# Title: Epigenomic profiling of primate LCLs reveals the coordinated evolution of gene expression and epigenetic signals in regulatory architectures

**Authors: Raquel García-Pérez[1*#], Paula Esteller-Cucala[1#], Glòria Mas[2,3], Irene Lobón[1], Valerio Di Carlo[2,3], Meritxell Riera[1], Martin Kuhlwilm[1], Arcadi Navarro[1,4,5], Antoine Blancher[6,7], Luciano Di Croce[2,3,5], José Luis Gómez-Skarmeta[8], David Juan[1*], Tomàs Marquès-Bonet[1,4,9,10*]**

**Affiliations:**

[1] Institute of Evolutionary Biology (UPF-CSIC), PRBB, Barcelona, Spain

[2] Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Spain

[3] Universitat Pompeu Fabra (UPF), Barcelona, Spain

[4] National Institute for Bioinformatics (INB), PRBB, Barcelona, Spain

[5] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

[6] Laboratoire d'immunologie, CHU de Toulouse, Institut Fédératif de Biologie, hôpital Purpan, Toulouse, France

[7] Centre de Physiopathologie Toulouse-Purpan (CPTP), Université de Toulouse, Centre National de la Recherche Scientifique (CNRS), Institut National de la Santé et de la Recherche Médicale (Inserm), Université Paul Sabatier (UPS), Toulouse, France

[8] Centro Andaluz de Biología del Desarrollo (CABD), Consejo Superior de Investigaciones Científicas-Universidad Pablo de Olavide-Junta de Andalucía, Seville, Spain

[9] CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain

[10] Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona, Spain

[#] Contributed equally to this work

*Corresponding author. Email: tomas.marques@upf.edu (T.M.-B); david.juan@upf.edu (D.J.); raquel.garcia@bsc.es (R.G.-P.)

**Summary**

To gain insight into the evolution of the epigenetic regulation of gene expression in primates, we extensively profiled a new panel of human, chimpanzee, gorilla, orangutan and macaque lymphoblastoid cell lines, using ChIP-seq for five histone marks, ATAC-seq and RNA-seq, further complemented with WGS and WGBS. We annotated regulatory elements and integrated chromatin contact maps to define gene regulatory architectures, creating the largest catalog of regulatory elements in primates to date. We highlight the role of promoters and intragenic enhancers in epigenetically coordinated gene regulatory architectures. We also observe that epigenetic conservation and its correlation with sequence conservation depends on the activity state of the regulatory element. Remarkably, we find that novel human-specific intragenic enhancers with weak activities are enriched in human-specific mutations and appear in genes with signals of positive selection, tissue-specific expression and specific functional enrichments, suggesting that these genes may have contributed to important human adaptations.

**Keywords:** Epigenomics, gene regulation, evolution.

## Introduction

Changes in chromatin structure and gene regulation play a crucial role in evolution[1,2]. Gene expression differences have been extensively studied in a variety of species and conditions[3–6]. However, there is still much unknown about how regulatory landscapes evolve, even in closely related species. Previous work has focused on the dynamics of the establishment and removal of strongly active regulatory elements during the evolution of mammals –mainly defined from ChIP-seq experiments on a few histone marks[7–10]. These analyses suggested that enhancers evolve faster than promoters[8,11]. It has also been highlighted that the regulatory complexity of a gene defined as the number of strongly active enhancers located near a gene is important for the conservation of gene expression[9]. Moreover, in a selected group of primates –mostly chimpanzees and macaques– changes in histone mark enrichments are associated with gene expression differences[12]. Several studies have also targeted the appearance of human-specific methylation patterns[13,14] and strongly active promoters and enhancers in different anatomical structures and cell types[8,10]. All these studies have proven that comparative epigenomics is a powerful tool to investigate the evolution of regulatory elements[15,16]. Nevertheless, a deeper understanding of the evolution of gene regulation requires the integration of multi-layered epigenome data. Only such integration can provide the necessary resolution for investigating recent evolutionary time frames, for example, within human evolutionary relatives. Here, we provide an in-depth comparison of the recent evolution of gene regulatory architectures using a homologous cellular model system in human and non-human primates.

## Results

### Comprehensive profiling of primate lymphoblastoid cell lines (LCLs)

We have extensively characterized a panel of lymphoblastoid cell lines (LCLs) from human, chimpanzee, gorilla, orangutan and macaque, including two independent biological replicates for each species. This characterization includes chromatin immunoprecipitation data (ChIP-seq) from five key histone modifications (H3K4me1, H3K4me3, H3K36me3, H3K27ac, and H3K27me3) and deep-transcriptome sequencing (RNA-seq) (Fig. 1). We integrate these datasets into gene regulatory architectures (Fig. 2) to (1) better understand how primate gene expression levels are controlled and how expression changes occur between species and to (2) study patterns of evolutionary conservation of regulatory elements in primates. As part of this primate epigenomic resource, we have also processed high coverage whole-genome and whole-genome bisulfite sequencing data, as well as

chromatin accessibility data (Supplementary Fig. 1, Supplementary Tables 1-10 and Additional files 1-5). This catalog is the most extensive collection of great apes and macaque transcriptomic and epigenomic data to date.

**Annotation and classification of regulatory elements**

We used the signal of the ChIP-seq experiments from the five histone marks to identify regulatory regions with characteristic marks of promoters or enhancers (Supplementary Fig. 1) and defined regulatory regions for each cell line as those containing chromatin states enriched in any regulatory-related histone mark (Fig. 2a). Overlapping regulatory regions in the two replicates of every species were merged to define species regulatory elements.

We classified regulatory elements based on their epigenetic state (Fig. 2b). We established a hierarchy of epigenetic states that differentiates regulatory elements into those showing epigenetic promoter (P) and enhancer (E) states, with three different levels of activity: strong (s), poised (p) or weak (w) (Methods). Then, to improve the robustness of the assignments, we applied a linear discriminative analysis (LDA) using normalized histone and open chromatin enrichments (Methods and Supplementary Figs. 2 and 3). This resulted in significantly more similar regulatory landscapes between biological replicates (Wilcoxon signed rank-test: $P < 0.05$ in all species; Supplementary Figs. 4 and 5) which translated into a higher number of regulatory elements with the same epigenetic state in all species (Wilcoxon signed rank-test: $P = 0.03$; Supplementary Figs. 6 and 7).

On average, we found ~11,000 and ~76,000 regulatory elements with promoter and enhancer states per species, respectively (Fig. 3a-b), of which 69% and 33% were strong, 8% and 4% were poised, and 14% and 45% were weak, respectively (Supplementary Fig. 8 and Supplementary Table 1). Active and poised activities were significantly more associated with promoter states, whereas weak activities were more frequently associated with enhancer states (Chi-square test: $P < 2.2 \times 10^{-16}$ in all species; Supplementary Fig. 9).

We also classified regulatory elements into five different types of gene regulatory components according to the role they had in their target genes (Fig. 2, Methods). We first classified regulatory elements based on their proximity to a gene as genic promoters (gP), intragenic enhancers (gE) and proximal enhancers (prE) based on the corresponding species gene annotation[17]. Given that gene expression is controlled by a combination of short- and long-distance regulatory interactions[18], we used available 3D contact maps for human LCLs[19–21] to link interacting regulatory elements to their

target gene/s. This way, we defined promoter-interacting enhancers (PiE), regulatory elements that interact with genic promoters, and enhancer-interacting enhancers (EiE), regulatory elements that interact with any other type of gene associated enhancer (Fig. 3e). This approach allowed us to link, on average, nearly 3,500 distal regulatory elements to genes, which would have remained orphan otherwise (Supplementary Fig. 10). The set of regulatory elements associated with a gene defines its regulatory architecture.

On average, 70% of the regulatory elements were associated with genes, of which 93% were protein-coding and 61% were 1-to-1 orthologous protein-coding genes in all primate species (Fig. 3c-d). We annotated ~12,500 genic promoters, ~35,000 intragenic enhancers, ~6,700 proximal enhancers, ~6,200 promoter-interacting enhancers and ~1,800 enhancer-interacting enhancers per species (Supplementary Fig. 11), of which 48%, 69%, 40%, 62%, and 61% were associated with 1-to-1 orthologous protein-coding genes in all primate species (Fig. 3e-f).

Altogether, this catalog of regulatory elements provides a comprehensive view of the regulatory landscape both in humans and non-human primates. Moreover, in contrast to other commonly used definitions of promoters and enhancers limited to strongly active regions, our multi-layered integration approach allows the additional annotation of weak and poised activities[7,8]. These activities are of particular relevance to improve the definition of regulatory elements and explore the regulatory potential of regions whose activity can differ in other cell types or conditions.

**Coordinated epigenetic signals at regulatory components explain gene expression levels**

To study the coordination of regulatory signals and gene expression in a comparable set of genes between all species, we focused on 1-to-1 orthologous protein-coding genes. The epigenetic state and the component assigned to regulatory elements were highly concordant. On average, 75% of genic promoters had a promoter regulatory state, and 90% of gene-associated enhancers had an enhancer state (Fisher's exact test: $P < 2.2 \times 10^{-16}$ in all species, average $OR = 64$; Supplementary Fig. 12). More specifically, 98% of intragenic enhancers, 95% of proximal enhancers, 81% of promoter-interacting enhancers and 89% of enhancer-interacting enhancers showed an enhancer epigenetic state (Supplementary Fig. 12). The activity state of promoters and enhancers in each type of regulatory component was also consistent across species (Chi-square test: $P < 2.2 \times 10^{-16}$ in all species; Fig. 3g and Supplementary Fig. 13). Genic promoters were strongly enriched in regulatory elements with strong promoter and poised promoter and enhancer states. Strong enhancer states were

mostly enriched at intragenic and promoter-interacting enhancers, whereas weak enhancer states were strongly associated with proximal enhancers (Supplementary Figs. 12 and 13).

Gene expression levels are positively associated with the presence of elements with strong activities and negatively associated with the presence of elements with poised or weak activities in their regulatory architectures (Kruskal-Wallis test: $P < 0.05$ in all species and regulatory components; Supplementary Fig. 14), with particularly strong associations in genic promoters and intragenic enhancers (Dwass-Steel-Critchlow-Fligner test; Supplementary Fig. 15).

To disentangle the underlying network of regulatory interactions between the different regulatory components and gene expression in primates, we use Sparse Partial Correlation Analysis (SPCA)[22] of the normalized RNA-seq and histone mark enrichments (aggregated by promoter and enhancer state in every type of regulatory component) (Methods). This approach establishes a stringent protocol (Benjamini-Hochberg's correction, $P < 1.8 \times 10^{-22}$ for all selected partial correlations) that selects informative partial correlations based on the increase they confer to the explained variance of the co-dependent variables[22].

SPCAs performed independently for every histone mark showed a highly consistent global structure of regulatory interactions, with genic promoters and intragenic enhancers directly regulating gene expression and co-regulated between them, proximal and promoter-interacting enhancers connected with promoters and enhancer-interacting enhancers connected with intragenic and promoter-interacting enhancers (Fig. 4a-b, Supplementary Figs. 16-18 and Supplementary Table 11).

A global SPCA using all the histone marks and gene expression showed that this global structure of inter-component regulatory interactions implies strong coordination between the different histone marks within the regulatory components (Supplementary Fig. 19 and Supplementary Table 12). To unravel the contribution of each of these components to gene expression[23], we defined the consensus signal of each type of regulatory component (or eigencomponents) inspired by the notion of eigengenes[23]. A SPCA based on the eigencomponents clearly shows the robust structure of the different regulatory components and gene expression coordination (Fig. 4c and Supplementary Table 13). Interestingly, the network of Sparse Partial Correlations for the residuals of the histone mark signals consistently points to the same global structure (Supplementary Fig. 20 and Supplementary Table 14), suggesting that those histone mark signals that are not explained by the common coordination of the histone marks captured by the components still contribute to the coordination of the same inter-component regulatory network. To account for the possibility of incompleteness in our

architectures, we replicated all the analyses using only genes with full regulatory architectures (i.e., genes associated with regulatory components of every type) obtaining consistent results (Supplementary Figs. 21-28 and Supplementary Tables 11-14).

In agreement with the structure of regulatory interactions recovered by our SPCAs, a generalized linear model of gene expression based on H3K27ac, H3K27me3 and H3K36me3 signals at genic promoters and intragenic enhancers and their interactions (15 variables) explained ~67% of gene expression variability (Supplementary Table 15). Remarkably, this is only a 6% lower than an exhaustive naive model, including the signal from all histone marks at all types of regulatory components with all possible interactions (1,225 variables) (Supplementary Table 16). These results confirm that the epigenetic activities of genic promoters, intragenic enhancers, and their interactions are likely the most direct determinants of gene expression regulation in our regulatory architectures. However, their co-dependency with the other components suggests that they are dependent, in turn, on the coordination of the whole architecture.

The number of regulatory elements in a gene regulatory architecture can be considered a proxy of its regulatory complexity[9]. Genes are often associated with multiple intragenic enhancers, as well as several proximal, promoter-interacting and enhancer-interacting enhancers (Supplementary Fig. 29). Therefore, we evaluated the relationship between the complexity of each component-state combination and gene expression. Consistently with the SPCAs, the number of strong enhancers in intragenic and promoter-interacting enhancers was positively correlated with the gene expression level in all species (average Spearman's rank correlation coefficient ($\rho$) = 0.21, $P < 7.6 \times 10^{-26}$ in all species, and $\rho = 0.12$, $P < 0.0012$ in all species, respectively; Supplementary Figs. 30 and 31).

We investigated if evolutionary differences in the complexity or number of the regulatory elements associated with a gene were related to changes in expression between species (Supplementary Fig. 32). We found that the effect on gene expression levels depends on the epigenetic state gained or lost at each type of regulatory component (Supplementary Fig. 33). Evolutionary changes that alter the epigenetic state at genic promoters, specifically the presence of either a strong promoter or poised enhancer state, as well as the number of intragenic enhancers with either strong or poised enhancer states, showed the most robust associations with gene expression differences (Supplementary Fig. 33). The number of proximal enhancers in any enhancer epigenetic state and of strong promoter states and strong and poised enhancer states in promoter-interacting enhancers also showed significant though modest effects (Supplementary Fig. 33). These results highlight that the additive nature of gene regulation depends on the regulatory architectures. This dependency can be captured either by

the aggregation of histone enrichment signals into regulatory components (as in our SPCAs) or by the quantification of the evolution of gene complexities as changes in the number of regulatory components with specific activities.

**Patterns of evolutionary conservation of epigenetic regulatory states in primates**

To investigate the evolutionary conservation of the different epigenetic states throughout primate evolution, we focused on the 28,703 1-to-1 orthologous regions with a promoter or enhancer state in at least one species (Supplementary Fig. 34). Most of these orthologous regulatory regions (~76%) were associated with genes (Methods), a proportion larger than expected by chance (Binomial test: $P < 2.2 \times 10^{-16}$). Of the latter, ~96% could be associated with human genes and were significantly enriched in protein-coding genes (81% of orthologous regions associated with at least one protein-coding gene) (Fisher's exact test: $P < 2.2 \times 10^{-16}$, $OR = 7$). We defined the conservation of an epigenetic state in an orthologous regulatory region as the number of primate species showing that epigenetic state. We observed that promoter states were more conserved than enhancer states when associated with protein-coding genes (Supplementary Figs. 35-37), with 73% and 60% of orthologous regulatory regions with a promoter or enhancer state being fully conserved across primates (Fisher's exact test: $P < 2.2 \times 10^{-16}$, $OR = 1.84$; Supplementary Fig. 36). On the contrary, less than 14% and 8% of orthologous regulatory regions with a promoter or enhancer state, respectively, were specific to a primate species (Supplementary Fig. 36). Remarkably, enhancer states associated with non-coding genes were more conserved than the corresponding promoter states (Fisher's exact test: $P < 2.2 \times 10^{-16}$, $OR = 0.39$; Supplementary Fig. 37), with 46% and 69% of fully conserved and 26% and 3% of species-specific elements, respectively.

We use the term repurposed promoters or enhancers to refer to those orthologous regulatory regions where one species showed a promoter state and all the rest showed an enhancer state or vice versa. We observed that most (93%) recently evolved regulatory elements with promoter states were acquired through repurposing events, whereas the majority (90%) of recently evolved regulatory elements with enhancer states were gained at orthologous regions with non-regulatory states in the other species (Chi-square test: $P < 2.2 \times 10^{-16}$; Methods and Supplementary Figs. 38 and 39). Interestingly, we observed the same pattern for those orthologous regions associated with protein-coding (Fisher's exact test: $P < 2.2 \times 10^{-16}$, $OR = \text{Inf}$) and with non-coding genes (Fisher's exact test: $P = 6.2 \times 10^{-16}$, $OR = 138$; Supplementary Fig. 38, evaluated in human due to underrepresentation of non-coding gene annotations in non-human species). Our results are in agreement with previous studies in more distant species that also included regulatory changes associated with major genetic

277 alterations in their analysis[24]. These patterns suggest that the different dynamics observed for recently
278 evolved promoter and enhancer states are independent of the coding potential of the associated genes.
279
280 We further explored the patterns of evolutionary conservation of the different activity states. Globally,
281 orthologous regulatory regions conserve their regulatory state more often than expected
282 (Randomization analyses: 1,000 simulations, $P < 0.05$; Supplementary Figs. 40-42 and
283 Supplementary Table 17). Remarkably, we found that different promoter and enhancer activities have
284 characteristic patterns of conservation (Kruskal-Wallis test: $P < 2.2 \times 10^{-16}$; Supplementary Figs. 43-
285 45). Most pairwise comparisons of the conservation distribution of different epigenetic states were
286 significantly different (Dwass-Steel-Critchlow-Fligner test, Fig. 5a and Supplementary Figs. 43-45)
287 except for the poised-weak enhancer comparison which showed a different distribution but a very
288 similar average value.
289
290 Promoter epigenetic states with strong activity were the most conserved: 80% of the orthologous
291 regulatory regions in which we detected a strong promoter state showed a fully conserved activity in
292 primates. On the contrary, we observed poor conservation of poised and weak promoter activities in
293 human and non-human primates (Fig. 5a). Contrary to promoter activities, all enhancer activities
294 showed the same pattern of evolutionary conservation (Fig. 5b). Enhancer states with strong activities
295 were second in levels of evolutionary conservation. Nearly 40% of the orthologous regulatory regions
296 with a strong enhancer showed a fully conserved activity in primates. However, poised enhancers
297 followed closely, 36% of the orthologous regulatory regions with a poised enhancer state had a
298 conserved activity in the five species. Lastly, ~21% of the orthologous regulatory regions in which
299 we detected a weak enhancer conserved their activity across primates. We observed the same
300 evolutionary conservation trends for regulatory regions associated with protein-coding and non-
301 coding genes (Supplementary Figs. 46-47). Interestingly, fully conserved epigenetic states (hereafter
302 referred as conserved epigenetic states) were enriched in particular types of regulatory components
303 (Fisher's exact test: $P = 0.0005$). Conserved strong and poised promoter and enhancer states were
304 significantly enriched in genic promoters. Strong enhancer states fully conserved among primates
305 were enriched in intragenic enhancers and promoter-interacting-enhancers, whereas conserved weak
306 enhancer states were associated with intragenic enhancers, proximal enhancers, and enhancer-
307 interacting-enhancers (Supplementary Fig. 48).
308
309 The evolutionary conservation of the epigenetic state was positively correlated with the conservation
310 of the underlying sequence z-scores of background normalized PhastCons values (Fig. 5c-d, Methods
311 and Supplementary Fig. 49) for all epigenetic states but for weak promoters in regulatory regions

associated with protein-coding genes (Randomization analyses: 1,000 simulations; Fig. 5b, Supplementary Figs. 50-54). Orthologous regulatory regions with fully conserved epigenetic states showed significant differences in sequence conservation (Kruskal-Wallis test: $P < 2.2 \times 10^{-16}$; Supplementary Fig. 55). In particular, we found that strong promoter states were associated with significantly higher sequence conservation scores than weak promoter states and any other type of enhancer state, whereas weak enhancer states have significantly less conserved sequences than any other epigenetic state but weak promoter states (Dwass-Steel-Critchlow-Fligner test, Supplementary Fig. 55). The sequence conservation scores associated with strong and poised enhancer states were not significantly different. Note also that conserved poised promoter states were associated with very high conservation z-scores, which probably did not reach significance due to their low number (n = 9 pP). Orthologous regions associated with non-coding genes were fewer and less conserved (Supplementary Figs. 47 and 50), what could be related to the interesting lack of correlation between the conservation of the sequence and the epigenetic state observed in all but in strong enhancer states (Supplementary Fig. 54).

Then, we sought to characterize the expression patterns of the 1-to-1 orthologous protein-coding genes regulated by evolutionarily conserved regulatory regions. Not surprisingly, the expression levels of genes associated with genic promoters, intragenic enhancers and promoters-interacting enhancers with fully conserved strong activities were significantly higher than those associated with evolutionarily conserved weak enhancers (Kruskal-Wallis test: $P < 2.2 \times 10^{-16}$; Dwass-Steel-Critchlow-Fligner test; Supplementary Fig. 56). On the contrary, genes associated with poised promoter or enhancer states at genic promoters were repressed. Remarkably, genes associated with intragenic enhancers with weak activities had very low expression levels, significantly lower than those genes associated with proximal enhancers and enhancer-interacting-enhancers with weak enhancer states.

**Novel human-specific intragenic enhancers appear in lowly expressed tissue-specific genes with signals of positive selection**

To investigate to which extent these conservation patterns of regulatory states translate into functional outcomes, we first examined the relationship between epigenetic conservation and the patterns of gene expression across tissues (Methods). Genes associated with genic promoters and intragenic enhancers with fully conserved strong promoter and enhancer activities, respectively, were highly expressed in LCLs (Fig. 6a, Supplementary Fig. 57 and Supplementary Table 18). As would be expected from genes associated with strong-activity elements, these genes were highly expressed in

most tissues (median expression > 10 TPMs in 23 out of 29 tissues for both types of components) and enriched in several functions associated with metabolism, chromatin organization and regulation of the cell cycle (Fisher's exact test: Benjamini-Hochberg's correction, *FDR* < 0.05; Fig. 6b, Methods, Supplementary Fig. 58 and Supplementary Tables 19-21). Genes with conserved promoter-interacting enhancers and strong enhancer activities also showed high expression levels in many tissues (with LCLs as the fourth higher expressing tissue; Supplementary Fig. 57) but did not show significant functional enrichments (Supplementary Fig. 59). These three sets of genes with conserved strong regulatory activities, regardless of the functional enrichments, showed wide expression breadth (median tissue specificity index (τ, Tau) < 0.24 in all three; Methods and Supplementary Fig. 60).

On the contrary, genes with intragenic enhancers with fully conserved weak activities were lowly expressed in LCLs and showed their highest expression levels in testis, thyroid and pituitary (Fig. 6a). This group of genes showed uneven expression levels (Supplementary Fig. 57), consistent with higher tissue-specificity compared to that of conserved strong regulatory activities both in promoters and enhancers (median τ = 0.72, Dwass-Steel-Critchlow-Fligner test: $P < 2.2 \times 10^{-16}$ in the three tests; Supplementary Fig. 60). These genes were expressed at intermediate-low levels (median expression < 5 TPMs in 28 out of the 29 tissues) and protein-coding genes within this set were enriched in various annotations, including neuronal-specific ones, such as cell projection and synapse (Fisher's exact test: Benjamini-Hochberg's correction, *FDR* < 0.05; Fig. 6b, Supplementary Fig. 58 and Supplementary Table 22). We then focused on the 134 genes associated with novel human-specific intragenic enhancers with weak activities (Methods). Similarly to genes associated with intragenic enhancers with fully conserved weak enhancer states, these genes were typically expressed at low levels (median expression < 5 TPMs in all tissues). This group of genes showed their highest expression in tissues unrelated to LCLs, such as brain, tibial nerve and testis, and have marginal or no expression in numerous tissues including LCLs (Wilcoxon-Nemenyi-McDonald-Thompson test: $P < 1 \times 10^{-4}$; Rank-biserial correlation effect size between brain and LCLs = 0.633; Fig. 6a, Supplementary Fig. 57 and Supplementary Table 23). Remarkably, these genes have higher tissue-specific expression than those with conserved strong but not weak activities in their components (median τ = 0.843, Dwass-Steel-Critchlow-Fligner test: $P < 4.5 \times 10^{-14}$ when compared to strong activities and $P = 0.06$ compared to genes with weak enhancer states; Supplementary Fig. 60) and were enriched in neuron parts and synapse (Fisher's exact test: Benjamini-Hochberg's correction, *FDR* < 0.05; Fig. 6a and Supplementary Table 24).

We then sought to identify the particular contribution of the tissues driving the tissue-specific expression patterns seen in genes with conserved and human-specific intragenic enhancers with weak

enhancer states. We found that testis and brain were the tissues with the highest number of tissue-specific genes ($\tau_{Tissue} > 0.8$). Interestingly, whereas the fraction of testis-specific genes was comparable between gene sets (Two-tailed Fisher's exact test: $P = 0.54$, $OR = 1.20$; Supplementary Fig. 60), brain-specific genes were more than 2-fold enriched in genes with human-specific intragenic enhancers (Two-tailed Fisher's exact test: $P = 0.02$, $OR = 2.29$; Supplementary Fig. 60).

Among the genes associated with novel human-specific intragenic enhancers with weak activities, we found several genes previously proposed as candidates for positive selection in humans[25–28]. Some of these genes were *FOXP2*, *PALMD* and *ROBO1*, which have known brain-related functions[29–32] or *ADAM18*[33], *CFTR*[34,35] and *TBX15*[36]. To explore whether human-specific enhancers have been targeted by recent human adaptation, we investigated their co-occurrence in genes associated with signals of positive selection[25–28] (Methods and Supplementary Table 25). We found that more than one third (38%) of the genes with novel weak intragenic enhancers were associated with genes targeted by positive selection (Fisher's exact test: $P = 6.52 \times 10^{-18}$, $OR = 5.69$). Using a hierarchical strategy, we observed that this enrichment can be partially attributed to the human-specificity of these enhancers when compared to other genes with weak intragenic enhancers (Fisher's exact test: $P = 8.24 \times 10^{-7}$, $OR = 2.61$; Fig. 6c).

Finally, we explored whether these recently evolved human-specific intragenic enhancers were associated with human-specific mutations. For this, we collected a set of over 2.8 million single nucleotide changes fixed in humans (hSNCs) that differ from fixed variants in the genomes of the remaining non-human primates (Methods and Supplementary Table 25). We observed that the hSNCs density was higher in human-specific intragenic enhancers (Mann-Whitney U test: $P = 0.01$; Methods and Supplementary Fig. 61). More than one-third of the genes with novel human-specific intragenic enhancers with weak enhancers states and with hSNCs also have signals of selection (Fig. 5d), a proportion very similar to the expected 38% (see above). This result suggests that although human-specific mutations and positive selection signals are both associated with the presence of human-specific intragenic enhancers with weak activities, they are not mutually conditioned. As such, it implies that none of these signals is necessary (nor sufficient) to explain the appearance of human-specific intragenic enhancers with weak activities. Among the 11 genes with both signals of positive selection and hSNCs (Fig. 6d), there are several interesting candidates for adaptive evolution of different traits. Many of these genes are associated with neuronal functions (*ROBO1, CLVS1, SEMA5A, KCNH7, SDK1* and *ADGRL2*), but also with pigmentation (*LRMDA*) or actin organization in cardiomyocytes (*FHOD3*). Other genes that include human-specific weak intragenic enhancers are not associated with either hSNCs in these enhancers (*FOXP2*, *TNIK*, *ASTN2*, *NPAS3* or *NTM*) or

signals of human selection (*PALMD*, *VPS13C*, *IGSF21* or *CADM2*). Interestingly, we found only one antisense RNA gene, *MEF2C-AS1* showing both signals of positive selection and a human-specific enhancer with hSNCs (Supplementary Fig. 62). This gene has been associated with ADHD[37] and its target gene *MEF2C* is a very well known target of genetic alterations (many of them also affecting *MEF2C-AS1*) associated with severe intellectual disability[38], cerebral malformation[38] or depression[38,39].

Remarkably, three human-specific intragenic enhancers accumulated more hSNCs than expected (Randomization test: 10,000 simulations, Bonferroni correction, $P < 0.02$ in all cases; Methods and Supplementary Figs. 62 and 63), a number which is also significantly higher than expected (Randomization test: 10,000 simulations, $P = 8 \times 10^{-4}$; Supplementary Fig. 64). Two of these genes are protein-coding genes with known functions in brain cell types and with signals of positive selection. *CLVS1* is a protein-coding gene with brain-specific expression ($\tau_{Brain} = 0.964$) required for the normal morphology of endosomes and lysosomes in neurons[40]. *ROBO1* is a broadly expressed integral membrane protein that participates in axon guidance and neuronal migration ($\tau = 0.388$)[41,42] that has also been associated with human speech and language acquisition since the split from chimpanzees[30]. The third enhancer is included in *AC005906.2*, a long intergenic non-protein-coding gene specifically expressed in brain ($\tau_{Brain} = 1$). Interestingly, this gene overlaps with *KCNA1*, a voltage-gated potassium channel with the same brain-specific expression pattern ($\tau_{Brain} = 0.995$) and for which mutations have been associated with neurological malfunctions[43].

## Discussion

The evolution of human and non-human primates is an area of major interest, but access to direct biological material is often limited by ethical, legal and practical constraints. In this study, we have generated a unique, comprehensive and unified dataset of epigenomic landscapes in LCLs for human and four non-human primate species. Despite the artificial nature of our cellular model[44–46], previous studies have shown the value of LCLs as an experimentally convenient model of somatic cells that accurately resembles the phenotype of its cell type of origin[47] and which can be robustly used for comparative studies in humans and primates[12,48–50]. Moreover, its clonality ensures a cell type-specific experimental system reducing the confounding factors associated with cell population diversity in bulk tissue samples. With this cell model, we could reproduce biological observations about the dynamics of the evolution of regulatory elements previously obtained in more distant species using liver samples[7,9,24]. Moreover, we have expanded these observations to explain how these dynamics are a consequence of the different evolutionary constraints associated with their epigenetic

activities. Therefore, we prove that taking weak and poised activities into account is of major relevance to fully understand the evolution of regulatory regions.

The network of regulatory co-dependencies reveals that the epigenetic activities of each type of regulatory component influence gene expression levels and their evolution differently (Fig. 4). In brief, coordinated epigenetic activities in genic promoters and intragenic enhancers form the core of these architectures and explain gene expression levels. Regulatory activities in proximal and promoter-interacting enhancers are coordinated with promoter components, and activities in enhancer-interacting enhancers are associated with promoter-interacting enhancers. These results show that the influence of regulatory components on gene expression reflect the structure of the regulatory architecture.

The evolution of regulatory complexities, assessed as the number of elements associated with the gene, also reflects how each type of component influences the regulation of gene expression. Acquisition or removal of strong promoter activities in promoter components or strong and poised enhancer activities in intragenic enhancers consistently co-occurs with gene expression changes between primate species. The remaining components show fewer changes linked to expression differences, but they can still be instrumental for gene expression evolution, probably through their influence on promoters and intragenic enhancers. Our conceptual framework provides a starting point for future in-depth investigations on the inter-dependence of different regulatory regions and mechanisms in the evolution of gene regulation. In this sense, we stress the importance of considering promoter and enhancer activity states in the different types of gene components to achieve a more detailed description of the regulatory processes.

We also observed that different epigenetic activities have characteristic evolutionary patterns in primates that are likely the result of their different influence on gene expression. The correlations between epigenetic and sequence conservations are also different for each epigenetic state. Interestingly, similar epigenetic conservation patterns have lower or no correlation with sequence conservation for those orthologous regulatory regions associated with non-coding genes, compared to protein-coding genes. We hypothesize that the association of the former regulatory elements with different gene architectures in other cell types could explain better their sequence conservation in primates.

Despite the larger influence of strong and poised activities on gene regulation, our results in LCLs suggest that major insights can arise from the analysis of the elements with a negligible regulatory role in our cell model. Intragenic enhancers with weak enhancer activities seem to carry information about the degree of regulatory innovation in unrelated cell types. We report recently evolved

intragenic enhancers in the human lineage in genes that show signals of positive selection, patterns of tissue-specific gene expression and brain-related functions, suggesting that these genes may have contributed to human adaptation in several traits. Our findings suggest that the appearance of novel intragenic enhancers with tissue-specific and functionally relevant implications is bound to the co-appearance of weaker activity signals that can be detected in other cell types. These echoes that we detect as human-specific weak enhancer activity seem to provide an unexpected window to the study of regulatory evolution in the human lineage. Further research will be needed to clarify the specific role of these elements in different tissues and cell types.

Taken together, our results show that the evolution of gene regulation is deeply influenced by the coordination of epigenetic activities in gene regulatory architectures. Our insights call for the incorporation of better integrative datasets and refined definitions of regulatory architectures in comparative evolutionary studies to fully understand the interplay between epigenetic regulation and gene expression.

## Methods

### Definition of regulatory elements

We used ChromHMM to jointly learn chromatin states across samples and segment the genome of each sample[51]. ChromHMM implements a multivariate Hidden Markov Model aiming to summarize the combinatorial interactions between multiple chromatin datasets. Bam files from the five histone modifications profiled were binarized into 200 bp density maps. Each bin was discretized in two levels, 0 or 1, depending on their enrichment computed by comparing immunoprecipitated (IP) versus background noise (input) signal within each bin and using a Poisson distribution. Binarization was performed using the BinarizeBam function of the ChromHMM software[51]. A common model across species was learned with the LearnModel ChromHMM function for the concatenated genomes of all samples but O1 (orangutan sample 1) (Supplementary Fig. 75). Several models were trained with a number of chromatin states ranging from 8 to 20. To evaluate the different n-state models, for every sample, the overlap and neighborhood enrichments of each state in a series of functional annotations were explored. A 16-state model was selected for further analysis based on the resolution provided by the defined chromatin states, which capture the most significant interactions between histone marks and the state enrichments in function-annotated datasets (Supplementary Fig. 1). The genomic coordinates of regulatory elements (RE) were defined for each sample by merging all consecutive 200 bp bins excluding elongating (E1 and E2), repressed heterochromatin (E16) and low signal (E15) chromatin states. Species regulatory elements were defined as the union of sample regulatory elements. For orangutan we did not include regulatory elements specific to O1.

## Assignment of a regulatory state to regulatory elements

Regulatory elements were assigned a chromatin-state based annotation. Combining the information gathered through the overlap and neighborhood enrichment analyses in functionally defined regions, we established a hierarchy to designate poised (p), strong (s) and weak (w) promoter and enhancer states. Chromatin states E8, E9 and E11 defined promoter states (P); E8 and E9 were strongly enriched at TSSs, CGI, UMR (unmethylated regions) and open chromatin regions, while E11 was mostly located downstream the TSS; the presence of E14 defined poised promoter states (pP); absence of E14 and presence of E9 or E11 defined strong promoter states (sP); remaining P were classified as weak promoter states (wP). Non-promoter regulatory elements were assigned an enhancer state (E). The presence of E14 defined poised enhancer states (pE); absence of E14 and presence of E3, E4, E5, E6 and E12 defined strong enhancer states (sE): E5 and E6 were strongly enriched LMRs (low methylated regions) whereas E3, E4 and E12 were highly abundant at introns; remaining E were classified as weak enhancer states (wE) (Supplementary Figs. 1 and 65).

One of the limitations of chromatin states is that bin assignments are based on the presence or absence of particular epigenetic marks. However, oftentimes, the lines separating different regulatory elements are blurry: e.g., the distinction between promoter and enhancer states generally resides in the H3K4me3/H3K4me1 balance. Hence, some misclassifications are expected due to insufficient precision of the qualitative classification. Considering the quantitative relationship between co-existing histone modifications can help to accurately annotate epigenetic states in regulatory elements. We used linear discriminant analysis (LDA)[52] to refine chromatin-state based annotations. This method is commonly applied to pattern recognition and category prediction. LDA is a technique developed to transform the features into a lower-dimensional space, which maximizes the ratio of between-class variance to the within-class variance, thereby granting maximum class separation. We performed LDA analysis using the lda function in the R package MASS (version 7.3-47)[53]. The predictor variables were the background-noise normalized IP signals from the five different histone modifications profiled and chromatin accessibility signal at species regulatory elements. The categorical variable to be predicted based on the underlying enrichments was the chromatin-state based annotation. The regulatory state at the species level was determined based on the regulatory state in each of the biological replicates. Thus, the regulatory state of a regulatory element with different epigenetic states in the two replicates, could be aP or aE, when both samples of a given species were annotated as either P or E but differ in their activity; P/E, when a regulatory element was classified as P in one biological replicate and E in the other one; and P/Non-RE or E/Non-RE, when the regulatory elements was so only in one replicate (Supplementary Fig. 6 and Supplementary Table 1).

## Classification of regulatory elements in different types of components of gene regulatory architectures

We first pre-classified each regulatory element into gene regulatory component based on their genomic location with respect to their corresponding species ENSEMBL release 91[17] gene annotations. Regulatory elements found up to 5Kb upstream to the nearest TSS were classified as genic promoters (gP). Additional regulatory elements located up to 10Kb to the nearest TSS were classified as proximal enhancers (prE). Regulatory elements that overlapped a gene were classified as intragenic enhancers (gE). Other regulatory elements that could not be linked to a gene based on their genomic proximity were initially classified as distal enhancers (dE).

Then, we made use of available interaction data for the cell line GM12878 (HiC[19], HiChIP-H3K27ac[20] and ChIA-PET[21]) to map interactions between regulatory elements. Each interacting pair was mapped independently to hg38 coordinates using the liftOver tool from the UCSCTOOLS/331 suite[54], and only interactions for which both pairs could be mapped were kept. Subsequently, interactions were mapped to the non-human primate reference genome assemblies. For inter-species mappings, coordinates were mapped twice, going forward and backward, and only pairs that could be mapped in both directions were kept. Interacting regulatory elements were defined as those that overlapped with each pair of any given interaction. First-order interactions were annotated between promoters and enhancers, allowing the definition of promoter-interacting enhancers (PiE). Second-order interactions were annotated between enhancer components (gE, prE or PiE), allowing the definition of enhancer-interacting enhancers (EiE) (Fig. 2).

## Gene expression levels and regulatory states in gene components

To investigate the influence of the activity state of regulatory elements in each type of component on gene expression levels, we classified 1-to-1 orthologous protein-coding genes, separately for each species, into six mutually excluding categories, one for each regulatory state within each type of component (component-state combinations). Whereas genes can only be associated with one genic promoter and hence, they can only be classified into one category for genic promoters depending on the corresponding epigenetic state of the regulatory element, genes can be associated with more than enhancer component (gE, prE, PiE and EiE). In those cases we classified genes into a given component-state category accordingly to the presence of at least one regulatory element with a given epigenetic state in that component using the following state hierarchy: pE > pP > sE > sP > wE > wP (Supplementary Fig. 14). To statistically assess the influence of each state in each component we used (1) Kruskal-Wallis test (kruskal.test function as implemented in R)[55] to test whether the distributions of the expression levels of genes associated with each component-state combination were different for the different regulatory states, (2) Dwass-Steel-Critchlow-Fligner test to assess the significance of every pairwise comparison (dscfAllPairsTest function from the R package PCMRplus version 1.4.4)[56] and (3) Glass rank biserial correlation coefficient effect size for Mann-Whitney U test to compute the effect sizes

associated with all statistically significant pairwise comparisons (wilcoxonRG function from the R package rcompanion version 2.3.25)[57] (Supplementary Fig. 15). To explore whether expression levels were correlated with the number of regulatory elements with a given state in each enhancer component (gE, prEm PiE and EiE), for each component we calculated the number of elements in each component-state combination and calculated the Spearman rho correlation with the gene expression level (Supplementary Fig. 30). We corrected P-values obtained with the cor.test function implemented in R[55] using the Benjamini-Hochberg procedure[58].

## Partial correlation analysis

To disentangle the network of direct co-dependencies between the different components, regulatory states, histone marks and gene expression, we performed a series of partial correlation analyses[22,59]. To tackle the diversity of architectures detected for the different genes, we added up the calibrated signal of all the regulatory elements with a given regulatory state (promoter or enhancer) in a given type of component for any gene architecture. This decision was based on the observed relationship between the number of strong elements in a gene architecture and the expression level of its target gene. Separation of histone signals in each type of component between those contributing to a promoter or to an enhancer was intended to reflect the potential differences in their role in gene expression regulation. As a result of this design, our system has 51 variables (RNA-seq signal + 5 histone mark signals x 2 regulatory states x 5 components) and 57,370 cases (5,737 genes x 5 species x 2 samples).

All partial correlation analyses were performed using an adaptation of a recently published Sparse Partial Correlation Analysis protocol[22] based on the continuous values of the accumulated ChIP-seq signals (instead of their ranks) to take advantage of their pseudo-quantitative nature. This protocol combines the recovery of statistically significant partial correlations with a cross-validation process to filter out those relationships leading to overfitted reciprocal linear LASSO models (significant partial correlations unlikely to be biologically meaningful). In our case, in every analysis, we recovered those partial correlations recovered in at least four of the five species without leading to overfitting when determining the reciprocal explanatory power in the remaining species. This protocol is intended to detect biologically relevant co-dependences out of the set of significant partial correlations and as a result, this approach filters out many significant partial correlations with very low explanatory power. In fact, all the partial correlations recovered in any of the analyses performed showed very low P-values (Benjamini-Hochberg's correction[58], $P < 1.8 \times 10^{-22}$). In our case, given the relatively small amount of data, we focused on recovering those partial correlations that are likely to be relevant in any species. For these analyses, we used a modified version of the R code provided by the authors (http://spcn.molgen.mpg.de/code/sparse_pcor.R/) to perform 5-fold cross-validation analyses separati by species instead of the original 10-fold cross-validation protocol suitable for larger datasets. Network visualizations were performed with Cytospace[60].

Using this approach, we first performed independent histone analyses to determine the Sparse Partial Correlation Network of each of the histone marks and RNA-seq without considering the possible influence of

the remaining histone marks (Fig. 4a-b, Supplementary Figs. 16-18 and Supplementary Table 11). The similarity of these networks points to a common backbone of inter-component co-dependences reflected in every histone mark. A global partial correlation analysis considering all 51 variables shows a very clear structure of direct co-dependencies with a strong intra-component contribution for the two states of every single component and a clear but more modest exclusive inter-component contribution (Supplementary Fig. 19 and Supplementary Table 12).

In a partial correlation model, direct co-dependencies are established between individual variables. However, the strong intra-component contribution to the network suggests that coordination of the different histone marks within components is important to define the global epigenetic configuration of a component, which itself could be considered the relevant variable for this analysis. To better address this situation in our analysis, we defined a consensus signal for every component following the same approach established by WGCNA[23] to define eigengenes as representative variables of clusters of co-expressed genes. In brief, we defined eigencomponents as the variables summarizing the common signals of the different histone marks in a component (actually calculated as the first PCA component of these five variables). So that eigencomponents keep the meaning of the activities, they were defined as codirectional with H3K27ac signals in each component (eigenvectors negatively correlated with H3K27ac signals were multiplied by -1). We performed a Sparse Partial Correlation Analysis of these 10 eigencomponents and RNA-seq that recovers very clearly the structure of direct co-dependecies between the epigenetic configuration of the different components and gene expression (Fig. 4c).

Finally, we defined the remaining unexplained signal of every histone mark by its eigencomponent as the residuals of a linear model of the original variables and the corresponding eigencomponent. A Sparse Partial Correlation Analysis of these residuals (Supplementary Fig. 20 and Supplementary Table 13) shows that even these residuals reflect the same inter-component structure and highlights that our eigencomponents miss some relevant information for the definition of this regulatory coordination (mainly weaker co-dependencies involving promoter states in intragenic and promoter-interacting enhancers and enhancer states in promoters).

Our dataset of regulatory components shows a quite unbalanced contribution of the components to the architectures, with intragenic enhancers being the most abundant type of component and promoter-interacting and enhancer-interacting enhancers being the least abundant (Supplementary Fig. 11). These differences could be at least partially related to our inability to recover some of the chromatin interaction-mediated regulatory associations. More importantly, this imbalance, if not real, could affect the ability of our partial correlation networks to reflect the contribution of those components less represented in our datasets. To explore this point, we recovered the subset of genes (an average of 1068 genes per sample) with full architectures (those with at least one element in every type of component) and repeated all the Sparse Partial Correlation Analyses explained above with this dataset of genes. In all the cases, we obtained very similar results, recovering fewer relevant partial correlations due to the smaller number of genes, but with no signal of any relevant difference in the global structure of the coordinated network of components and gene expression (Supplementary Figs. 21-28 and Supplementary Tables 11-14).

All the components of the connected network can be very influential in gene expression through their direct or indirect connection with gene expression. However, our Sparse Partial Correlation Networks point consistently to the direct co-dependency of RNA-seq with the genic promoter and intragenic enhancer components and the co-dependency between them. To quantify the explanatory power of these dependencies for gene expression, we performed a simple generalized linear model (glm function as implemented in R[55]) for RNA-seq using H3K27ac, H3K27me3 and H3K36me3 signals in genic promoters and intragenic enhancers and the interactions between them. This model was able to explain 67% of the gene expression variance (Supplementary Table 15), a percentage 5% higher than the 62% explained by a naïve model including the signals of all histone marks in all the components but no interaction between them (Supplementary Table 16), supporting that genic promoters and intragenic enhancers contained nearly all the epigenetic information needed to define gene expression levels in our data.

## Differential gene expression analyses

We identified genes with differential expression levels across species using the iDEGES/edgeR pipeline in the R package TCC (version 1.12.1)[61,62] at an FDR of 0.1 and testing all species pairwise comparisons. Then, we determined the patterns of differential expression, species and direction of the gene expression change, using a two-step approach. For every gene, the Q-values obtained in species pairwise comparisons were ordered from smallest to largest. Different expression labels were then assigned to each species according to the ordered Q-values. Once all species had an assigned label, the average normalized expression values between groups were compared to determine the directionality of the change. We separate differentially expressed genes into two categories: genes with species-specific expression changes and gene with non-species-specific expression changes.

To investigate the relationship between changes in gene expression and changes in the regulatory architecture of a gene, for every type of regulatory component we run a Wilcoxon signed-rank test evaluating whether the number of regulatory elements with a given regulatory state in that particular regulatory component was significantly associated with higher expression levels, for strong and weak activities, or lower expression levels, for poised activities. P-values obtained for each regulatory role were corrected for multiple testing using the Benjamini–Hochberg procedure[58].

## Analysis of evolutionary conservation at orthologous regulatory regions

We studied patterns of evolutionary conservation of promoter and enhancer states using a set of 21,753 one-to-one orthologous regions associated with genes in which at least one species showed a promoter or enhancer epigenetic state. We use the term *repurposed promoters* to refer to those orthologous regulatory regions where one species showed a promoter state and all the rest showed an enhancer state or vice versa. We use the term

693 *novel promoter or enhancer states* to refer to those orthologous regulatory regions where a given species
694 showed a promoter or enhancer state and all other species showed no evidence of regulatory activity (classified
695 as *non-regulatory*).

696 To study the patterns of evolutionary conservation of regulatory states, we focused on the subset of 10,641
697 one-to-one orthologous regions in which at least one species showed a strong, poised or weak regulatory state
698 (we do not include orthologous regions including elements with different activities between biological
699 replicates). To statistically assess the different evolutionary dynamics observed for the different regulatory
700 states we first run randomization analysis. We randomized (1,000 randomizations) the regulatory states
701 associated with each species in orthologous regulatory regions. We determined the P-value as the number of
702 randomizations with an average conservation equal to or above the observed conservation for each regulatory
703 state. We further explored the different patterns of conservation combining: (1) Kruskal-Wallis test
704 (kruskal.test R function)[55] to test whether the global distributions of the number of species in which each
705 particular state was conserved were different for the different regulatory states and (2) Dwass-Steel-Critchlow-
706 Fligner test to assess the significance of every pairwise comparison (dscfAllPairsTest function from the R
707 package PMCMRplus version 1.4.4)[56] and (3) Glass rank biserial correlation coefficient for Mann-Whitney U
708 test to compute the effect sizes associated with all statistically significant pairwise comparisons (wilcoxonRG
709 function from the R package rcompanion version 2.3.25)[57].

710 To study the patterns of evolutionary conservation of the sequence underlying orthologous regulatory regions,
711 we first assigned each orthologous regulatory region a conservation score. We computed this score based on
712 the phastCons30way sequence conservation track[63]. To control for background sequence conservation levels,
713 we first computed the average and standard deviations phastCons30way in TADs defined in the cell line
714 GM12878[64] (Supplementary Fig. 49). Then, we used these summary statistics to calculate the z-score for each
715 bp in every orthologous regulatory region, using the average and standard deviations values of the TAD in
716 which each orthologous regulatory region was found. We averaged the z-scores within each orthologous
717 regulatory regions in bins of 200 bp that overlap 50 bp with the next bin and assign each orthologous regulatory
718 region the maximum z-score values associated with its bins. We computed the Spearman rho correlation
719 between the z-scores and the number of species in which each orthologous regulatory region was conserved,
720 separately for each regulatory state. To determine the statistical significance of these correlations we used
721 randomization analysis. For each regulatory state we created 1,000 sets randomizing the z-score associated
722 with each orthologous regulatory region and calculated the Spearman correlation in each randomization. We
723 determined the P-value as the number of randomizations with a Spearman rho correlation value equal to or
724 above the observed correlation (Supplementary Figs. 52-54).

725 We used a Chi-square test to identify the component-state combinations enriched in fully conserved
726 orthologous regulatory regions (Supplementary Fig. 48). To explore the expression patterns of genes regulated
727 by evolutionarily conserved component-state combinations, for all positively enriched component-state
728 combinations, we recovered the associated orthologous protein-coding genes and computed their average

729 expression across species. We excluded those genes associated with more than one component-state
730 combination.

731

732 **Analysis of tissue-specific gene expression patterns**

733 We defined sets of human genes associated with fully conserved component-state combinations, and human
734 genes associated with human-specific gains/losses of regulatory elements. Note that these gene lists are not
735 mutually exclusive since a gene can we associated with different types of conserved or species-specific
736 component-state combinations (e.g., a gene with both a human-specific intragenic enhancer with weak activity
737 and a fully conserved intragenic enhancer with a strong activity). We obtained expression levels (median TPM
738 values) across a collection of different tissues from the latest GTEx release (v8)[65]. We only included tissues
739 with at least 70 samples and grouped tissue subregions into the same tissue category, as stated in
740 Supplementary Table 18. For each component-state combination we followed a two-step approach to remove
741 consistently low-expressed genes across tissues. For that we first assigned a value of 0 to all genes with a
742 median expression level below 0.1 TPM and then we excluded from the analyses those genes that had an
743 accumulated expression value in all tissues below 0.1xNumber of tissues (n = 29 tissues). For each component-
744 state combination, differences in median expression across tissues were assessed with the Friedman test using
745 the friedman.test function as implemented in R[55]. We used the Wilcoxon-Nemenyi-McDonald-Thompson test
746 implemented in the pWNMT function of the R package NSM3 (version 1.14)[66] to assess whether expression
747 levels were significantly different for all pairwise tissue combinations. Then, we made use of the rank-biserial
748 correlation to calculate the effect sizes for all statistically significant pairwise tests with the wilcoxonPairedRC
749 function of the R package rcompanion (version 2.3.25)[57].

750 We then evaluated the tissue-specificity of the genes associated with the different component-state
751 combinations. For this we calculated the tissue specificity index[67] (τ, tau) for each gene, which is defined as:

752 $$\tau = \frac{\sum_{i=1}^{N} 1-x_i}{N-1}$$

753 where $N$ is the number of tissues and $x_i$ is the expression value normalised by the maximum expression value.
754 This value ranges from 0, for housekeeping genes, to 1, for tissue-specific genes (values above 0.8 are used to
755 identify tissue-specific genes)[68]. Tissue-specificity indices were calculated for all genes included in the latest
756 GTEx release[65]. Gene expression levels (median TMP) of grouped tissue categories (Supplementary Table 18)
757 were normalised within and across tissues before calculating τ as implemented in the R package tispec (version
758 0.99.0)[69]. The calcTau function from this package provides a tau value for each gene and also a tau expression
759 fraction for each tissue (which also ranges from 0 to 1) that indicates the specificity of a given gene for that
760 tissue.

761 After calculating τ values, we compared their distributions between gene datasets with the Kruskall-Wallis test
762 and assessed the significance of every pairwise comparison with the Dwass-Steel-Critchlow-Fligner test

(dscfAllPairsTest function from the R package PMCMRplus version 1.4.4)[56]. Glass rank biserial correlation coefficient was used to compute the effect sizes associated with all statistically significant pairwise comparisons using the wilcoxonRG function from the R package rcompanion version 2.3.25[57] ($P < 0.05$, Supplementary Fig. 60).

## Over-representation analyses (ORA) of functional annotations

To ensure the representativeness of the functional enrichments, we removed from the gene sets defined in the tissue-specificity gene expression analyses those genes associated with conserved components with different epigenetic states activities (strong, poised or weak promoter or enhancer) or conservation levels (fully conserved and species-specific) and kept those gene lists with a minimum of 15 genes for enrichment analyses.

Over-representation of Gene Ontology (GO) terms was performed using the WebGestaltR function from the R package WebGestaltR (version 0.4.3)[70] with minNum = 25 and remaining default options. This function controls the false discovery rate (FDR) by applying Benjamini-Hochberg procedure (default threshold FDR = 0.05)[58,71]. Previous analyses have shown that recent enhancers tend to occur in the same genes that already have highly conserved enhancers[9]. To avoid biases due to the presence of a gene in different gene sets, we filtered out those genes associated with both conserved and species-specific component-state combinations. To control for the particular background of each component, we built different background gene sets including the set of human genes associated with at least one-to-one orthologous regulatory regions of each type of component, hence we have specific and different backgrounds for genic promoters, intragenic enhancers and promoter-interacting enhancers. To represent and compare enriched GO terms between component-state combinations, we performed a clustering of all significantly enriched GO terms using REVIGO[72]. We associated each GO term with the proportion of genes from each component-state combination that overlapped that GO term. In the case of GO terms enriched in more than one gene set, we chose the highest proportion of genes. We used this list as input for REVIGO. Given that REVIGO only reports the clustering of approximately 350 GO terms and our input list was larger than that, we used the R package GofuncR (version 1.8.0)[73] to retrieve the parent GO terms of the remaining unassigned GO terms and add them to the corresponding group as defined by REVIGO. REVIGO group names were manually assigned, taking into account the most representative parent term (Supplementary Table 19).

## Association of genes containing intragenic enhancers with signals of positive selection in humans

First, we recovered a dataset of human genomic regions with previously detected signals of positive selection in humans[25–27] and selective sweeps in modern compared to archaic humans[28]. BEDtools[74] was used to assign these regions to both protein-coding and non-coding genes following similar criteria to those used for building

the gene regulatory architectures (Methods' section *Classification of regulatory elements in different types of components of gene regulatory architectures*). We assigned these regions to a protein-coding gene if they were located within the gene or up to 5 Kb upstream of its TSS. Then, we made use of available interaction data for the cell line GM12878 (HiC[19], HiChIP-H3K27ac[20] and ChIA-PET[21]) to assign positively selected regions to their interacting protein-coding genes. We defined the 2,004 genes associated with at least one positively selected region as the set of genes with signals of positive selection in the human lineage. We computed the overlaps between this gene list and the lists of genes associated with the different component-state combinations. We used one or two-tailed Fisher's exact test to assess the enrichment significance.

## Analyses of the density of human-fixed single nucleotide changes (hSNCs) in intragenic enhancers with weak enhancer states

In order to study the distribution of human-fixed changes in a specific type of regulatory element, we first built a dataset with human-specific changes. We used sequencing data from a diversity panel of 27 orangutans, 42 gorillas, 11 bonobos and 61 chimpanzees[75–77], as well as 19 modern humans from the 1000 genomes project[78], all mapped to the human reference assembly hg19. We applied a basic filtering for each site in each individual (sequencing coverage >3 and <100), and kept sites where at least half of the individuals in a given species had sufficient data. Furthermore, at least 90% of the kept individuals at a given site in a given species had to share the same allele, otherwise the site was labeled as polymorphic in the population. Indels and triallelic sites were removed, and only biallelic sites were kept. We used data from a macaque diversity panel[79], applying the same filters described above. The allele at monomorphic sites was added using bedtools getfasta[74] from the macaque reference genome rheMac8. Since this panel uses the macaque reference genome, we performed a liftover to hg19 using the R package rtracklayer[80] and merged the data with the great ape diversity panel.

Lineage-specific changes were retrieved as polymorphisms with sufficient information. Hence, human-specific changes (hSNCs) were defined as positions where each species carry only or mostly one allele within their respective population, the majority of individuals in each population have a genotype call at sufficient coverage, and the human allele differs from the allele in the other populations.

BEDtools[74] was used to annotate those hSNCs in conserved or human-specific weak intragenic enhancers and the density of changes was calculated as the number of hSNCs present in each enhancer divided by the length of the enhancer.

To determine which human-specific intragenic weak enhancers were enriched in human-specific changes, we compared their density to what would be expected at random. For that, we first established the number of hSNCs that fall in human intragenic enhancers with weak enhancer states associated with 1-to-1 orthologous regulatory regions (our universe of enhancers). In each simulation, this number of mutations was randomly placed in this universe and we computed the density for each of the human-specific weak intragenic enhancers (10,000 simulations). With this approach, we corrected for the differences in the length of the enhancers. The

P-value for each enhancer was computed as the number of simulations with a density equal to or above the observed density for that particular enhancer. All P-values were corrected by multiple testing using the Bonferroni method with the number of tests equal to the number of human-specific weak intragenic enhancers.

We then assessed whether the number of enhancers that were statistically enriched in hSNCs (or number of *hits*) was greater than what would be expected at random. In order to do that, for each enhancer we defined its mutation density critical value adjusting by multiple testing and using the simulated values. For example, in a hypothetical case of 100 enhancers and 10,000 simulations, for each enhancer we would order its simulated density of hSNCs from smallest to largest and take the 5th value as the critical one (given that our chosen alpha equals 5%, but it has to be corrected by 100 tests; therefore it becomes 0.05%). Once we established a critical value for each human-specific intragenic weak enhancer, we determined, for each simulation, how many enhancers had a density equal to or above their corresponding critical value. Finally, we computed the P-value comparing the the number of artificial *hits* in each simulation with the number of observed hits.

**Acknowledgments:** R.G.-P. was supported by a fellowship from MICINN (FPU13/01823). P.E.-C. was supported by a Formació de Personal Investigador fellowship from Generalitat de Catalunya (FI_B00122). M.K. was supported by a Deutsche Forschungsgemeinschaft (DFG) fellowship (KU 3467/1-1) and the Postdoctoral Junior Leader Fellowship Programme from “la Caixa” Banking Foundation (LCF/BQ/PR19/11700002). D.J. was supported by a Juan de la Cierva fellowship (FJCI2016-29558) from MICINN. T.M-B. is supported by funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 864203), BFU2017-86471-P (MINECO/FEDER, UE), “Unidad de Excelencia María de Maeztu”, funded by the AEI (CEX2018-000792-M), Howard Hughes International Early Career, Obra Social "La Caixa" and Secretaria d’Universitats i Recerca and CERCA Programme del Departament d’Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 880). G.M., V.D.C. and L.D.C. were supported by grants from the Spanish of Economy, Industry and Competitiveness (MEIC) (BFU2016-75008-P) and G.M. was also supported by the “Convocatoria de Ayudas Fundación BBVA a Investigadores, Innovadores y Creadores Culturales”. J.L.G.-S. was supported by the Spanish government (grants BFU2016-74961-P), an institutional grant Unidad de Excelencia María de Maeztu (MDM-2016-0687) and the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 740041). A.N. was supported by Fondo Europeo de Desarrollo Regional (FEDER) with project grants BFU2016-77961-P and PGC2018- 101927-B-I00 and by the Spanish National Institute of Bioinformatics (PT17/0009/0020).

865 **Author contributions**: T.M.-B. and J.L.G.-S. conceived the study; D.J. designed and supervised the
866 analyses; L.D.C. supervised the work of G.M. and V.D.C.; A.B. procured non-human great ape cell
867 lines; A.N. provided helpful insights; R.G.-P., G.M. and V.D.C. performed the experimental work;
868 R.G.-P., P.E.-C., D.J., I.L., M.R. and M.K analyzed the data; D.J., R.G.-P., and P.E.-C. wrote the
869 manuscript with input from all co-authors.

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889 **Figures**:

890 **Figure. 1 Overview of the study design and data generated. a,** One human and eight non-human
891 primate lymphoblastoid cell lines were cultured to perform a variety of high-throughput techniques
892 including whole genome sequencing (WGS), whole genome bisulfite sequencing (WGBS),
893 chromatin-accessibility sequencing (ATAC-seq), chromatin immunoprecipitation sequencing (ChIP-
894 seq) targeting five different histone modifications (H3K27me3, H3K4me1, H3K27ac, H3K4me3 and
895 H3K36me3) and transcriptome sequencing (RNA-seq). We integrated previously published datasets
896 from an extensively profiled human lymphoblastoid cell line (GM12878) to balance the number of
897 human samples (Supplementary Methods). **b,** Number of sequencing reads generated per sample and
898 experiment. Striped lines indicate data retrieved from previously published experiments[81,82].
899
900 **Figure 2. Schematic illustration of the approach followed to annotate and classify regulatory**
901 **elements. a,** DNA strand represents a gene annotation track, wherein dark grey regions correspond
902 to coding annotated regions. The second row represents the binarized output from ChromHMM[51],
903 wherein each box corresponds to a 200 bp bin. Light grey indicates bins without evidence of promoter
904 or enhancer states, whereas the different colors represent the different learned chromatin states
905 (Methods, Supplementary Fig. 1). Shorter DNA strands represent the genomic coordinates defined
906 for regulatory elements that result from merging adjacent 200 bp bins with epigenetic signals
907 associated with promoter or enhancer states. We defined species regulatory elements from the union
908 of the regulatory elements detected in each biological replicate (Methods). **b,** We established a
909 hierarchy between chromatin states based on the combination of chromatin marks found within each
910 regulatory region and classified regulatory elements into epigenetic promoter (P) and enhancer (E)
911 states with three different activity levels: strong (s), weak (w) or poised (p) (Methods). Then, we
912 applied a linear discriminative analysis (LDA) (Methods) using normalized histone and open
913 chromatin enrichments to refine this epigenetic classification (Supplementary Methods). **c,** We linked
914 regulatory elements to genes and assigned them to a type of regulatory component. We first classified
915 regulatory elements into genic promoters (gP), genic enhancers (gE) and proximal enhancers (prE)
916 based on their linear proximity to annotated genes. Then, we used previously published 3D chromatin
917 maps in GM12878 cells[19–21] to recover physical interactions between regulatory elements adding
918 promoter-interacting enhancers (PiE) and enhancer-interacting enhancers (EiE), to the list of gene
919 regulatory components. This approach allowed us to link distal regulatory elements, that otherwise
920 would have remained orphan, to their target genes.
921
922 **Figure 3. Epigenetic and regulatory characterization of regulatory elements annotated in**
923 **primates.** Number of regulatory elements with **a,** promoter and **b,** enhancer epigenetic states in each

924 species. **c,** Number of regulatory elements associated with genes in each species. Dark, medium and
925 light shades correspond to 1-to-1 orthologous protein-coding, protein-coding and non-protein-coding
926 genes, respectively. **d,** Number of orphan regulatory elements (not associated with any gene, Fig. 2c)
927 in each species. **e,** Average number of regulatory elements across species associated with 1-to-1
928 orthologous protein-coding genes classified as gP, gE, prE, PiE and EiE. **f,** Average number of
929 orthologous protein-coding genes associated with each type of regulatory element. In **e,** and **f,** error
930 bars show the standard deviation across species and differently shaped points show the values for
931 each species as indicated in the legend. **g,** Proportion of regulatory elements with a given epigenetic
932 state associated with 1-to-1 orthologous protein-coding genes for each type of regulatory component.
933 Dots and error bars show the average proportion and standard deviation across species, respectively.
934
935 **Figure 4. Epigenetic signals in gene regulatory architectures explain gene expression levels.**
936 Sparse Partial Correlation Networks showing the statistical co-dependence of the RNA-seq (Gene
937 expression) and the ChIP-seq signals for the histone marks in the different components (segregated
938 by their promoter or enhancer epigenetic state) of the gene regulatory architectures. ChIP-seq
939 enrichment values of different elements classified as the same component in the same gene were
940 aggregated. Sparse Partial Correlation Network for **a,** H3K27ac (minimal partial correlation = -0.4;
941 maximal partial correlation = 0.49; all partial correlations Benjamini-Hochberg's $P < 1.4 \times 10^{-109}$), **b,**
942 H3K27me3 (minimal partial correlation = -0.4; maximal partial correlation = 0.2; all partial
943 correlations Benjamini-Hochberg's $P < 3.9 \times 10^{-57}$) and **c,** the eigencomponents representing the
944 consensus ChIP-seq signals of the five histone marks in every component (minimal partial correlation
945 = -0.41; maximal partial correlation = 0.33; all partial correlations Benjamini-Hochberg's $P < 4.1 \times$
946 $10^{-303}$). Blue edges represent positive partial correlations and red edges negative ones. Edge widths
947 are proportional to absolute partial correlation values within each network. All the networks are based
948 on the 5,737 1-to-1 orthologous protein-coding genes associated with at least one regulatory element
949 in all species. Only nodes for values with significant and relevant partial correlations (Methods) are
950 represented.
951
952 **Figure 5. Different regulatory activities have different patterns of epigenetic and sequence**
953 **conservation. a,** Barplots show the average number of orthologous regulatory regions across species
954 with the corresponding color-coded epigenetic state conserved in 1, 2, 3, 4 or 5 species. Error bars
955 show the standard deviation across species and differently shaped dots show the number of regulatory
956 regions with this conservation for each species, as in Fig, 3e-f. **b,** Distribution of the sequence
957 conservation scores (calculated as z-scores of the distribution of phastCons30way[63] values for non-

coding regions in the same Topologically Associated Domain[64]; Methods) of human orthologous regulatory regions with different epigenetic states conserved in 1, 2, 3, 4 or 5 of our primate species.

**Figure 6. Intragenic enhancers with weak activities echo brain-specific regulation and co-localize with signals of recent human selection. a,** Median expression levels of genes associated with intragenic enhancers with strong conserved (763 genes), weak conserved (528 genes) and weak human-specific (105 genes) enhancer activities in their three least and three most expressed tissues from GTEx data. **b,** Functional enrichment of conserved and human-specific activities (strong promoter -sP- and strong and weak enhancer activities -sE and wE, respectively-) in specific regulatory components (genic promoters -gP- and intragenic enhancers -gE-). Circles denote significant enrichment of conserved elements, whereas diamonds refer to human-specific elements (Fisher's exact test: Benjamini-Hochberg's correction, FDR < 0.05). The size of the circles/diamonds shows the proportion of genes included in that functional category out of the total number of genes contained in the corresponding regulatory group. The number of genes with conserved elements is 1372, 730 and 445 genes for sP, sE and wE, respectively, and 78 genes that include human-specific gains of wE. Terms associated with molecular functions in Supplementary Fig. 58. **c,** Hierarchical strategy to assess the specific enrichment of signals of positive selection in human-specific intragenic enhancers (gE) with a weak enhancer state (wE). We first tested the enrichment in the set of genes with intragenic enhancers with weak activities (1740 genes) compared to the genes containing intragenic enhancers from any other activity (3608 genes) (One-tailed Fisher's exact test: $P = 1.75 \times 10^{-14}$, $OR = 2.03$). To test whether this activity-associated enrichment is specific of the conservation level of the enhancer, we considered the enrichment in intragenic enhancers with human-specific (134 genes) or with conserved weak enhancer activities (600 genes) compared to the complementary sets of genes with intragenic weak enhancers (One-tailed Fisher's exact test: $P = 8.24 \times 10^{-7}$, $OR = 2.61$ and $P = 0.38$ $OR = 1.05$, respectively) and the enrichment in genes with human-specific enhancers compared to genes with conserved enhancers (One-tailed Fisher's exact test: $P = 0.0013$, $OR = 2.1$). H, C, G, O, M are used to refer to each species and non-RE is used to define a non-regulatory element. **d,** Top: Schematic representation of a human-specific intragenic weak enhancer with a hSNC (nucleotide change in humans shown in red) contained in a gene with signals of selection (orange peaks). Bottom: Venn diagram illustrating the overlap between the 41 genes containing human-specific weak intragenic enhancers with signals of selection (orange) and the 30 genes with these enhancers and with human single nucleotide changes (hSNCs) fixed in humans and distinct from other non-human primates (red).

## References

1. Britten, R. J. & Davidson, E. H. Gene Regulation for Higher Cells: A Theory. *Science* vol. 165 349–357 (1969).
2. Britten, R. J. & Davidson, E. H. Repetitive and Non-Repetitive DNA Sequences and a Speculation on the Origins of Evolutionary Novelty. *The Quarterly Review of Biology* vol. 46 111–138 (1971).
3. Zhu, Y. *et al.* Spatiotemporal transcriptomic divergence across human and macaque brain development. *Science* **362**, (2018).
4. Cardoso-Moreira, M. *et al.* Gene expression across mammalian organ development. *Nature* **571**, 505–509 (2019).
5. Xu, C. *et al.* Human-specific features of spatial gene expression and regulation in eight brain regions. *Genome Research* vol. 28 1097–1110 (2018).
6. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* vol. 478 343–348 (2011).
7. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).
8. Vermunt, M. W. *et al.* Epigenomic annotation of gene regulatory alterations during evolution of the primate brain. *Nat. Neurosci.* **19**, 494–503 (2016).
9. Berthelot, C., Villar, D., Horvath, J. E., Odom, D. T. & Flicek, P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat Ecol Evol* **2**, 152–163 (2018).
10. Reilly, S. K. *et al.* Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**, 1155–1159 (2015).
11. Prescott, S. L. *et al.* Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. *Cell* **163**, 68–83 (2015).
12. Zhou, X. *et al.* Epigenetic modifications are associated with inter-species gene expression variation in primates. *Genome Biol.* **15**, 547 (2014).
13. Zeng, J. *et al.* Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. *Am. J. Hum. Genet.* **91**, 455–465 (2012).
14. Hernando-Herraez, I. *et al.* The interplay between DNA methylation and sequence divergence in recent human evolution. *Nucleic Acids Res.* **43**, 8204–8214 (2015).

15. Hernando-Herraez, I. *et al.* Dynamics of DNA Methylation in Recent Human and Great Ape Evolution. *PLoS Genetics* vol. 9 e1003763 (2013).

16. Lowdon, R. F., Jang, H. S. & Wang, T. Evolution of Epigenetic Regulation in Vertebrate Genomes. *Trends in Genetics* vol. 32 269–283 (2016).

17. Aken, B. L. *et al.* Ensembl 2017. *Nucleic Acids Res.* **45**, D635–D642 (2017).

18. Ong, C.-T. & Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.* **12**, 283–293 (2011).

19. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* vol. 159 1665–1680 (2014).

20. Mumbach, M. R. *et al.* Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nature Genetics* vol. 49 1602–1612 (2017).

21. Tang, Z. *et al.* CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* vol. 163 1611–1627 (2015).

22. Lasserre, J., Chung, H.-R. & Vingron, M. Finding Associations among Histone Modifications Using Sparse Partial Correlation Networks. *PLoS Comput. Biol.* **9**, e1003168 (2013).

23. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 1–13 (2008).

24. Carelli, F. N., Liechti, A., Halbert, J., Warnefors, M. & Kaessmann, H. Repurposing of promoters and enhancers during mammalian evolution. *Nat. Commun.* **9**, 4066 (2018).

25. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* vol. 478 476–482 (2011).

26. Prabhakar, S., Noonan, J. P., Pääbo, S. & Rubin, E. M. Accelerated evolution of conserved noncoding sequences in humans. *Science* **314**, 786 (2006).

27. Gittelman, R. M. *et al.* Comprehensive identification and analysis of human accelerated regulatory DNA. *Genome Research* vol. 25 1245–1255 (2015).

28. Peyrégne, S., Boyle, M. J., Dannemann, M. & Prüfer, K. Detecting ancient positive selection in humans using extended lineage sorting. *Genome Research* vol. 27 1563–1572 (2017).

29. Enard, W. *et al.* Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* vol. 418 869–872 (2002).

30. Mozzi, A. *et al.* The evolutionary history of genes involved in spoken and written language: beyond FOXP2. *Sci. Rep.* **6**, 22157 (2016).

31. Kalebic, N. *et al.* Neocortical Expansion Due to Increased Proliferation of Basal Progenitors Is Linked to Changes in Their Morphology. *Cell Stem Cell* vol. 24 535–550.e9 (2019).

32. Kuhlwilm, M. & Boeckx, C. A catalog of single nucleotide changes distinguishing modern humans from archaic hominins. *Scientific Reports* vol. 9 (2019).

33. Finn, S. & Civetta, A. Sexual Selection and the Molecular Evolution of ADAM Proteins. *Journal of Molecular Evolution* vol. 71 231–240 (2010).

34. Riordan, J. R. Identification of the cystic fibrosis gene: Cloning and characterization of complementary DNA. *Trends in Genetics* vol. 5 363 (1989).

35. Poolman, E. M. & Galvani, A. P. Evaluating candidate agents of selective pressure for cystic fibrosis. *Journal of The Royal Society Interface* vol. 4 91–98 (2007).

36. Racimo, F. *et al.* Archaic adaptive introgression in TBX15/WARS2. *Molecular Biology and Evolution* msw283 (2016) doi:10.1093/molbev/msw283.

37. Demontis, D. *et al.* Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Genet.* **51**, 63–75 (2019).

38. Meur, N. L. *et al.* MEF2C haploinsufficiency caused by either microdeletion of the 5q14.3 region or mutation is responsible for severe mental retardation with stereotypic movements, epilepsy and/or cerebral malformations. *Journal of Medical Genetics* vol. 47 22–29 (2010).

39. Hyde, C. L. *et al.* Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nature Genetics* vol. 48 1031–1036 (2016).

40. Katoh, Y. *et al.* The clavesin family, neuron-specific lipid- and clathrin-binding Sec14 proteins regulating lysosomal morphology. *J. Biol. Chem.* **284**, 27646–27654 (2009).

41. Long, H. *et al.* Conserved roles for Slit and Robo proteins in midline commissural axon guidance. *Neuron* **42**, 213–223 (2004).

42. Andrews, W. *et al.* Robo1 regulates the development of major axon tracts and interneuron migration in the forebrain. *Development* **133**, 2243–2252 (2006).

43. Yin, X.-M. *et al.* Familial paroxysmal kinesigenic dyskinesia is associated with mutations in the KCNA1 gene. *Hum. Mol. Genet.* **27**, 757–758 (2018).

1080 44. Carter, K. L., Cahir-McFarland, E. & Kieff, E. Epstein-Barr Virus-Induced Changes in B-Lymphocyte
1081 Gene Expression. *Journal of Virology* vol. 76 10427–10436 (2002).
1082 45. Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat.*
1083 *Genet.* **43**, 768–775 (2011).
1084 46. Sugawara, H. *et al.* Comprehensive DNA methylation analysis of human peripheral blood leukocytes
1085 and lymphoblastoid cell lines. *Epigenetics* **6**, 508–515 (2011).
1086 47. Hussain, T. & Mulherkar, R. Lymphoblastoid Cell lines: a Continuous in Vitro Source of Cells to Study
1087 Carcinogen Sensitivity and DNA Repair. *Int J Mol Cell Med* **1**, 75–87 (2012).
1088 48. Khaitovich, P., Enard, W., Lachmann, M. & Pääbo, S. Evolution of primate gene expression. *Nature*
1089 *Reviews Genetics* vol. 7 693–702 (2006).
1090 49. Pai, A. A., Bell, J. T., Marioni, J. C., Pritchard, J. K. & Gilad, Y. A genome-wide study of DNA
1091 methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS*
1092 *Genet.* **7**, e1001316 (2011).
1093 50. Shibata, Y. *et al.* Extensive evolutionary changes in regulatory element activity during human origins
1094 are associated with altered gene expression and positive selection. *PLoS Genet.* **8**, e1002789 (2012).
1095 51. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat.*
1096 *Protoc.* **12**, 2478–2492 (2017).
1097 52. Tharwat, A., Gaber, T., Ibrahim, A. & Hassanien, A. E. Linear discriminant analysis: A detailed
1098 tutorial. *AI Communications* vol. 30 169–190 (2017).
1099 53. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S*. (Springer, New York, 2002).
1100 54. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Briefings in*
1101 *Bioinformatics* vol. 14 144–161 (2013).
1102 55. R Core Team. R: A language and environment for statistical computing.
1103 56. Pohlert, T. PMCMRplus: Calculate Pairwise Multiple Comparisons of Mean Rank Sums Extended.
1104 (2020).
1105 57. Mangiafico, S. rcompanion: Functions to Support Extension Education Program Evaluation. (2020).
1106 58. Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. & Golani, I. Controlling the false discovery rate in
1107 behavior genetics research. *Behavioural brain research* vol. 125 279–284 (2001).
1108 59. Juan, D. *et al.* Epigenomic Co-localization and Co-evolution Reveal a Key Role for 5hmC as a

Communication Hub in the Chromatin Network of ESCs. *Cell Rep.* **14**, 1246–1257 (2016).

60. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
61. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* vol. 19 185–193 (2003).
62. Sun, J., Nishiyama, T., Shimizu, K. & Kadota, K. TCC: an R package for comparing tag count data with robust normalization strategies. *BMC Bioinformatics* vol. 14 219 (2013).
63. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
64. Wang, Y. *et al.* The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biology* vol. 19 (2018).
65. Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. doi:10.1101/787903.
66. Grant Schneider, E. C. A. R. B. NSM3: Functions and Datasets to Accompany. (2020).
67. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
68. Kryuchkova, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. doi:10.1101/027755.
69. Condon, K. tispec: Calculates tissue specificity from RNA-seq data. (2020).
70. Wang, J. & Liao, Y. *WebGestaltR: Gene Set Analysis Toolkit WebGestaltR*. (2020).
71. Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **47**, W199–W205 (2019).
72. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS One* **6**, e21800 (2011).
73. Grote, S. GOfuncR: Gene ontology enrichment using FUNC. (2020).
74. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
75. Prado-Martinez, J. *et al.* Great ape genetic diversity and population history. *Nature* **499**, 471–475

(2013).

76. de Manuel, M. *et al.* Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science* **354**, 477–481 (2016).

77. Nater, A. *et al.* Morphometric, Behavioral, and Genomic Evidence for a New Orangutan Species. *Curr. Biol.* **27**, 3576–3577 (2017).

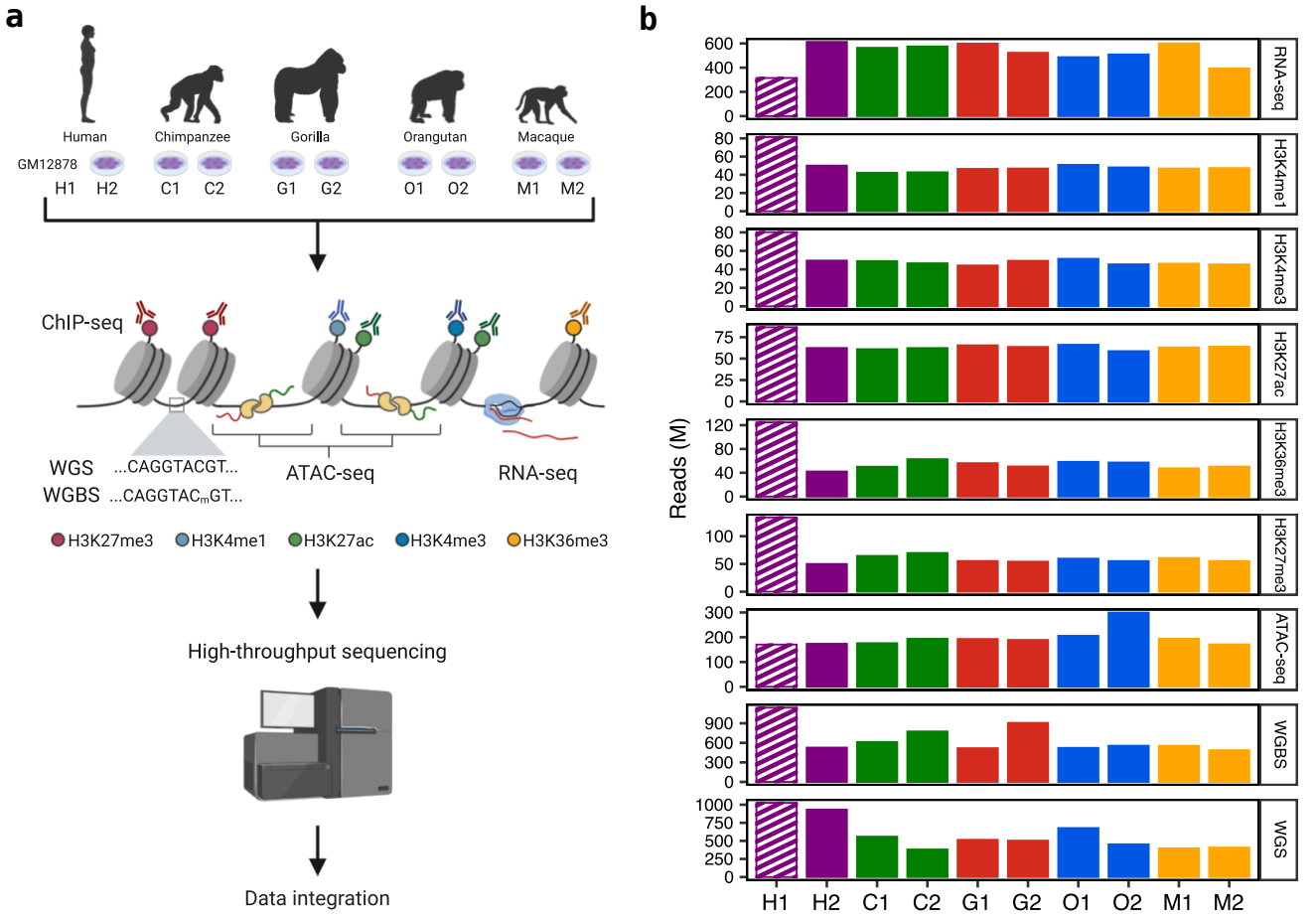
78. Consortium, T. 1000 G. P. & The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* vol. 526 68–74 (2015).

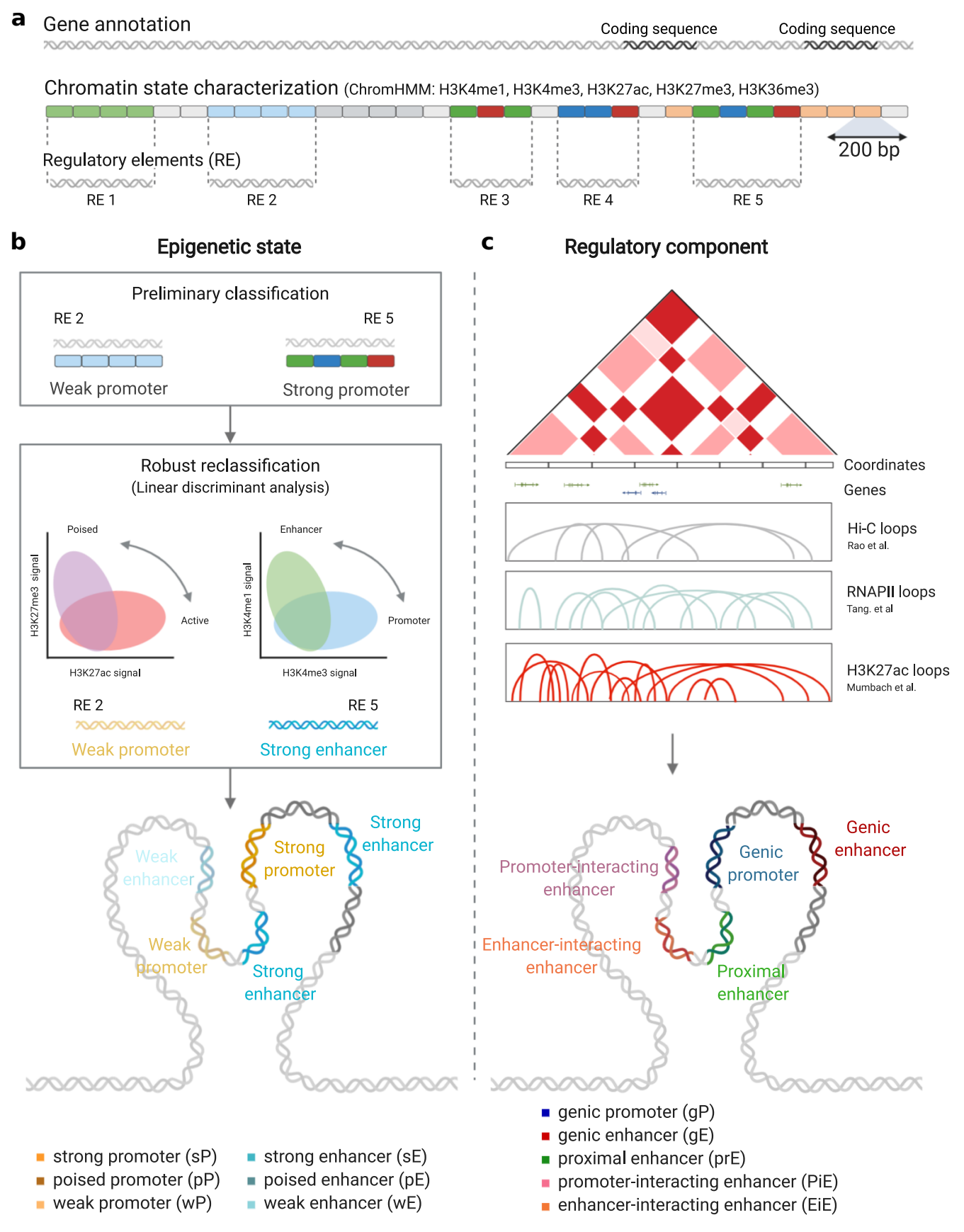
79. Xue, C. *et al.* The population genomics of rhesus macaques (Macaca mulatta) based on whole-genome sequences. *Genome Res.* **26**, 1651–1662 (2016).

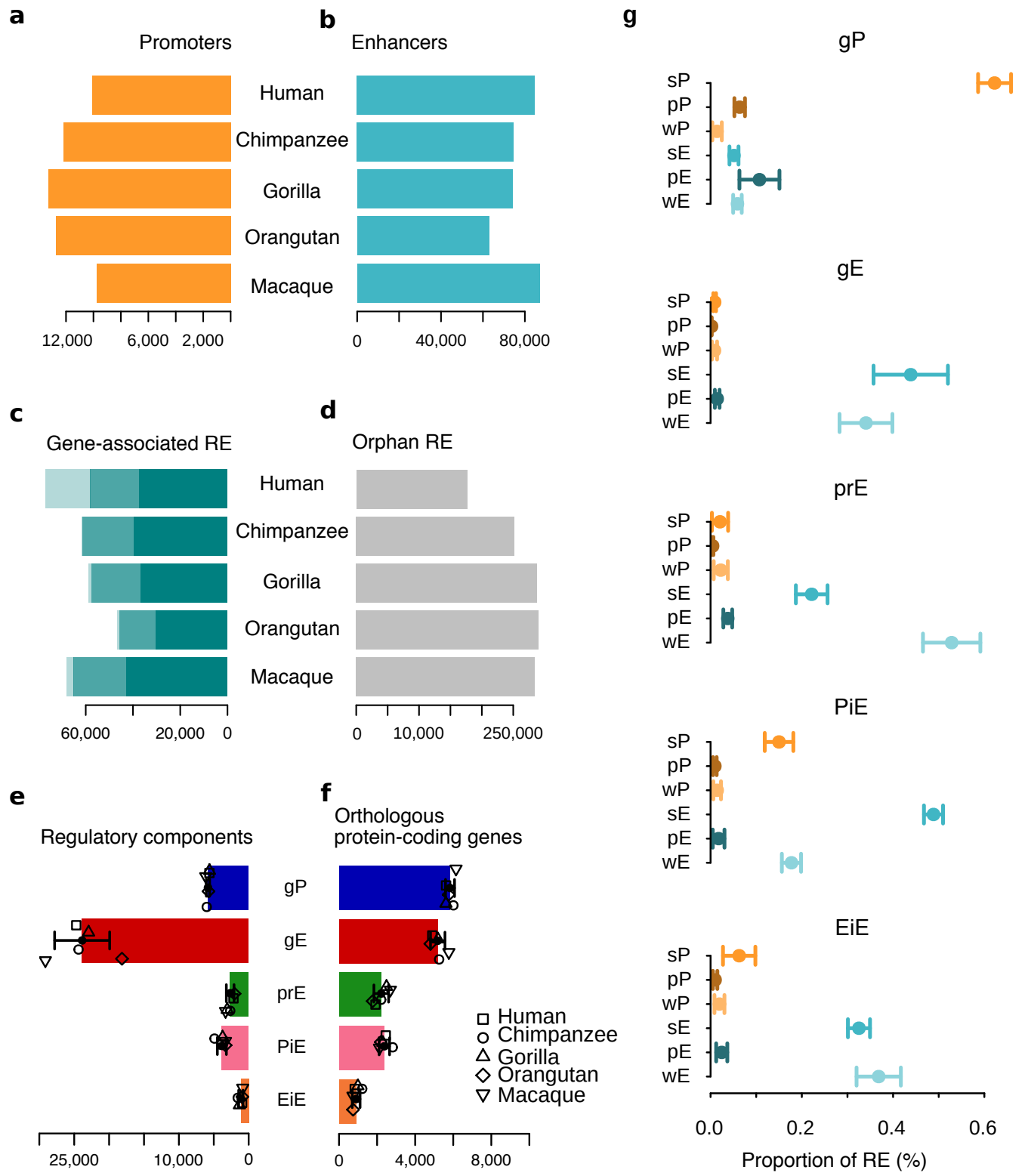
80. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome ebrowsers. *Bioinformatics* vol. 25 1841–1842 (2009).

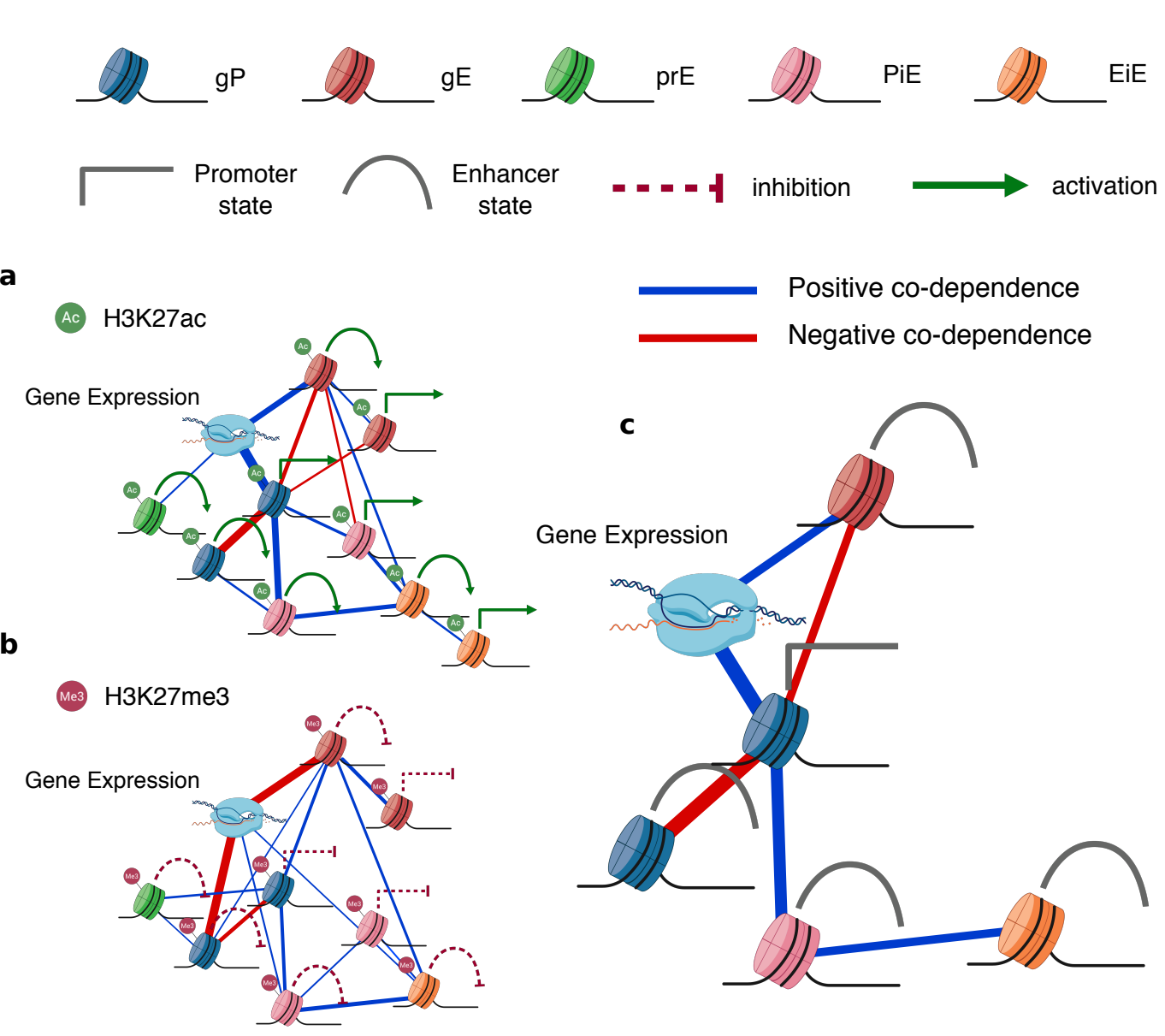
81. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science* **342**, 750–752 (2013).

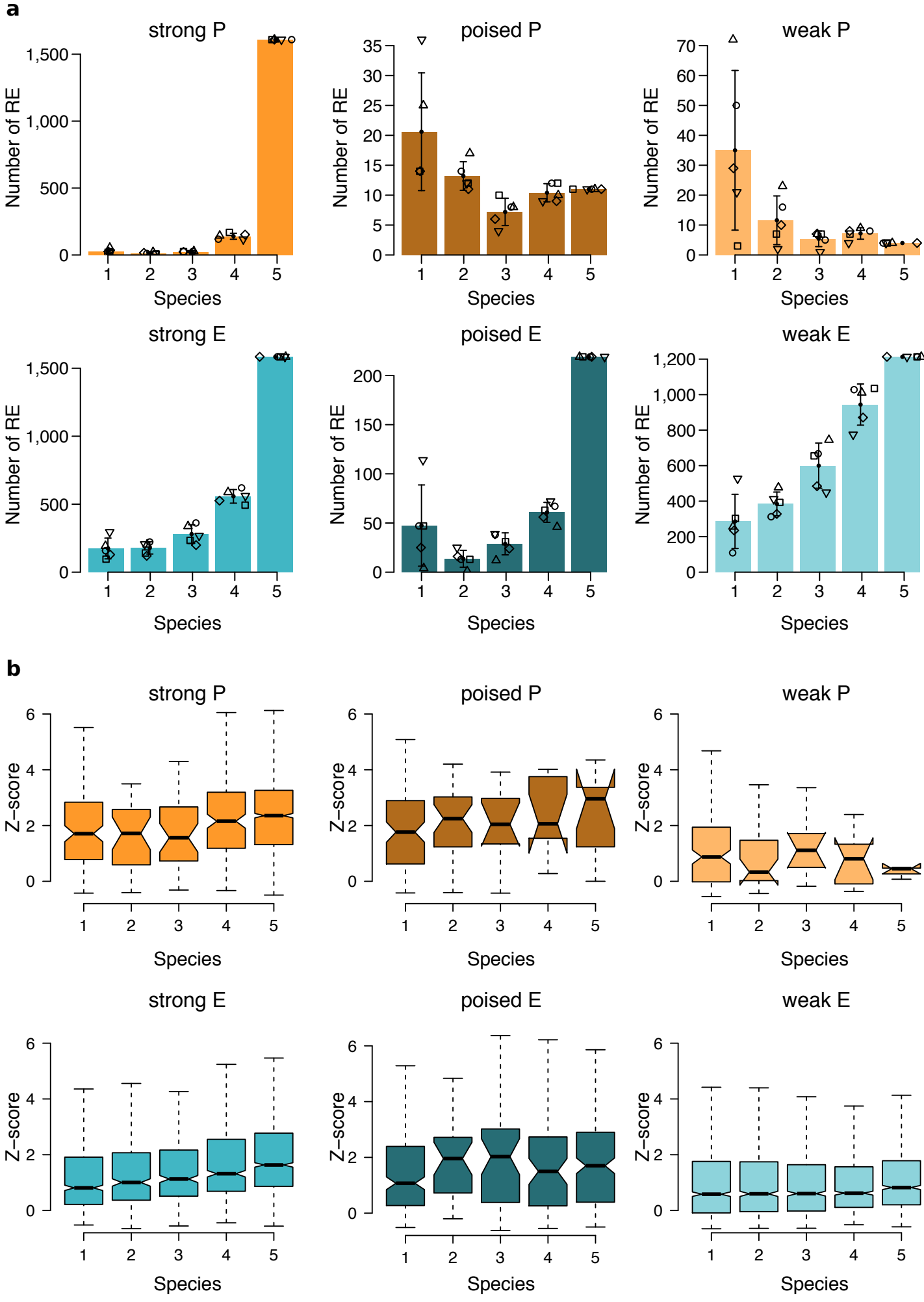
82. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

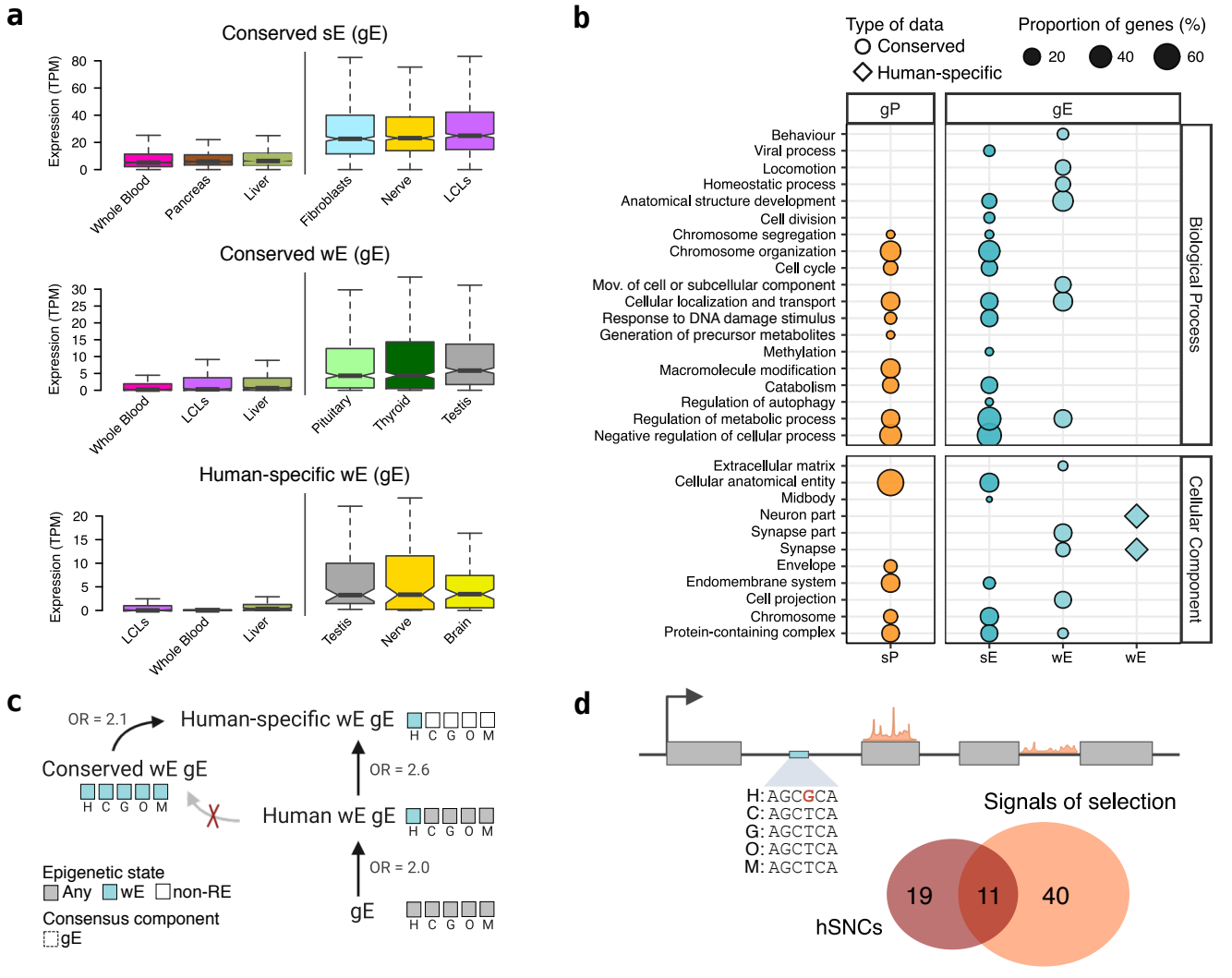

a
Human
Chimpanzee
Gorilla
Orangutan
Macaque
GM12878
H1 H2 C1 C2 G1 G2 O1 O2 M1 M2
ChIP-seq
WGS ...CAGGTACGT...
WGBS ...CAGGTAC$_m$GT...
ATAC-seq
RNA-seq
H3K27me3 H3K4me1 H3K27ac H3K4me3 H3K36me3
High-throughput sequencing
Data integration
b
Reads (M)
RNA-seq
H3K4me1
H3K4me3
H3K27ac
H3K36me3
H3K27me3
ATAC-seq
WGBS
WGS
H1 H2 C1 C2 G1 G2 O1 O2 M1 M2


a
Gene annotation
Coding sequence
Coding sequence
Chromatin state characterization (ChromHMM: H3K4me1, H3K4me3, H3K27ac, H3K27me3, H3K36me3)
200 bp
Regulatory elements (RE)
RE 1
RE 2
RE 3
RE 4
RE 5
b
Epigenetic state
Preliminary classification
RE 2
RE 5
Weak promoter
Strong promoter
Robust reclassification
(Linear discriminant analysis)
Poised
Active
H3K27me3 signal
H3K27ac signal
Enhancer
Promoter
H3K4me1 signal
H3K4me3 signal
RE 2
RE 5
Weak promoter
Strong enhancer
Weak enhancer
Strong promoter
Strong enhancer
Weak promoter
Strong enhancer
c
Regulatory component
Coordinates
Genes
Hi-C loops
Rao et al.
RNAPII loops
Tang. et al
H3K27ac loops
Mumbach et al.
Promoter-interacting enhancer
Genic promoter
Genic enhancer
Enhancer-interacting enhancer
Proximal enhancer
strong promoter (sP)
poised promoter (pP)
weak promoter (wP)
strong enhancer (sE)
poised enhancer (pE)
weak enhancer (wE)
genic promoter (gP)
genic enhancer (gE)
proximal enhancer (prE)
promoter-interacting enhancer (PiE)
enhancer-interacting enhancer (EiE)


a
Promoters
Human
Chimpanzee
Gorilla
Orangutan
Macaque
12,000
6,000
2,000
b
Enhancers
0
40,000
80,000
c
Gene-associated RE
60,000
20,000
0
d
Orphan RE
0
10,000
250,000
e
Regulatory components
gP
gE
prE
PiE
EiE
25,000
10,000
0
f
Orthologous
protein-coding genes
0
4,000
8,000
Human
Chimpanzee
Gorilla
Orangutan
Macaque
g
gP
gE
prE
PiE
EiE
sP
pP
wP
sE
pE
wE
0.0
0.2
0.4
0.6
Proportion of RE (%)


gP
gE
prE
PiE
EiE
Promoter state
Enhancer state
inhibition
activation
a
Ac
H3K27ac
Gene Expression
Positive co-dependence
Negative co-dependence
b
Me3
H3K27me3
Gene Expression
c
Gene Expression


a
strong P
poised P
weak P
strong E
poised E
weak E
Number of RE
Species
b
Z−score


a
Conserved sE (gE)
Expression (TPM)
Whole Blood
Pancreas
Liver
Fibroblasts
Nerve
LCLs
Conserved wE (gE)
Whole Blood
LCLs
Liver
Pituitary
Thyroid
Testis
Human-specific wE (gE)
LCLs
Whole Blood
Liver
Testis
Nerve
Brain
b
Type of data
Conserved
Human-specific
Proportion of genes (%)
20
40
60
gP
gE
Behaviour
Viral process
Locomotion
Homeostatic process
Anatomical structure development
Cell division
Chromosome segregation
Chromosome organization
Cell cycle
Mov. of cell or subcellular component
Cellular localization and transport
Response to DNA damage stimulus
Generation of precursor metabolites
Methylation
Macromolecule modification
Catabolism
Regulation of autophagy
Regulation of metabolic process
Negative regulation of cellular process
Biological Process
Extracellular matrix
Cellular anatomical entity
Midbody
Neuron part
Synapse part
Synapse
Envelope
Endomembrane system
Cell projection
Chromosome
Protein-containing complex
Cellular Component
sP
sE
wE
wE
c
OR = 2.1
Human-specific wE gE
H C G O M
Conserved wE gE
H C G O M
OR = 2.6
Human wE gE
H C G O M
OR = 2.0
gE
H C G O M
Epigenetic state
Any
wE
non-RE
Consensus component
gE
d
H: AGCGCA
C: AGCTCA
G: AGCTCA
O: AGCTCA
M: AGCTCA
Signals of selection
19
11
40
hSNCs