Title: Gene regulatory architectures dissect the evolutionary dynamics of regulatory elements in humans and non-human primates

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

† Current address: Life Sciences Department, Barcelona Supercomputing Center (BSC), Barcelona, Spain

*Corresponding author. Email: tomas.marques@upf.edu (T.M.-B); david.juan@upf.edu (D.J.)

Summary

Genes undergoing substantial evolutionary shifts in their expression profiles are often modulated by critical epigenomic changes that are among the primary targets of selection in evolution. Here, we investigate the evolution of epigenetic regulatory activities and their interplay with gene expression in human and non-human primate lineages. We extensively profiled a new panel of human and non-human primate lymphoblastoid cell lines using a variety of NGS techniques and integrated genome-wide chromatin contact maps to define gene regulatory architectures. We observe that epigenetic and sequence conservation are coupled in regulatory elements and reflect the impact of their activity on gene expression. The addition or removal of strong and poised promoters and intragenic enhancers is frequent in gene expression changes during recent primate evolution. In contrast, novel human-specific weak intragenic enhancers, dormant in our cell lines, have emerged in genes showing signals of recent adaptive selection, suggesting that they echo important regulatory innovations in other cell types. Among the genes targeted by these regulatory innovations, we find key candidate drivers of recently evolved human traits, such as *FOXP2* or *ROBO1* for speech and language acquisition, and *PALMD* for neocortex expansion, thus highlighting the importance of regulatory changes in human evolution.

Keywords: Epigenomics, gene regulation, evolution.

Introduction

Changes in chromatin structure and gene regulation are thought to 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, little is known about how fine-tuning regulatory changes evolved in closely related species, even from a human perspective. 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 few histone marks[7–10]. These analyses suggested that enhancers have evolved faster than promoters[8,11]. It has also

been highlighted that 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 shown that comparative epigenomics is a powerful tool to investigate the evolution of regulatory elements[15,16]. Yet, the integration of multi-layered coherent epigenome data is essential for investigating recent evolutionary time frames, for example within human relatives.

Here, we provide an in-depth view of the recent evolution of gene regulatory architectures using a homologous cellular model system in human and non-human primates. For this, we extensively profiled and characterized lymphoblastoid cell lines (LCLs) from human, chimpanzee, gorilla, orangutan and macaque (Supplementary Materials). This characterization includes whole-genome sequencing at high coverage (WGS, Supplementary Figs. 1-4 and Supplementary Tables 1-2), whole-genome bisulfite (WGBS, Supplementary Figs. 5-7), deep-transcriptome sequencing (RNA-seq, Supplementary Figs. 8-9), chromatin accessibility (ATAC-seq) and ChIP-seq data (Supplementary Figs. 10-13) from five key histone modifications (H3K4me1, H3K4me3, H3K36me3, H3K27ac and H3K27me3). This results in the most extensive collection of great apes and macaque transcriptomic and epigenomic data to date.

Landscapes of chromatin states were robustly defined for all samples by the integration of multivariate HMM-based combinatorial analysis of ChIP-seq peaks co-localization information[17] (Supplementary Figs. 14-21 and Supplementary Tables 3) and Linear Discriminative Analysis of normalized histone enrichments (Supplementary Materials, Supplementary Figs. 22-46). Chromatin states were hierarchically grouped according to their epigenomic state (promoter, enhancer or non-regulatory) and activity (strong, weak or poised).

In contrast to other commonly used definitions of promoter and enhancers limited to strongly active regions[7,8], the multi-layered integration of this epigenetic resource allows the additional definition of weak and poised activities. These activities are of particular relevance to improve the definition of regulatory regions and explore the regulatory potential of regions whose activity can differ in other cell types or conditions. Hence, regulatory elements in each sample were identified as genomic regions displaying such regulatory states (Supplementary Materials). Altogether, this catalog of regulatory elements provides a comprehensive view of the regulatory landscape both in humans and in our closest relatives.

Results

Evolution of promoter and enhancer epigenetic states in human and non-human primates

We identified 29,693 clusters of one-to-one orthologous regions present in all species (Figure S47) where a promoter or enhancer state was detected in at least one species –hereinafter referred to as 'regulatory regions' (Supplementary Materials, Supplementary Table 4). The presence of regulatory states in these regions is highly conserved, with 61% of them having a detectable regulatory state in all five species (Fig. 1a). Consistent with previous studies in more distant species[9], we observed that the presence of promoter states in regulatory regions is more conserved than that of enhancer states (68% and 56% of promoters and enhancers are fully conserved, respectively, Chi-square test, $P < 2.2 \times 10^{-16}$). However, the high conservation values of enhancer states indicates that a great amount of them have conserved their regulatory potential –regardless of their activity– during a recent evolutionary time frame.

Next, we investigated changes in the regulatory state during the evolution of human and non-human primates (Fig. 1a). About 97% of the regions undergoing either gains or losses of regulatory states correspond to enhancers recently established or removed in primates (Supplementary Table 5-7). Most gains/losses are species-specific (63% in enhancers and 91% in promoters). We observed a preferential loss of conserved enhancers over promoters

(22% and 3% of the regions with the corresponding state conserved in the remaining species respectively, Chi-square test, $P < 2.2 \times 10^{-16}$). The human lineage shows higher rates of both gains and losses of enhancer states than the chimpanzee lineage (Chi-square test, $P < 10^{-12}$ in both cases), while it has accumulated fewer gains and losses in promoter states than the latter (Chi-square test, $P = 1.5 \times 10^{-3}$ and 4.3×10^{-12}, respectively, Supplementary Fig. 48).

In addition, we found 721 regulatory regions showing signals of robust repurposing (Supplementary Table 8). Most of these cases (72%) reflect recent species-specific events in regions with conserved states. 347 promoter states are repurposed from conserved enhancer states and 175 enhancer states from conserved promoter states, with a significant enrichment in promoter to enhancer repurposing (Chi-square test, $P < 2.2 \times 10^{-16}$). However, the lower number of promoters in the genome limits the number of cases of promoter-to-enhancer repurposing, leading to most (92%) species-specific enhancer states being gained from regions with non-regulatory states in all the other species. In contrast, the higher number of enhancers allows most (53%) species-specific promoters to arise from conserved enhancer repurposing (in agreement with previous observations in vertebrates[18]. Taken together, 99% of the changing regulatory regions and 88% of the fully conserved regions display enhancer configurations, highlighting their fundamental role in the recent evolution of regulatory landscapes in human and non-human primates.

Enhancer and promoter regulatory activities show specific evolutionary dynamics

While the study of the evolution of enhancer and promoter states provides a global perspective, a detailed understanding of the underlying evolutionary dynamics requires the consideration of their activities. The different enhancer and promoter activities show characteristic conservation patterns (Supplementary Materials and Supplementary Fig. 1b). Strong promoter activities are highly conserved, whereas poised and weak promoter activities show poor conservation in human and non-human primates. As most of the detected promoter states (85.9%) are strongly active (Fig. 1c), there is a relatively small number of regions (686

regions) changing from/to/between promoter activities. Strong enhancer activities are also more conserved than poised and weak activities, but the three of them show similar conservation patterns (Fig. 1d). All enhancer activities in primates display a U-shaped conservation pattern, reflecting their intermediate levels of epigenetic conservation. This highlights the importance of enhancer activities in defining both common and divergent cellular configurations in every lineage (Supplementary Figs. 49-53).

We observed that most gains/losses of enhancer states involve strong and weak activities (Supplementary Fig. 48). Strong enhancer activities are rarely gained whereas weak enhancer activities are both gained and lost at higher rates. The smaller number of gains and losses of strong enhancers in the human lineage contrasts with a previous study targeting gains of strong enhancers in brain[8], probably reflecting tissue-specific differences. Promoter activities are gained and lost at very different rates. Losses correspond exclusively to strong promoter activities, while weak activities are preferentially gained. Consequently, the comparatively higher rates in chimpanzee-specific changes imply a substitution of strong with weak promoter activities in different regions. These gains and losses in promoters, though potentially relevant for gene expression, are infrequent in primate evolution (4.4% of the regions with annotated promoter states in primates but only 4 human-specific cases and none associated with protein coding genes, Supplementary Tables 5-6). Taken together, the observed numbers of conserved and changing enhancers support the prevalent role of enhancer activities both in regulatory conservation and innovation in recent human evolution.

We next evaluated the sequence conservation of the different activities. The sequences of strong promoter activities are highly conserved, and the more conserved the state, the higher the sequence conservation (Fig. 1c and Supplementary Materials). This indicates that their incorporation or removal implies radical changes in the evolutionary constraint of the region. On the other hand, sequence and epigenomic conservation of poised promoters are not linked, but their high sequence conservation suggests a possible strong activity in other cell types or

conditions. Finally, sequences of weak promoters are poorly conserved suggesting a less relevant regulatory role. Like promoters, strong and poised enhancers show high levels of sequence conservation in human and non-human primates, while weak enhancers are much less conserved (Fig. 1d). However, enhancers show a direct association of activity conservation and sequence conservation for all the activity types, which is consistent with corresponding differences in evolutionary constraint. This observation also indicates that the activity conservation of enhancers and strong promoters in our cellular model is a good proxy of their functional importance during human and non-human primate evolution.

Definition of gene regulatory architectures

We have shown that the regulatory state and activity of a region strongly conditions its genomic and epigenomic conservation in human and non-human primates. However, these activities are defined without considering their interaction with their target genes. We defined gene regulatory architectures by linking the regulatory elements with their putative target genes. We retrieved over 350,000 (69.2% of the regulatory elements) gene-element assignments for all five species based on a combination of genome proximity and available 3D contact maps for human LCLs[19–21] (chromatin contacts were projected to non-human primates based on the orthology of the interacting regions, Supplementary Materials, Fig. 2a, Supplementary Figs. 54-58 and Supplementary Tables 9-15). The remaining unassigned orphan regions are depleted in strong and poised activities (Chi-square test, $P < 2.2 \times 10^{-16}$) and show a poor sequence conservation (Mann-Whitney U test, $P < 2.2 \times 10^{-16}$; Fig. 2b). The higher evolutionary constraint in the regulatory regions linked to genes is reflected also in the higher epigenomic conservation of the weak enhancer activities (Fig. 2c), suggesting that we were able to assign target genes for the most relevant regulatory regions in our system.

Given that gene expression is controlled by a combination of short- and long-distance regulatory interactions[22], elements in our gene regulatory architectures were classified in five regulatory components according to the nature of their association with their target genes (3D

contact and/or genomic position relative to the gene). We defined promoters, intragenic enhancers, promoter-interacting enhancers, proximal enhancers and enhancers-interacting enhancers for every gene, regardless of their actual epigenomic state. It is important to note that the same gene-architectural component can display enhancer or promoter epigenetic states in different conditions. For this reason, we decided to define our components independently of their regulatory states. However, regulatory activities are in strong agreement with our regulatory components (Fig. 2d and Supplementary Fig. 55), with regulatory activities being globally enriched in their analogous regulatory components (Chi-square test, $P < 2.2 \times 10^{-16}$).

Role of the gene-architectural components in gene expression and its evolution in human and non-human primates

Our observations suggest that the evolutionary conservation of an element reflects its importance in the regulation of its target gene. However, the actual importance of each type of component and regulatory state in gene regulation and in its evolutionary changes remains to be elucidated. Previous analyses have shown that gene expression can be predicted based on the pseudo-quantitative ChIP-seq signals from informative marks in regulatory regions, mostly promoters and gene surroundings[12,23,24]. We reasoned that the relevance of the different gene-architectural components in gene regulation could be deduced from the strength of these co-dependencies. In this way, types of regulatory components important for regulating gene expression are expected to show histone enrichments coordinated with gene expression levels along all the genes in human and non-human species. Covariations in tightly interdependent multivariate systems are the result of the complex network of dependencies and often offer a distorted view of their actual underlying causal relationships[25–27]. To unravel this scenario, we used partial correlation analyses to define the common network of direct co-dependencies between RNA-seq and ChIP-seq signals for protein-coding genes. We also used generalized linear models to determine the ability of key components of our regulatory architectures to explain gene expression (Supplementary Materials).

Protein-coding genes show a high variety of regulatory architectures (Figure S54) and previous studies have shown that conservation in the number of strong enhancers is important for the evolution of gene expression in more distant species[9]. Thus, for simplicity, we considered an additive scenario in which ChIP-seq signals of all elements in each gene-architectural component were aggregated for promoter and enhancer states separately. This approach accommodates all the different combinations of components and elements found in our regulatory architectures in 50 regulatory variables (2 states x 5 components x 5 histone marks). We performed a partial correlation analysis of gene expression and these regulatory variables (Supplementary Materials) to elucidate the relevance of the different types of regulatory components and states for explaining gene expression levels in human and non-human primate species. The network of partial correlations shows that the RNA-seq signal is specifically explained by the combination of promoters and intragenic enhancers (Fig. 3a and Supplementary Fig. 59). Interestingly, we also observed co-dependencies between the elements of these two components indicating that their interdependence can contribute to gene regulation. Promoters and intragenic enhancers also show negative Pearson's correlations between their histone mark signals (Supplementary Fig. 60), suggesting that promoters and intragenic enhancers could be part of different complementary regulatory mechanisms.

To evaluate the strength of the co-dependence of the transcriptional output with promoters and intragenic enhancers, we predicted protein-coding gene expression levels from ChIP-seq signals in these core regulatory regions. For this, we fitted generalized linear models based only on the normalized enrichments of H3K27ac, H3K27me3 and H3K36me3 in promoters and intragenic enhancers, considering first-order interactions between them (Supplementary Materials). This multivariate model explains 72% of gene expression variability (Supplementary Fig. 61, Supplementary Materials), outperforming a model including all histone marks (and ATAC-seq) in all the elements without first-order interactions (65%,

Supplementary Fig. 62). These results confirm the high influence of both genic promoters and intragenic enhancers on gene regulation and support the previously unknown interdependence between them.

We then investigated the contribution of the different components to gene expression changesn. The specific contribution of strong enhancers to gene expression evolution can be explained by the number of enhancers in the genomic neighborhood of the gene[9]. We dissected the different effect of regulatory states and activities for each gene-architectural component in gene expression changes, in terms of their changes in number in the regulatory architectures of orthologous genes (Fig. 3b, Supplementary Materials). Consistent with all components being (directly or indirectly) connected to gene expression in our partial correlation network (Fig. 3a), differences in the number of every regulatory component are significantly associated with inter-species gene expression differences. However, the contribution to this effect of each component depends on its regulatory state and activity. The presence of promoter components (for strong promoter and poised enhancer activities) and the number of intragenic enhancers (for strong enhancer and poised enhancer activities) show the most robust associations with gene expression differences. Proximal enhancers (for strong, weak and poised activities) also show significant, although less supported associations that according to our partial correlation analysis could occur through promoter activities in promoter components (Fig. 3a). Enhancers interacting with promoters (for strong promoter and enhancer activities) and with other enhancers associated with the gene (for weak enhancers and the combination of both poised activities) also show significant but modest effects (Fig. 3b).

Weak enhancers echo the regulatory activity of different cell types

Next we assessed the functional profiles of the genes targeted by conserved and human-specific promoter and intragenic enhancer components (Supplementary Tables 16-19). The small number of genes carrying human-specific strong promoters and enhancers show no

significant enrichments. Fully conserved strong promoter activities in promoter components and strong enhancers in intragenic enhancers show overlapping enrichments for housekeeping intracellular functions, associated with metabolism, chromatin organization or regulation of the cell cycle (Fisher's exact test, BH correction FDR<0.05, Supplementary Tables 20-23). These enrichments are coherent with their essential roles and reflect the proliferative state of these cell lines.

We explored the role of weak enhancers in our architectures, since their functional interpretation is not obvious. Weak enhancers are more conserved when they are associated with the regulatory architectures (Fig. 2c). However, they seem not to be very relevant for gene expression changes in our primate cell lines (Fig. 3b). Weak enhancers are characterized by the presence of H3K4me1 in the absence of H3K27ac and H3K27me3 (Supplementary Figs. 16-17, Supplementary Materials). Intronic H3K4me1 sites are specifically enriched in brain[28] and alterations in the regulation of H3K4 methylation have been associated with a variety of neurodevelopmental disorders[29]. Therefore, intragenic enhancers may have a particularly relevant role in the epigenetic regulation of the central nervous system. The exact function of H3K4me1 in enhancers remains unclear[30] but in the absence of H3K27ac they have been proposed to mark 'primed' enhancers[31,32] or even to be involved in expression fine-tuning[30].

We hypothesized that weak intragenic enhancers could reflect the degree of regulatory conservation in genes active in other cell types or conditions. For this reason, we analyzed conserved and human-specific weak intragenic enhancers as a proxy of regulatory elements potentially relevant to the evolution of other cell types. We observed that genes with conserved weak intragenic enhancers are highly enriched in functions related to ion transmembrane transport, neuronal genes and blood vessel development (Fisher's exact test, BH correction, FDR < 0.05; Supplementary Tables 24-25, Supplementary Materials). In fact, we found that they were enriched in genes with cerebral cortex- and kidney-specific gene expression (hypergeometric test, BH correction, 62 genes and $P = 1.3 \times 10^{-4}$; 18 genes and $P = 1.3 \times 10^{-}$

^{5}, respectively; Fig. 4a and Supplementary Fig. 63). Similarly, genes with human-specific weak intragenic enhancers are enriched in neuronal and membrane genes (Fisher's exact test, BH correction, FDR < 0.05; Supplementary Table 26-28, Supplementary Materials), reinforcing the involvement of weak intragenic enhancers in the regulation of genes associated with transmembrane transport, especially in synapsis. This is consistent with their enrichment in genes with cerebral cortex-specific gene expression (hypergeometric test, BH correction, 26 genes and $P = 3.5 \times 10^{-6}$; Fig. 4b).

Novel weak intragenic enhancers mark regulatory innovations in candidate driver genes of human adaptation

Although the direct role of human-specific weak intragenic enhancers in the regulation of neuronal processes remains to be elucidated, they point towards the acquisition of regulatory innovations in a small set of genes. Among the 77 genes with human-specific weak intragenic enhancers, we found some particularly interesting cases (detailed list in Supplementary Table 28). For these instances we explored their epigenetic context in other cell types and tissues[33] finding strong or weak enhancer activities in most of the cases with cell types matching their functions (Fig. 4 and Supplementary Figs. 64-75).

The presence of human-specific weak intragenic enhancers in these examples is associated with two main regulatory scenarios (Fig. 4c,d and Supplementary Figs. 64-75), according to an independent analysis in human cell lines[33]. First, we found cases as *FOXP2* (Fig. 4c and Supplementary Fig. 64), where our human-specific intragenic enhancers typically show heterochromatin or elongation states in most cell types, but display weak enhancers (or it is surrounded by such) in more specific tissues (often brain, lung and/or aorta). Second, we detected cases as *PALMD* (Fig. 4d and Supplementary Fig. 65) showing weak or strong enhancer states in more tissues. These two scenarios might imply the presence of two levels of specificity. One of them associated with activation in very specific tissue regions, moments or conditions and a second scenario reflecting a more global activation in the targeted tissues.

Two of the genes with human-specific acquisition of weak intragenic enhancers are *FOXP2* and *ROBO1* (Fig. 4a and Supplementary Figs. 66-68), both of which are involved in human speech and language acquisition[34,35] and may have been important during the evolution of the human lineage since the split from chimpanzees[34,35]. The *SORCS3*, *ADGRL2* and *PTPRG* genes (Supplementary Figs. 69-71), like *FOXP2*, are associated with human-accelerated conserved non-coding sequences and show differential expression in brain areas involved in speech and language processing[36]. *SYBU* also shows signals of adaptive selection in the human lineage[37] and has been associated with cognitive decline in neurodegenerative diseases[38]. *PRSS12* (Supplementary Fig. 72) shows a putative signal of positive selection in humans[39] and modulates hippocampal function and social interaction in mice[40].

PALMD (Fig. 4b and Supplementary Fig. 65) has been recently proposed as a driver of the evolutionary expansion of the neocortex in mammals[41] and, in addition to present a human-specific weak intragenic enhancer, it contains a large number of non-synonymous changes fixed in modern humans after the split from Neanderthals[42]. This suggests that *PALMD* might also have a role in the expansion of the neocortex in humans, maybe in coordination with other genes, such as *ARHGAP11B*[43]. *ADAM18* (Supplementary Fig. 73) is involved in spermatogenesis and also carries non-synonymous changes fixed in modern humans[42]. Selection on *ADAM18* has also been associated with the evolution of promiscuity in primates[44].

Besides these genes, we found many other interesting cases both related and unrelated to neuronal functions. For instance, the *TBX15* gene (Supplementary Fig. 74), which is associated with adipose tissue differentiation and body-fat distribution, contains a Denisovan-like haplotype subject to adaptive introgression in modern humans from Greenland[45]. *CFTR* (Supplementary Fig. 75) is another interesting case carrying a human-specific weak intragenic enhancer. Mutations in *CFTR* are responsible for cystic fibrosis[46] and the high allele frequency

of its pathological allele in European populations suggests the existence of a heterozygous adaptive advantage[47]. However, given that one of the human cell lines used in this study is of Yoruban origin (GM19150 cell line, see Supplementary Fig. 2) and also shows the weak enhancer linked to *CFTR*, the acquisition of this regulatory element probably precedes the introduction of this allele. Taken together, our results show that human-specific acquisition of weak intragenic enhancers in LCLs points to genes that were potentially subject to adaptation in the human lineage at different timescales with tissue-specific activation and expression patterns.

Discussion

The evolution of human and non-human primates is an area of major interest, in which the 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 primates. Despite the artificial nature of our cell model[48–50], 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[51] and which can be robustly used for comparative studies in humans and primates[12,52–54]. 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[7,9,18]. Moreover, we have expanded these observations to explain how these dynamics are a consequence of the different evolutionary constraints associated with their regulatory activities. Therefore, we prove that considering weak and poised activities is of major relevance to better understand the evolution of regulatory regions.

In LCLs, the human lineage shows higher rates of incorporation and removal of strong enhancers, but lower rates for strong promoters than the chimpanzee lineage. These rates

are likely to differ between different cell types, as they convey information about the phenotypic changes and the functional profiles associated with each cell type. In fact, a recent work focused on strong activities in bulk brain samples showed a higher number of changes in human promoters compared to chimpanzee[8]. These observations suggest that there is room for defining cell type-specific epigenomic evolutionary signatures based on the changes in strong regulatory activities. We and others have shown that cell lines provide an experimentally sound and biologically informative resource for this research, even more in the context of endangered species. Future studies performing cell-type-aware comparative epigenomics will provide additional insights into the dynamics of the evolution of the regulatory landscapes and their integration will help broaden the understanding of the evolution of more complex phenotypic traits.

Our results show that the association of regulatory components with gene expression reflects the logic of the structural configuration of the regulatory architecture and influences the evolution of the regulatory landscape in human and non-human primates. In brief, promoter and intragenic enhancer components constitute the interdependent core of these architectures explaining gene expression levels. Proximal and promoter-interacting enhancers are codependent with promoter components, and enhancer-interacting enhancers are associated with promoter interacting enhancers. We observed that the evolutionary behavior of the regulatory components is highly conditioned by its association with gene expression. Acquisition or removal of these 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 and affects the evolutionary constraint of the component. Despite the weaker and indirect co-dependencies of the remaining components, they can still be instrumental for gene expression evolution through their influence on promoters and intragenic enhancers. Our analyses demonstrate that for understanding the evolution of regulatory landscapes, it is fundamental to unravel their actual role in gene regulation. This conceptual framework provides a starting point for future in-depth

investigations on the interdependence of different regulatory regions and mechanisms in the evolution of gene regulation. In this sense, we stress the importance of embracing higher levels of complexity in order to achieve a more detailed description of the regulatory processes.

Interestingly, major insights about this process can arise from the analysis of the regulatory elements with a negligible regulatory role in our system. Weak intragenic enhancers seem to carry information about the degree of regulatory innovation in a broader context than the studied cell type (mostly in transmembrane transporters and neuronal functions). Interestingly, gains of these elements in the human lineage are associated with candidate genes that may have driven human adaptation in several important traits at different timescales. This observation suggests that changes in the regulatory potential of intragenic enhancers lead to conformational epigenetic changes that can be observed in cell types where they are not active. These echoing regulatory states provide an unexpected window to the evolution of regulatory landscapes in the human lineage. Further research will be needed to clarify the actual role of these elements in the differential regulation of these genes. We conclude that differences in the regulatory roles and activities deeply condition the evolutionary dynamics of epigenomic landscapes and their association with gene expression changes. Our insights call for the incorporation of better integrative datasets and key molecular regulatory details in comparative evolutionary studies to better understand the interplay between epigenetic regulation and gene expression in recent human evolution.

Data availability

The raw fastq files from the genomic, transcriptomic and epigenomic data generated and used for the analyses in this study were uploaded to the Sequence Read Archive (SRA) with the BioProject accession number PRJNA563344.

References

1. Britten, R. J. & Davidson, E. H. Gene Regulation for Higher Cells: A Theory. *Science* **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* **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* **28**, 1097–1110 (2018).
6. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **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* **9**, e1003763 (2013).
16. Lowdon, R. F., Jang, H. S. & Wang, T. Evolution of Epigenetic Regulation in Vertebrate Genomes. *Trends in Genetics* **32**, 269–283 (2016).
17. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
18. Carelli, F. N., Liechti, A., Halbert, J., Warnefors, M. & Kaessmann, H. Repurposing of promoters and enhancers during mammalian evolution. *Nature Communications* **9**, (2018).
19. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **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* **49**, 1602–1612 (2017).
21. Tang, Z. *et al.* CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163**, 1611–1627 (2015).
22. 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).
23. Karlic, R., -R. Chung, H., Lasserre, J., Vlahovicek, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences* **107**, 2926–2931 (2010).
24. Lasserre, J., Chung, H.-R. & Vingron, M. Finding associations among histone modifications using sparse partial correlation networks. *PLoS Comput. Biol.* **9**,

e1003168 (2013).

25. Chen, B. & Pearl, J. Graphical Tools for Linear Structural Equation Modeling. (2014). doi:10.21236/ada609131
26. Stein, R. R., Marks, D. S. & Sander, C. Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. *PLoS Comput. Biol.* **11**, e1004182 (2015).
27. de Juan, D., Pazos, F. & Valencia, A. Emerging methods in protein co-evolution. *Nat. Rev. Genet.* **14**, 249–261 (2013).
28. Zhu, J. *et al.* Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642–654 (2013).
29. Shen, E., Shulha, H., Weng, Z. & Akbarian, S. Regulation of histone H3K4 methylation in brain development and disease. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **369**, (2014).
30. Rada-Iglesias, A. Is H3K4me1 at enhancers correlative or causative? *Nature Genetics* **50**, 4–5 (2018).
31. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).
32. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21931–21936 (2010).
33. Libbrecht, M. W. *et al.* A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types. *Genome Biol.* **20**, 180 (2019).
34. Enard, W. *et al.* Molecular evolution of FOXP2 , a gene involved in speech and language. *Nature* **418**, 869–872 (2002).
35. Mozzi, A. *et al.* The evolutionary history of genes involved in spoken and written language: beyond FOXP2. *Sci. Rep.* **6**, 22157 (2016).
36. Johnson, M. B. *et al.* Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron* **62**, 494–509 (2009).
37. Gayà-Vidal, M. & Albà, M. M. Uncovering adaptive evolution in the human lineage. *BMC*

Genomics **15**, 599 (2014).

38. Bereczki, E. *et al.* Synaptic markers of cognitive decline in neurodegenerative diseases: a proteomic approach. *Brain* **141**, 582–595 (2018).
39. Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K.-D. & Wray, G. A. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat. Genet.* **39**, 1140–1144 (2007).
40. Mitsui, S. *et al.* A mental retardation gene, motopsin/neurotrypsin/prss12, modulates hippocampal function and social interaction. *Eur. J. Neurosci.* **30**, 2368–2378 (2009).
41. Kalebic, N. *et al.* Neocortical Expansion Due to Increased Proliferation of Basal Progenitors Is Linked to Changes in Their Morphology. *Cell Stem Cell* **24**, 535–550.e9 (2019).
42. Kuhlwilm, M. & Boeckx, C. A catalog of single nucleotide changes distinguishing modern humans from archaic hominins. *Sci. Rep.* **9**, 8463 (2019).
43. Florio, M. *et al.* Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science* **347**, 1465–1470 (2015).
44. Finn, S. & Civetta, A. Sexual selection and the molecular evolution of ADAM proteins. *J. Mol. Evol.* **71**, 231–240 (2010).
45. Racimo, F. *et al.* Archaic Adaptive Introgression in TBX15/WARS2. *Mol. Biol. Evol.* **34**, 509–524 (2017).
46. Riordan, J. R. Identification of the cystic fibrosis gene: Cloning and characterization of complementary DNA. *Trends in Genetics* **5**, 363 (1989).
47. Poolman, E. M. & Galvani, A. P. Evaluating candidate agents of selective pressure for cystic fibrosis. *J. R. Soc. Interface* **4**, 91–98 (2007).
48. Carter, K. L., Cahir-McFarland, E. & Kieff, E. Epstein-Barr Virus-Induced Changes in B-Lymphocyte Gene Expression. *Journal of Virology* **76**, 10427–10436 (2002).
49. Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.* **43**, 768–775 (2011).
50. Sugawara, H. *et al.* Comprehensive DNA methylation analysis of human peripheral

blood leukocytes and lymphoblastoid cell lines. *Epigenetics* **6**, 508–515 (2011).

51. Hussain, T. & Mulherkar, R. Lymphoblastoid Cell lines: a Continuous in Vitro Source of Cells to Study Carcinogen Sensitivity and DNA Repair. *Int J Mol Cell Med* **1**, 75–87 (2012).
52. Khaitovich, P., Enard, W., Lachmann, M. & Pääbo, S. Evolution of primate gene expression. *Nature Reviews Genetics* **7**, 693–702 (2006).
53. Pai, A. A., Bell, J. T., Marioni, J. C., Pritchard, J. K. & Gilad, Y. A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS Genet.* **7**, e1001316 (2011).
54. Shibata, Y. *et al.* Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. *PLoS Genet.* **8**, e1002789 (2012).
55. Siepel, A. & Haussler, D. Phylogenetic Hidden Markov Models. *Statistical Methods in Molecular Evolution* 325–351 doi:10.1007/0-387-27733-1_12
56. Jain, A. & Tuteja, G. TissueEnrich: Tissue-specific gene enrichment analysis. *Bioinformatics* **35**, 1966–1967 (2019).
57. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

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. TMB is supported by BFU2017-86471-P (MINECO/FEDER, UE), U01 MH106874 grant, 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).

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

Figure Captions

Fig. 1 | Evolutionary dynamics of epigenetic states and activities. a, Evolutionary stability of regulatory states and **b**, activities in orthologous regions. Cell values represent the percentage of regions showing a regulatory state in a species (rows) whose orthologous regions display a given regulatory state in other species (columns). **c**, Promoters and **d**, enhancers epigenomic (top) and sequence conservation (bottom). X axis represents conservation in 1 to 5 primates. U-shaped patterns of epigenomic conservation highlight the accumulation of species-specific activities (each species contributes with an independent set of regions). Sequence conservation corresponds to the most conserved 200-bp long region in each element. Conservation is estimated as phastCons[55] values for the alignments including

30 primate species (retrieved from http://hgdownload.soe.ucsc.edu/goldenPath/hg38/phastCons30way/).

Fig. 2 | Characterization of gene regulatory architectures. a, Annotation of interactions between regulatory elements. intragenic, proximal and distal enhancers (gE, prE and dE, respectively) are reannotated as promoter-interacting-enhancers when interacting with promoters (PiE, first-order interactions) and enhancer-interacting-enhancers (EiE) when interacting with enhancers already assigned to the architecture (second-order interactions). **b**, Sequence conservation of unassigned orphan elements vs. elements assigned to regulatory architectures. **c**, Epigenomic composition of gene-architectural components of autosomal protein-coding genes. **d**, Epigenetic conservation of the regulatory activity in elements assigned to regulatory architectures.

Fig. 3 | Interplay between gene regulatory architectures and gene expression. a, Partial correlation network for gene expression and histone modification signals across primates. Partial correlations between variables are shown as edges between nodes. Edge width is proportional to absolute values of partial correlations (partial correlations with $P < 10^{-40}$ are shown, Supplementary Materials). Blue and red edges for positive and negative correlation values, respectively. Histone modification labels lack *H3* prefix. **b**, Inter-primate expression differences depend on the number of regulatory elements at given architectural components (y axis) showing specific epigenomic activities (x axis). Orthologous genes showing gene expression changes were grouped according to their normalized gene expression values and the differences in the mean number of each type of element between species with higher and lower gene expression were assessed (Supplementary Materials). Values are exact -log *P* of the corresponding paired Wilcoxon signed rank test. Colors indicate the direction of the association (blue = positive, red = negative). * indicates associations with $P < 10^{-3}$.

Fig. 4 | Weak enhancers echo brain-specific regulation. Expression profiles of cerebral cortex-specific genes in (**a**) conserved and (**b**) human specific weak intragenic enhancers. Both gene sets were evaluated for tissue-specific gene expression enrichment in RNA-seq data[56] from the Human Protein Atlas[57]. Genes with intragenic enhancers were used as background. Only the genes enriched in cerebral cortex compared to non-brain regions are represented in the heatmap. Regulatory annotation of human-specific weak enhancers in the brain-associated genes: (**c**) *FOXP2* and (**d**) *PALMD.* Gene diagram with intronic location of human-specific enhancers (brown, top). Epigenetic annotation of the intragenic enhancer and surrounding regions for selected cell types and tissues (box top). For simplicity, tissue annotations were collapsed prioritizing the visualization of promoter and enhancer states (for uncollapsed annotations see Figures S64 and S65). Correspondence of these annotations with the analogous regulatory activities defined in this study is indicated in the legend. Conservation-associated activity plot (box bottom). Labels are vertically scaled by their conservation-associated activity score (CAAS), reflecting the prevalence of regulatory states established in 164 human cell types[33]. Positive height corresponds to a position's conservation-associated activity score and it is colored proportionally to the fraction of the score for each chromatin state. Negative light grey distribution of phyloP area indicates the 75th percentile of phyloP scores within 100 bp of a given genomic position. Genome coordinates are relative to genome assembly hg19.

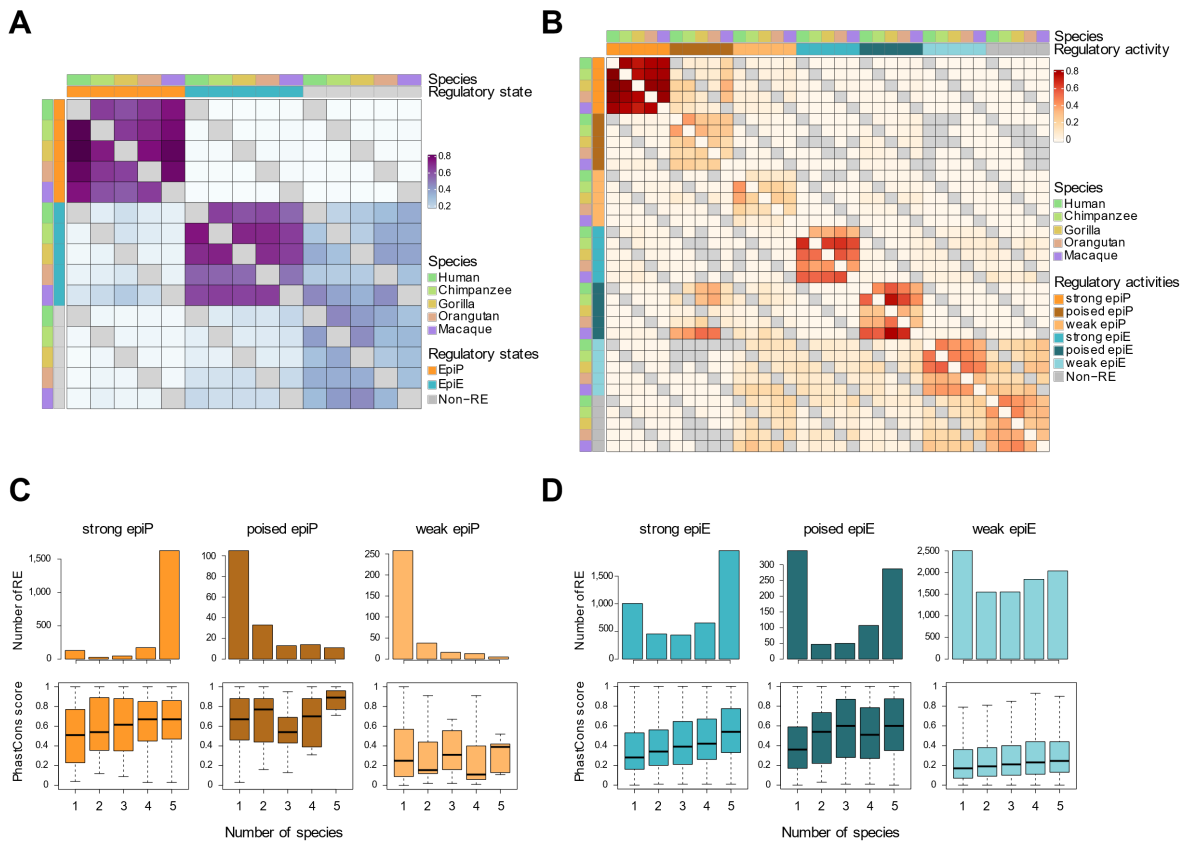

Fig. 1

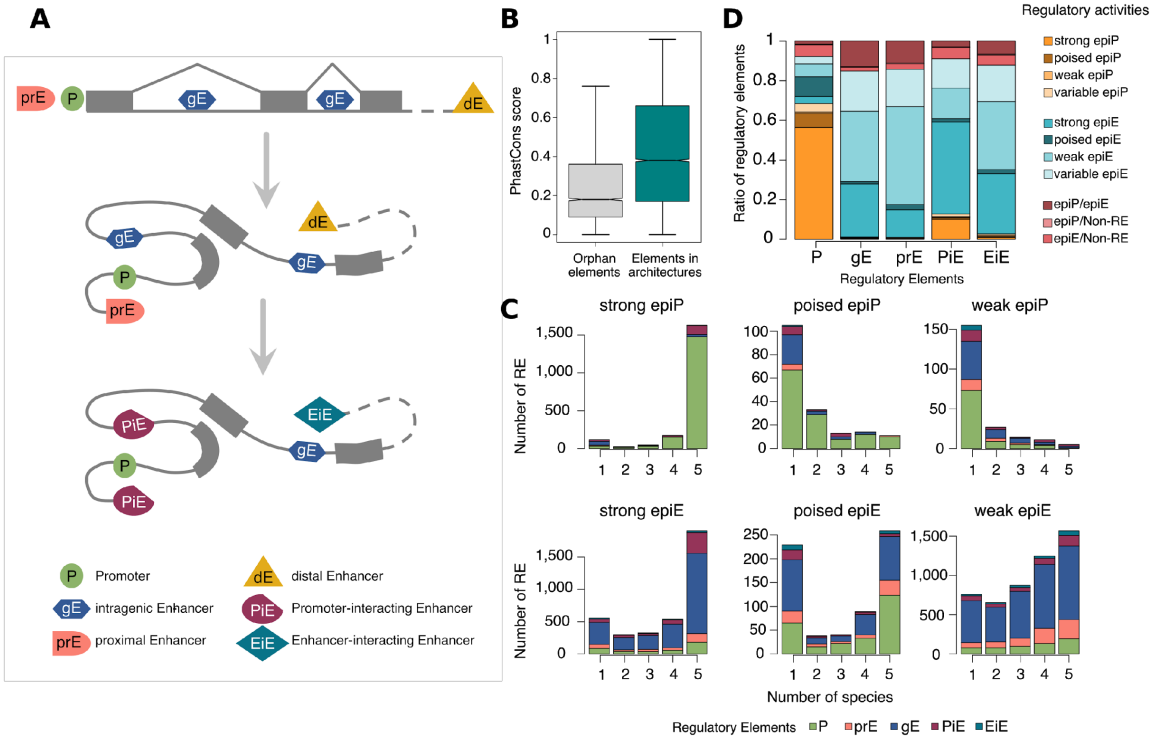

Fig. 2

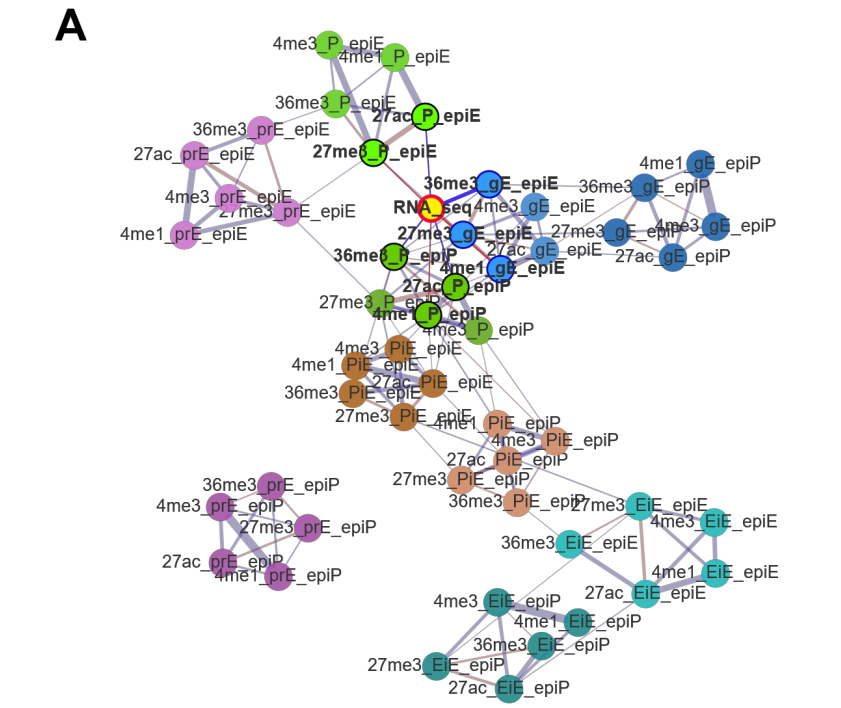

Fig. 3

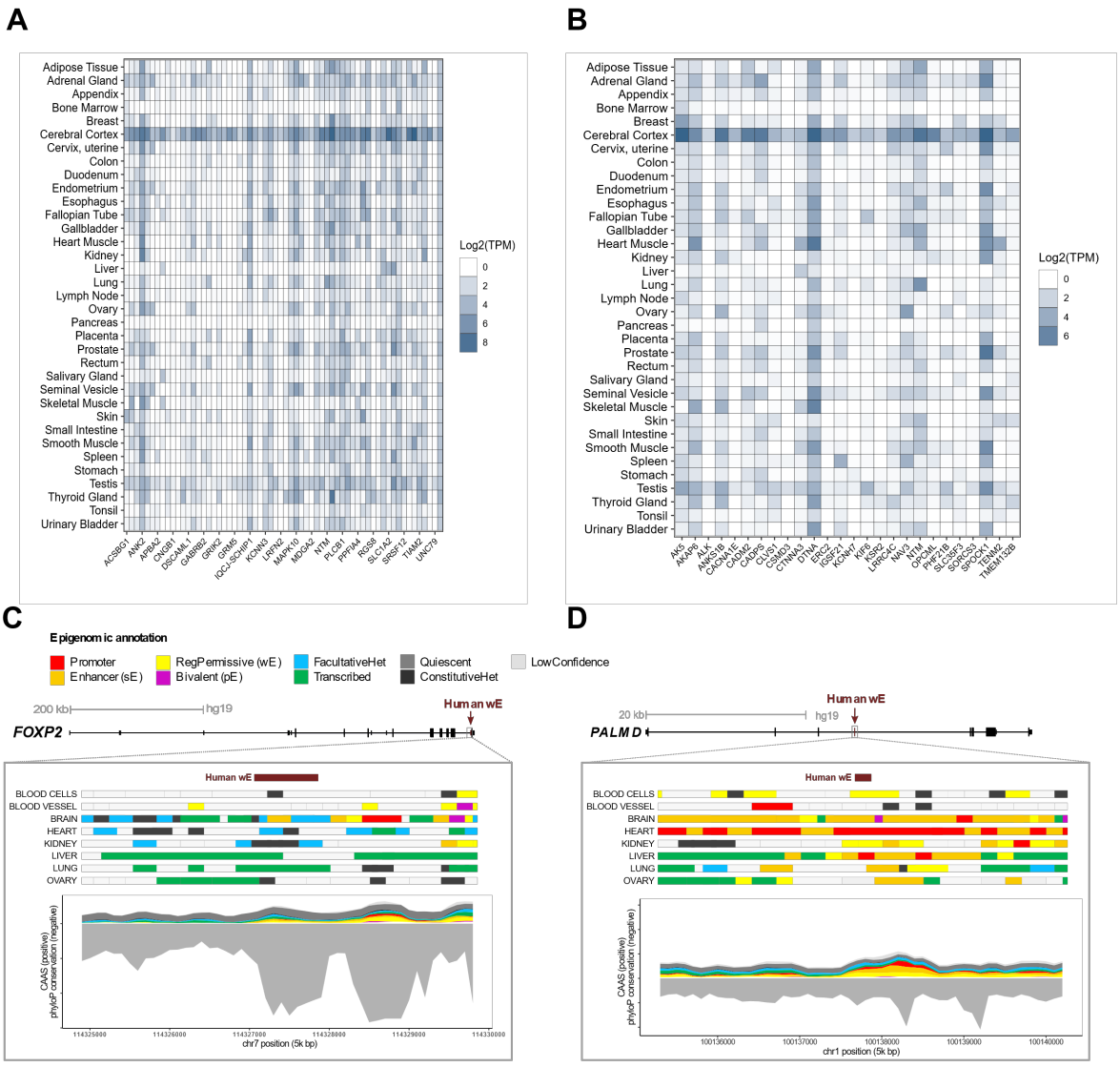

Fig. 4