among all three species, we only estimated potential human gains of H3K27ac or H3K4me2 at three-way orthologous sites. Regions of ambiguous orthology and lineage-specific sequences were excluded from our analysis.

H3K27ac and H3K4me2 signals were compared between species to identify human gain regions as previously described (*9*). The epigenetic signal at each human site in a single time point and tissue was compared in a pairwise manner to the corresponding signal obtained at the orthologous site in rhesus or mouse at all tissues and time points examined in those species. In each comparison, human read counts were unscaled, while mouse or rhesus read counts were scaled by a ratio of (human region length/orthologous mouse or rhesus region length). We did not attempt to identify epigenetic gains at a specific developmental stage in human compared to one specific stage in rhesus or mouse, due to inherent uncertainties in precisely identifying equivalent developmental stages at a fine scale across species.

To identify human gains at a single time point compared to time points in mouse and rhesus with replicates (Table S1), we used a previously described log-linear model with a Poisson link and likelihood ratio test based on model fitting (*9*, *31*). The number of reads in a region p in sample i is denoted as X_{pi}, and is modeled as: $\log\left(E\left[X_{pi}\middle|X_{.i}\right]\right) = \log X_{.i} + \lambda_{pj(i)} + \theta_{pi}$, where $X_{.i}$ is the total number of mapped reads for sample i, $\lambda_{pj(i)}$ is the species-specific signal level, and θ_{pi} is the individual replicate effect. A likelihood ratio test is used to assess the significance of a difference in signal at each human site compared to rhesus or mouse. Resulting P values were corrected for multiple testing using the Benjamini-Hochberg (BH) method (*32*). For pairwise comparisons to the single rhesus time point without replication (8 p.c.w.) (Table S1), we employed a two-sided Fisher's exact test. Human read counts were used directly, while rhesus read counts were scaled as described above. A two-way contingency table was constructed for each region. P values were subsequently corrected using the BH method (*32*).

Human gains at each time point were identified by jointly ranking all sites based on fold change and P value, and then thresholding on both values. To be called as a human gain, a site had to show a strong increase in human signal (which we define as BH P value ≤ 0.001 and fold increase ≥ 1.5) compared to all time points and tissues in rhesus and mouse (seven comparisons in all). Read counts, fold change values and P values for all human sites are available for download here: http://130.132.212.207/~sreilly/pvalues.xlsx.

Our objective in choosing this approach was to identify a limited set of high-confidence sites exhibiting gains of epigenetic marking on the human lineage for downstream analysis. The high validation rate we obtained using ChIP-qPCR supports the robustness of our filtering scheme and the thresholds we chose for identifying human gains. We note that we may be overestimating gains at 7 p.c.w., due to the lack of a comparable developmental stage in rhesus. However, this concern is mitigated by our inclusion of a comparable mouse time point and our requirement that each site exhibit an epigenetic gain compared to all mouse and rhesus time points and tissues.

We did not consider losses of epigenetic marking in human, as many losses on the human lineage cannot be detected in the phylogeny we examined (*9*, *33*). Our test is not well powered to detect losses, as it only allows us to identify human losses based on reduced H3K27ac compared to both rhesus and mouse. These sites are more likely to include *cis*-regulatory functions conserved in both species and thus under deep evolutionary constraint in mammals. In contrast, human gain of H3K27ac may in principle occur at any site whether or not that site exhibits H3K27ac marking in any other species or has a conserved regulatory function. Robust identification of losses would require a phylogeny inclusive of additional non-human primates, including apes.

Validation of human lineage epigenetic gains by ChIP-qPCR

For validation, we chose sites exhibiting human lineage H3K27ac or H3K4me2 at each of the four developmental stages we analyzed. We selected sites showing a range of positive fold changes in human. Due to the limited availability of rhesus tissues, we compared epigenetic marking in human and mouse. Only orthologous sites with <10% change in length between human and mouse were considered to ensure mouse and human qPCR amplicons would be of comparable sizes. Human primers were designed using Primer3, and amplicons were targeted close to the H3K27ac or H3K4me2 signal peak (*34*). Mouse primers were designed based on the orthologous mouse sequence corresponding to the human amplicon. All primers are listed in Table S1. H3K27ac and H3K4me2 ChIP-qPCR analysis was performed on independent biological replicates in human and mouse from those used in the genome-wide studies. Fold enrichment in a single species was determined by comparing ChIP C_t values versus input C_t values. The magnitude of human gain at each site was determined by dividing human fold enrichment by mouse fold enrichment, analogous to $\Delta\Delta C_t$ values typically utilized in detecting gene expression differences by RT-qPCR.

Substitution Analysis

We conducted two pairwise comparisons of substitution events in H3K27ac gains: human-specific versus chimpanzee-specific substitutions, and ape-specific versus rhesus-specific substitutions. Since human gains are ascertained relative to rhesus and mouse, both human-specific and ape-specific changes may contribute to gain events. Human-specific substitutions were defined as human nucleotides that differ at an otherwise conserved nucleotide in chimpanzee, orangutan, rhesus, and marmoset. Chimpanzee-specific substitutions were defined as chimpanzee nucleotides that differ from a conserved human, orangutan, rhesus, and marmoset nucleotide. Ape-specific substitutions were defined as nucleotides in an ape consensus sequence we constructed that differ from conserved nucleotides in rhesus, marmoset, and mouse. Rhesus-

specific substitutions were defined as rhesus positions that differ from a conserved human, chimp, orangutan, and marmoset nucleotide.

In each H3K27ac region we calculated a human, chimpanzee, ape or rhesus-specific substitution rate defined as (# species-specific substitutions)/(# species-specific substitutions + # conserved nucleotides), where "conserved nucleotide" is defined as sites that are invariant across all 5 primate species in the analysis. We observed increased substitution rates for all four categories of substitution in human gains compared to all three-way orthologous promoters or enhancers. Therefore, we compared the ratio of (human-specific substitution rate)/(chimpanzee-specific substitution rate) or (ape-specific substitution rate)/(rhesus-specific substitution rate) for human gains and three-way orthologous sites. Using a Wilcoxon rank-sum test, we observed no consistently significant differences between the distribution of substitution rates on the human and chimpanzee, or ape and rhesus, lineages. P-Values and rates are listed in Table S1.

Comparison to known enhancer elements and accelerated regions

Human sequences with known enhancer activity (as of February 18, 2014) were obtained from the VISTA Enhancer Browser (enhancer.lbl.gov) (*19*). Reproducible human H3K27ac or H3K4me2 peaks from all time points were intersected with forebrain enhancers as well as all positive enhancers that did not exhibit activity in the forebrain. Overrepresentation of forebrain enhancers compared to other positive enhancers was assessed using Fisher's exact test. Mouse H3K27ac and H3K4me2 peaks were compared for peak overlap to mouse E11.5 p300 sites obtained from reference 20. Human accelerated regions were obtained from references (*5*, *6*) 14 and 15. Human accelerated regions exhibiting a human lineage gain at any time point are listed in Table S1.

LacZ mouse transgenic enhancer assay

The enhancer gain tested was chosen for two reasons: 1) It shows robust gains of both H3K27ac and H3K4me2 in humans; and 2) it encompasses a human sequence that had been previously shown to have strong and reproducible forebrain activity in the VISTA Enhancer Browser, simplifying interpretation of any differences we might see compared to rhesus. Orthologous 5.1kb sequences, chosen based on the H3K27ac signal encompassing the VISTA element hs754 (chr5:3,197,865-3,198,942 in hg19), were amplified from human and rhesus macaque genomic DNA using PCR. These sequences were cloned into a previously described *Hsp68-lacZ* reporter vector. Generation of transgenic mice and embryo staining was performed as described (*35*). One injection series was performed for the human construct and two independent injection series were performed for the rhesus construct. At least three embryos were required to exhibit reporter gene expression in a given domain for the expression pattern to be considered reproducible. Embryo images and annotation are summarized in Figure S7. Sequence coordinates and primers are listed in Table S1. Coronal sections of one human embryo revealed that the dorsal domain was restricted to the neocortex. The ventral domain consisted of the caudal ganglionic eminence, which produces GABAergic interneurons destined for superficial layers of the neocortex *(36,37)*.

Identification of topological domains enriched for epigenetic gains

We performed permutation tests to identify topological domains significantly enriched in human gains. In each permutation, human enhancer or promoter gains were randomly reassigned

within the superset of three-way orthologous enhancers or promoters. Permutations were carried out 20,000 times. The topological domains were obtained from previous studies, and include three-dimensional interactions detected by Hi-C in human ESCs and IMR90 cells (*21*), RNA PolII ChIA-PET in mouse NPC *(38*), and CTCF ChIA-PET in human cell lines (*39*). The significance of enrichment of human gain promoters or enhancers in a topological domain was calculated as the fraction of permuted datasets exhibiting an equal or greater number of gains as actually observed in the domain. The resulting permutation P values were then corrected for multiple testing using the BH method (*32*).

WGCNA network construction and visualization

RNA-seq data from multiple cortical regions spanning 12 consecutive periods of human brain development were obtained from the BrainSpan resource (www.brainspan.org). Gene expression levels were summarized by BrainSpan as RPKM values using the Gencode (v10) annotation. We divided these data into overlapping temporal windows, each spanning three consecutive periods of development as defined by BrainSpan, starting with period 2 and ending at period 13 (e.g. periods 2-4, 3-5, 4-6…) This provided temporal resolution for the co-expression analysis, and retained a sufficient number of data points for network construction as described below. Genes with an RPKM standard deviation of 0 in a given temporal window were excluded from network construction.

Weighted gene co-expression network analysis (WGCNA) was performed using the Bioconductor package *WGCNA* (*22*). Networks were independently constructed for each temporal window. A soft threshold power (the power correlations are raised to, to emphasize high correlations at the expense of low correlations) was selected to be the smallest threshold resulting in a scale free topology with R^2 fit >0.8. This power was applied to the correlation scores in order to calculate the model of correlations, known as the topological overlap matrix (TOM). Because of the large number of genes included in the network, the correlation matrix was split into three blocks in order to make network reconstruction computationally feasible (*22*). The function *blockwiseModules* was used to automate network construction and module detection in a "block-wise" manner. The *blockwiseModules* function was used with default parameters, except as follows: *maxBlockSize* was set to 20,000, *minModuleSize* was set to 30, *reassignThreshold* was set to 0, and *mergeCutHeight* was set to 0.25. Pairwise Pearson correlations of all genes in each block were calculated. Modules were dynamically cut from the TOM dendrogram, and merged into final modules by combining modules with similar expression profiles. The full WGCNA network construction code we used is available upon request. A table of all modules is available for download here:

http://130.132.212.207/~sreilly/WGCNA_mods.xlsx.

In order to evaluate the specificity of the network built at 8-15 p.c.w. (periods 2-4 in BrainSpan), we evaluated the preservation of each module in this network across all temporal windows of human cortical development using the Z-summary score calculated by the *WGCNA modulePreservation* function. This score summarizes the module density (the mean adjacency among all the nodes) and connectivity (the sum of connection strength based on correlations of each node to all the other nodes). A Z-summary score >10 supports that the module is preserved. Z-summary scores decrease for modules that are poorly preserved across time points (Figure S12A).

Hub genes were defined by evaluating the intra-module connectivity of all the genes in a module. Intra-module connectivity was calculated using the *intramodularConnectivity* function in the WGCNA package. Genes with intra module connectivity within the top 5% of a module were defined as hub genes. The 8-15 p.c.w. network had the highest number of hub genes associated with human lineage epigenetic gains (Fig. S12B).

Modules were visualized for figures using multidimensional scaling (MDS). MDS was carried out using the Pearson correlation matrix for each module. The first and second dimensions of MDS were used to plot the gene network in Cytoscape (*40*). Only genes with an absolute correlation coefficient > .7 with any other gene in the module were plotted on the figure. We used a correlation coefficient of .95 for Module 3 (Fig. 2B) due to its large size.

The expression correlation of all the modules in the network was then determined by correlating the module eigengene expression. The eigengene expression, corresponding to the first principal component of the expression matrix of the module, was calculated using the *moduleEigengenes* function. MDS was applied on the correlation matrix of module eigengenes. The first and second dimensions from MDS were used to plot the network of modules shown in Fig. 4.

Identification of gain-enriched modules

We used permutation analysis to identify modules significantly enriched in genes associated with human gains. Independently, for each developmental stage and histone modification, human gain assignments at promoters or enhancers were randomly shuffled among three way orthologous sites marked at that stage. We performed 20,000 iterations. Human gain promoters were assigned to their parent gene in Gencode (v10). Human gain enhancers were assigned to genes using regulatory domains defined by GREAT (*18*).

The number of human gain promoters or enhancers assigned to each module was then counted. The significance of human gain promoter or enhancer enrichment for each module was assessed based on the fraction of permuted datasets with an equal or greater number of gains as the observed number of gains assigned to the module. Resulting permutation test P values were corrected for multiple testing using the BH method (*32*).

Transcription factor motif enrichment analysis

We collected 591 transcription factor binding motifs from JASPAR (vertebrate core motifs) (*41*), and Uniprobe *(42*, *43*). All three-way orthologous H3K27ac and H3K4me2 regions in human (including gains) were scanned for the presence of each motif using FIMO with default settings (*44*) To identify enriched motifs in each module, we conducted permutation tests using custom Perl scripts. The permutation was conducted for each of the motifs independently 100,000 times. The total number of instances for each motif was counted in all promoters or enhancers assigned to each module. In each permutation, this number was compared to the number of motifs called in an equally sized set of enhancers or promoters derived from all three-way orthologous regions. The test was conducted using only regions from the same time point (e.g. 8.5 p.c.w. enhancers assigned to Module 6, vs all 8.5 p.c.w. three-way orthologous enhancers assigned to any module in the network). The raw permutation P value assessing the

enrichment of each TF motif in each module was defined as the fraction of permuted datasets exhibiting an equal or greater number of motif instances as observed in the actual set of promoters or enhancers in the module. P values were corrected using the BH method (*32*). All P values are in Table S5.

Identifying putative ancestral promoter and enhancer functions at human lineage gains

Raw ChIP-seq reads were obtained for 18 mouse tissues/cell-lines (bone marrow, cerebellum, cortex, embryonic brain, adult heart, embryonic heart, adult liver, embryonic liver, intestine, kidney, lung, mMEF, mESC, olfactory bulb, placenta, spleen, testis, and thymus)(*30*). H3K27ac enriched regions were identified as described above. H3K27ac data obtained in this study from rhesus and mouse cortex and previously ascertained H3K27ac data from rhesus and mouse embryonic limb (*9*) were also included. Human lineage gains were assigned three potential, mutually exclusive ancestral states: modification, co-option, or *de novo* events. Modification events were defined as peaks exhibiting a 1bp overlap with a mouse or rhesus cortical H3K27ac peak (identified at any developmental stage, and in any replicate) at the orthologous site in either species. Co-option events were defined as human peaks overlapping with any H3K27ac peak in any other tissue (i.e., non-cortex) at the orthologous site in mouse or rhesus. *De novo* events were defined as human peaks for which no peak was called in any tissue at the orthologous location in rhesus and mouse.

Author Contributions:

S.K.R., P.R. and J.P.N. conceived and designed the study. S.K.R. and R.S. conducted ChIP-seq experiments. S.K.R. and A.E.A. collected, staged, and dissected cortical tissues. S.K.R., J.L. and J.C. designed and implemented the human gain analysis pipeline. S.K.R. generated constructs for the transgenic reporter assay with embryo staining and scoring help from D.E. and J.P.N. with sectioning done by A.E.A. S.K.R. and J.Y. conducted co-expression network analyses and permutation tests for module gain enrichments.. S.K.R. performed TF enrichment analyses. S.K.R., J.Y., A.E.A., P.R. and J.P.N. wrote the paper with input from all authors.

Table S1.
Table of samples used in study, sequencing alignment statistics, Liftover statistics, overlap with human accelerated regions, overlap with VISTA elements, enhancer cloning primers, substitution analysis, and PCR primers.

Table S2.
Gain enrichment in topological domains analysis.

Table S3.
Module assignments of genes, with connectivity and hub gene information.

Table S4.
Enrichment summary and gain permutation results.

T**able S5.**
DAVID results.

Supplemental references

27. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg,*Genome Biol.* **10**, 25 (2009).
28. M. Garber *et al.*, *Mol Cell* **47**, 810–822 (2012).
29. A. R. Quinlan, I. M. Hall, *Bioinformatics* **26**, 841–842 (2010).
30. Y. Shen *et al.*, *Nature* **488**, 116–120 (2012).
31. J. H. Bullard, E. Purdom, K. D. Hansen, S. Dudoit, *BMC Bioinformatics* **11**, 94 (2010).
32. Y. Benjamini, Y. Hochberg, *J. R. Statis. Soc.B* **57,** 289-300 (1995).
33. Y. Shibata *et al.*, J. M. Akey, Ed. *PLoS Genet* **8**, 1002789 (2012).
34. A. Untergasser *et al.*, *Nucleic Acids Research* **35**, 71–74 (2007).
35. R. Kothary *et al.*, *Development* **105**, 707–714 (1989).
36. K. Tang, J. L. R. Rubenstein, S. Y. Tsai, M.J. Tsai, *Development* **139**, 1630-1639 (2012).
37. G. Miyoshi *et al.*, *Journal of Neuroscience* **30**, 1582-1594 (2010).
38. Y. Zhang *et al.*, *Nature* **504**, 306–310 (2013).
39. G. Li *et al.*, *Cell* **148**, 84–98 (2012).
40. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, *Genome Res*. **13,** 2498-2504 (2003).
41. E. Portales-Casamar *et al.*, *Nucleic Acids Res.* **38**, 105–110 (2010).
42. A. Mathelier *et al.*, *Nucleic Acids Res.* **42**, 1-6 (2013).
43. K. Robasky, M. L. Bulyk, *Nucleic Acids Res.* **39**, 124–128 (2011).
44. C. E. Grant, T. L. Bailey, W. S. Noble, *Bioinformatics* **27,** 1017-1018 (2011).

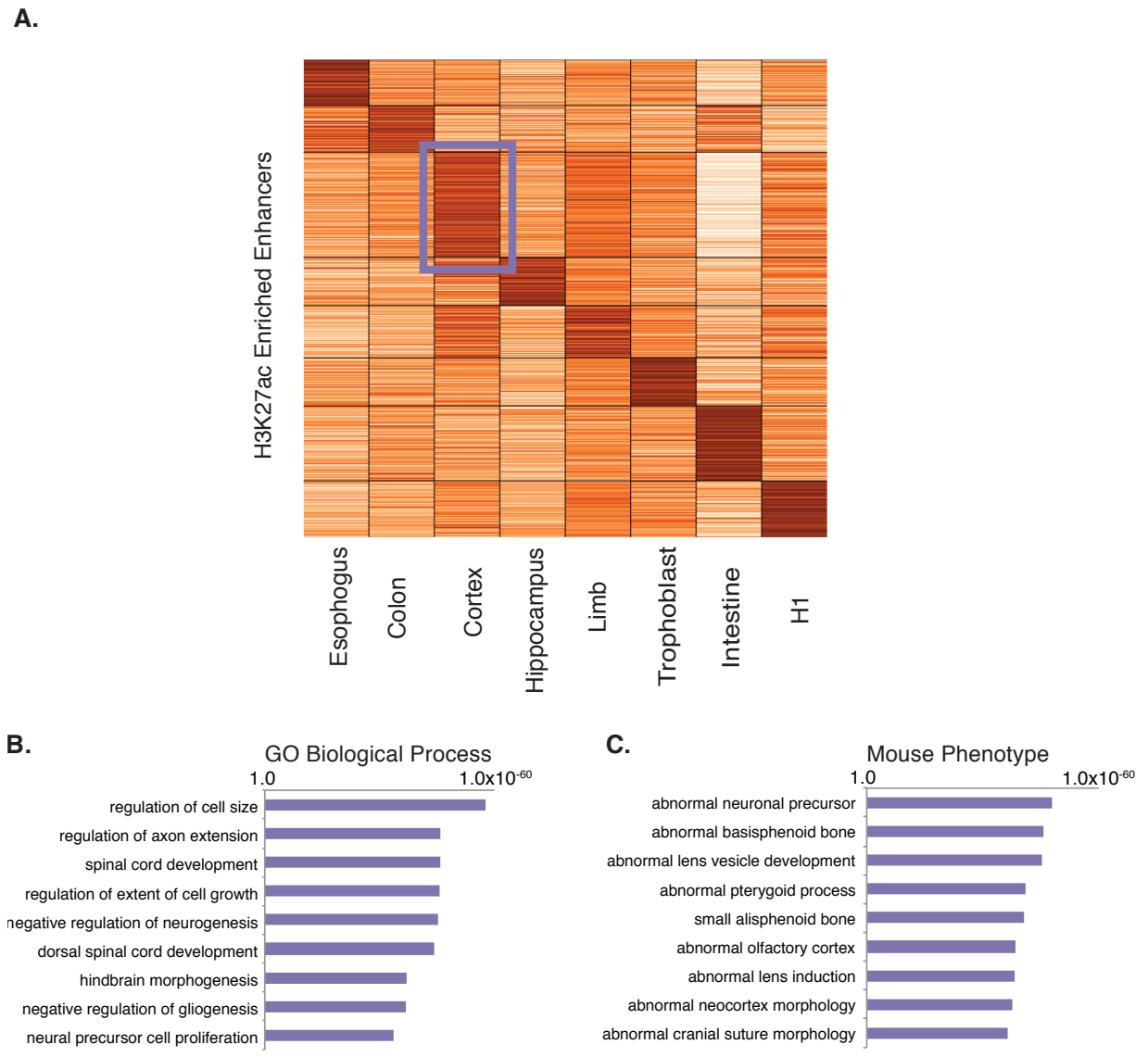


Fig. S1. Predicted human cortex enhancers show strong tissue specificity and are associated with genes implicated in cortical development. A) K-means clustering of H3K27ac signals across 137,408 putative enhancers from human tissues profiled in this study and by the Roadmap Epigenomics Project. A cluster of 16,473 enhancers strongly marked in the cortex compared to other tissues is indicated by a purple box. B-C) Gene Ontology enrichments identified by GREAT for the strong cortex enhancers in A). P values were calculated using the binomial test implemented in DAVID.

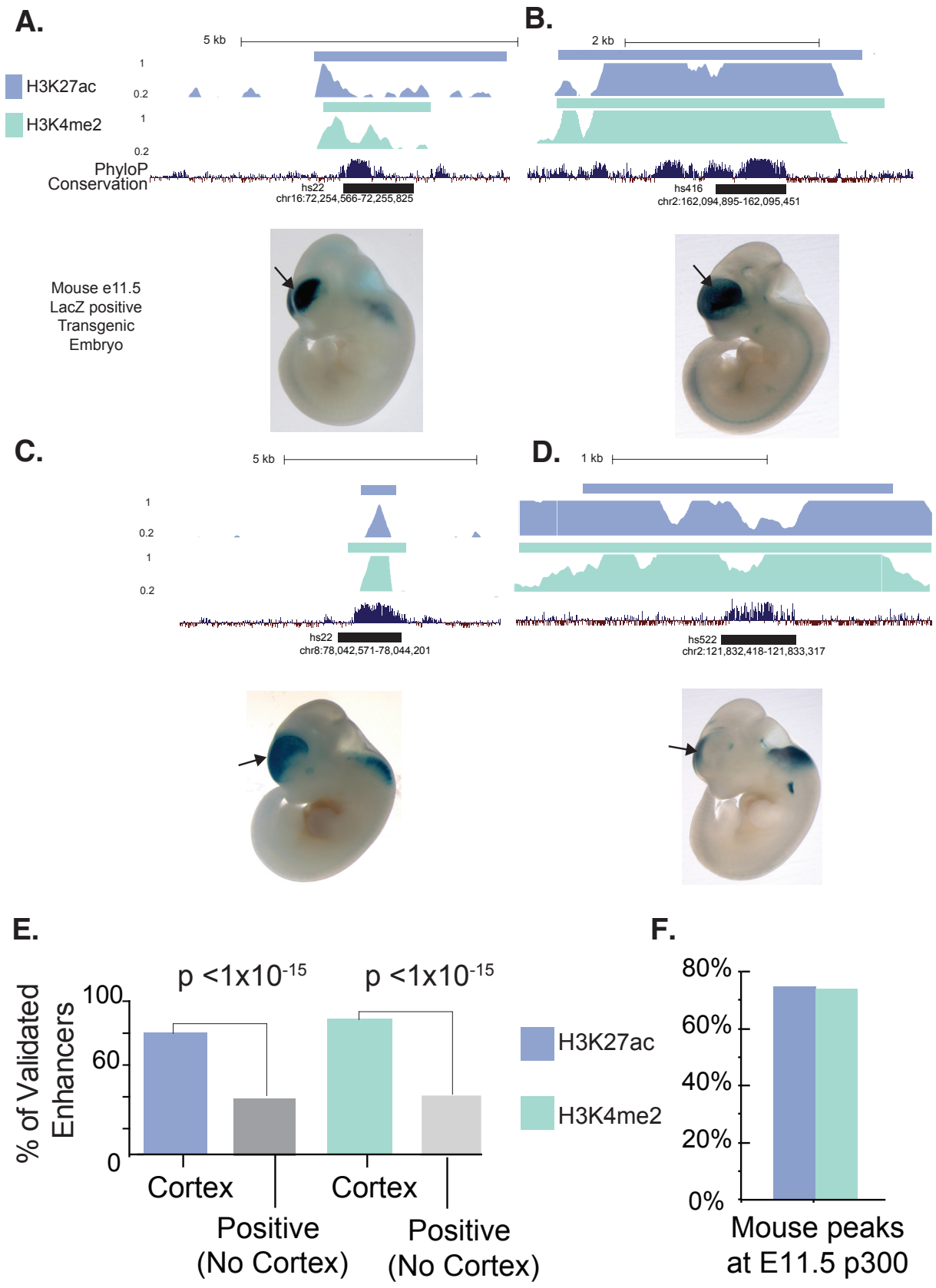

A.
5 kb
H3K27ac
H3K4me2
PhyloP Conservation
hs22
chr16:72,254,566-72,255,825
Mouse e11.5 LacZ positive Transgenic Embryo
B.
2 kb
hs416
chr2:162,094,895-162,095,451
C.
5 kb
hs22
chr8:78,042,571-78,044,201
D.
1 kb
hs522
chr2:121,832,418-121,833,317
E.
p <1x10-15
p <1x10-15
% of Validated Enhancers
100
60
0
Cortex
Positive (No Cortex)
Cortex
Positive (No Cortex)
H3K27ac
H3K4me2
F.
80%
60%
40%
20%
0%
Mouse peaks at E11.5 p300

Fig. S2. H3K27ac and H3K4me2 signal profiles at human enhancer sequences with known activity in forebrain. A-D) Four examples of human enhancers that drive LacZ reporter gene expression in the embryonic forebrain in a mouse E11.5 transgenic enhancer assay. The images, tested sequence coordinates (black bar; hg19 assembly), and VISTA ID for each enhancer were obtained from the VISTA Enhancer Browser (enhancer.lbl.gov). Normalized H3K27ac and H3K4me2 signal profiles, enriched regions (indicated by a bar above each signal profile), and phyloP conservation in placental mammals are shown for each element. E) Percentage of experimentally validated human enhancers in the VISTA Enhancer Browser active in forebrain or other tissues that are reproducibly marked by H3K27ac or H3K4me2. P values were calculated using Fisher's exact test. F) Percentage of mouse P300 peaks identified by overlaps with an H3K27ac or H3K4me2 peak in mouse.

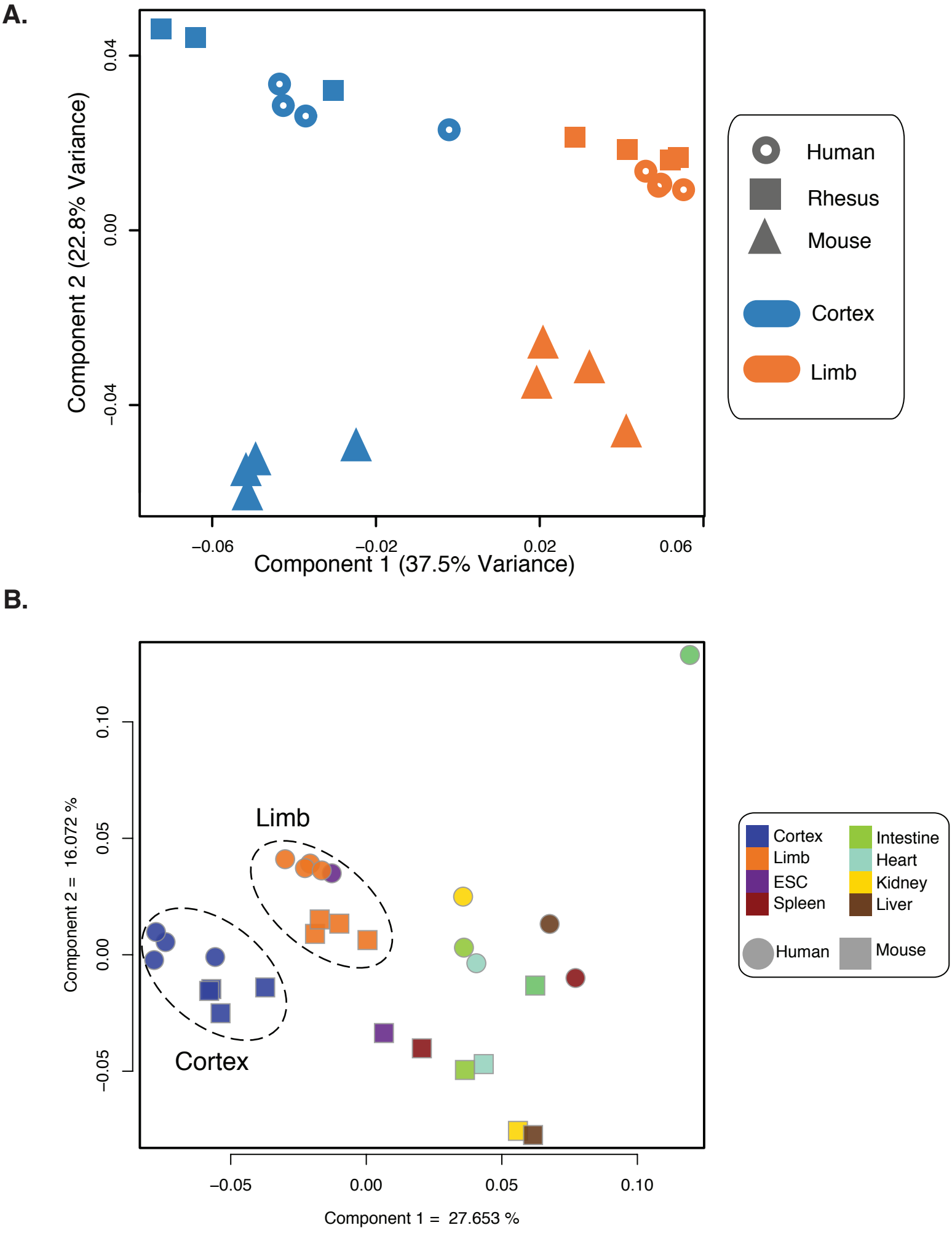
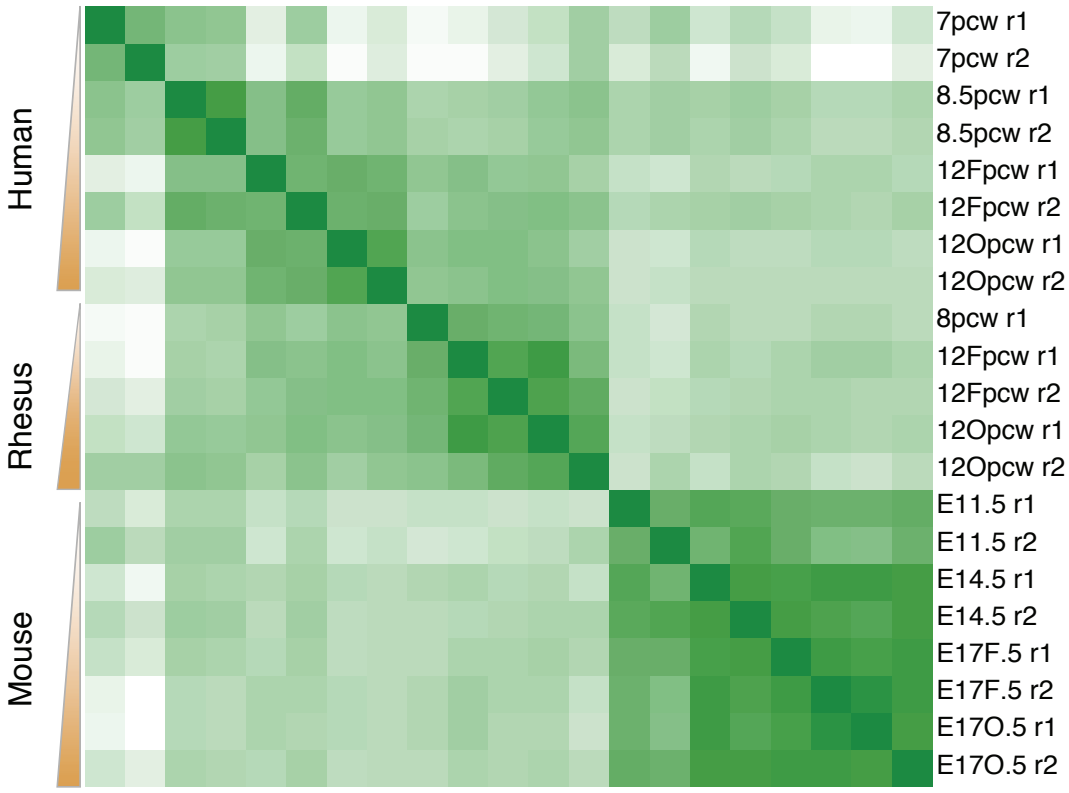
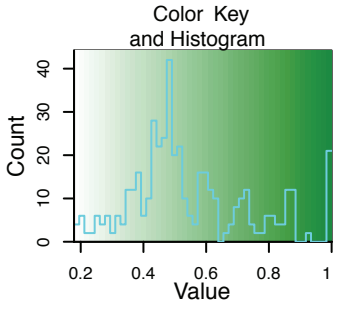

A.
Component 2 (22.8% Variance)
0.04
0.00
−0.04
−0.06
−0.02
0.02
0.06
Component 1 (37.5% Variance)
Human
Rhesus
Mouse
Cortex
Limb
B.
Component 2 = 16.072 %
0.10
0.05
0.00
−0.05
Limb
Cortex
−0.05
0.00
0.05
0.10
Component 1 = 27.653 %
Cortex
Limb
ESC
Spleen
Intestine
Heart
Kidney
Liver
Human
Mouse

Fig. S3. Principal component analysis of H3K27ac marking in human and mouse tissues.
A) Principal component analysis of H3K27ac signals at orthologous sites in developing human, rhesus, and mouse cortex and limb. B) PCA of H3K27ac signals in human and mouse in developing cortex, developing limb, embryonic stems cells, and adult spleen, liver, kidney, intestine, and heart. Dashed ellipses are added for emphasis, to show close clustering in the PCA of the human and mouse limb or cortex tissues, compared to the adult tissues for each species.

A.

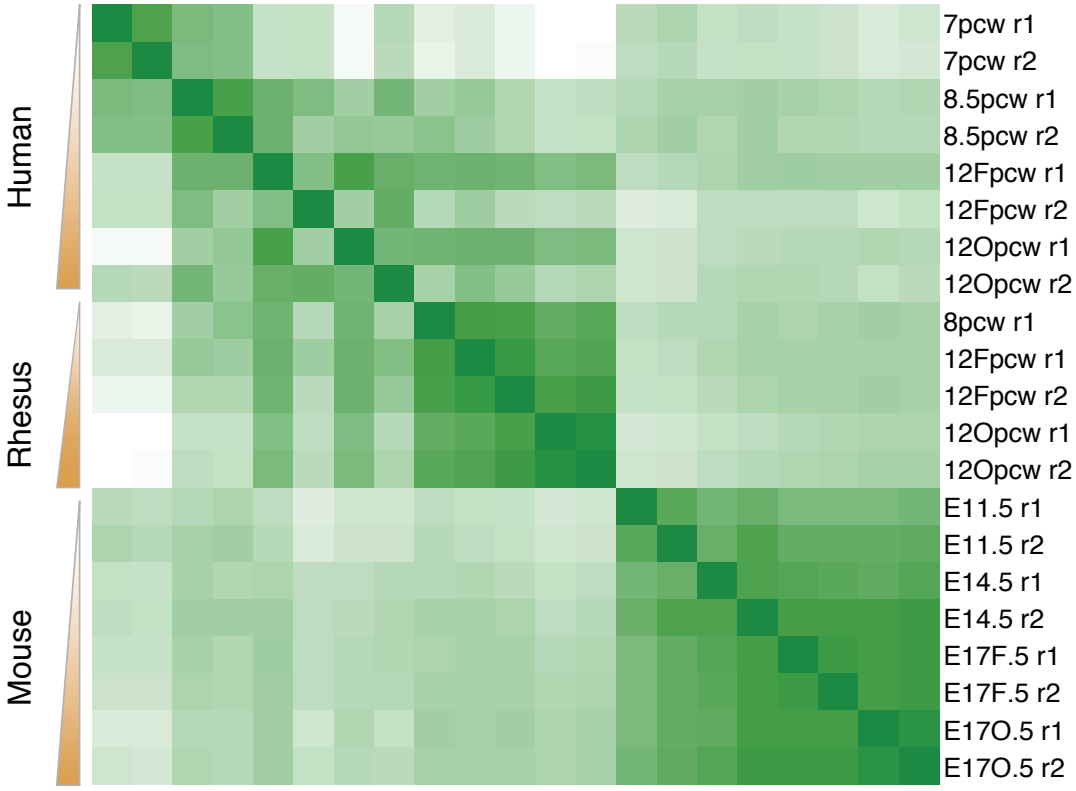


Cross Species Correlation of H3K27ac Enriched Regions in Developing Cortex

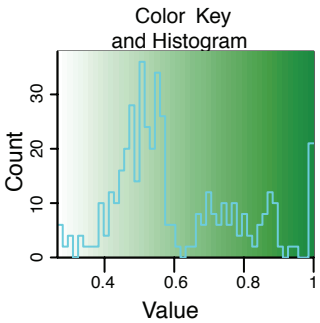


Increasing Developmental Age

B.



Cross Species Correlation of H3K4me2 Enriched Regions in Developing Cortex



Increasing Developmental Age

Fig. S4. Spearman correlation analysis of H3K27ac and H3K4me2 in the developing cortex. *Left.* Spearman correlation matrix for A) H3K27ac and B) H3K4me2 signals at all 3-way orthologous sites in each ChIP-seq replicate from each species in the study. Replicates are denoted as *r1* and *r2*. *Right*. Color key and histogram for all pairwise correlation values.

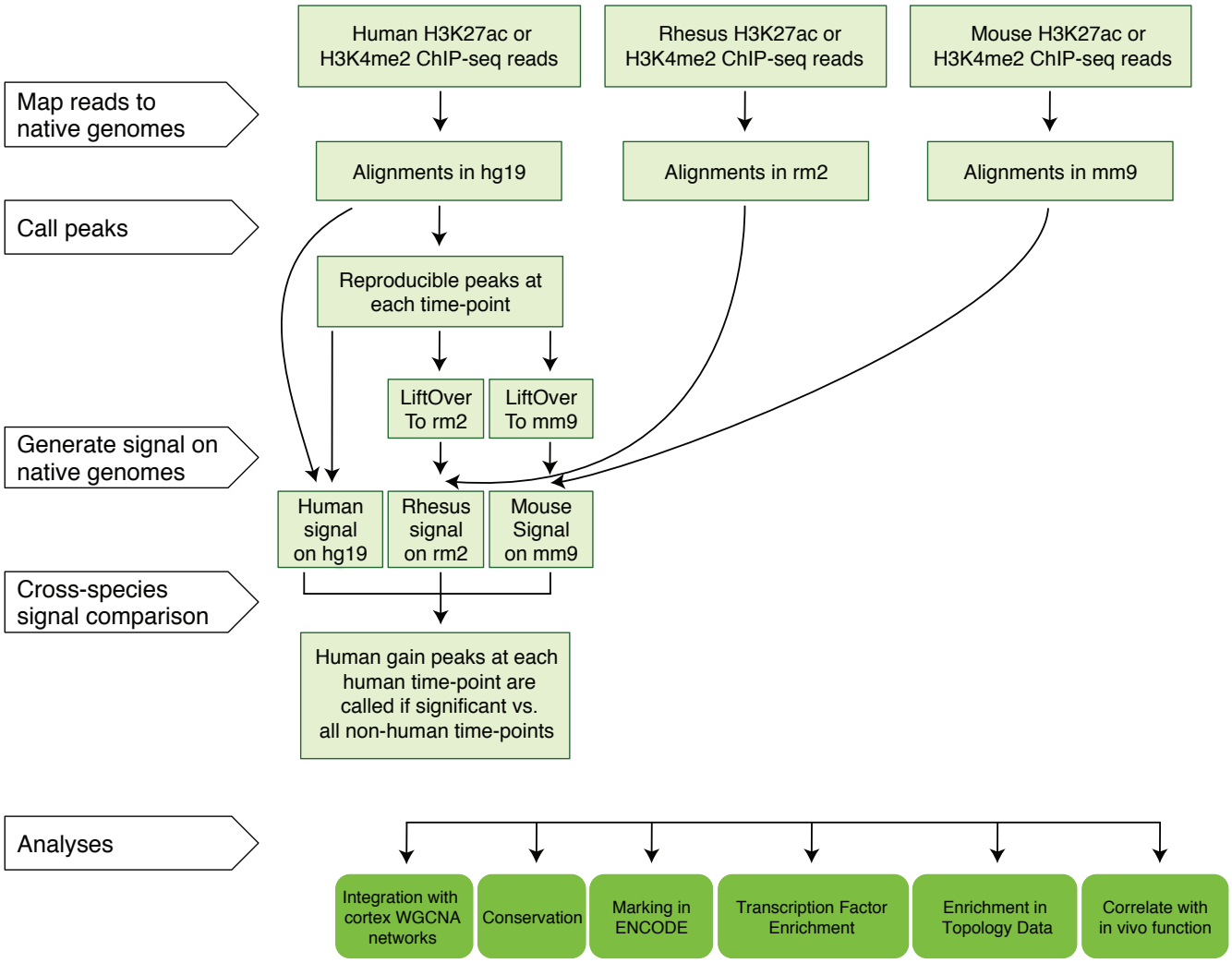
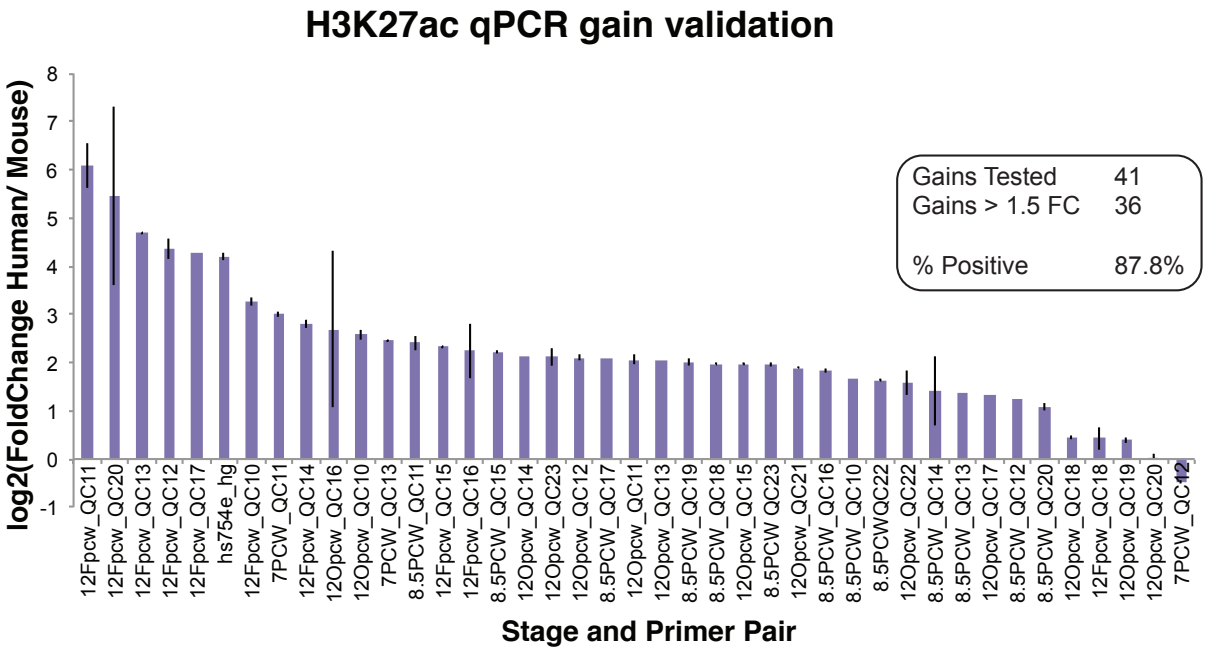


Fig. S5. Workflow for the identification and analysis of human lineage epigenetic gains.

A.



B.

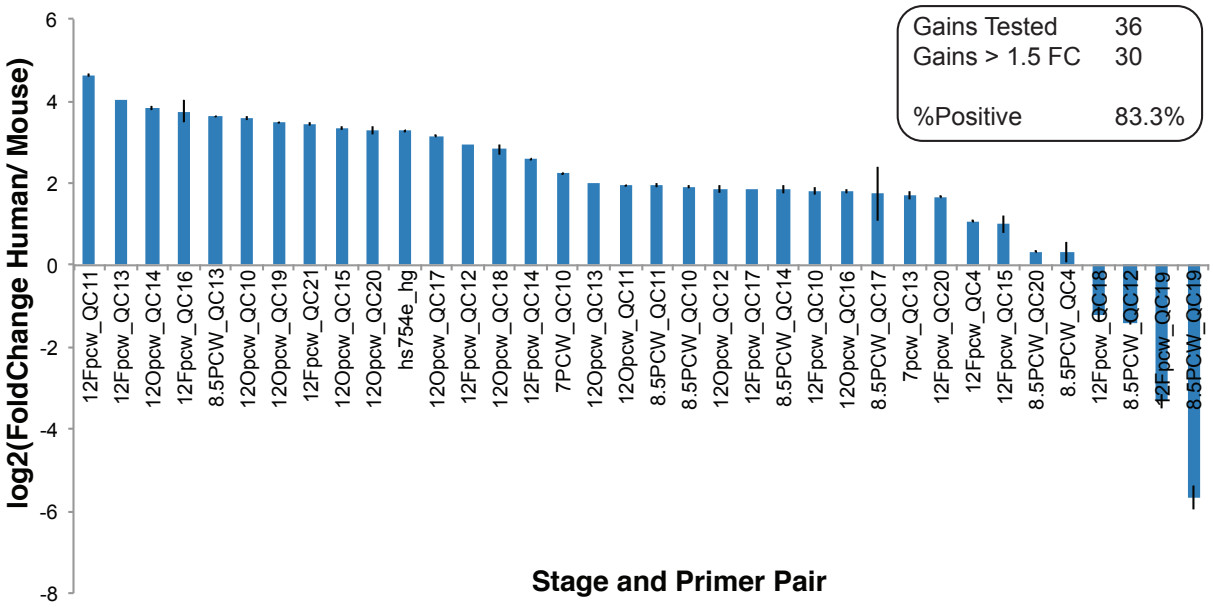
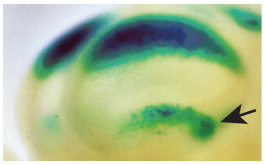


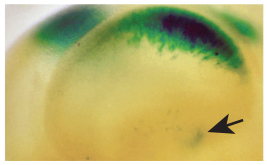
Fig. S6. Validation of human lineage epigenetic gains by ChIP-qPCR.
A) H3K27ac and B) H3K4me2 fold enrichments identified by qPCR at gains identified in our genome-wide screen. The human developmental stage at which human-mouse differences were assessed is shown for each target. Fold enrichment (FE) was calculated as: FE = (human signal(ChIP/input))/(mouse signal(ChIP/input). Error bars indicate FE +/- standard deviation across three replicates.

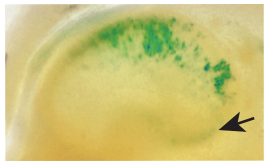
Human

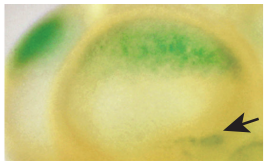
A.

Human Reproducibility
Forebrain:
Rostral Ventral (RV): 4/8
Rostral Dorsal (RD): 7/8
Caudal Ventral (CV): 7/8
Caudal Dorsal (CD): 8/8
Midbrain:
Rostral (R): 3/4
Caudal (C): 1/4
Hindbrain: 3/4

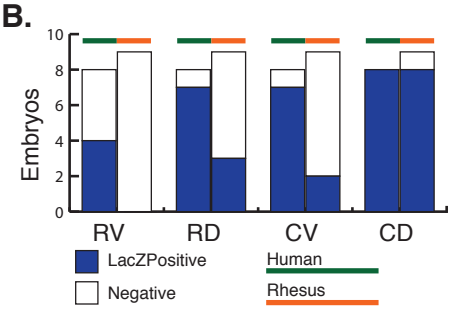

Forebrain: RV: 1; RD: 1; CV: 1; CD:1
Midbrain: R: 1; C: 0
Hindbrain: 1


Forebrain: RV: 0; RD: 1; CV: 1; CD:1
Midbrain: R: 1; C: 0
Hindbrain: 1

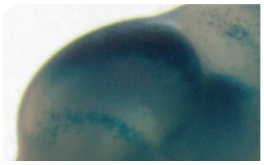

Forebrain: RV: 0; RD: 1; CV: 1; CD:1
Midbrain: R: 1; C: 0
Hindbrain: 0

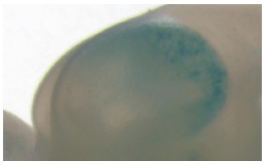

Forebrain: RV: 0; RD: 1; CV: 1; CD:1
Midbrain: R: 0; C: 1
Hindbrain: 1

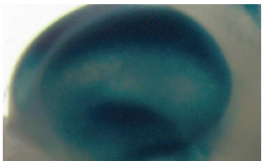
B.

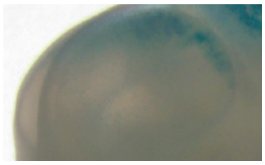


VISTA


Forebrain: RV: 1; RD: 1; CV: 1; CD:1
Midbrain: R: 1; C: 1
Hindbrain: 1

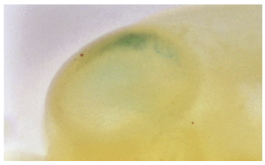

Forebrain: RV: 0; RD: 0; CV: 1; CD:1
Midbrain: R: 0; C: 0
Hindbrain: 0

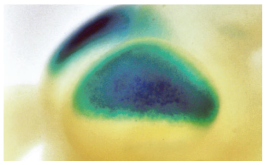

Forebrain: RV: 1; RD: 1; CV: 1; CD:1
Midbrain: R: 1; C: 1
Hindbrain: 1

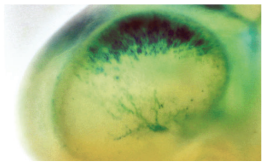

Forebrain: RV: 1; RD: 1; CV: 0; CD:1
Midbrain: R: 1; C: 1
Hindbrain: 1

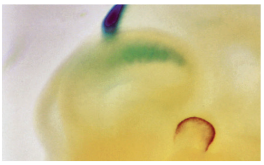
Rhesus

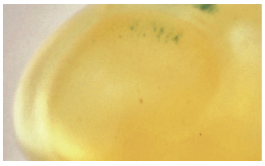
Rhesus Reproducibility
Forebrain:
Rostral Ventral (RV): 0/9
Rostral Dorsal (RD): 3/9
Caudal Ventral (CV): 2/9
Caudal Dorsal (CD): 8/9
Midbrain:
Rostral (R): 7/9
Caudal (C): 2/9
Hindbrain: 3/9

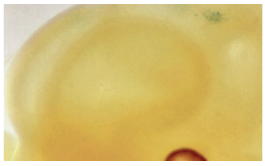

Forebrain: RV: 0; RD: 0; CV: 0; CD:1
Midbrain: R: 0; C: 0
Hindbrain: 0

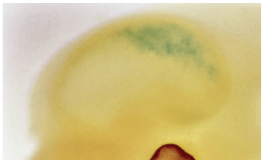

Forebrain: RV: 0; RD: 1; CV: 1; CD:1
Midbrain: R: 1; C: 0
Hindbrain: 0

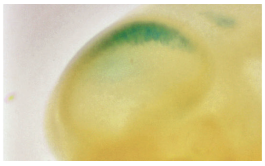

Forebrain: RV: 0; RD: 1; CV: 1; CD:1
Midbrain: R: 1; C: 0
Hindbrain: 1

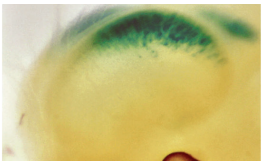

Forebrain: RV: 0; RD: 0; CV: 0; CD:1
Midbrain: R: 1; C: 1
Hindbrain: 0


Forebrain: RV: 0; RD: 0; CV: 0; CD:1
Midbrain: R: 1; C: 0
Hindbrain: 1

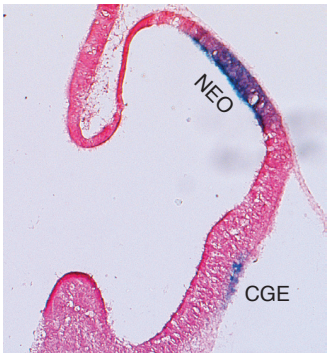

Forebrain: RV: 0; RD: 0; CV: 0; CD:0
Midbrain: R: 1; C: 0
Hindbrain: 0


Forebrain: RV: 0; RD: 0; CV: 0; CD:1
Midbrain: R: 0; C: 0
Hindbrain: 0


Forebrain: RV: 0; RD: 0; CV: 0; CD:1
Midbrain: R: 1; C: 1
Hindbrain: 0


Forebrain: RV: 0; RD: 1; CV: 0; CD:1
Midbrain: R: 1; C: 0
Hindbrain: 1

C.



D.

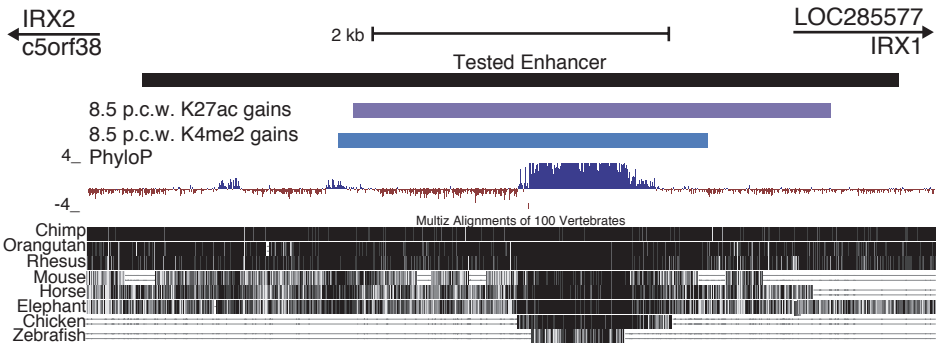


Fig. S7. Photo array of LacZ forebrain expression in all LacZ-positive embryos obtained for the human and rhesus orthologous enhancers shown in Fig. 1C. A). Embryos illustrating LacZ activity in cortex for the human (*top*) and rhesus (*bottom*) enhancer sequences tested in our laboratory, as well as annotated activity of the corresponding hs754 element in the VISTA Enhancer Browser. The number of embryos exhibiting activity in the indicated anatomical regions of the forebrain are listed to the left of the human and rhesus photo arrays. The human-specific increase in caudal-ventral LacZ expression in each embryo is indicated by an arrow. B) Graphical summary of LacZ expression in the indicated forebrain regions for the orthologous human and rhesus enhancers. C) Coronal section from a representative embryo transgenic for the human enhancer, with expression in the neocortex *(NEO)* and caudal ganglionic eminence *(CGE)* indicated. D) UCSC Genome Browser image (hg19) showing the location and evolutionary conservation of the human enhancer sequence. The two closest upstream and downstream genes are listed.

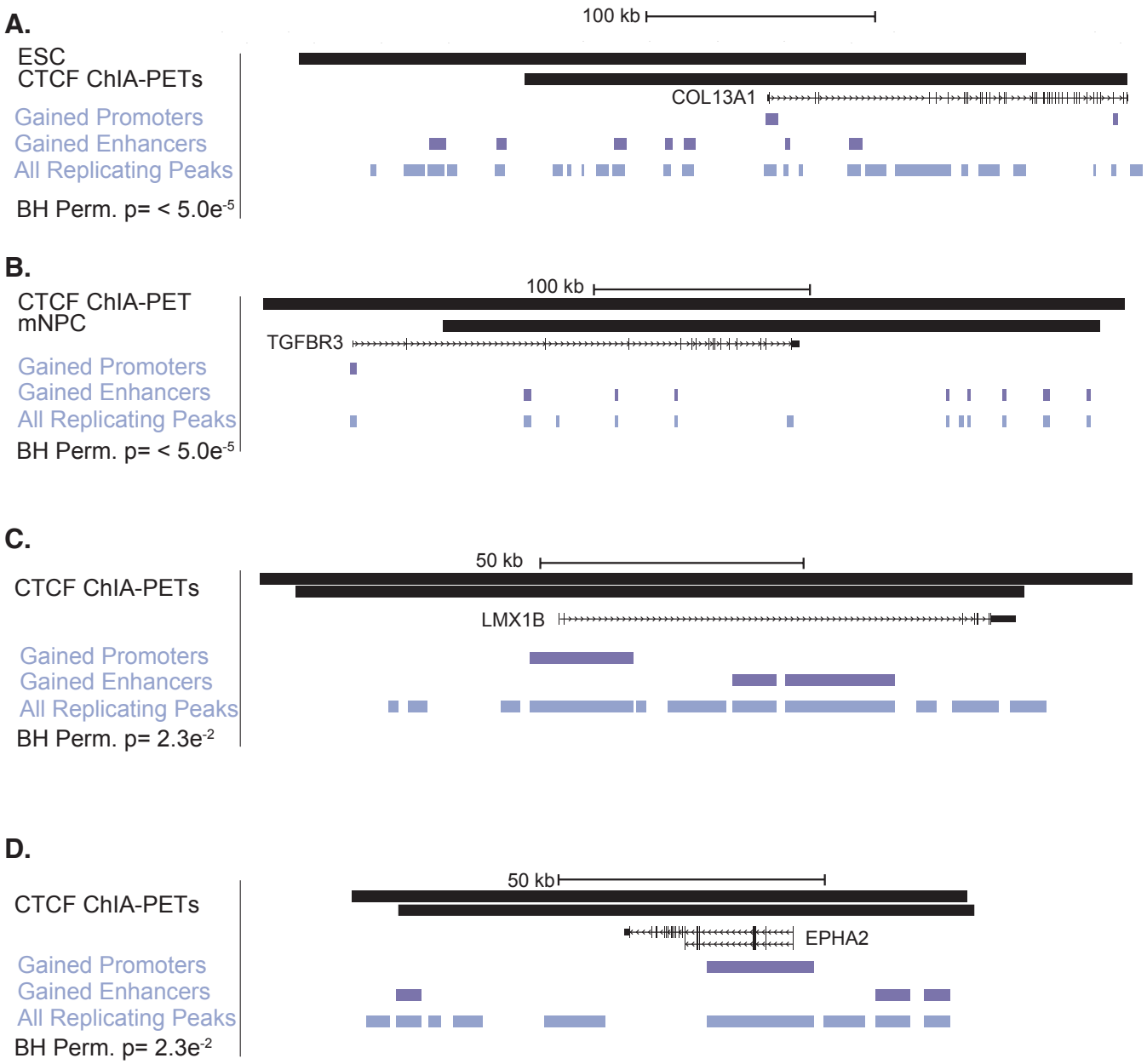


Fig. S8. Topological domains enriched in human gains. Four topological domains (A-D) significantly enriched in human gains based on permutation (see Methods). Representative looping events enriched for promoter or enhancer gains are shown. Only 7 p.c.w. peak calls are shown for clarity (UCSC hg19). The BH corrected permutation P-value for the most enriched loop in each panel is shown.

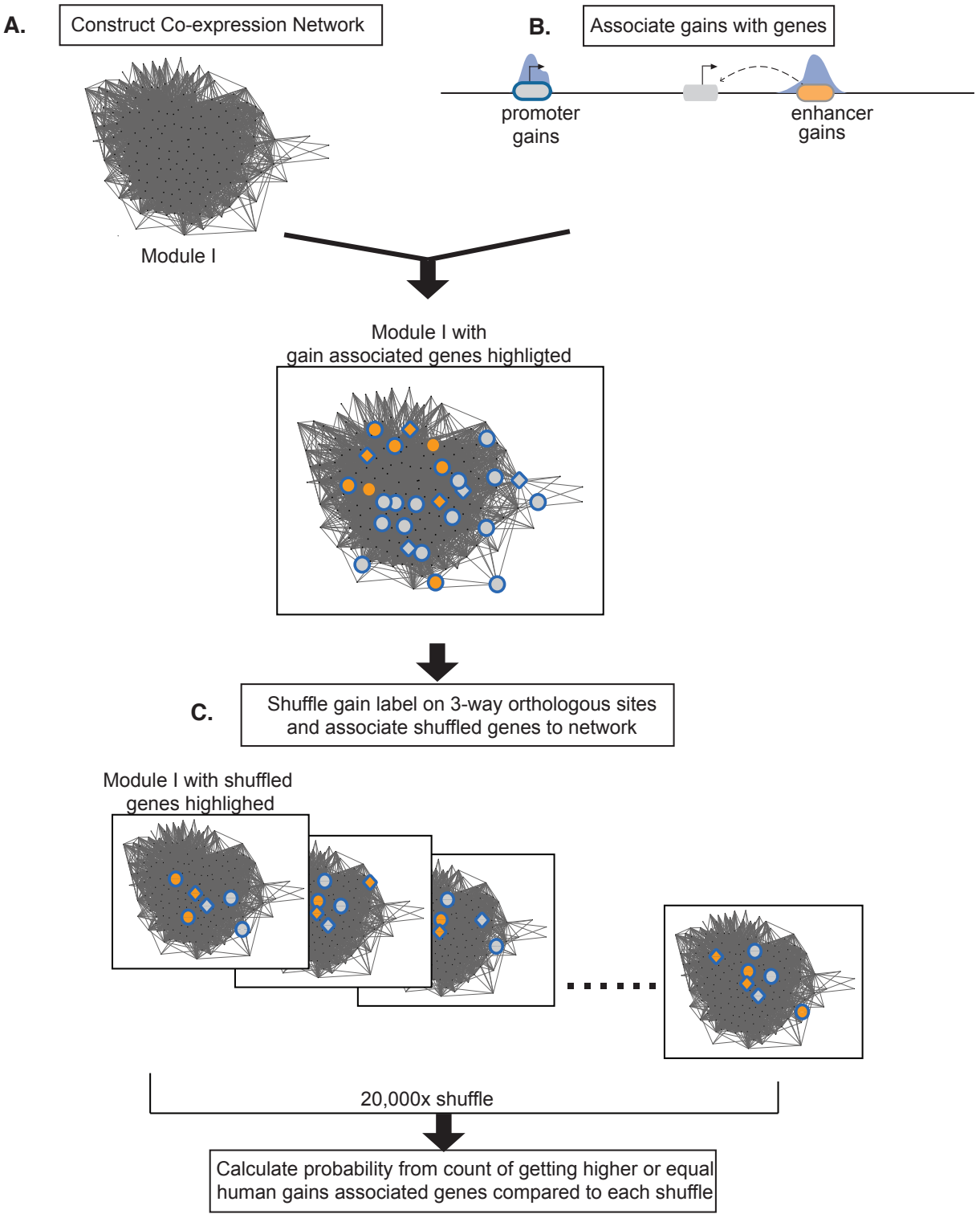


Fig. S9. Workflow for integrating epigenetic gains into gene co-expression modules.

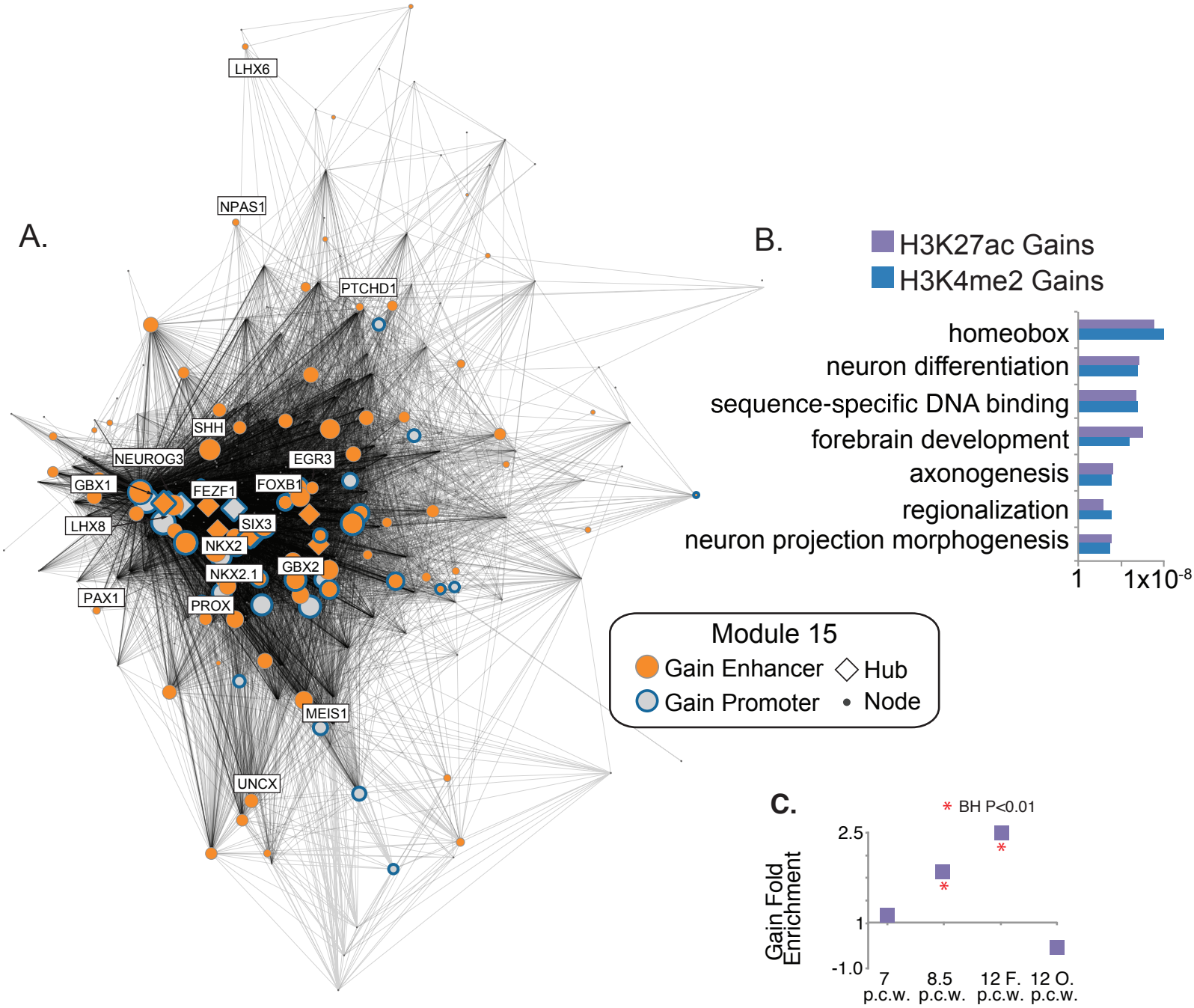
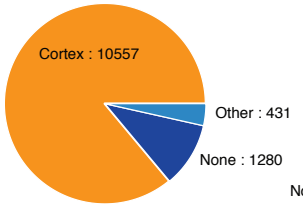
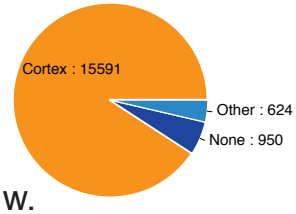

A.
LHX6
NPAS1
PTCHD1
SHH
NEUROG3
EGR3
GBX1
FEZF1
FOXB1
LHX8
SIX3
NKX2
NKX2.1
GBX2
PAX1
PROX
MEIS1
UNCX
B.
H3K27ac Gains
H3K4me2 Gains
homeobox
neuron differentiation
sequence-specific DNA binding
forebrain development
axonogenesis
regionalization
neuron projection morphogenesis
1
1x10-8
Module 15
Gain Enhancer
Hub
Gain Promoter
Node
C.
* BH P<0.01
Gain Fold Enrichment
2.5
1
-1.0
7 p.c.w.
8.5 p.c.w.
12 F. p.c.w.
12 O. p.c.w.

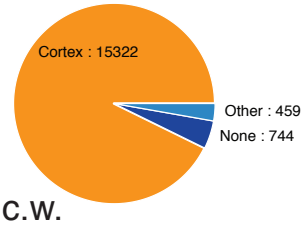
Fig. S10. Enrichment of epigenetic gains in Module 15 associated with cortical patterning. A) Epigenetic gains mapped onto Module 15, with genes associated with gains highlighted as in Fig. 2B and Fig. 3A. B) Gene Ontology enrichments obtained for genes associated with H3K27ac (purple) and H3K4me2 (blue) gains. P values were calculated using the binomial test implemented in DAVID. C) Fold enrichment of H3K27ac promoter gains at each human time point in this module (* = BH corrected permutation P value < 0.01).

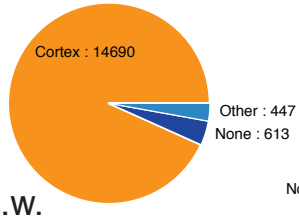
Enhancers

All 3-way Orthologous

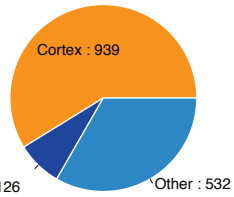


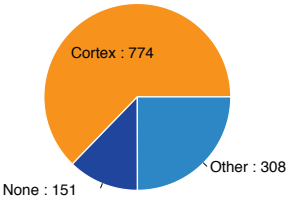


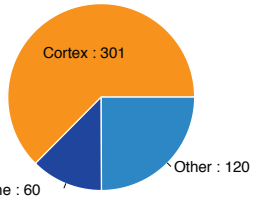


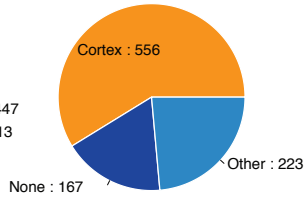


Human Gains



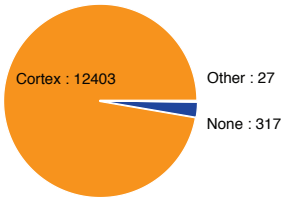


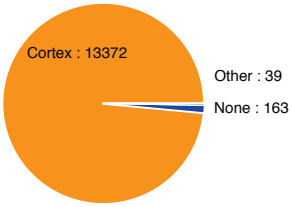


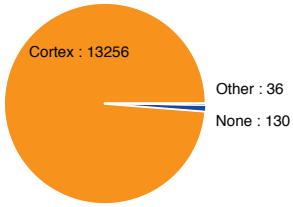


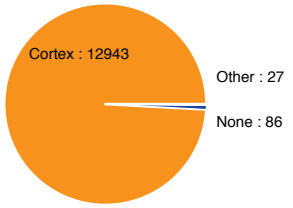
Promoters

All 3-way Orthologous

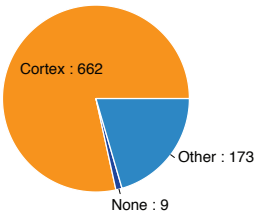


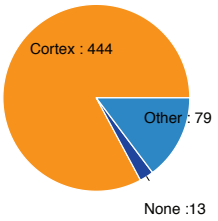


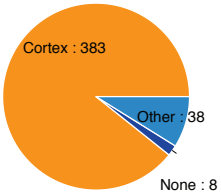


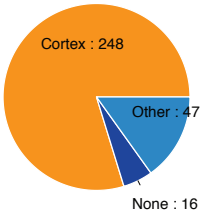


Human Gains









7 p.c.w.

8.5 p.c.w.

12F p.c.w.

12O p.c.w.

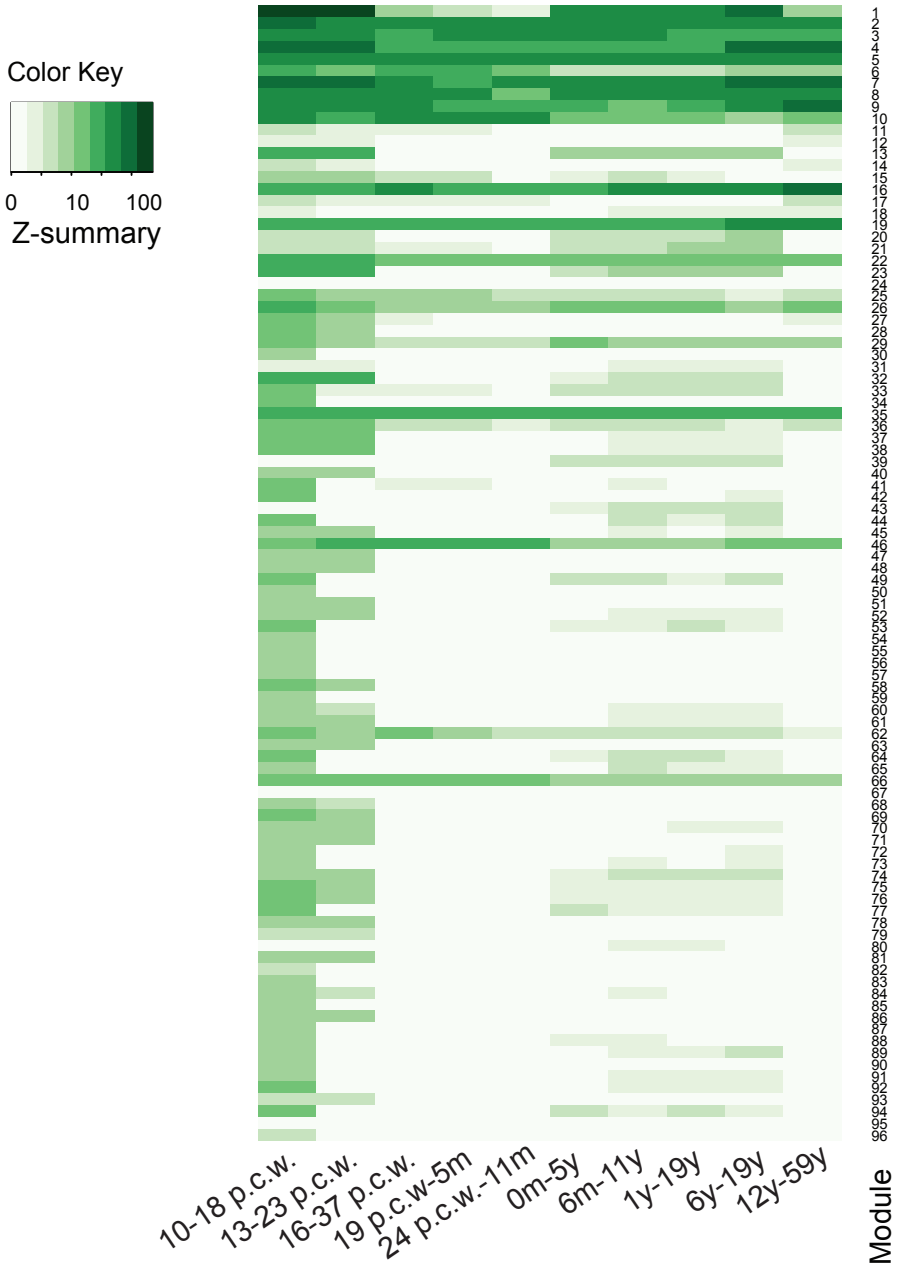
Cortex

Other Tissues

None

Fig. S11. Putative ancestral regulatory activities associated with human lineage gains.
A) The proportion of human lineage H3K27ac gains at enhancers (*left*) or promoters (*right*) that were marked by H3K27ac in rhesus or mouse cortex ("Cortex"), at least one non-cortical rhesus or mouse tissue ("Other Tissues"), or that were not marked by H3K27ac in any of the 22 rhesus and mouse datasets that were analyzed ("None"). These categories respectively correspond to potential modification, co-option, or *de novo* events as described in the text.

A.


Color Key
0
10
100
Z-summary
10-18 p.c.w.
13-23 p.c.w.
16-37 p.c.w.
19 p.c.w-5m
24 p.c.w.-11m
0m-5y
6m-11y
1y-19y
6y-19y
12y-59y
Module

B.

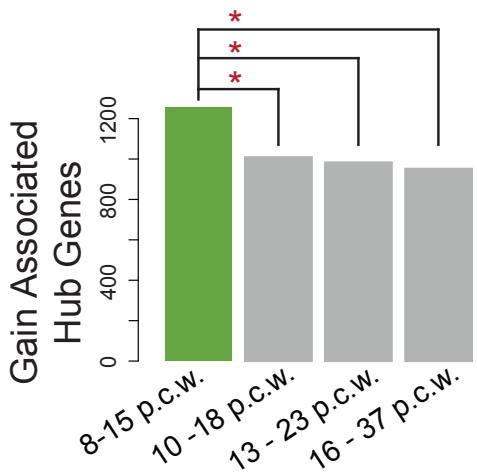

*
*
*
Gain Associated Hub Genes
0
400
800
1200
8-15 p.c.w.
10 -18 p.c.w.
13 - 23 p.c.w.
16 - 37 p.c.w.

Fig. S12. Temporal preservation of gene co-expression modules. A) Temporal preservation of modules within the 8-15 p.c.w. network compared to subsequent developmental windows evaluated using Z-summary scores. The Z-summary score for each module incorporates the module density (the mean adjacency among all the nodes) and connectivity (the sum of connection strength of each node to all the other nodes). If the Z-summary score is >10, there is a strong evidence that the module is preserved. Z-summary decreases when the module is poorly preserved. The figure shows a heatmap of Z-summary scores for each module (y-axis) in each subsequent developmental window (x-axis). The color key scale for the heatmap is shown at the upper left. Most modules identified at 8-15 p.c.w. are not well preserved at later developmental stages, indicating they are unique to this time frame. (B) Human lineage epigenetic gains are associated with a significantly larger number of hub genes identified at 8-15 p.c.w. than hub genes identified in subsequent developmental windows. The number of hub genes associated with gains is highest at 8-15 p.c.w., and decreases at later time points. P values were calculated using Fisher's exact test (*= $P < 0.05$).