Transposable elements are the primary source of novelty in primate gene regulation

Marco Trizzino, YoSon Park, Marcia Holsbach-Beltrame, et al.

Genome Res. published online August 30, 2017
Access the most recent version at doi:10.1101/gr.218149.116

Transposable elements are the primary source of novelty in primate gene regulation

Marco Trizzino[1,2,#,*], YoSon Park[1,2,*], Marcia Holsbach-Beltrame[1], Katherine Aracena[1], Katelyn Mika[3], Minal Caliskan[1,2], George H. Perry[4], Vincent J. Lynch[3] and Christopher D. Brown[1,2,#]

1. Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104

2. Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, 19104

3. Department of Human Genetics, University of Chicago, Chicago, IL, 60637

4. Departments of Anthropology and Biology, Pennsylvania State University, University Park, PA, 16802

* Co-first author; equally contributed

[#] Correspondence: chrbro@upenn.edu (C.D.B.), marco.trizzino83@gmail.com (M.T.)

Running title: Regulatory evolution in primates

Keywords: Enhancers, promoters, evolution, primates, transposons, liver

M.T. Current affiliation: Gene Expression and Regulation Program, The Wistar Institute, Philadelphia, PA, 19104

Abstract

Gene regulation shapes the evolution of phenotypic diversity. We investigated the evolution of liver promoters and enhancers in six primate species using ChIP-seq (H3K27ac and H3K4me1) to profile cis-regulatory elements (CREs), and RNA-seq to characterize gene expression in the same individuals. To quantify regulatory divergence, we compared CRE activity across species by testing differential ChIP-seq read depths directly measured for orthologous sequences. We show that the primate regulatory landscape is largely conserved across the lineage, with 63% of the tested human liver CREs showing similar activity across species. Conserved CRE function is associated with sequence conservation, proximity to coding genes, cell-type specificity, and transcription factor binding. Newly evolved CREs are enriched in immune response and neurodevelopmental functions. We further demonstrate that conserved CREs bind master regulators, suggesting that while CREs contribute to species adaptation to the environment, core functions remain intact. Newly evolved CREs are enriched in young transposable elements, including Long-Terminal-Repeats (LTR) and SINE-VNTR-*Alu*s (SVAs), that significantly affect gene expression. Conversely, only 16% of conserved CREs overlap TEs. We tested the cis-regulatory activity of 69 TE subfamilies by luciferase reporter assays, spanning all major TE classes, and showed that 95.6% of tested TEs can function as either transcriptional activators or repressors. In conclusion, we demonstrated the critical role of TEs in the primate gene regulation, and illustrated the potential mechanisms underlying evolutionary divergence among the primate species through the non-coding genome.

Introduction

The contribution of cis-regulatory elements (CREs) to phenotypic and behavioral evolution has been shown in many taxa (King and Wilson, 1975; Prabhakar et al., 2008; Schmidt et al., 2010; Cain et al., 2011; Marnetto et al., 2014; Zhou et al., 2014; Prescott et al., 2015; Railey et al., 2015; Villar et al., 2015; Emera et al., 2016; Berthelot et al., 2017). Although previous studies have suggested a role for transposable elements (TEs) in the evolution of gene regulation, validating the functional contribution of TEs in mammalian gene regulation remains a challenge (McClintock, 1950, 1984; Britten and Davidson, 1969; Davidson and Britten, 1979; Jordan et al., 2003; Bejerano et al., 2006; Wang et al., 2007; Bourque et al., 2008; Sasaki et al., 2008; Markljung et al., 2009; Kunarso et al., 2010; Lynch et al., 2011, 2015; Schmidt et al., 2012; Chuong et al., 2013, 2016; Jacques et al., 2013; Xie et al., 2013; del Rosario et al., 2014; Sundaram et al., 2014; Du et al., 2016; Rayan et al., 2016; Ward et al., 2017).

A significant fraction of the accessible regions in primate genomes overlap a TE (Jacques et al 2013). Similarly, the recruitment of novel regulatory networks in the uterus was likely mediated by ancient mammalian TEs (Lynch et al. 2011, 2015). Conversely, neocortical enhancers do not exhibit strong evidence of transposon co-option (Emera et al. 2016).

Many important questions remain unanswered: to what extent are regulatory elements functionally conserved across primates? Are specific genomic features predictive of CRE conservation? To what extent have TEs driven the evolution of gene regulation?

To address these questions, we investigated the evolution of gene expression and regulation in the primate liver. While liver functions are largely conserved across primates, different environmental exposures, diets, and lifestyles shape the adaptation of liver functions, making this tissue a suitable model in which to explore the conservation and divergence of gene regulation.

We performed ChIP-seq for Histone H3 Lysine 27 acetylation (H3K27ac) and Histone H3 Lysine 4 mono-methylation (H3K4me1), which mark functional and poised regulatory elements, in the liver of six primate species, including at least one species from each major primate clade (Perelman et al., 2012). We generated RNA-seq data from the same specimens and estimated the degree of evolutionary conservation of regulatory activity and gene expression levels across the entire lineage. We identified genomic features associated with the evolutionary conservation of gene regulation. Finally, to functionally characterize the contribution of TEs to gene expression divergence, we performed extensive experimental validation on TE-derived CREs.

Results

Data generation, quality assessment, and validation

We generated RNA-seq and ChIP-seq data from *post mortem* livers of three or four individuals per species of mouse lemur (*Microcebus murinus*), bushbaby (*Otolemur garnettii*), marmoset (*Callithrix jacchus*), rhesus macaque (*Macaca mulatta*), chimpanzee (*Pan troglodytes*), and human (*Homo sapiens*) (Fig. 1). The samples were from young adults, and, with the exception of the bushbaby, included both males and females. After stringent quality control, a total of 18 RNA-seq and 14 ChIP-seq samples remained post-QC and were used for analyses (Table S1). We identified H3K27ac and H3K4me1 peaks in the human liver, treating all human individuals as replicates in the peak calling procedure with MACS2 (FDR < 5%; Zhang et al., 2008). Overlapping peaks from the two histone marks were merged.

Several lines of evidence indicate that the regions of histone modification we have identified represent active CREs. First, 66.1% of ENCODE HepG2 H3K27ac regions overlapped one of our human peaks (the ENCODE Project Consortium, 2012). Moreover, 71.6% of the 244,269 Roadmap liver H3K27ac regions, and 49.7% of the 233,386 Roadmap liver H3K4me1 regions overlapped one of our human peaks (Roadmap Epigenomic Mapping

Consortium, 2015). The vast majority of predicted CREs do not exhibit significant sex-biased regulatory activity (Table S8). Finally, we tested the regulatory activity of 1 kb fragments from 276 predicted CREs in HepG2 cells using a novel parallelized reporter assay (Melnikov et al., 2012; Patwardhan et al., 2012; Sharon et al., 2012; File S4): 191 drove significant reporter activity (69.2%; Table S6), demonstrating that the majority of the CREs predicted from our ChIP-seq data are likely functional regulatory elements in the human liver.

The majority of human six-way-alignable CREs are functionally conserved across primates

We identified regions orthologous to human peaks from the genomes of non-human primates using the Ensembl multiple sequence alignment (MSA) database (Flicek et al., 2014). Excluding the sex chromosomes, we catalogued 39,710 total human CREs with orthologs in all six species: 32,759 enhancers (distance from nearest TSS > 1kb) and 6,951 promoters (distance from nearest TSS < 1kb).

After extracting ChIP-seq read counts for the six species from the 39,710 regions, we assessed evidence of differential histone modification with DESeq2 (Love et al., 2014), using ChIP-seq input data as a covariate. We compared the read counts across all possible human-centric species × species and group × group pairwise comparisons. This approach provides a quantitative assessment of histone modification profiles across species, while avoiding issues arising from experimental variables that may confound peak calling (Waszak et al., 2015). An analysis of human and marmoset, the latter being the species with the smallest number of peaks called in this study, strongly supported the robustness of this approach (detailed analysis described in File S4; Fig. S1). Additionally, we compared our H3K27ac data to a recent study of liver CREs in mammals (Villar et al., 2015). Only 10.9% of human H3K27ac peaks at FDR < 5% (5.8% with FDR <1%) exhibited evidence of differential histone modification between two studies (File S4; Figure S7).

The majority of the 39,710 human CREs (25,067 CREs [63.1%]; FDR < 10%) did not exhibit significant differential histone modification in any of the tested pairwise comparisons (Fig. 2). This suggests that these regulatory regions are consistently active across the primate lineage, and thus may represent evolutionarily conserved primate CREs. This conclusion was robust to changes to the FDR threshold (Table S7). As an additional control, we performed a chimpanzee-centric analysis for regions orthologous to chimpanzee H3K27ac consensus peaks, and demonstrated that 54.1% of these regions were not differentially histone modified in any of the pairwise comparisons, indicating that species-specific bias is unlikely.

Primate promoters were more conserved than enhancers (75.1% and 60.5%, respectively; Fisher's exact test $p < 2.2\times10^{-16}$; Fig. 2), as observed previously in mammals (Villar et al., 2015). On the other hand, 36.9% of orthologous CREs exhibited differential histone modification state across species (Fig. 2). We detected 57 human-specific CREs (0.14%) and 2,259 ape-specific CREs (5.7%; Fig. 3; Table S2). Together, our differential histone state analysis results are broadly supported by several studies that have consistently suggested a high degree of regulatory element conservation between closely related species in metazoans (Cotney et al., 2013; Boyle et al., 2014; Prescott et al., 2015; Emera et al., 2016). We note that the estimated fraction of conserved CRE is lower (36.2%: 39.6% of promoters and 24.3% of enhancers) when analyses are not restricted to 6-way orthologous regions (i.e. treating all human CREs lacking an ortholog in any of the other species as not conserved). However, we note that some primate genome assemblies, particularly mouse lemur and bushbaby, are largely incomplete, and the lack of orthology is most likely a consequence of assembly quality. We thus limited our analyses to the six-way-alignable CREs.

To assess the extent to which restricting this analysis to the set of six-way-orthologous CREs could affect our results, we re-performed the same analysis for H3K27ac, only considering human, chimpanzee, and rhesus macaque. 86.9% of human H3K27ac

CREs have an aligned ortholog in both species. In total, 98.3% and 97.5% of the tested CREs were conserved between human and chimpanzee (99.0% in the six species analysis), and human and rhesus macaque (98.6% in the six species analysis), respectively. Results between the two analyses are thus comparable (Fisher's exact test $p > 0.05$ for both of the comparisons), suggesting that the restriction of our analyses to six-way orthologous CREs did not significantly bias the results.

Conservation of the nucleotide sequence is associated with conservation of regulatory activity

For each human-centric species × species comparison, we estimated: i) the fraction of differentially modified CREs; ii) the fraction of differentially expressed genes from a set of 10,243 genes with six way orthologs (Table S3); and iii) the per-nucleotide pairwise sequence divergence for each species with respect to humans for each of the 39,710 unique orthologous CREs.

Differential histone modification ranged from 0.79% in the human × chimpanzee, to 28.4% in the human × mouse lemur comparisons (Fig. 3a). Similarly, differential gene expression ranged from 5.93% in the human × chimpanzee to 16.0% in the human × mouse lemur comparisons (Fig. 3b). Both differential histone modification and differential gene expression reflected phylogenetic distance between humans and other tested species, and differentially expressed genes were significantly more likely to be associated with a differentially modified CRE than expected by chance (Cochran–Mantel–Haenszel test Odds Ratio (OR) = 1.45; $p < 2.2\times10^{-16}$).

Sequence divergence was significantly associated with differential histone modification (human × chimpanzee, logistic regression $p = 6.8\times10^{-8}$; human × rhesus macaque, $p < 2.2\times10^{-16}$; human × marmoset, $p = 1.3\times10^{-8}$; human × bushbaby, $p = 3.3\times10^{-4}$; human × mouse lemur, $p = 2.4\text{x}10^{-4}$; Table S9). Moreover, functionally conserved CREs were significantly more likely to overlap a placental mammal phastCons element (Siepel et

al., 2005) than expected by chance (Fisher's exact test $p < 2.2\times10^{-16}$). Together, these data demonstrate that CREs with conserved nucleotide sequence are significantly more likely to have conserved regulatory activity and are associated with conserved gene expression, as previously suggested (Brown et al., 2007; Cooper and Brown, 2008; Pollard et al., 2010; Gittelman et al., 2015; Holloway et al., 2015; Yang et al., 2015; Dong et al., 2016; Lewis at al., 2016).

Genomic features associated with CRE conservation and rapid evolution

To understand the mechanisms responsible for CRE conservation and turnover, we identified genomic features associated with conserved regulatory activity. CREs associated with protein-coding genes were significantly more conserved than CREs associated with either pseudogenes (Fisher's exact test $p = 1.4\times10^{-5}$) or lincRNAs (Fisher's exact test $p = 8.9\times10^{-14}$; Fig. 4a). For closely related species, regulatory activity was conserved, regardless of the distance to the nearest TSS (human × chimpanzee, logistic regression $p = 0.30$; Fig. 4b). However, for more distantly related species pairs, the evolutionary conservation of the CRE activity was significantly lower in regions more distant from TSSs (human × rhesus macaque, logistic regression $p < 2.2\times10^{-16}$; human × marmoset, $p < 2.2\times10^{-16}$; human × bushbaby, $p < 2.2\times10^{-16}$, human × mouse lemur, $p < 2.2\times10^{-16}$; Fig. 4b). Intronic enhancers were significantly more conserved than intergenic enhancers (64.2% and 55.8% respectively; Fisher's exact test $p < 2.2\times10^{-16}$). These data demonstrate increased selective pressure to maintain regulatory activity in the vicinity of protein-coding genes.

Multiple genomic features indicative of broad regulatory element activity were significantly associated with regulatory conservation. Promoters and enhancers overlapping regions of chromatin accessibility in many cell types (the ENCODE Project Consortium, 2012) were significantly more conserved than those that are functional in only a small number of cell types (logistic regression $p = 6.2\times10^{-15}$; Fig. 4c). Similarly, conservation of CRE activity was significantly correlated to the number of TFBS, as identified by ENCODE

ChIP-seq in HepG2 cells (logistic regression $p < 2.2\times10^{-16}$; Fig. 4d). A Gene Ontology analysis for genes associated with conserved CREs revealed enrichment for regulation of cellular, transcriptional, and developmental processes (Table S4).

Specific transcription factor motifs are associated with regulatory conservation and turnover

We used the MEME Suite (Bailey et al., 2009) to identify sequence motifs enriched in human-specific, ape-specific, and evolutionarily conserved liver CREs. Human-specific CREs are enriched with motifs for TFs associated with immune response and hematopoietic maintenance (Fig. 4e; Supplemental File S2), such as RFX5, SMAD1 and EOMES. The rapid evolution of immune response genes and TFs is supported by many studies in vertebrates and in *Drosophila melanogaster*, demonstrating that while the central machinery of immune responses is strongly conserved, components of the extended molecular networks can evolve rapidly or diversify as a consequence of evolutionary competition between hosts and pathogens (Jansa et al., 2003; Vallender 2004; Sackton et al., 2007; Obbard et al., 2009; Schadt et al., 2009; Grueber et al., 2014; Lazzaro and Schneider 2014; Salazar-Jaramillo et al., 2014; Zak et al., 2014; Sironi et al., 2015; Wertheim, 2015). Ape-specific CREs are enriched for binding sites of TFs involved in liver function but also in brain and neural system proliferation and development (Fig. 4e).

Evolutionarily conserved CREs are enriched with motifs for master regulators and homeobox genes that establish cell-type identity in liver cells (Fig. 4e; Supplemental File S2). Among these master regulators, HNF4A is essential for the differentiation of human hepatic progenitor cells (DeLaForest et al., 2011). Likewise, CEBPA is required for the liver cell specification and gene function, and the associated TFBSs are highly conserved across mammals (Ballester et al., 2014). Both CEBPA and HNF4A have conserved cis-regulatory activity and a large number of shared TF binding events across distant vertebrates (Schmidt et al., 2010). These results demonstrate that evolution shapes the regulatory landscape by

preserving the regulatory activity in essential metabolic and developmental pathways, while permitting incessant renovation of specific networks that are under strong selective pressures.

TE derived CREs are pervasive in the primate genomes

To quantify the contribution of TEs to the regulation of liver gene expression, we annotated each liver CRE based on overlap with RepeatMasker elements (Smit et al., 2013-2015; http://www.repeatmasker.org). 9,877 of the 39,710 six-way-alignable CREs, (24.9%) overlapped a TE for at least 20% of their length. A total of 24 TE families were significantly enriched in CREs (FDR < 1%; Table S5), nearly all of which were SINE-VNTR-*Alu*s (SVAs), and LTRs (mostly ERV1; Fig. 5). As we filtered our ChIP-seq data for high confidence alignments (see Methods), reads mapping to young TE families are likely underrepresented among our identified CREs (Fig. S6). While only 0.01% of the human TEs overlap a human CRE, when restricting the analysis to the 39,710 six-way-alignable CREs, 73.5% of TE insertions are found within a differentially modified CRE (Fisher's exact test $p < 2.2\times10^{-16}$; Table S10).

The majority (75.0%) of enriched TE families were relatively young, and specific to humans (SVA-F), Hominidae (SVA-B, SVA-C, and SVA-D), Hominoidea (the LTR12 subfamily), Simiiformes (LTRs), or primates (*Alu* elements), whereas the remaining 25.0% were Eutherian-specific or older (Fig. 5, Fig. S5, and Table S5). Given that young TE families are likely depleted from our alignments due alignment limitations, we believe these estimates are likely conservative. TE enrichment analyses were based on the expected proportion of CRE that contain each kind of TE. Thus, SVAs, and also LTRs, were significantly more abundant than expected, regardless the TE lineage specificity, despite being among the least common classes of repeats in the human genome (15.9% and 0.69% of the total TEs respectively; Fisher's exact test $p < 2.2\times10^{-16}$ for both of the TE categories).

57.3% of CREs harboring TEs were differentially active in at least one pairwise comparison. Among these, SVAs (2.5%) were over-represented (Fisher's exact test $p < 2.2\times10^{-16}$). We therefore investigated whether these recent TE insertions altered the expression patterns of nearby genes in primates. We focused on SVAs and LTR12-C,D,E TE categories present only in human and chimpanzee. 17.6% of genes associated with CREs overlapping either SVAs or LTR12-C,D,Es were differentially expressed between apes and the other tested primates. In contrast, genes associated with primate-specific TEs or not associated with any TE are significantly less likely to be differentially expressed between apes and non-ape primates (Fisher's exact test $p < 0.002$). Genes whose CREs recently acquired TE insertions did not exhibit greater within species expression variability than genes without TE insertions. This suggests that genes that acquire TE insertions are not simply more tolerant to variable expression (File S4).

The vast majority of recently evolved CREs are derived from TE insertions

Overall, 77.1% of ape-specific CREs and nearly all human-specific CREs overlap a TE (Fig. 5; Fig. S5). In contrast, only 16.0% of evolutionarily conserved CREs contain an annotated TE. LTRs (in particular LTR-12C) and SVAs are the most common TEs overlapping newly evolved CREs (LTR = 40.1% of the recruited TEs in ape-specific CREs; SVA = 75.3% of the recruited TEs in the human-specific CREs; Fig. 5). The regulatory function of 9 CREs overlapping a TE were validated in our MPRA experiment, including 5 with ape-specific function.

The contribution of LTRs to gene regulation has been proposed in previous studies (Wang et al., 2007; Cohen et al., 2009; Sundaram et al., 2014; Chuong et al., 2016; Janoušek et al., 2016). An example of an ape-specific CRE derived from LTR insertion is an enhancer at the gene *GRIN3A*. This gene is involved in physiological and pathological processes in the central nervous system and has been associated with several complex human diseases, including schizophrenia (Takata et al., 2013). *GRIN3A* is upregulated in

apes compared to other primates (log_2 Fold Change = 2.00; FDR < 0.02; Fig. S3). Further, our differential histone modification analysis identified an ape-specific ChIP-seq peak overlapping a 1-kb long ape-specific insertion (present also in orangutan and gorilla, but not in other primates; GRCh38 Chr9:101,723,127-101,724,197). This insertion, located 13 kb from the TSS of *GRIN3A*, is entirely derived from an LTR-12C. The insertion drove strong enhancer activity upon transfection into HepG2 cells (Wilcoxon's rank sum test p = 0.00017; Fig. S3), suggesting that the TE insertion results in a functional enhancer at the *GRIN3A* locus.

SVAs are a hominid-specific family of composite retrotransposons active in humans (Hancks and Kazazian, 2010), with more than 3,500 annotated copies. SVAs that overlap a liver CRE are significantly closer to the TSS of the associated gene than those that do not overlap a CRE (Wilcoxon rank-sum test p = 0.00117), suggesting that an SVA has a higher probability of becoming a CRE if it inserts near gene promoters (Fig. S2).

Among the SVAs with significant histone modification, we identified an intronic CRE for the gene *JARID2.* This gene is an accessory component of Polycomb Repressive Complex-2 (PRC2), recruits PRC2 to chromatin, and is involved in liver, brain, neural tube development, and embryonic stem cell differentiation (Kaneko et al., 2014). Our differential histone state analysis identified a human-specific ChIP-seq peak overlapping a human-specific 1.9 kb-long insertion, entirely derived from an SVA-F human specific retrotransposon. *JARID2* is significantly downregulated in humans compared to all the other primates (log_2 Fold Change = -3.33; FDR < 0.02; Fig. S4). SVAs-Fs overlapping a CRE exhibit significant enrichment for binding sites of known transcriptional repressors such as PAX5, FEV, and SREBF1 (Maurer et al., 2003; Fazio et al., 2008; Lecomte et al., 2010). Indeed, the JARID2 SVA-F insertion leads to significantly decreased expression in HepG2 reporter assays (Wilcoxon rank sum test p = 0.00275; Fig. S4). Furthermore, 78.3% of the genes associated to CREs with human specific SVA insertions exhibit significantly decreased expression relative to their non-human orthologs (Fisher's exact test $p < 1.6x10^{-}$

[6]). This, along with additional validation assays presented below, supports the role of SVA insertions as transcriptional repressors.

Broad regulatory activity of TE insertions in the primate liver

Our findings strongly suggest that the majority of novel CREs in primates are derived from TE insertions. To validate the predicted regulatory activity of recent TE insertions, we tested the cis-regulatory activity of 69 TE subfamilies, covering all major classes and families of primate TEs (Table S6). TEs from these families overlap 3,897 of our predicted CREs. We synthesized the mammalian consensus sequence for 69 different TE families, cloned them into a luciferase reporter vector with a minimal promoter, and transfected into HepG2 cells to perform luciferase reporter assays. Luciferase expression levels for 66 of the 69 (95.6%) tested TE families were significantly different from the negative control (Fig. 6a; Wilcoxon's rank sum test p-values in Table S6). Strikingly, only 17 (25.7%) of these, mostly LTRs and DNA transposons, increased gene expression (Fig. 6a), whereas the remaining 49 (74.3%), mostly LINEs, repressed transcription. SVAs were confirmed as transcriptional repressors.

These findings demonstrate that LTRs, among the most enriched TEs in our peak set, and the most common TE with a signature of regulatory function in apes, frequently result in increased regulatory activity. LTRs are known to have strong regulatory elements (Chuong et al., 2016). While only 2% of all human LTRs are marked by active histone modification, 25.5% of genes with an LTR insertion in an associated CRE are differentially expressed, suggesting that the effects of LTR insertions on local gene expression are strongly context specific.

The consensus sequences for the 66 TEs that significantly affected reporter expression were analyzed with MEME to identify enriched motifs. Motifs for known master regulators of liver cell identity, including FOX, USF2, GABP, and HNF4A (Wallerman et al., 2009), were significantly enriched within the sequences of the 66 TE families with significant regulatory activity (Fig. 6b). In summary, most TE families are capable of functioning as

337 CREs in the primate liver, either as enhancers or repressors, further supporting our findings
338 on the pervasive involvement of TEs in the primate gene regulation.
339

340 ## Discussion

341 Less than 1% of tested CREs resulted as differentially active between humans and
342 chimpanzees. This suggests that even modest changes in gene regulation produce
343 significant differences, and confirms that cis-regulatory evolution plays a central role in
344 primate diversification (Davidson 2001, 2006; Wray, 2007; Ho et al., 2009; Tsankov et al.,
345 2010; Martin et al., 2012; Coolon et al., 2014; Martin and Reed, 2014; Guo et al., 2015;
346 Lynch et al., 2015; Villar et al., 2015; Adachi et al., 2016; Landeen et al., 2016; Lesch et al.,
347 2016; Zhang and Reed, 2016).

348 Our approach for the comparison of CREs across species, based on the analysis of
349 differential histone modification state in orthologous regions, demonstrated that cis-
350 regulatory divergence across species may be overestimated when assessed based on
351 binary peak overlap. The increased density of primate sampling in our dataset, compared to
352 previous studies, allowed us to address the timing of primate regulatory divergence, and
353 improved the interpretation of when, and how, different TE families have been recruited for
354 regulatory function. Generating transcriptome data from the same specimens that were
355 epigenetically profiled allowed us to directly measure the effects of conserved and recently
356 evolved CREs on the expression levels of the primary liver tissue.

357 With a combination of three different reporter assay techniques, we have tested and
358 validated the regulatory activity of thousands of predicted CREs and TE insertions. Although
359 previous studies have suggested the recruitment of TEs as functional elements (Huda et al.,
360 2010; Jacques et al., 2013; Lynch et al., 2011, 2015; Chuong et al., 2016), we
361 demonstrated, and functionally validated, the extent of this phenomenon in primates,
362 demonstrating that LTRs and SVAs have played an important role in rewiring ape gene

363 regulation. In contrast, only a minor fraction of evolutionarily conserved CREs overlap an
364 annotated TE. Together, our data suggest that the core regulatory network that establishes
365 liver cell-type identity in primates is conserved, whereas rapid evolution occurs on the
366 periphery of the network, where TEs have the most impact on gene regulatory evolution.
367

368 ## Materials and Methods

369 **Tissue sampling**

370 We obtained liver tissue samples for three/four individuals belonging to each of the
371 studied species (Table S1) from Texas Biomedical Research Institute (San Antonio, TX) and
372 from Duke University Lemur Center (Durham, NC). Samples were collected and flash-frozen
373 immediately.
374

375 **RNA-seq sample processing**

376 We processed samples from all species in random batches of four to minimize batch
377 effects. We used 4µg of total RNA to produce barcoded RNA sequencing libraries using the
378 Illumina TruSeq Stranded mRNA kit (File S4). Libraries were pooled in two different pools
379 based on barcode compatibility, and each pool was sequenced on two Illumina HiSeq 2500
380 lanes, producing on an average of 42.1 million single end 100-bp reads per sample.
381

382 **ChIP-seq sample processing**

383 We processed samples in six randomly assigned groups in order to minimize batch
384 effects (File S4). We used 5 to 15 ng of input and immunoprecipitated DNA to generate
385 sequencing libraries using the NEBNext Ultra ChIPseq library kit. Libraries were multiplexed,
386 pooled, and sequenced on a total of 16 Illumina HiSeq 2500 lanes, producing on an average
387 of 40.6 million SE 100-bp reads per sample.
388

Sequence QC: ChIP-seq and RNA-seq

We assessed standard quality control (QC) measures on FASTQ files using FASTQC v0.11.3 (Andrews, 2010). We trimmed sequencing adapters and low quality base calls using TrimGalore! v0.4.1 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/).

RNA-seq alignment and gene expression quantification

We aligned all sequences that passed QC to the reference genomes from the Ensembl database Release 87 (bushbaby: otoGar3; chimp: CHIMP2.1.4; humans: GRCh38; rhesus macaque: Mmul1; marmoset: C_jacchus3.2.1; mouse lemur: Mmur1) using STAR v2.5, in 2-pass mode (Dobin et al., 2013; File S4). We used FeatureCounts (Liao et al., 2014) to count reads mapping to each gene, according to Ensembl annotations for the six studied species.

Differential gene expression analysis

We analyzed differential gene expression levels using read counts, normalized by feature length with DESeq2 (Love et al., 2014), with the following model: *design* = ~*condition,* where condition indicates the species or the group of species (e.g. apes).

We used a set of 10,243 genes annotated as orthologs in the six species according to Ensembl (BioMart v. 0.9; Smedley et al., 2015; Table S3) and used 5% False Discovery Rate (FDR; Benjamini-Hochberg; Benjamini and Hochberg, 1995) as our multiple-testing-corrected significance threshold. The overall analysis included five comparisons: 1) human vs each of the other five species; 2) human-specific differential expression (human vs other five primates grouped together); 3) ape-specific differential expression (human + chimpanzee vs other four primates). 4) Catarrhini-specific differential expression (human + chimpanzee + rhesus macaque vs other primates). 5) Comparison between Haplorrhini

(human, chimpanzee, rhesus macaque and marmoset) and Strepsirrhini (mouse lemur and bushbaby).

ChIP-seq QC and alignment

We aligned the sequences that passed QC to the reference genomes from the Ensembl database (bushbaby: otoGar3; chimp: CHIMP2.1.4; humans: GRCh38; rhesus macaque: Mmul8.0.1; marmoset: C_jacchus3.2.1; mouse lemur: Mmur2), using Burrows Wheeler Alignment tool (BWA), with the MEM algorithm (Li, 2013). Aligned reads were filtered based on mapping quality (MAPQ > 10) to restrict our analysis to higher quality and likely uniquely mapped reads, and PCR duplicates were removed.

ChIP-seq peak calling and QC

We called peaks for each individual using MACS2, at 5% FDR, with parameters recommended for histone modifications (Liu, 2014): --no model --ext size 147 -B. We performed QC on peaks called for each specimen using metrics recommended by ENCODE (Landt et al., 2012; File S4). Samples that did not pass the three main QC metrics (FRiP, NSC, RSC) were excluded for any downstream analysis. We called human consensus peaks for H3K27ac and H3K4me1 using MACS2 and the above-described parameters. All human samples passing QC were considered as replicates for the consensus peak calling. The human consensus H3K27ac and H3K4me1 peaks were used to perform all human-centric downstream analyses.

Parallelized reporter assay

We obtained a list of 334 putative 1-kb long CREs overlapping liver eQTLs (Brown et al., 2013). 276 out of these 334 CREs overlapped one of our human peaks (96 enhancers and 180 promoters; Table S6). Within each of the loci defined by the 276 liver eQTLs, we predicted a 1-kb CRE, and tested their functionality as described in File S4.

Detection of orthologous regions for human peaks in each primate

We mapped orthologous sequences using all identified human consensus ChIP-seq peak regions in both H3K27ac and H3K4me1 experiments, using the 40 Eutherian mammals Ensembl multiple sequence alignment (MSA). The detailed pipeline is illustrated in File S4.

Differential histone modification analysis

Using the above-described procedure, for both H3K27ac and H3K4me1, we produced a single matrix including the human peaks having an ortholog in each of the studied species, and the associated read count for each histone mark plus the input in all of the six species. Read counts were used for differential ChIP-seq analysis with DESeq2, performing an interaction analysis between the histone marks read counts and their associated input values, using the Wald statistic: *design = ~assay + condition + assay:condition*, where the assay indicates either IP data or input data, and condition indicates the species or the group of species.

Differential histone mark analysis included the same species × species and group × group comparisons described for RNA-seq (FDR <10%). Further, different FDRs (up to 50%) were tested to assess the robustness of our approach. We initially analyzed differential histone modifications for the two marks independently. Then, overlapping CREs were merged.

Sequence conservation

We estimated per-nucleotide pairwise divergence for all five species in comparison to humans using the MSA aligned sequences of orthologous regions for consensus peaks +/- 500 bps. All gaps in human were excluded from analysis. Regions not included in the set of six-way orthologous CREs were pruned. Finally, we removed outliers - with respect to the

distribution of the genetic distances in the given pairwise comparison - using the R package *outliers* (Komsta, 2006).

Transposable elements enrichment

TE enrichment analysis was performed using the *TEAnalysis* pipeline with TE-analysis_Shuffle_bed v. 2.0, setting 1000 replicates (https://github.com/4ureliek/TEanalysis; Kapusta et al., 2013).

Luciferase reporter assay validation of *GRIN3A* and *JARID2*

We compared activity of two predicted functional CREs with the empty pGL4.23 vector as a negative control, as described in File S4.

Validation of the gene regulatory functionality of TE families

Transposable element constructs were built by synthesizing (GenScript) the Dfam (Hubley et al., 2016) consensus sequence for 69 TE subfamilies, representing all of the major TE classes and families. Each element was cloned into pGL3 Basic vector (Promega) with an added minimal promoter (pGL3 Basic[minP]). Luciferase assays were performed as described in File S4.

Statistical and genomic analyses

All statistical analyses were performed using R v3.3.1. Figures were made with the package ggplot2 (Wickham, 2009). BEDTools v2.25.0 (Quinlan et al., 2010) was used for genomic analyses. Scripts and pipelines are available online (https://github.com/ypar/cre_evo_primates.git), and in Supplemental File 5.

Data Access

All raw sequence data from this study have been submitted to the NCBI BioProject database (BioProject; https://www.ncbi.nlm.nih.gov/bioproject/) under accession numbers PRJNA349047 (RNA-seq) and PRJNA349046 (ChIP-seq).

Acknowledgements

We thank Texas Biomedical Research Institute, Duke University Lemur Center and the Hospital of University of Pennsylvania for providing the samples. We thank Jonathan Schug (UPenn NGS Core Facility) for help with data production. We are grateful to Y. Barash, C. Berthelot, B. Engelhardt, C. Feschotte, P. Flicek, E. Fry, Y. Kamberov, S. Tishkoff, and B. Voight for providing insightful comments on preliminary versions of the results and manuscript.

Disclosure declaration

Authors declare no conflicts of interests.

Author contributions

MT, MC, and CDB conceived the project. MT, YP, GHP, and CDB designed the taxon sampling and experiments. MT performed ChIP-seq and RNA-seq experiments. MT and MHB performed the parallelized reporter assays. MHB and KA produced luciferase assay data on *GRIN3A* and *JARID2*. KM and VJL designed and performed the luciferase assays on the 69 TE families. MT, VJL, and KM performed the TE enrichment analysis. YP designed computational pipelines for the detection of orthologous regions in the Ensembl MSA alignment and related analyses. MT, YP, and CDB analyzed the data and wrote the manuscript. All authors read and approved the manuscript.

519 ## Figure legends

520

521 **Figure 1** - **Experimental design and analytical pipeline**. (A) Sampling included three to
522 four specimens from six species representing all major primate clades. (B) ChIP-seq and
523 RNA-seq profiles were produced from the liver samples. Differential histone modification and
524 gene expression analyses were performed on the orthologous CREs and genes in each
525 species, respectively.

526

527 **Figure 2** - **Primates CREs are evolutionarily conserved**. (A) Examples of human-specific,
528 ape-specific and conserved CREs. (B) Fraction of conserved and recently evolved primate
529 CREs, with breakdown of enhancers and promoters.

530

531 **Figure 3** - **Differential histone mark and gene expression across species**. (A) Human-
532 centric pairwise comparisons for differential histone modification states on 39,710
533 orthologous CREs. (B) Human-centric pairwise comparisons for differential gene expression
534 of 10,243 orthologous genes. (C) Number of lineage-specific CREs and genes across the
535 primate phylogeny.

536

537 **Figure 4** - **Genomic features associated with CRE conservation**. (A) Fraction of
538 conserved and recently evolved CREs associated with protein-coding genes, lincRNAs, and
539 pseudogenes. (B) CRE conservation (y-axis) as a function of distance to the nearest gene
540 start (quantiles on the x-axis). (C) CRE conservation (y-axis) as function of cell-type
541 specificity (quantiles on the x-axis) based on ENCODE data. (D) CRE conservation (y-axis)
542 and number of ENCODE HepG2 TFBSs overlapping each CRE (quantiles on the x-axis). (E)
543 Examples of enriched motifs in human-specific CREs; ape-specific CREs; conserved CREs.

544

545 **Figure 5** - **Newly evolved CREs are enriched in TEs**. (A) Proportion of CREs that overlap

546 TEs in the different primate lineages. (B) Number of enriched TE families within CREs in the
547 different primate lineages. (C) Most enriched TE families in primates.
548
549 **Figure 6** - **Regulatory ability of TE families found in poised and active regulatory**
550 **elements in HepG2 cells.** (A) The p-value (Wilcoxon Rank sum test), class, and lineage
551 specificity for the 69 TE families tested in HepG2 cells for regulatory ability, the empty vector
552 control (Basic[minP]), and the positive control (TAP2_C) are all shown above the 6 luciferase
553 assay replicates conducted and the average regulatory ability found across the replicates.
554 Red indicates luciferase expression higher than the empty vector control; blue indicates
555 luciferase expression less than Basic[minP]. (B) Motifs enriched in the sequences of the 66
556 TE families that drive expression significantly different from background.
557
558
559
560
561

562 **References**

563 Adachi N, Robinson M, Goolsbee A, Shubin NH. 2016. Regulatory evolution of Tbx5 and the
564 origin of paired appendages. *Proc Natl Acad Sci* **113:** 10115-10120.

565 Andrews S. 2010. FASTQC: a quality control tool for high throughput sequence data.
566 Available online at: http://www.bioinformatics.babraham.ac.uk/projects/FASTQc

567 Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS.
568 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37:** W202–
569 208.

570 Ballester B, Medina-Rivera A, Schmidt D, Gonzàlez-Porta M, Carlucci M, Chen X,
571 Chessman K, Faure AJ, Funnell AP, Goncalves A, et al. 2014. Multi-species, multi-
572 transcription factor binding highlights conserved control of tissue-specific biological
573 pathways. *eLife* **3:** e02626.

574 Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, James Kent W,
575 Haussler D. 2006. A distal enhancer and an ultraconserved exon are derived from a novel
576 retroposon. *Nature* **441:** 87–90.

577 Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful
578 Approach to Multiple Testing. 1995. *J Royal Stat Soc*, B **57(1)**: 289–300.

579 Berthelot C, Villar D, Horvath J, Odom DT, Flicek P. 2017. Complexity and conservation of
580 regulatory landscapes underlie evolutionary resilience of mammalian gene expression.
581 *Biorxive* doi: https://doi.org/10.1101/125435.

582 Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew J-L, Ruan Y, Wei C-
583 L, Ng HH, et al. 2008. Evolution of the mammalian transcription factor binding repertoire via
584 transposable elements. *Genome Res* **18:** 1752–1762.

585 Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, Gardner K, Hillier LW, Janette
586 J, Jiang L, et al. 2014. Comparative analysis of regulatory information and circuits across
587 distant species. *Nature* **512:** 453–456.

588 Britten RJ, Davidson EH. 1969. Gene Regulation for Higher Cells: A Theory. *Science* **165:**
589 349–357.

590 Brown CD, Johnson DS, Sidow A. 2007. Functional Architecture and Evolution of
591 Transcriptional Elements That Drive Gene Coexpression. *Science* **317:** 1557.

592 Brown CD, Mangravite L, Engelhardt B. 2013. Integrative Modeling of eQTLs and Cis-
593 Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs.
594 *PLoS Genetics* **9:** e1003649.

595 Cain CE, Blekhman R, Marioni JC, Gilad Y. 2011. Gene Expression Differences Among
596 Primates Are Associated With Changes in a Histone Epigenetic Modification. *Genetics* **187:**
597 1225–1234.

598 Chuong EB, Rumi MAK, Soares MJ, Baker JC. 2013. Endogenous retroviruses function as
599 species-specific enhancer elements in the placenta. *Nat Genet* **45:** 325–329.

600 Chuong EB, Elde NC, Feschotte C. 2016. Regulatory evolution of innate immunity through
601 co-option of endogenous retroviruses. *Science* **351:** 1083–1087.

602 Cohen CJ, Lock WM, Mager DL. 2009. Endogenous retroviral LTRs as promoters for human
603 genes: A critical assessment. *Gene* **448:** 105–114.

604 Coolon JD, McManus CJ, Stevenson KR, Graveley BR, Wittkopp PJ. 2014. Tempo and
605 mode of regulatory evolution in Drosophila. *Genome Res* **24***(5)***:** 797–808.

606 Cooper GM, Brown CD. 2008. Qualifying the relationship between sequence conservation
607 and molecular function. *Genome Res* **18:** 201–205.

608 Cotney J, Leng J, Yin J, Reilly SK, DeMare LE, Emera D, Ayoub AE, Rakic P, Noonan JP.
609 2013. The Evolution of Lineage-Specific Regulatory Activities in the Human Embryonic Limb.
610 Cell **154:** 185–196.

611 Davidson E, Britten R. 1979. Regulation of gene expression: possible role of repetitive
612 sequences. *Science* **204:** 1052–1059.

613 DeLaForest A, Nagaoka M, Si-Tayeb K, Noto FK, Konopka G, Battle MA, Duncan SA. 2011.
614 HNF4A is essential for specification of hepatic progenitors from human pluripotent stem
615 cells. *Development* **138:** 4143–4153.

616 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M,
617 Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**(1)**:** 15–21.

618 Dong X, Wang X, Zhang F, Tian W. 2016. Genome-Wide Identification of Regulatory
619 Sequences Undergoing Accelerated Evolution in the Human Genome. *Mol Biol Evol* **33:**
620 2565–2575.

621 Du J, Leung A, Trac C, Lee M, Parks BW, Lusis AJ, Natarajan R, Schones DE. 2016.
622 Chromatin variation associated with liver metabolism is mediated by transposable elements.
623 *Epigen Chrom* **9:** 28.

624 Emera D, Yin J, Reilly SK, Gockley J, Noonan JP. 2016. Origin and evolution of
625 developmental enhancers in the mammalian neocortex. *Proc Natl Acad Sci* **113**: 2617–
626 2626.

627 ENCODE Project Consortium 2012. An integrated encyclopedia of DNA elements in the
628 human genome. *Nature* **489**: 57–74.

629 Fazio G, Palmi C, Rolink A, Biondi A, Cazzaniga G. 2008. PAX5/TEL acts as a
630 transcriptional repressor causing down-modulation of CD19, enhances migration to CXCL12,
631 and confers survival advantage in pre-BI cells. *Cancer Res* **68**(1): 181–189.

632 Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P,
633 Coates G, Fitzgerald S, et al. 2014. Ensembl 2014. *Nucleic Acids Res* **42**: D749–D755.

634 Gittelman RM, Hun E, Ay F, Madeoy J, Pennacchio L, Noble WS, Hawkins RD, Akey JM.
635 2015. Comprehensive identification and analysis of human accelerated regulatory DNA.
636 *Genome Res* **25:** 1245–1255.

637 Grueber CE, Wallis GP, Jamieson IG. 2014. Episodic Positive Selection in the Evolution of
638 Avian Toll-Like Receptor Innate Immunity Genes. *PLoS ONE* **9:** e89632.

639 Guo C, Ludvik AE, Arlotto ME, Hayes MG, Armstrong LL, Scholtens DM, Brown CD,
640 Newgard CB, Becker TC, Layden BT, et al. 2015. Coordinated regulatory variation
641 associated with gestational hyperglycaemia regulates expression of the novel hexokinase
642 HKDC1. *Nat Commun* **6:** 6069.

643 Hancks DC, Kazazian JHH. 2010. SVA retrotransposons: Evolution and genetic instability.
644 *Sem Cancer Biol* **20**: 234–245.

645 Ho MCW, Johnsen H, Goetz SE, Schiller BJ, Bae E, Tran DA, Shur AS, Allen JM, Rau C,
646 Bender W,, et al. 2009. Functional Evolution of cis-Regulatory Modules at a Homeotic Gene
647 in Drosophila. *PLoS Genet* **5**: e1000709.

648 Holloway AK, Bruneau BG, Sukonnik T, Rubenstein JL, Pollard KS. 2016. Accelerated
649 Evolution of Enhancer Hotspots in the Mammal Ancestor. *Mol Biol Evo* **33:** 1008–1018.

650 Huda A, Marino-Ramirez L, Jordan IK. 2010. Epigenetic histone modifications of human
651 transposable elements: genome defense versus exaptation. *Mobile DNA* **1**: 2.

652 Jacques PE, Jeyakani J, Bourque G. 2013. The majority of primate-specific regulatory
653 sequences are derived from transposable elements. *PLoS Genet* **9**: e1003504.

654 Janoušek V, Laukaitis CM, Yanchukov A, Karn R. 2016. The Role of Retrotransposons in
655 Gene Family Expansions in the Human and Mouse Genomes. *Genom Biol Evol*, 8(9), 2632-
656 2650.

657 Jansa AS, Lundrigan LB, Tucker KP. 2003. Tests for Positive Selection on Immune and
658 Reproductive Genes in Closely Related Species of the Murine Genus Mus. *J Mol Evol* **56:**
659 294–307.

660 Jordan IK, Rogozin IB, Glazko GV, Koonin EV. 2003. Origin of a substantial fraction of
661 human regulatory sequences from transposable elements. *Trends Genet* **19**: 68–72.

662 Kaneko S, Bonasio R, Saldaña-Meyer R, Yoshida T, Son J, Nishino K, Umezawa A,
663 Reinberg D. 2013. Interactions between JARID2 and Noncoding RNAs Regulate PRC2
664 Recruitment to Chromatin. *Mol Cell* ***53*****:** 290–300.

665 Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte
666 C. 2013. Transposable Elements Are Major Contributors to the Origin, Diversification, and
667 Regulation of Vertebrate Long Noncoding RNAs. *PLoS Genet* **9:** e1003470.

668 King M, Wilson A. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**:
669 107–116.

670 Komsta L. 2006. Processing data for outliers. *R News*, **6**(2): 10-13.

671 Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G. 2010.
672 Transposable elements have rewired the core regulatory network of human embryonic stem
673 cells. *Nat Genet* **42:** 631–634.

674 Landeen EL, Muirhead CA, Wright L, Meiklejohn CD, Presgraves DC. 2016. Sex
675 Chromosome-wide Transcriptional Suppression and Compensatory Cis-Regulatory
676 Evolution Mediate Gene Expression in the Drosophila Male Germline. *PLoS Biol* **14:**
677 e1002499.

678 Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE,
679 Bickel P, Brown JB, Cayting P, et al. 2012. ChIP-seq guidelines and practices of the
680 ENCODE and modENCODE consortia. *Genome Res* **22:** 1813–1831.

681 Lazzaro BP, Schneider DS. 2014. The Genetics of Immunity. *G3* **4**, 943–945.

682 Lecomte V, Maugnier E, Euthine V, Durand C, Freyssenet D, Nemoz G, Rome S, Vidal H,
683 Lefai E. 2010. A new role for sterol regulatory element binding protein 1 transcription factors
684 in the regulation of muscle mass and muscle cell differentiation. *Mol Cell Biol* **30**(5)**:** 1182–
685 1189.

686 Lesch BJ, Silber SJ, McCarrey JR, Page DC. 2016. Parallel evolution of male germline
687 epigenetic poising and somatic development in animals. *Nat Genet* **48:** 888–894.

688 Lewis JJ, van der Burg KRL, Mazo-Vargas A, Reed, RD. 2016. ChIP-Seq-Annotated
689 Heliconius erato Genome Highlights Patterns of cis-Regulatory Evolution in Lepidoptera. *Cell*
690 *Rep* **16:** 2855–2863.

691 Li H. 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-
692 MEM. arXiv:1303.3997v1

693 Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for
694 assigning sequence reads to genomic features. *Bioinformatics* **30:** 923–930.

695 Liu T. 2014. https://github.com/taoliu/MACS/wiki/Call-differential-binding-events.

696 Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for
697 RNA-seq data with DESeq2. *Genome Biol* **15:** 1–21.

698 Lynch VJ, Leclerc RD, May G, Wagner GP. 2011. Transposon-mediated rewiring of gene
699 regulatory networks contributed to the evolution of pregnancy in mammals. *Nat Genet* **43**:
700 1154–1159.

Lynch VJ, Nnamani MC, Kapusta A, Brayer K, Plaza SL, Mazur EC, Emera D, Sheikh SZ, Grützner F, Bauersachs S, et al. 2015. Ancient Transposable Elements Transformed the Uterine Regulatory Landscape and Transcriptome during the Evolution of Mammalian Pregnancy. *Cell Rep* **10:** 551–561.

Markljung E, Jiang L, Jaffe JD, Mikkelsen TS, Wallerman O, Larhammar M, Zhang X, Wang L, Saenz-Vash V, Gnirke A, et al. 2009. ZBED6, a Novel Transcription Factor Derived from a Domesticated DNA Transposon Regulates IGF2 Expression and Muscle Growth. *PLoS Biol* **7:** e1000256.

Marnetto D, Molineris I, Grassi E, Provero P. 2014. Genome-wide Identification and Characterization of Fixed Human-Specific Regulatory Regions. *Am J Hum Genet* **95**: 39–48.

Martin A, Reed RD. 2014. Wnt signaling underlies evolution and development of the butterfly wing pattern symmetry systems. *Dev Biol* **395**: 367–378.

Martin A, Papa R, Nadeau NJ, Hill RI, Counterman BA, Halder G, Jiggins CD, Kronforst, MR, Long AD, McMillan WO, et al. 2012. Diversification of complex butterfly wing patterns by repeated regulatory evolution of a Wnt ligand. *Proc Natl Acad Sci* **109:** 12632–12637.

Maurer P, Tsas F, Coutte L, Callens N, Brenner C, Van Lint C, de Launoit Y, Baert JL. 2003. FEV acts as a transcriptional repressor through its DNA-binding ETS domain and alanine-rich domain. *Oncogene* **22**(21): 3319–3329.

McClintock B. 1950. The Origin and Behavior of Mutable Loci in Maize. *Proc Natl Acad Sci* **36**: 344–355.

McClintock B. 1984. The significance of responses of the genome to challenge. *Science* **226:** 792–801.

723 Melnikov A, Murugan A., Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan
724 CG, Kinney JB, et al. 2012. Systematic dissection and optimization of inducible enhancers in
725 human cells using a massively parallel reporter assay. *Nat Biotech* **30:** 271–277.

726 Obbard DJ, Welch JJ, Kim KW, and Jiggins FM. 2009. Quantifying Adaptive Evolution in the
727 Drosophila Immune System. *PLoS Genet* **5**: e1000698.

728 Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee S.-
729 I., Cooper GM, et al. (2012). Massively parallel functional dissection of mammalian
730 enhancers in vivo. *Nat Biotech* **30**: 265–270.

731 Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MAM, Kessing B,
732 Pontius J, Roelke M, Rumpler Y, et al. 2011. A Molecular Phylogeny of Living Primates.
733 *PLoS Genet* **7**: e1001342.

734 Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral
735 substitution rates on mammalian phylogenies. *Genome Res* **20**: 110–121.

736 Prabhakar S, Visel A, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Morrison H,
737 FitzPatrick DR, Afzal V, et al. 2008. Human-Specific Gain of Function in a Developmental
738 Enhancer. *Science* **321**: 1346–1350.

739 Prescott SL, Srinivasan R, Marchetto MC, Grishina I, Narvaiza I, Selleri L, Gage FH, Swigut
740 T, Wysocka J. 2015. Enhancer Divergence and cis-Regulatory Evolution in the Human and
741 chimpanzee Neural Crest. *Cell* **163**: 68–83.

742 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic
743 features. *Bioinformatics* **26**: 841–842.

744 R Core Team 2016. R: A language and environment for statistical computing. R. Foundation
745 for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

746 Rayan NA, Del Rosario RCH, Prabhakar S. 2016. Massive contribution of transposable
747 elements to mammalian regulatory sequences. *Semin Cell Dev Biol* **57**: 51–56.

748 Reilly SK, Yin J, Ayoub AE, Emera D, Leng J, Cotney J, Sarro R, Rakic P, Noonan JP. 2015.
749 Evolutionary changes in promoter and enhancer activity during human corticogenesis.
750 *Science* **347**: 1155–1159.

751 Roadmap Epigenomic Mapping Consortium. 2015. Integrative analysis of 111 reference
752 human epigenomes. *Nature* **518**: 317–330.

753 del Rosario RCH, Rayan NA, Prabhakar S. 2014. Noncoding origins of anthropoid traits and
754 a new null model of transposon functionalization. *Genome Res* **24**: 1469–1484.
755 Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG. 2007. Dynamic
756 evolution of the innate immune system in Drosophila. *Nat Genet* **39**: 1461–1468.

757 Salazar-Jaramillo L, Paspati A, van de Zande L, Vermeulen CJ, Schwander T, Wertheim B.
758 2014. Evolution of a cellular immune response in Drosophila: a phenotypic and genomic
759 comparative analysis. *Genome Biol Evol* **6**: 273–289.

760 Sasaki T, Nishihara H, Hirakawa M, Fujimura K, Tanaka M, Kokubo N, Kimura-Yoshida C,
761 Matsuo I, Sumiyama K, Saitou N, et al. 2008. Possible involvement of SINEs in mammalian-
762 specific brain formation. *Proc Natl Acad Sci* **105**: 4220–4225.

763 Schadt EE. 2009. Molecular networks as sensors and drivers of common human diseases.
764 *Nature* **461**: 218–223.

765 Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S,
766 Martinez-Jimenez CP, Mackay S, et al. 2010. Five-Vertebrate ChIP-seq Reveals the
767 Evolutionary Dynamics of Transcription Factor Binding. *Science* **328**: 1036.

768 Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves Â, Kutter C, Brown GD,
769 Marshall A, Flicek P, Odom DT. 2012. Waves of Retrotransposon Expansion Remodel

770 Genome Organization and CTCF Binding in Multiple Mammalian Lineages. *Cell* **148**: 335–
771 348.

772 Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z,
773 Weinberger A, Segal E. 2012. Inferring gene regulatory logic from high-throughput
774 measurements of thousands of systematically designed promoters. *Nat Biotech* **30**: 521–
775 530.

776 Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth
777 J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect,
778 worm, and yeast genomes. *Genome Res* **15**: 1034–1050.

779 Sironi M, Cagliani R, Forni D, Clerici M. 2015. Evolutionary insights into host-pathogen
780 interactions from mammalian sequence data. *Nat Rev Genet* **16**: 224–236.

781 Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, Arnaiz O, Awedh MH,
782 Baldock R, Barbiera G, et al. 2015. The BioMart community portal: an innovative alternative
783 to large, centralized data repositories. *Nucleic Acids Res* **43**: W589–W598.

784 Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, and Wang T. 2014.
785 Widespread contribution of transposable elements to the innovation of gene regulatory
786 networks. *Genome Res* **24**: 1963–1976.

787 Takata A, Iwayama Y, Fukuo Y, Ikeda M, Okochi T, Maekawa M, Toyota T, Yamada K,
788 Hattori E, Ohnishi T, et al. 2013. A Population-Specific Uncommon Variant in GRIN3A
789 Associated with Schizophrenia. *Biol Psychiatry* **73**(6): 532–539.

790 Tsankov AM, Thompson DA, Socha A, Regev A, Rando OJ. 2010. The Role of Nucleosome
791 Positioning in the Evolution of Gene Regulation. *PLoS Biol* **8**: e1000414.

792 Vallender EJ, and Lahn BT. 2004. Positive selection on the human genome. *Human Mol*
793 *Genet* **13**: R245–R254.

794 Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R,
795 Erichsen JT, Jasinska AJ, et al. Enhancer Evolution across 20 Mammalian Species. *Cell*
796 **160**: 554–566.

797 Wallerman O, Motallebipour M, Enroth A, Patra K, Bysani MSR, Komorowski J, Wadelius C.
798 2009. Molecular interactions between HNF4a, FOXA2 and GABP identified at regulatory
799 DNA elements through ChIP-sequencing. *Nucleic Acids Res* **37**: 7498–7508.

800 Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, Burgess SM, Brachmann RK,
801 Haussler D. 2007. Species-specific endogenous retroviruses shape the transcriptional
802 network of the human tumor suppressor protein p53. *Proc Natl Acad Sci* **104**: 18613–18618.

803 Ward M, Zhao S, Luo K, Pavlovic B, Karimi MM, Stephens M, Gilad Y. 2017. Silencing of
804 transposable elements may not be a major driver of regulatory evolution in primate induced
805 pluripotent stem cells. *Biorxive* doi: http://dx.doi.org/10.1101/142455.

806 Warner LR, Babbitt CC, Primus AE, Severson TF, Haygood R, Wray GA. 2009. Functional
807 consequences of genetic variation in primates on tyrosine hydroxylase (TH) expression in
808 vitro. *Brain Res* **1288**: 1–8.

809 Wertheim B. 2015. Genomic basis of evolutionary change: evolving immunity. *Front Genet* **6**:
810 222.

811 Wickham H. 2009. ggplot2: Elegant graphics for data analysis. Springer-Verlag New York,
812 2009.

813 Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* **8**:
814 206–216.

815 Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, Zhou X, Lee HJ, Maire CL, Ligon KL, et
816 al. 2013. DNA hypomethylation within specific transposable element families associates with
817 tissue-specific enhancer landscape. *Nat Genet* **45**: 836–841.

Yang S, Oksenberg N, Takayama S, Heo S-J, Poliakov A, Ahituv N, Dubchak I, Boffelli D. 2015. Functionally conserved enhancers with divergent sequences in distant vertebrates. *BMC Genomics* **16**: 1–13.

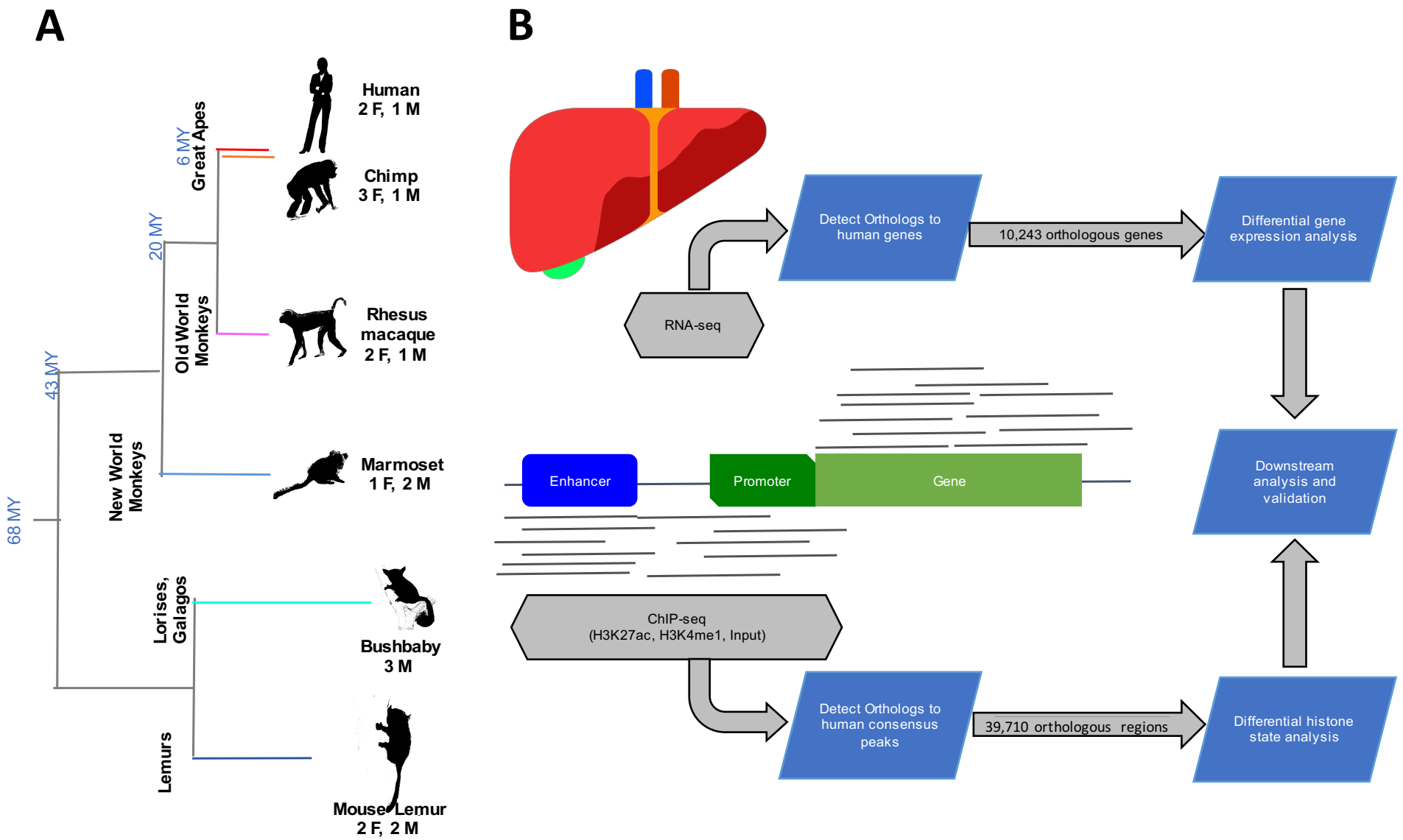
Yates A, Beal K, Keenan S, McLaren W, Pignatelli M, Ritchie GRS, Ruffier M, Taylor K, Vullo A, Flicek P. 2015. The Ensembl REST API: Ensembl Data for Any Language. *Bioinformatics* **31**: 143–145.

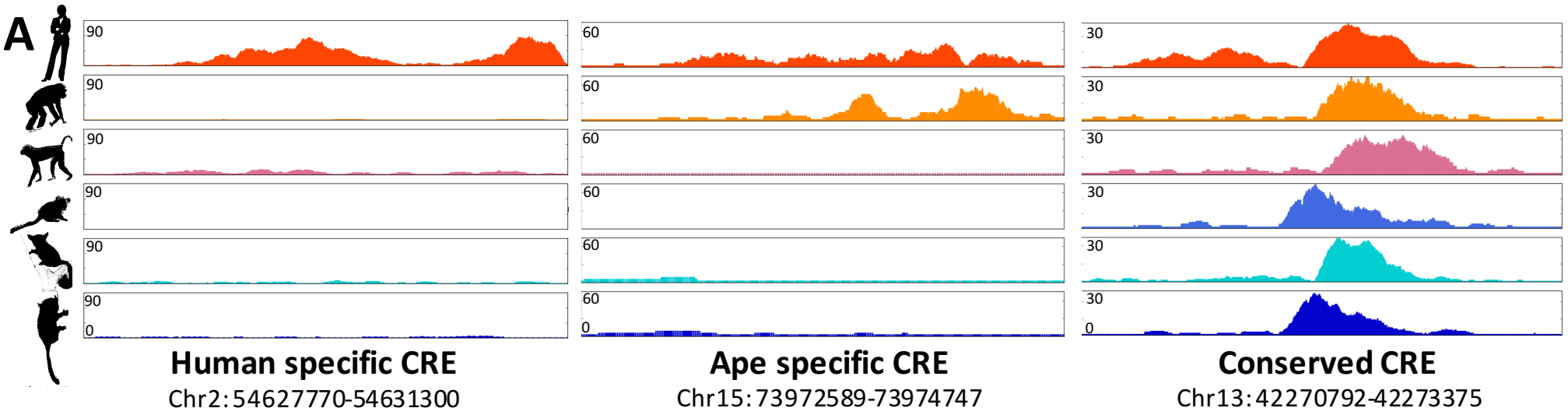
Zak DE, Tam VC, Aderem A. 2014. Systems-Level Analysis of Innate Immunity. *Annu Rev Immunol* ***32***: 547–577.

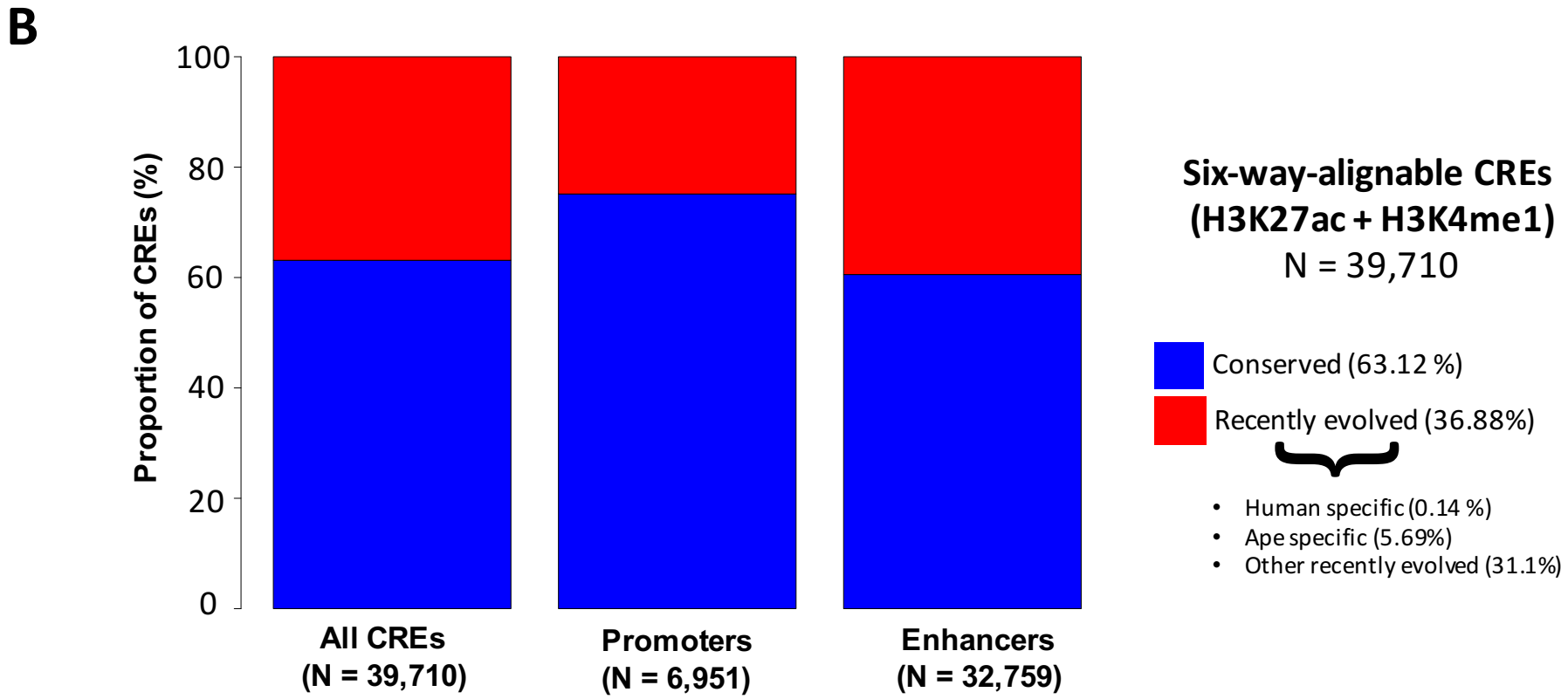
Zhang L, Reed RD 2016. Genome editing in butterflies reveals that spalt promotes and Distal-less represses eyespot colour patterns. *Nat Commun* **15***: 11769*.

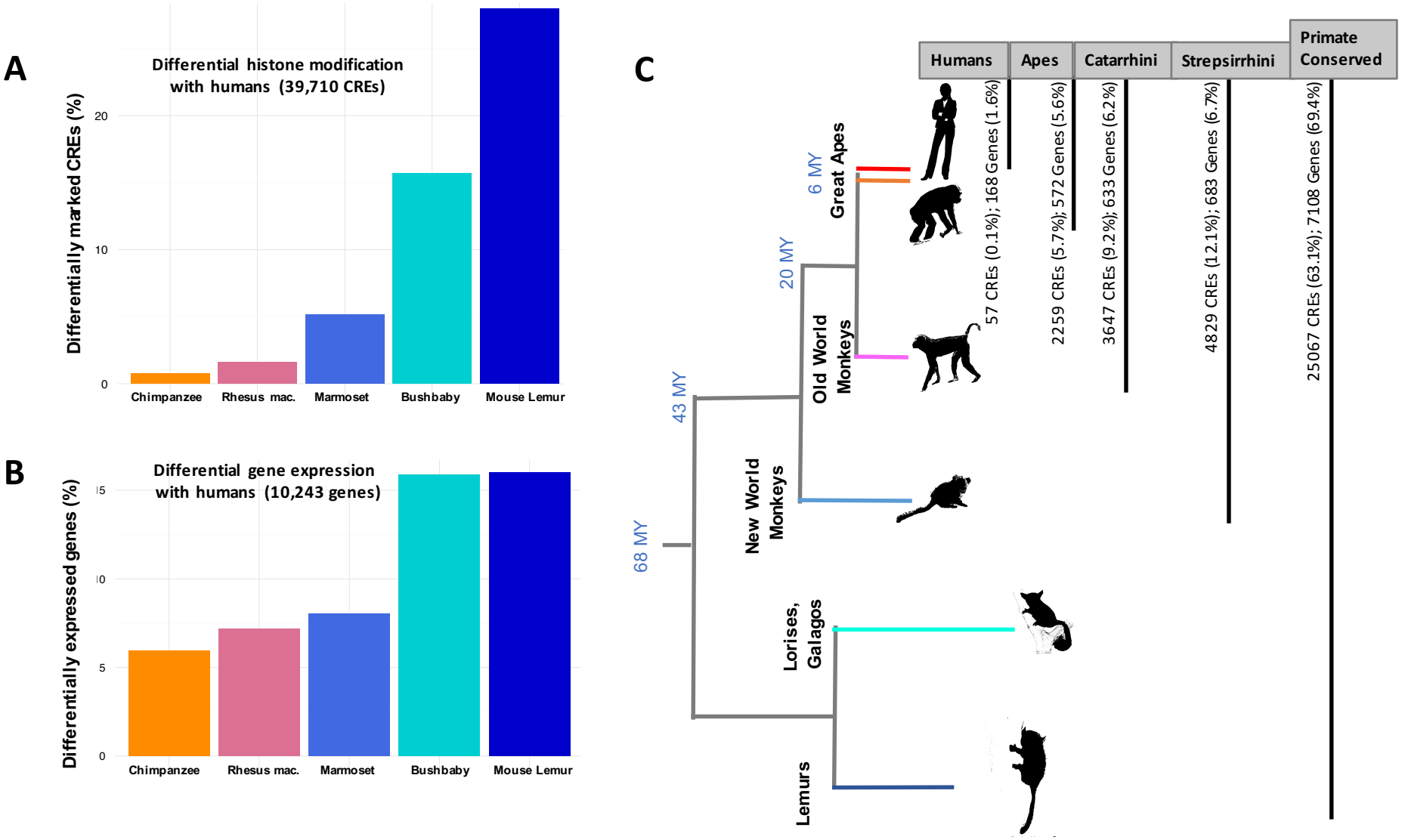
Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* **9**: 1–9.

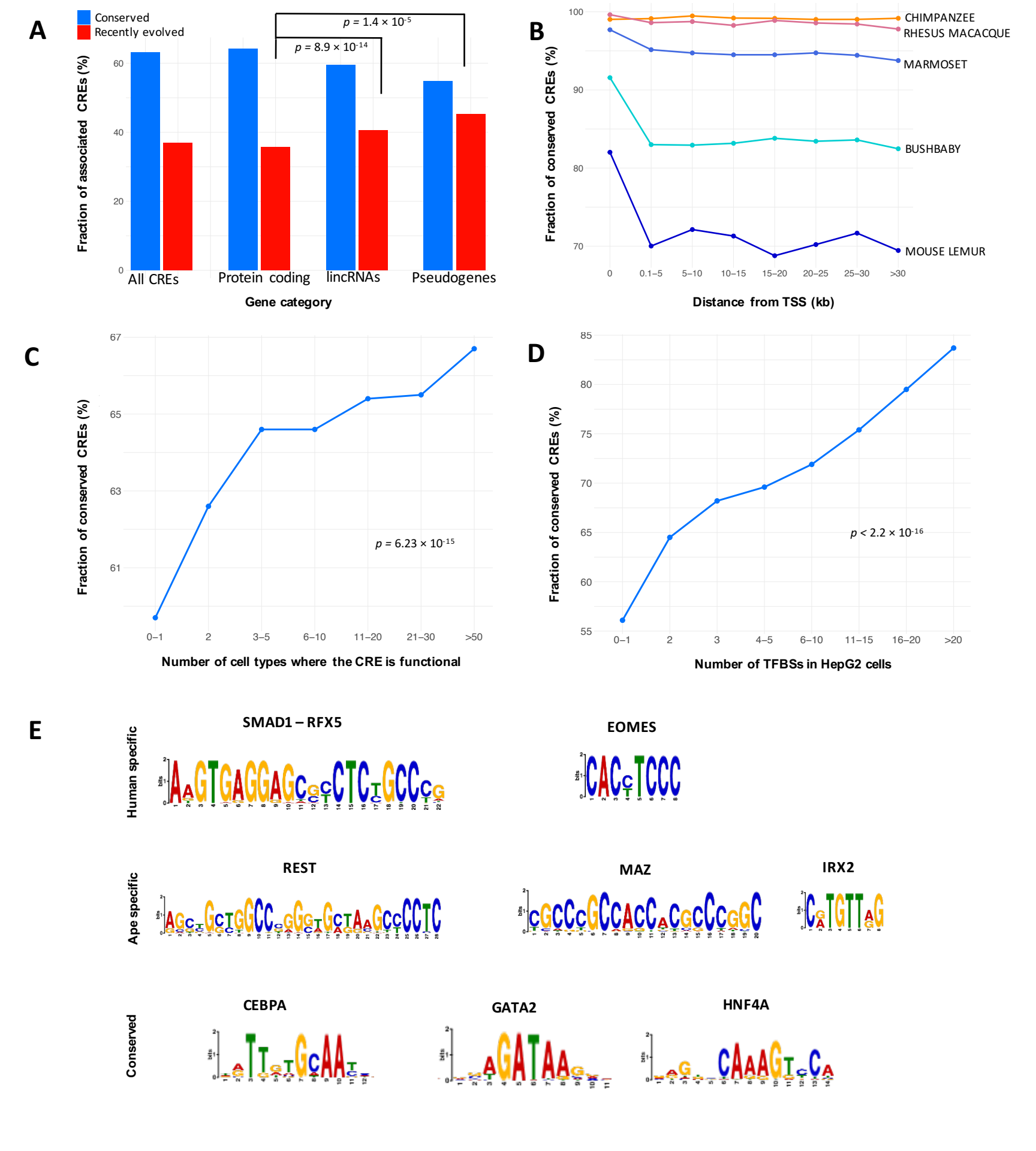
Zhou X, Cain CE, Myrthil M, Lewellen N, Michelini K, Davenport ER, Stephens M, Pritchard JK, Gilad Y 2014. Epigenetic modifications are associated with inter-species gene expression variation in primates. *Genome Biol* **15**: 1–19.

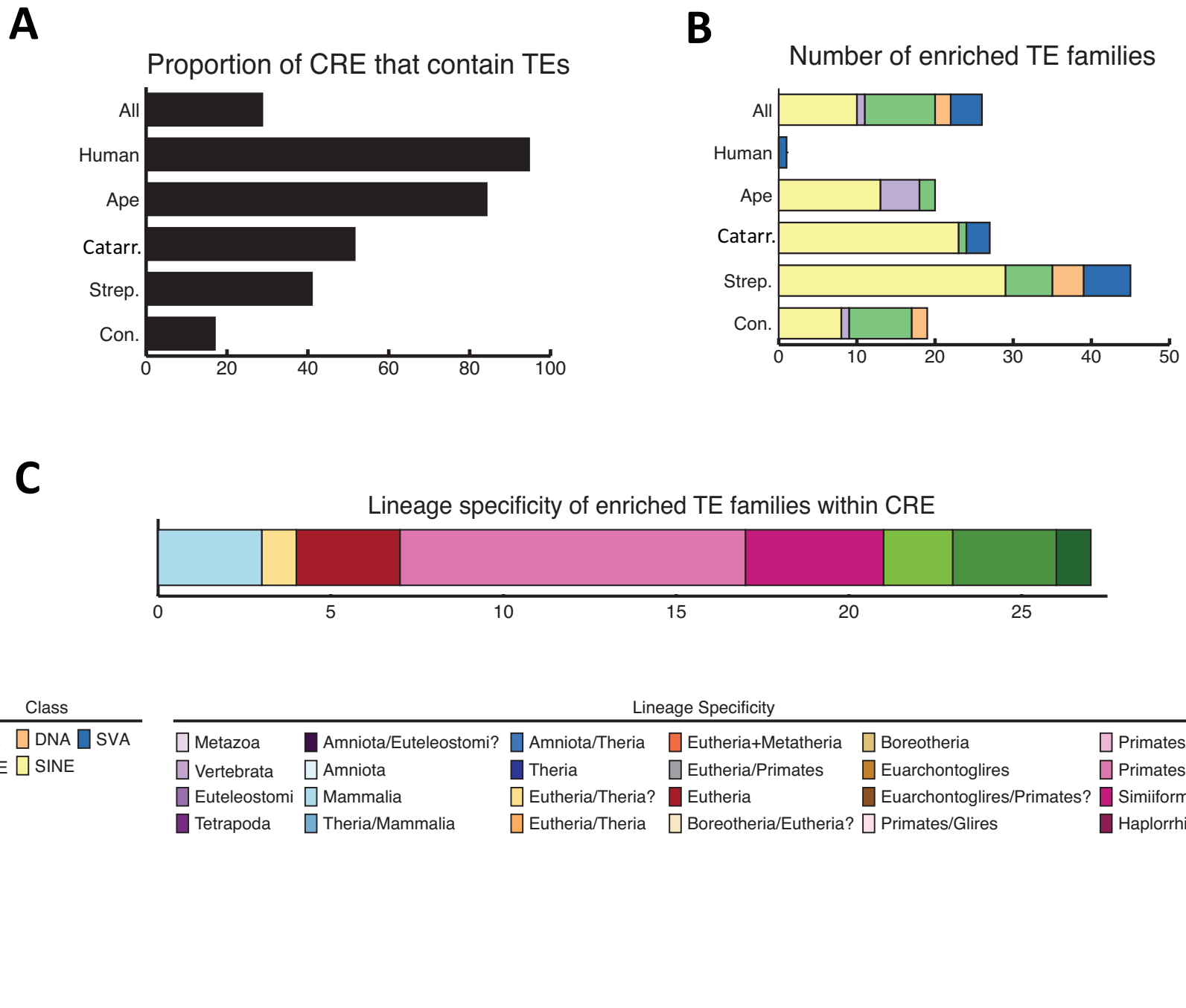

A
Human
2 F, 1 M
Chimp
3 F, 1 M
6 MY
Great Apes
20 MY
Old World
Monkeys
Rhesus
macaque
2 F, 1 M
43 MY
New World
Monkeys
Marmoset
1 F, 2 M
68 MY
Lorises,
Galagos
Bushbaby
3 M
Lemurs
Mouse Lemur
2 F, 2 M
B
Detect Orthologs to
human genes
10,243 orthologous genes
Differential gene
expression analysis
RNA-seq
Enhancer
Promoter
Gene
Downstream
analysis and
validation
ChIP-seq
(H3K27ac, H3K4me1, Input)
Detect Orthologs to
human consensus
peaks
39,710 orthologous regions
Differential histone
state analysis

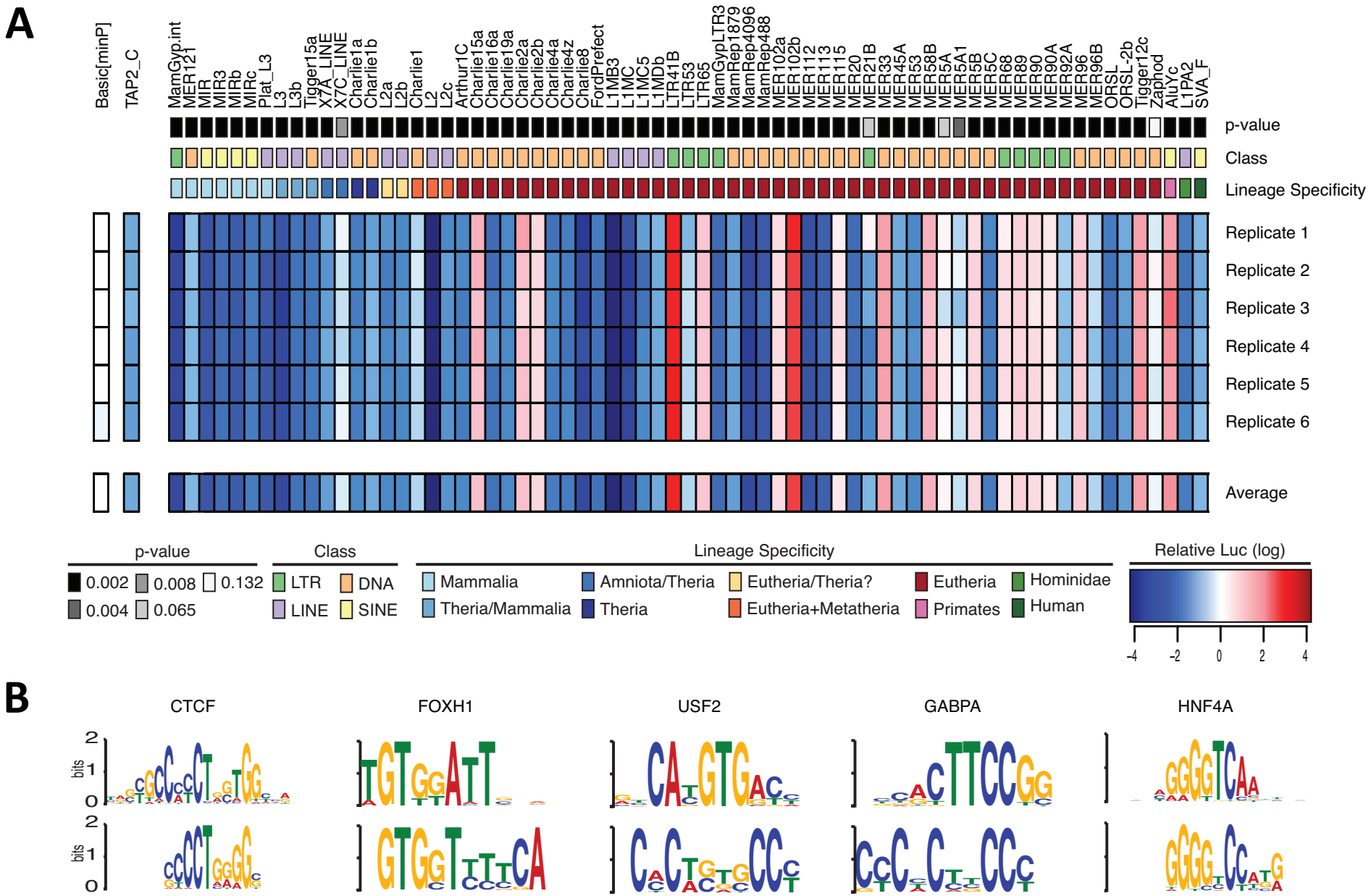

A
90
0
60
0
30
0
Human specific CRE
Chr2:54627770-54631300
Ape specific CRE
Chr15:73972589-73974747
Conserved CRE
Chr13:42270792-42273375


B
Proportion of CREs (%)
100
80
60
40
20
0
All CREs
(N = 39,710)
Promoters
(N = 6,951)
Enhancers
(N = 32,759)
Six-way-alignable CREs
(H3K27ac + H3K4me1)
N = 39,710
Conserved (63.12 %)
Recently evolved (36.88%)
Human specific (0.14 %)
Ape specific (5.69%)
Other recently evolved (31.1%)


A
Differential histone modification with humans (39,710 CREs)
Differentially marked CREs (%)
20
10
0
Chimpanzee
Rhesus mac.
Marmoset
Bushbaby
Mouse Lemur
B
Differential gene expression with humans (10,243 genes)
Differentially expressed genes (%)
15
10
5
0
Chimpanzee
Rhesus mac.
Marmoset
Bushbaby
Mouse Lemur
C
Humans
Apes
Catarrhini
Strepsirrhini
Primate Conserved
57 CREs (0.1%); 168 Genes (1.6%)
2259 CREs (5.7%); 572 Genes (5.6%)
3647 CREs (9.2%); 633 Genes (6.2%)
4829 CREs (12.1%); 683 Genes (6.7%)
25067 CREs (63.1%); 7108 Genes (69.4%)
6 MY
Great Apes
20 MY
Old World Monkeys
43 MY
New World Monkeys
68 MY
Lorises, Galagos
Lemurs


A
Conserved
Recently evolved
p = 1.4 × 10-5
p = 8.9 × 10-14
Fraction of associated CREs (%)
0
20
40
60
All CREs
Protein coding
lincRNAs
Pseudogenes
Gene category
B
CHIMPANZEE
RHESUS MACACQUE
MARMOSET
BUSHBABY
MOUSE LEMUR
Fraction of conserved CREs (%)
70
80
90
100
0
0.1–5
5–10
10–15
15–20
20–25
25–30
>30
Distance from TSS (kb)
C
Fraction of conserved CREs (%)
61
63
65
67
p = 6.23 × 10-15
0–1
2
3–5
6–10
11–20
21–30
>50
Number of cell types where the CRE is functional
D
Fraction of conserved CREs (%)
55
60
65
70
75
80
85
p < 2.2 × 10-16
0–1
2
3
4–5
6–10
11–15
16–20
>20
Number of TFBSs in HepG2 cells
E
Human specific
SMAD1 – RFX5
EOMES
Ape specific
REST
MAZ
IRX2
Conserved
CEBPA
GATA2
HNF4A


A
Proportion of CRE that contain TEs
All
Human
Ape
Catarr.
Strep.
Con.
0
20
40
60
80
100
B
Number of enriched TE families
All
Human
Ape
Catarr.
Strep.
Con.
0
10
20
30
40
50
C
Lineage specificity of enriched TE families within CRE
0
5
10
15
20
25
Class
LTR
DNA
SVA
LINE
SINE
Lineage Specificity
Metazoa
Vertebrata
Euteleostomi
Tetrapoda
Amniota/Euteleostomi?
Amniota
Mammalia
Theria/Mammalia
Amniota/Theria
Theria
Eutheria/Theria?
Eutheria/Theria
Eutheria+Metatheria
Eutheria/Primates
Eutheria
Boreotheria/Eutheria?
Boreotheria
Euarchontoglires
Euarchontoglires/Primates?
Primates/Glires
Primates/Indep
Primates
Simiiformes
Haplorrhini
Catarrhini
Hominoidea
Hominidae
Human


A
Basic[minP]
TAP2_C
MamGyp.int
MER121
MIR
MIR3
MIRb
MIRc
Plat_L3
L3
L3b
Tigger15a
X7A_LINE
X7C_LINE
Charlie1a
Charlie1b
L2a
L2b
Charlie1
L2
L2c
Arthur1C
Charlie15a
Charlie16a
Charlie19a
Charlie2a
Charlie2b
Charlie4a
Charlie4z
Charlie8
FordPrefect
L1MB3
L1MC
L1MC5
L1MDb
LTR41B
LTR53
LTR65
MamGypLTR3
MamRep1879
MamRep4096
MamRep488
MER102a
MER102b
MER112
MER113
MER115
MER20
MER21B
MER33
MER45A
MER53
MER58B
MER5A
MER5A1
MER5B
MER5C
MER68
MER89
MER90
MER90A
MER92A
MER96
MER96B
ORSL
ORSL-2b
Tigger12c
Zaphod
AluYc
L1PA2
SVA_F
p-value
Class
Lineage Specificity
Replicate 1
Replicate 2
Replicate 3
Replicate 4
Replicate 5
Replicate 6
Average
p-value
0.002
0.008
0.132
0.004
0.065
Class
LTR
DNA
LINE
SINE
Lineage Specificity
Mammalia
Amniota/Theria
Eutheria/Theria?
Eutheria
Hominidae
Theria/Mammalia
Theria
Eutheria+Metatheria
Primates
Human
Relative Luc (log)
-4
-2
0
2
4
B
CTCF
FOXH1
USF2
GABPA
HNF4A
bits
2
1
0