Transposable element exaptation is the primary source of novelty in the primate gene regulatory landscape

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

1. Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104
2. Department of Human Genetics, University of Chicago, Chicago, IL, 60637
3. Departments of Anthropology and Biology, Pennsylvania State University, University Park, PA, 16802

* Co-first author
# Correspondence: chrbro@upenn.edu (C.D.B.), marco.trizzino83@gmail.com (M.T.)

Abstract

Gene regulation plays a critical role in the evolution of phenotypic diversity. We investigated the evolution of liver promoters and enhancers in six primate species. We performed ChIP-seq for two histone modifications and RNA-seq to profile cis-regulatory element (CRE) activity and gene expression. The primate regulatory landscape is largely conserved across the lineage. Conserved CRE function is associated with sequence conservation, proximity to coding genes, cell type specificity of CRE function, and transcription factor binding. Newly evolved CREs are enriched in immune response and neurodevelopmental functions, while conserved CREs bind master regulators. Transposable elements (TEs) are the primary source of novelty in primate gene regulation. Newly evolved CREs are enriched in young TEs that affect gene expression. However, only 17% of conserved CREs overlap a TE, suggesting that target gene expression is under strong selection. Finally, we identified specific genomic features driving the functional recruitment of newly inserted TEs.

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

Introduction

The evolution of cis-regulatory elements (CREs) plays an important role in phenotypic and behavioral evolution (King and Wilson, 1975; Rockman et al., 2005; Loisel et al., 2006; Pollard et al., 2006; Prabhakar et al., 2008; Warner et al., 2009; Babbitt et al., 2010; Cain et al., 2011; Marnetto et al., 2014; Zhou et al., 2014; Villar et al., 2015). Many aspects of CRE evolution in mammals have been characterized (Schmidt et al., 2010; Cain et al., 2011; Zhou et al., 2014; Prescott et al., 2015; Reilly et al., 2015; Villar et al., 2015; Emera et al., 2016), and suggest a role for transposable elements (TEs) in the evolution of gene regulation (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; Chuong et al., 2013, 2016; Schmidt et al., 2012; Xie et al., 2013; del Rosario et al., 2014; Sundaram et al., 2014; Du et al., 2016; Rayan et al., 2016). However, validating the functional contribution of TEs in the mammalian gene regulation remains a challenge. Lynch et al. (2011, 2015) demonstrated that the recruitment of novel regulatory networks in the uterus was likely mediated by ancient mammalian TEs. However, Emera and colleagues (2016) suggested that neocortical enhancers do not exhibit strong evidence of transposon exaptation.

Many important questions remain unanswered: to what extent are poised and active regulatory elements functionally conserved? Are specific genomic features predictive of CRE conservation? To what extent have TEs driven the evolution of gene regulation? And finally, what determines which TE insertions are recruited as functional CREs? Establishing answers to these questions is critical for understanding how the evolution of regulatory elements contributes to the conservation and divergence of gene expression and complex traits.

With a goal of answering these questions, we collected liver samples from six primate species. Core liver functions are largely conserved across primate species. However, different environmental exposures, diets, and lifestyles likely directed the adaptation of liver functions, and associated regulatory evolution, making this tissue an optimal model in which explore the conservation and divergence of the gene regulation.

To characterize primate liver cis-regulatory evolution, we performed chromatin immunoprecipitation followed by sequencing (ChIP-seq) for Histone H3 Lysine 27 acetylation (H3K27ac), which marks active enhancers and promoters, and Histone H3 Lysine 4 mono-methylation (H3K4me1), which marks poised regulatory elements on liver tissues from six primate species, including at least one species from each major primate clade to maximize evolutionary diversity within primates (Perelman et al., 2012). We generated whole transcriptome sequencing (RNA-seq) data from the same specimens to quantify gene expression variation across species. We estimated the degree of evolutionary conservation of regulatory activity and gene expression levels across the entire lineage, and characterized the genomic features associated with evolutionary conservation of gene regulation, to understand why some enhancers and promoters are conserved across species whereas others are subject to rapid turnover.

The activity of the majority of human CREs is conserved across the entire primate lineage, and the differences in gene expression and regulation reflect the phylogenetic distance between species. Conservation of cis-regulatory activity is associated with nucleotide sequence conservation, gene function, gene distance, cell-specificity of CRE function, and transcription factor binding site (TFBS) density. Strikingly, human- and ape-specific enhancers and promoters are significantly enriched for evolutionarily young TEs. In particular, the majority of human- and ape-specific CREs are derived from SINE-VNTR-*Alus* (SVAs) and Long-Terminal-Repeats (LTRs), respectively. On the other hand, only a minor fraction of evolutionarily conserved CREs are derived from TEs, indicating that purifying selection on the associated genes likely preserve these CREs from being disrupted by TE insertions, thus conserving the expression of the associated genes. In addition, we characterized SVAs that evolved into regulatory elements, and estimated potential impacts of these SVAs on gene regulation across the lineage. We validated the regulatory activity of several TE families, and conclude with a new model describing specific genomic features that strengthen the potential adaptation of TEs into functional regulatory elements (exaptation; Brosius and Gould; 1992; de Souza et al., 2013).

Results

Data generation, quality assessment, and validation

We generated a total of 757 million RNA-seq reads and 1.70 billion ChIP-seq reads (H3K27ac, H3K4me1, and input) 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; Table S1). The six species were selected to include at least one species from each of the major primate clades, thus maximizing phylogenetic diversity within primates. The samples were all from young adults and, with the exception of the bushbaby, included both males and females. In total, 18 RNA-seq and 14 ChIP-seq samples remained post-QC and were used for analyses (Table S1). On average, we sequenced 42.1 million RNA-seq reads and 40.6 million ChIP-seq reads per sample (Table S1). We applied stringent quality control (QC) measures to assess library construction, sequencing, and peak-calling methods. Read mappability after the filtering steps was consistent across species and assays (77% in ChIP-seq, 78% in RNA-seq; Table S1), suggesting that differences in genome assembly quality do not introduce large biases.

ChIP-seq data were used to map the distribution of active (H3K27ac) and poised (H3K4me1) CREs in the six primate genomes (Fig. 1). We mapped ChIP-seq reads to their respective reference genomes with BWA-mem (Li, 2013) and identified regions of significant H3K27ac and H3K4me1 enrichment in the human liver, treating all human individuals as replicates in the peak calling procedure with MACS2 (Zhang et al., 2008). 85.6% of the regions marked by H3K4me1 overlapped regions marked by H3K27ac. A total of 84,253 human peaks remained after merging overlapping peaks from the two histone markers. Next we identified regions orthologous to the human consensus peaks from the genomes of non-human primates using the Ensembl multiple sequence alignment (MSA) database. We catalogued 47,673 total human CREs with orthologs in all six species: 40,527 enhancers (distance from transcription start site (TSS) > 1 kb) and 7,146 promoters (distance from TSS ≤ 1 kb).

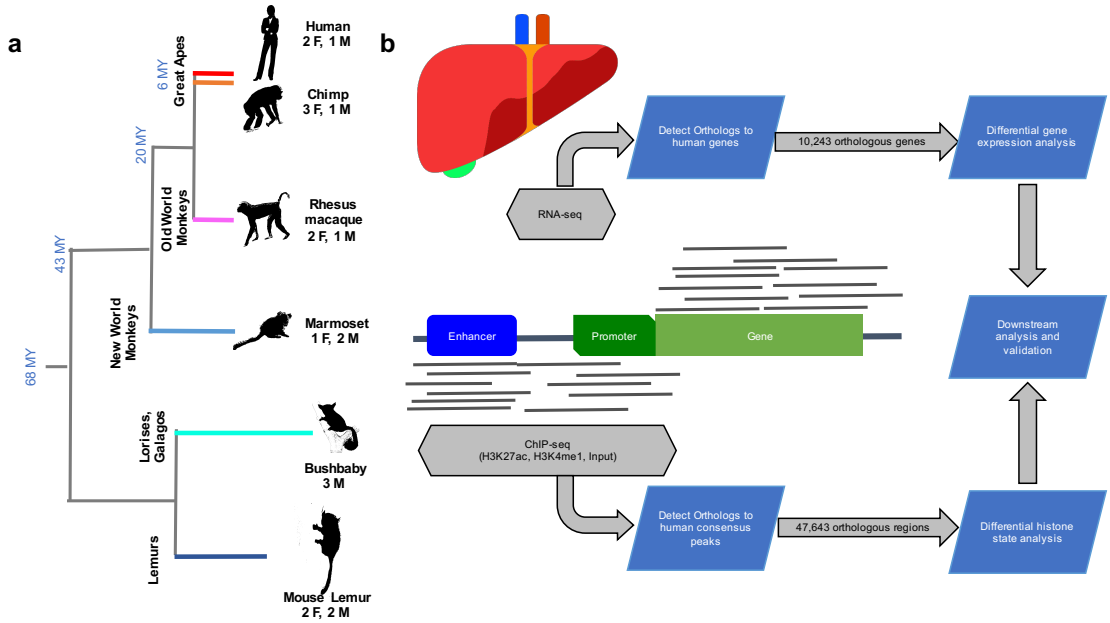


Figure 1 - Experimental design and analytical pipeline. (A) Sampling included three to four specimens from six species representing all of the major primate clades. (B) ChIP-seq and RNA-seq profiles were produced from the liver samples. Differential histone modification and gene expression analyses were performed on the orthologous CREs and genes in each species, respectively. Outputs shown in the diagram were used for downstream analyses and validations.

Several lines of evidence indicate that the regions of histone modification we have identified represent active CREs. First, 99.0% and 73.3% of ENCODE HepG2 H3K27ac and H3K4me1 regions (ENCODE Project Consortium, 2012), respectively, overlapped with one or more human peaks. Similarly, 63.2% of the 43,011 permissive enhancers that were predicted by the FANTOM5 Consortium based on enhancer RNA expression (Andersson et al., 2014) overlapped with one or more of human peaks identified in our study. Further, the promoters of 98.1% of genes expressed in the liver overlapped a region of histone modification (Fig. S1). Finally we compared our human H3K27ac data to a recent study focused on liver CREs in mammals (Villar et al., 2015) and demonstrated that peaks bearing signatures of robust and broad regulatory activities are largely reproducible across studies, despite variation attributable to different study designs (Fig. S2).

ChIP-seq experiments can identify regions of the genome bound by histones and other proteins that characterize regulatory elements, however, this does not guarantee these regions are functional regulatory elements (Pickrell et al., 2011;

156 Cusanovic et al., 2014; Jain et al., 2015). Therefore, we used a novel parallelized
157 reporter assay (Melnikov et al., 2012; Patwardhan et al., 2012; Sharon et al., 2012;
158 see methods) to validate the regulatory function of predicted human liver CREs.
159 Specifically, we tested the regulatory activity of 1 kb fragments from 122 putative
160 regulatory elements in HepG2 cells, including both evolutionarily conserved and
161 recently evolved CREs (see below). Among 122 tested elements, 79 drove
162 significant reporter gene expression levels (42/53 enhancers [79.2%] and 36/69
163 promoters [52.2%]; Table S6), suggesting that the majority of the CREs predicted in
164 our study based on the enrichment of active histone modification states are likely
165 functional regulatory elements in the human liver.
166
167 **The majority of human CREs are functionally conserved across primates**
168 After extracting ChIP-seq read counts for the six species from 47,673 orthologous
169 regions, we assessed evidence of differential histone modification between species
170 with DESeq2 (Love et al., 2014), using the ChIP-seq input data as a covariate. We
171 compared ChIP-seq read counts in the 47,673 regions by means of all possible
172 human-centric species × species and group × group pairwise comparisons (see
173 methods). This approach provides a quantitative assessment of histone modification
174 profiles across species, while avoiding issues arising from many potential
175 experimental variables that may confound peak calling (Waszak et al., 2015). A
176 specific analysis of human and marmoset, the latter being the species with the
177 smallest number of peaks called in this study (Supplemental File S1), strongly
178 supported the validity and robustness of our approach (Fig. S3).
179 The majority of the 47,673 human CREs (63.8%) did not show significant
180 differential histone modifications in any of the tested pairwise comparisons (FDR <
181 10%; Fig. 2). This suggests that these regulatory regions are consistently active
182 across the primate lineage and thus represent evolutionarily conserved primate
183 CREs. Although the absence of differential histone modifications in a pairwise
184 comparison between two species is not a direct proof of CRE conservation, we
185 demonstrated that the selected FDR threshold does not affect downstream
186 conclusions (Table S7). As an additional control, we performed a chimpanzee-centric
187 analysis for the regions orthologous to chimpanzee consensus peaks (hereafter,
188 chimpanzee CREs), and demonstrated that 62.5% of these regions were not
189 differentially histone modified in any of the pairwise comparisons. This observation is

consistent with 63.8% conserved CREs identified in the human-centric analysis, indicating that the differential histone modification analysis is robust and species-specific bias is unlikely.

Promoters are significantly more conserved than enhancers (68.9% and 62.8%, respectively; Fisher's exact test $p < 2.2\times10^{-16}$; Fig. 2), as observed previously (Villar et al., 2015). On the other hand, 36.2% of orthologous CREs exhibited differential histone modification state across species (Fig. 2). We detected 57 human-specific CREs (0.13%; Fig. 3) and 544 ape-specific CREs (1.42%; Fig. 3). 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; Reilly et al., 2015; Emera et al., 2016).

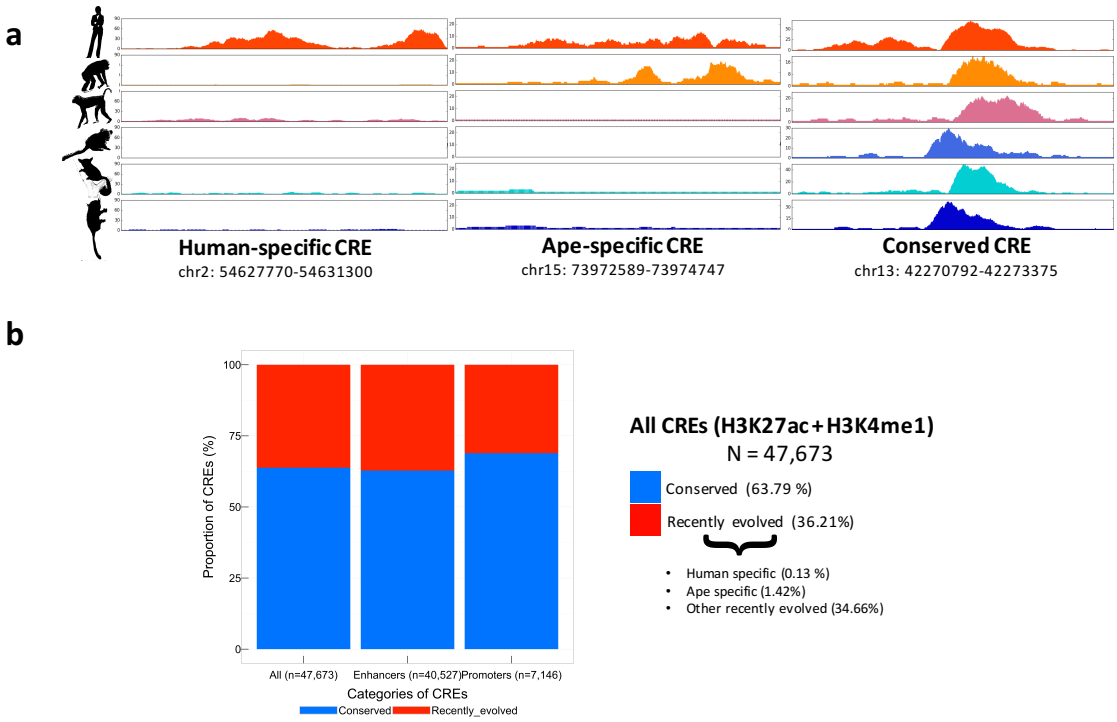
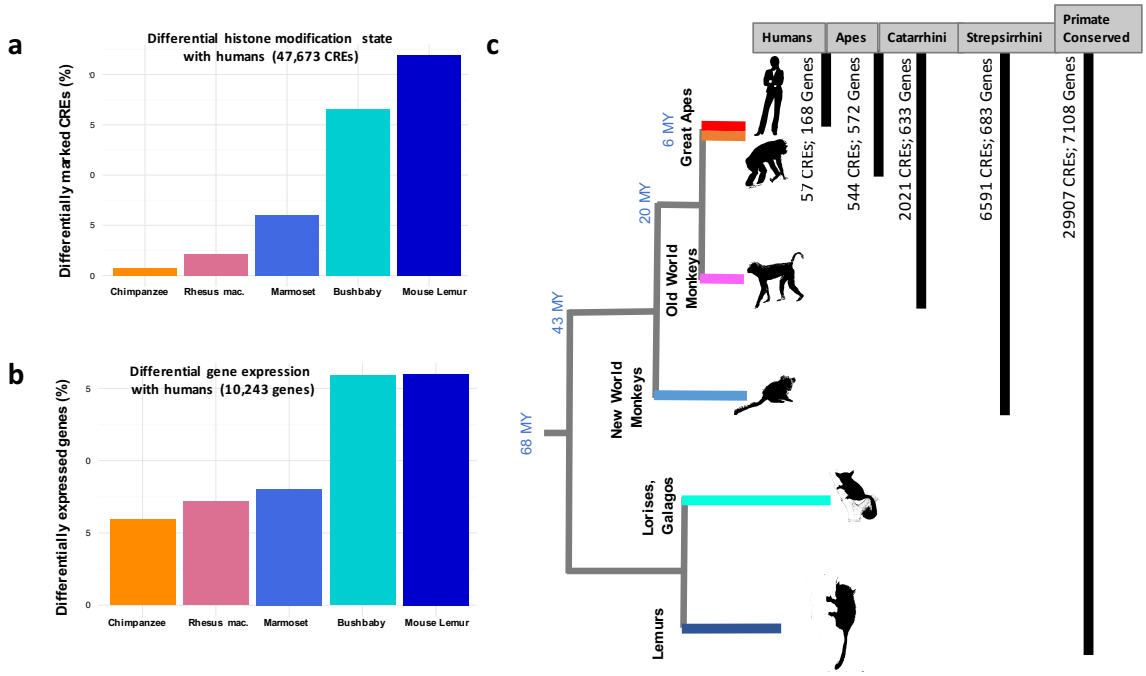


Figure 2 - Primates CREs are evolutionarily conserved. (A) Plots show examples of human-specific, ape-specific and conserved CREs. (B) Fractions of conserved and recently evolved primate CREs, with breakdown of enhancers and promoters.

Next, we assessed the extent to which primate-conserved CREs identified in this study are also evolutionarily conserved across a broader range of mammals. In particular, we compared our conserved H3K27ac CREs with the H3K27ac profile of the opossum, the species with the earliest divergence from humans (>180 million

211 years) in the Villar et al. (2015) dataset. 2,854 primate-conserved promoters and
212 9,456 primate-conserved enhancers have orthologous regions in the opossum
213 genome. Among these, 71.3% of the promoters and 19.1% of the enhancers had
214 significant H3K27ac enrichment in both species, supporting that most of the primate
215 conserved promoters show conserved activity in all of the mammalian clade,
216 whereas only a fraction of the primate conserved enhancers is also conserved
217 across mammals. Further, the two studies come to consistent estimates of the
218 fraction of differentially active CREs per million years: 0.06–0.12% in primates and
219 0.07% in mammals.
220
221 **The conservation of the nucleotide sequence is associated with conservation**
222 **of regulatory activity**
223 Previous studies have suggested that the sequence conservation is associated with
224 conservation of regulatory activity, especially in absence of comparative functional
225 assays (Brown et al., 2007; Cooper and Brown, 2008; Pollard et al., 2010; Gittelman
226 et al., 2015; Holloway et al., 2015; Villar et al., 2015; Yang et al., 2015; Dong et al.,
227 2016; Lewis at al., 2016). For each human-centric species × species comparison, we
228 estimated: i) the fraction of differentially modified CREs; ii) the fraction of
229 differentially expressed genes from a set of 10,243 genes with six way orthologs
230 (Table S3); and iii) the per-nucleotide pairwise sequence divergence for each
231 species with respect to humans for each of the 47,673 unique orthologous CREs
232 Differential histone state fractions ranged from 0.77% in the human ×
233 chimpanzee, to 21.9% in the human × mouse lemur comparisons (Fig. 3a). Similarly,
234 differential gene expression ranged from 5.93% in the human × chimpanzee to
235 16.0% in the human × mouse lemur comparisons (Fig. 3b). Both differential histone
236 modification and differential gene expression fractions reflected phylogenetic
237 distance between humans and other tested species. Differentially expressed genes
238 were significantly more likely to be associated with a differentially modified CRE than
239 expected by chance (9.90%; Fisher's exact test $p < 2.2\times10^{-16}$).
240 Sequence conservation was significantly correlated with regulatory activity
241 (human × chimpanzee, logistic regression $p = 5.7\times10^{-16}$; human × rhesus macaque,
242 $p = 6.7\times10^{-8}$; human × marmoset, $p = 4.0\times10^{-9}$; human × bushbaby, $p < 2.2\times10^{-16}$;
243 human × mouse lemur, $p < 2.2\times10^{-16}$; Fig. 3c). 24,691 CREs overlapped 94,578
244 placental mammal phastCons elements (i.e. regions of the genomes with consistent

245 nucleotide sequence conservation across species; Siepel et al., 2005). The fraction
246 of evolutionary conserved CREs overlapping these conserved elements was higher
247 than expected by chance (Fisher's exact test $p < 2.2\times10^{-16}$). Together, these data
248 demonstrate that CREs with conserved nucleotide sequence are significantly more
249 likely to have conserved regulatory activity and are associated with conserved gene
250 expression.
251



252

253 **Figure 3 - Differential histone mark and gene expression across species**. (A)
254 Human-centric pairwise comparisons for differential histone modification states on
255 47,673 orthologous CREs. (B) Human-centric pairwise comparisons for differential
256 gene expression of 10,243 orthologous CREs. (C) Number of lineage-specific CREs
257 (i.e. CREs significantly more active in each lineage compared to other primates) and
258 genes (i.e. genes upregulated in each lineage compared with other primates) in the
259 primate phylogeny.
260

261 **Genomic features associated with CRE conservation and rapid evolution**

262 To understand the mechanisms responsible for CRE conservation and turnover, we
263 identified genomic features associated with conserved regulatory activity. CREs
264 associated with protein coding genes were significantly more conserved than CREs
265 associated with either pseudogenes (Fisher's exact test $p = 3.3\times10^{-3}$) or lincRNAs
266 (Fisher's exact test $p = 5.2\times10^{-7}$; Fig. 4a). For closely related species, regulatory
267 activity was conserved, regardless of the distance to the nearest TSS (human ×
268 chimpanzee, logistic regression $p = 0.261$; human × rhesus macaque, $p = 0.336$; Fig.

269 4b). However, for more distantly related species pairs, the evolutionary conservation
270 of the CRE activity was significantly lower in regions more distant from TSSs (human
271 × marmoset, logistic regression $p = 6.0\times10^{-15}$; human × bushbaby, $p = 2.1\times10^{-5}$,
272 human × mouse lemur, $p = 2.8\times10^{-5}$; Fig. 4b). Intronic enhancers were significantly
273 more conserved than intergenic enhancers (63.0% and 55.2% respectively; Fisher's
274 exact test $p < 2.2\times10^{-16}$). These data demonstrate increased selective pressure to
275 maintain regulatory activity in the vicinity of protein coding genes.
276 Multiple genomic features indicative of broad regulatory element activity were
277 significantly associated with the conservation of regulatory activity across the primate
278 phylogeny. Promoters and enhancers overlapping regions of chromatin accessibility
279 in many cell types (ENCODE Project Consortium, 2012) were significantly more
280 functionally conserved than those that are functional in only a small number of
281 tissues (logistic regression $p < 2.2\times10^{-16}$; Fig. 4c). Similarly, CREs that overlapped
282 many TFBS, as identified by ENCODE ChIP-seq in HepG2 cells, were significantly
283 more evolutionarily conserved than those with fewer binding sites (logistic regression
284 $p < 2.2\times10^{-16}$; Fig. 4d). Further, liver enhancers that generate consistent enhancer
285 RNA transcription were significantly more evolutionarily conserved than the
286 untranscribed enhancers (Fig. S3; Fisher's exact test $p = 0.04613$), validating
287 previous findings (Andersson et al., 2014).
288 Finally, we used GOrilla (Eden et al., 2007; Eden et al., 2009) to identify
289 biological processes Gene Ontology terms that are enriched in genes found within
290 10-kb of evolutionarily conserved CREs, using as background all of the genes found
291 within 10 kb from any of the 47,673 orthologous CREs. We found an enrichment for
292 housekeeping functions involved in the regulation of cellular, transcriptional and
293 developmental processes (Table S4). These findings support previous observations
294 that conserved CREs are proximal to housekeeping genes (FANTOM5 Consortium,
295 2014; Villar et al., 2015).
296
297 **Specific transcription factor motifs are associated with regulatory**
298 **conservation and turnover**
299 We used the MEME Suite (Bailey et al., 2009) to identify sequence motifs enriched
300 in human-specific, ape-specific, and evolutionarily conserved liver CREs. Human-
301 specific CREs were enriched with motifs for TFs associated with immune response
302 and hematopoietic maintenance (Fig. 4e; Supplemental File S2), such as PRDM1

303 (BLIMP1) which is induced upon viral infection and represses beta-interferon (β-IFN)
304 gene expression. The rapid evolution of immune response genes and TFs is
305 supported by many studies in vertebrates and in *Drosophila melanogaster*,
306 demonstrating that while the central machinery of immune responses is strongly
307 conserved, several components of the extended molecular networks can evolve
308 rapidly or diversify as a consequence of evolutionary competition between hosts and
309 pathogens (Jansa et al., 2003; Vallender 2004; Sackton et al., 2007; Obbard et al.,
310 2009; Schadt et al., 2009; Grueber et al., 2014; Lazzaro and Schneider 2014;
311 Salazar-Jaramillo et al., 2014; Zak et al., 2014; Sironi et al., 2015; Wertheim, 2015;
312 Chuong et al., 2016). Similarly, recently evolved promoters and enhancers in
313 primates are enriched in functions associated to neuronal proliferation, migration and
314 cortical map organization (Boyd et al., 2015; Reilly et al., 2015; Emera et al., 2016).
315 In contrast, ape-specific CREs were instead enriched with motifs representing
316 binding sites for TFs involved in liver function but, remarkably, also in brain and
317 neural system proliferation and development (Fig. 4e; Supplemental File S2).
318 Evolutionarily conserved CREs were enriched in TFBSs for master regulators
319 and homeobox genes that establish cell-type identity in liver cells (Fig. 4e;
320 Supplemental File S2). Among these master regulators, HNF4A is essential for the
321 differentiation of human hepatic progenitor cells by establishing the expression of the
322 network of transcription factors that controls the onset of hepatocyte cell fate
323 (DeLaForest et al., 2011). Likewise, CEBPA is required for the liver cell specification
324 and gene function, and the associated TFBSs are highly conserved across mammals
325 (Ballester et al., 2014). Both CEBPA and HNF4A have conserved cis-regulatory
326 activity and significantly higher numbers of shared TF binding events than expected
327 by chance alone across distant vertebrates (Schmidt et al., 2010). These results
328 demonstrate that evolution shapes the regulatory landscape by preserving the
329 regulatory activity in essential metabolic and developmental pathways, while
330 permitting incessant renovation of specific networks that are under strong selective
331 pressures.
332
333
334

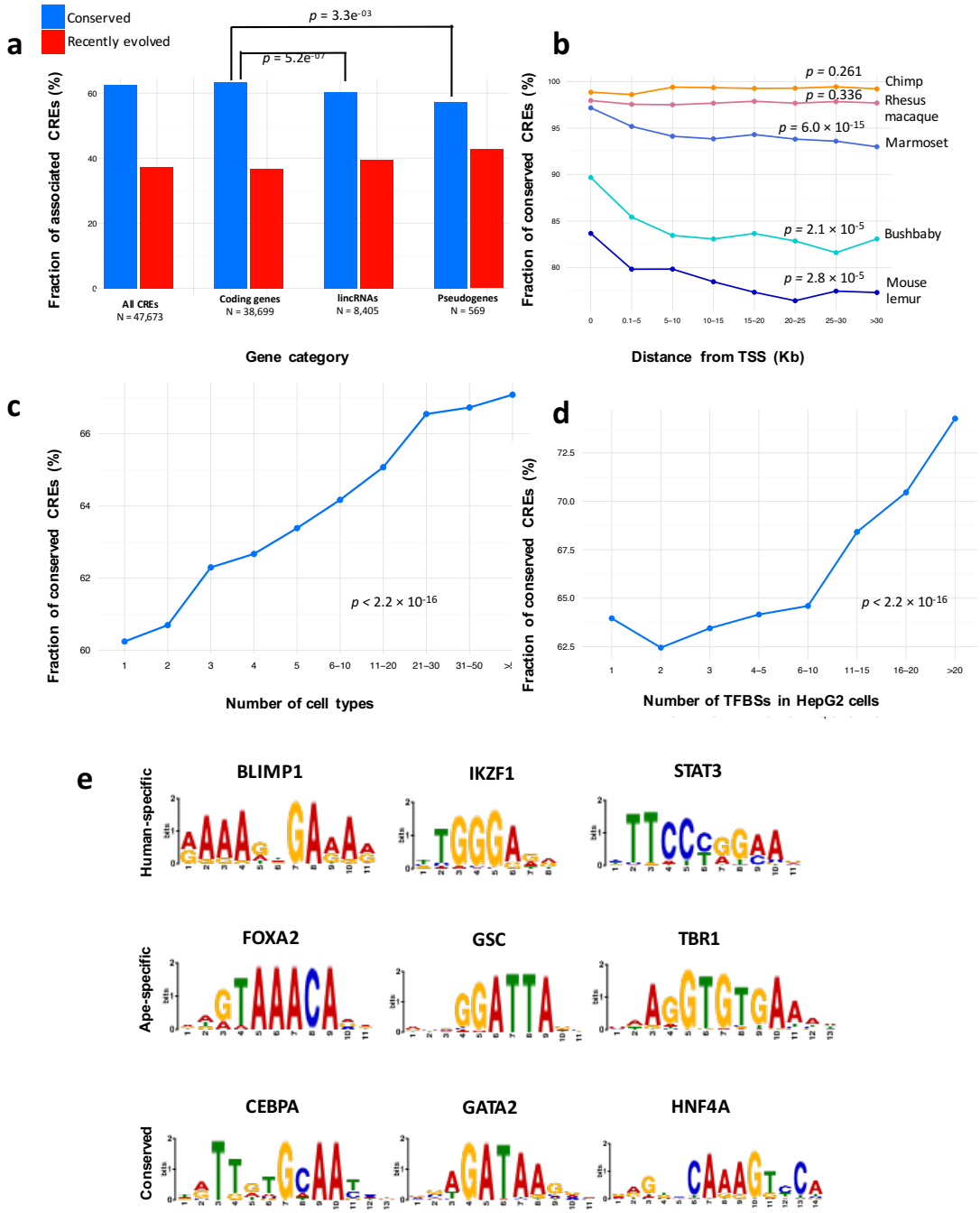


Figure 4 - Genomic features associated to CRE conservation. (A) Fractions of conserved and recently evolved CREs associated to protein coding genes, lincRNAs and pseudogenes. (B) Distribution of CRE conservation along the genome. (C) Correlation between CRE conservation and cell-type specificity based on ENCODE data. (D) Correlation between CRE conservation and number of ENCODE HepG2 TFBSs present of each CRE. (E) Examples of i) enriched motifs associated to immune response (BLIMP1, IZKF1, STAT3) in human-specific CREs; ii) enriched motifs associated to language and neural development (FOXA2, GSC, TBR1) in ape-specific CREs; iii) enriched motifs associated to master regulators (CEPBA, GATA2, HNF4A) in conserved CREs.

347 **Exaptation of TEs into functional CREs is pervasive in the primate genomes**
348 Previous studies have found that TEs can contribute to the origin of CREs. To
349 quantify the contribution of TEs in the liver gene expression regulation in primates,
350 we annotated each liver CRE based on overlap with RepeatMasker elements (Smit,
351 Hubley & Green, 2013-2015; http://www.repeatmasker.org). We found that 28.7% of
352 human liver CREs overlapped an annotated TE, most of which were SINEs (59.8%
353 of the total) and LINEs (21.2%), although LTRs (9.3%) and DNA transposons (8.4%)
354 were also abundant. 27 TE families were significantly enriched within the set of
355 47,673 orthologous CREs (FDR < 1%), nearly all of which were SINE-VNTR-*Alu*s
356 (SVAs), LTRs, and *Alus* (Figure 5), suggesting these TEs contributed to the
357 regulatory landscape in the primate liver (Jordan et al., 2003; Schmidt et al., 2012;
358 Sundaram et al., 2014; Lynch et al., 2015).
359 The majority (75.0%) of the enriched TE families were relatively young, and
360 specific to humans (SVA-F), Hominidae (SVA-B, SVA-C, and SVA-D), Hominoidea
361 (the LTR12 subfamily), Simiiformes (LTRs), or primates (Alu elements), whereas the
362 remaining 25.0% were Eutherian-specific or older (Fig. 5, Fig. S7, and Table S5). We
363 therefore investigated whether these recent TE insertions altered the expression
364 patterns of nearby genes in primates. Specifically, we focused on the enriched TE
365 families younger than the Strepsirrhini/Haplorrhini divergence, thus not present in
366 mouse lemur and bushbaby. We found that 22.6% of the CREs overlapping SVAs,
367 and 12.5% of the CREs overlapping LTRs were differentially expressed between the
368 Strepsirrhini (human, chimpanzee, rhesus macaque and marmoset) and the
369 Haplorrhini (mouse lemur and bushbaby). Both of these fractions were significantly
370 higher than expected by chance (binomial test $p < 2.2\times10^{-16}$). Together, these
371 findings indicate that TEs have played a key role in shaping primate gene regulation,
372 introducing novel gene expression patterns as a consequence of their recruitment as
373 functional CREs.
374
375 **The vast majority of newly evolved CREs are derived from TE insertions**
376 84.2% of ape-specific CREs and 94.7% of human-specific CREs overlap at least one
377 TE (Fig. 5; Fig. S7). In contrast, only 17.0% of evolutionarily conserved CREs
378 contain an annotated TE (Fisher's exact test $p < 2.2\times10^{-16}$). However, we
379 hypothesize that this may be an underestimate due to the inability to recognize
380 ancient TE insertions accumulating mutations over time. LTRs and SVAs were the

most frequently exapted TEs in newly evolved CREs (LTR = 40.1% of the exapted TEs in ape-specific CREs; SVA = 75.3% of the exapted TEs in the human-specific CREs; Fig. 5), 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).

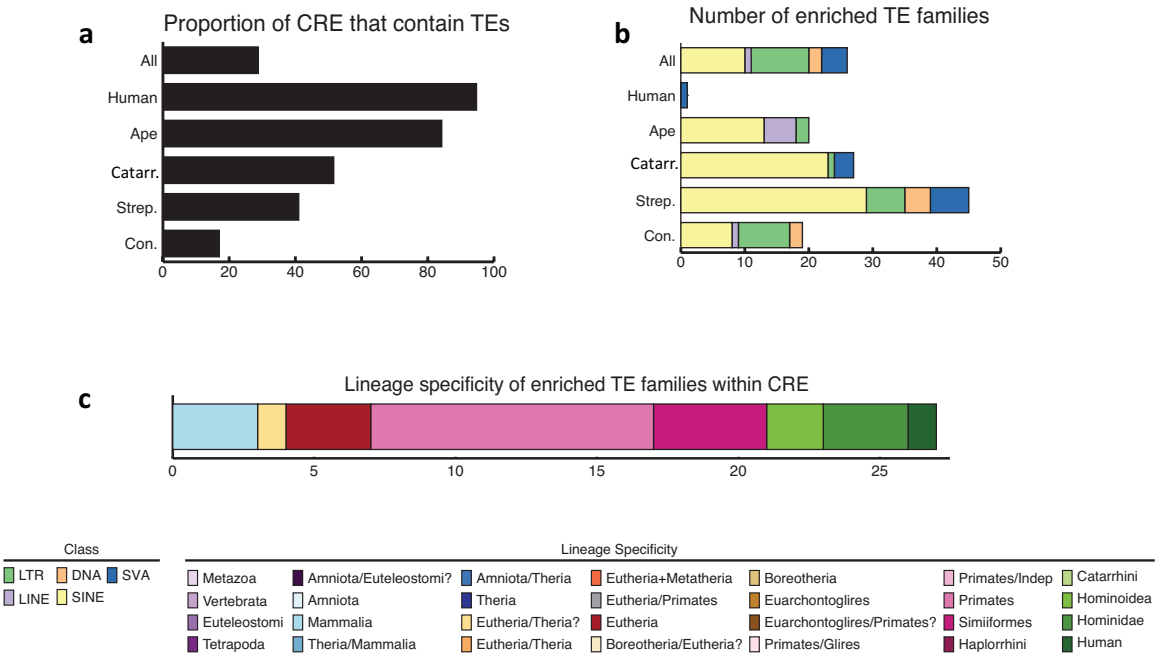


Figure 5 - Newly evolved CREs are enriched in TEs. (A) Proportion of CREs that overlap TEs in the different primate lineages. (B) Number of enriched TE families within CREs in the different primate lineages. (C) Most enriched TE families in primates.

The contribution of LTRs in primate gene regulation has been characterized in previous studies (Wang et al., 2007; Cohen et al., 2009; Sundaram et al., 2014; Janoušek et al., 2016). In our data, a remarkable 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). 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

403 activity upon transfection into HepG2 cells (Wilcoxon's rank sum test p = 0.00017;
404 Fig. S5), suggesting that the transposable element was recruited as functional
405 enhancer in the *GRIN3A* locus.

406 SVAs are instead a hominid-specific family of composite retrotransposons that
407 are active in humans (Hancks and Kazazian, 2010), with more than 3,500 annotated
408 copies. Given that nearly all human-specific liver CREs were derived from SVA
409 insertions in our analysis, we further investigated the genomic features of SVA
410 insertions that lead to exaptation (Table S5). 49.5% of human SVAs overlapped
411 regions of significant histone modification, and 97.8% of those were enhancers.
412 These exapted SVAs are significantly more likely to be associated with protein
413 coding genes than the non-exapted SVAs (Fisher's exact test p = 0.017) and are
414 significantly closer to the TSS of the associated gene (mean of 52.9 kb versus 64.1
415 kb; Wilcoxon rank-sum test $p < 2.2\times10^{-16}$). Exapted and non-exapted SVAs lie within
416 open chromatin regions in approximately the same number of cell types (3.44 and
417 3.94 respectively; logistic regression p = 0.827) and host on average a comparable
418 density of TFBSs (3.28 and 1.95; logistic regression p = 0.679). However, exapted
419 SVAs have a significantly higher number of TFBSs in the neighboring regions (8.12
420 versus 5.86 in +/- 10kb; logistic regression $p = 1.91\times10^{-5}$). Taken together, these
421 data suggest a model where an SVA has a higher probability of becoming a CRE if it
422 inserts in TFBS-dense regions near protein coding gene promoters. (Fig. 6).

423 Among the exapted SVAs, our data predicted an intronic CRE for the gene
424 *JARID2.* This gene is an accessory component of Polycomb Repressive Complex-2
425 (PRC2), recruits PRC2 to chromatin, and is involved in liver, brain, neural tube
426 development, and embryonic stem cell differentiation (Kaneko et al., 2014). Our
427 differential histone state analysis revealed a human-specific ChIP-seq peak
428 overlapping a human-specific 1.9 kb-long insertion, entirely derived from an SVA-F
429 retrotransposon. JARID2 is significantly downregulated in humans compared to all
430 the other primates (Benjamini–Hochberg p = 0.019). Exapted SVAs-Fs exhibit
431 significant enrichment for binding sites of known transcriptional repressors such as
432 PAX-5, FEV, and SREBF1 (Maurer et al., 2003; Fazio et al., 2008; Lecomte et al.,
433 2010). Indeed, the JARID2 SVA-F insertion leads to significantly decreased
434 expression in HepG2 reporter assays (Wilcoxon's rank sum test p = 0.00275; Fig.
435 S6), supporting the role of this SVA-F as a transcriptional repressor.
436

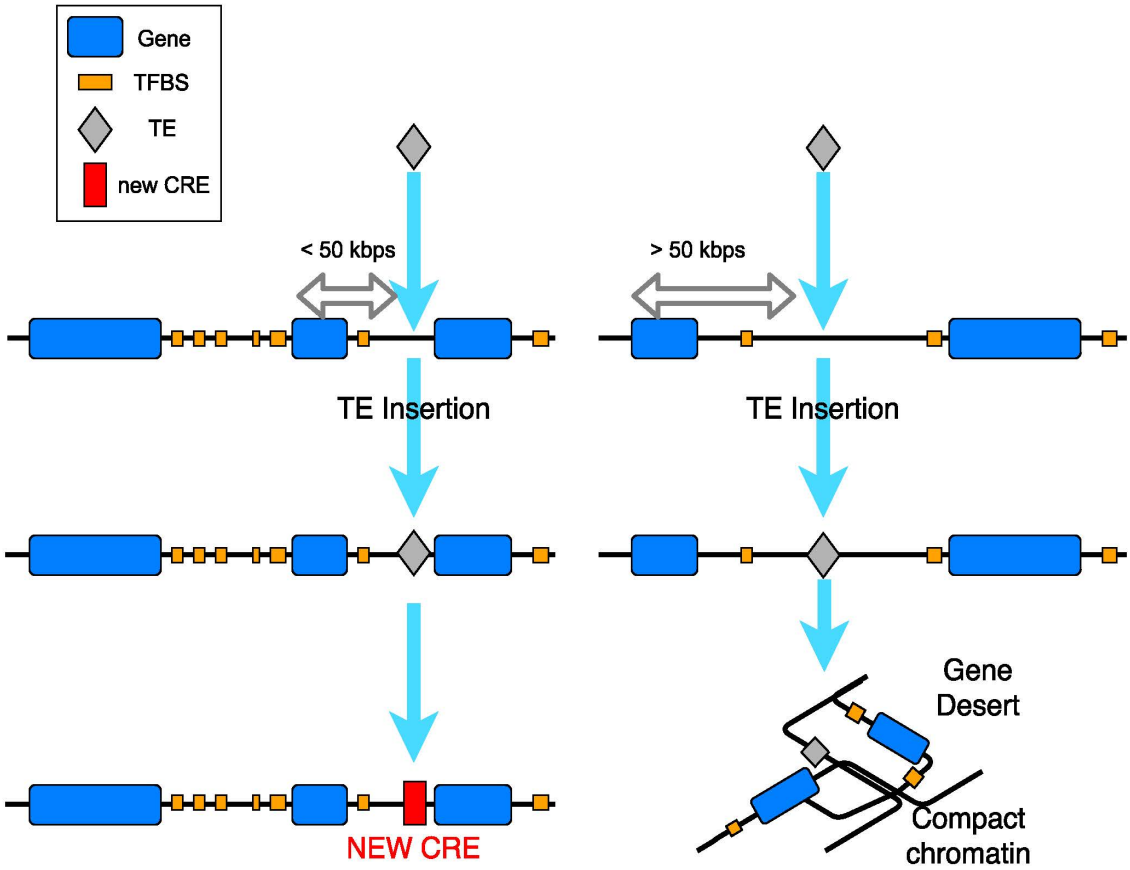


Figure 6 - Model for exaptation of SVAs into human functional CREs. A given SVA has a higher probability of being recruited as a functional cis-regulatory element if it is found within 50 kb from a protein coding gene, in a genomic region already enriched in TF binding motifs in the surrounding 10 kb up- and downstream.

Broad regulatory activity of TE insertions in the primate liver

Our findings strongly suggest that the majority of novel CREs in primates is 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 of the main classes and families of TEs (Table S6). TEs from these families overlap 3,897 of our predicted CREs. We synthesized the mammalian consensus sequence (see experimental procedures) for 69 different TE families, cloned them into a luciferase reporter vector with a minimal promoter, and transfected them into HepG2 cells to perform dual luciferase reporter assays. We found that luciferase expression for 66 of the 69 (95.6%) tested TE families was significantly different from the negative control (Fig. 7a; Wilcoxon's rank sum test p-values in Table S6). Strikingly, only 17 (25.7%) of these, mostly LTRs and DNA transposons, produced activity

significantly higher than the negative control (Fig. 7a), whereas the remaining 49 (74.3%), mostly LINEs, repressed transcription. Consistent with results from the JARID2 locus presented above, SVA-Fs were confirmed to function as transcriptional repressors.

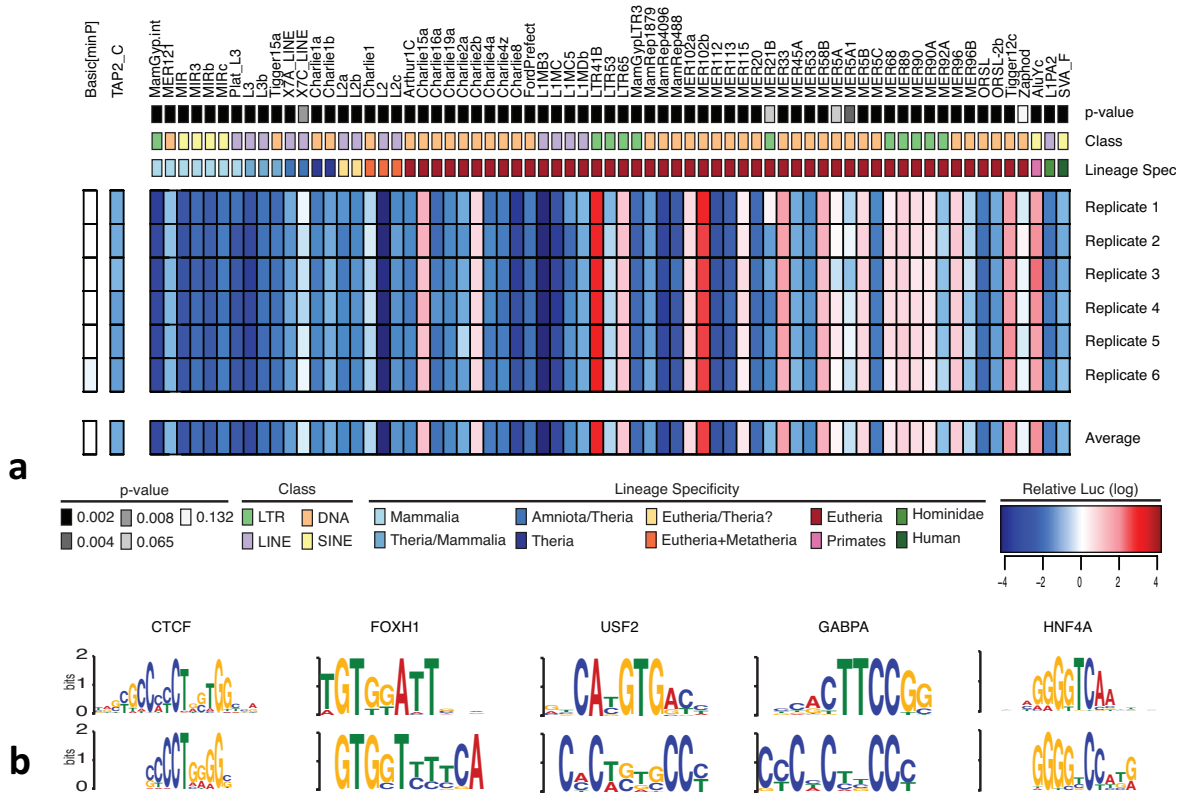


Figure 7 - Regulatory ability of TE families found in poised and active regulatory elements in HepG2 cells. (A) The p-value, class, and lineage specificity for the 69 TE families tested in HepG2 cells for regulatory ability, the empty vector control (Basic[minP]), and the positive control (TAP2_C) are all shown above the 6 luciferase assay replicates conducted and the average regulatory ability found across the replicates. Red indicates luciferase expression higher than the empty vector control; blue indicates luciferase expression less than Basic[minP]. (B) CTCF, FOX, USF2, GABP, and HNF4a binding motifs were found to be enriched in the sequences of the 66 TE families that drive expression significantly different from background. The top row shows the enriched sequence found while the bottom shows the Jaspar binding motifs recognized for each transcription factor.

These findings support that LTRs, among the most enriched TEs in our peak set, and the most common exapted TEs in apes, are likely co-opted as active enhancer elements. The putative repressor activity of LINEs is consistent with their underrepresentation in the human ChIP-seq peak set. However, we hypothesize that, at least for some of the TE families, such observed repressing activity levels

478 could be the byproduct of secondary biological mechanisms leading the cell to
479 recognize these elements as newly inserted TEs, and therefore silencing them via
480 epigenetic mechanisms. However, further tests will be needed to support this
481 hypothesis.

482 The consensus sequences for the 66 TEs that drove reporter expression
483 significantly different from background were analyzed with MEME to identify enriched
484 motifs. Motifs for known master regulators of liver cell identity, including FOX, USF2,
485 GABP, and HNF4A (Wallerman et al., 2009), were significantly enriched within the
486 sequences of the 66 TE families with significant regulatory activity (Fig. 7b). In
487 summary, most TE families function as CREs in the primate liver, either as strong
488 enhancers or as repressors. TEs are actively recruited into the regulatory landscape,
489 further supporting our findings on the pervasive involvement of TEs in the primate
490 gene regulation.

491

492 ## Discussion

493

494 **The primate regulatory landscape is evolutionarily conserved**

495 Only a small fraction (<1.50%) of the CREs were differentially active between
496 humans and chimpanzees. This suggests that even modest changes in gene
497 regulation produce observable phenotypic differentiation, and confirm that cis-
498 regulatory evolution plays a central role in primate diversification (Davidson 2001,
499 2006; Wray, 2007; Ho et al., 2009; Tsankov et al., 2010; Smith et al., 2013; Martin et
500 al., 2012; Coolon et al., 2014; Martin and Reed, 2014; Guo et al., 2015; Lynch et al.,
501 2015; Villar et al., 2015; Adachi et al., 2016; Landeen et al., 2016; Lesch et al., 2016;
502 Zhang and Reed, 2016). Our approach for the comparison of CREs across species,
503 based on the analysis of differential histone modification state in orthologous
504 regions, demonstrated that cis-regulatory divergence across species may be
505 overestimated when based on the peak overlap status as a binary variable.

506

507 **Specific genomic features are associated with CRE conservation**

508 Evolutionarily conserved promoters and enhancers have conserved nucleotide
509 sequence, are close to protein coding genes, are functional in many cell types, and
510 harbor many TFBSs. Many regulatory pathways, specifically those involved in the
511 regulation of liver function and housekeeping functions, are strongly conserved

512 across primates, while other pathways, such as immune response, are less
513 constrained and evolve more rapidly. This observation is consistent with the
514 expected arms-race in evolution between host and pathogens.
515
516 **Newly evolved CREs are derived from TE exaptation**
517 Based on our findings, exaptation of transposable elements into functional promoters
518 and enhancers is a pervasive phenomenon in primate genomes. LTRs and SVAs are
519 the most frequently exapted in humans and other apes, despite not being among the
520 most common transposable elements in the genome. Primate liver CREs are
521 enriched in young TEs. These young TEs, after being recruited into the primate
522 regulatory network, introduced novel gene expression in the associated species. To
523 our knowledge, this is the first study demonstrating how specific genomic features
524 are associated to the recruitment of TEs as functional elements in the primate
525 regulatory landscape. In contrast, only a minor fraction of the evolutionarily
526 conserved CREs overlap an annotated TE. This suggests the action of strong
527 selection against the disruption of these regulatory elements, in order to maintain
528 stable gene expression. Further, these data suggest that the core regulatory network
529 that establishes liver cell-type identity is very conserved, whereas adaptive evolution
530 occurs on the periphery of the network, where TEs have the most impact on gene
531 regulatory evolution.
532
533 **Data availability**
534 All non-human raw sequence data have been deposited in the Sequence Read
535 Archive under following BioProject IDs: PRJNA349047 for RNA-seq and
536 PRJNA349046 for ChIP-seq data.
537
538 **Supplemental information**
539 Supplemental can be found with this article online at **xxxxxxxx**
540
541 **Author contributions**
542 MT, MC and CDB conceived the project. MT, YP, GHP and CDB designed the taxon
543 sampling and experiments. MT performed ChIP-seq and RNA-seq experiments. MT
544 and MHB performed the parallelized reporter assays. MHB and KA produced
545 luciferase assay data on *GRIN3A* and *JARID2*. KM and VJL designed and performed

546 the luciferase assays on the 69 TE families. MT, VJL and KM performed the TE
547 enrichment analysis. YP designed computational pipelines for the detection of
548 orthologous regions in the Ensembl MSA alignment and all related analyses. MT, YP
549 and CDB analyzed the data and wrote the paper. All authors read and approved the
550 manuscript.
551

552 **Acknowledgements**
553 We thank Texas Biomedical Research Institute, Duke University Lemur Center and
554 the Hospital of University of Pennsylvania for providing the samples. We thank
555 Jonathan Schug and other members of UPenn NGS Core Facility. We are grateful to
556 Sarah Tishkoff, Yoseph Barash, Benjamin Voight, Yana Kamberov, Erin Fry, and
557 Barbara Engelhardt for providing insightful comments on the preliminary versions of
558 the results and of the manuscript.
559

560

561 **References**

562

563 Adachi, N., Robinson, M., Goolsbee, A., and Shubin, N.H. (2016). Regulatory
564 evolution of Tbx5 and the origin of paired appendages. Proceedings of the National
565 Academy of Sciences.

566 Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M.,
567 Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active
568 enhancers across human cell types and tissues. Nature *507*, 455–461.

569 Austenaa, L.M.I., Barozzi, I., Simonatto, M., Masella, S., Della Chiara, G., Ghisletti,
570 S., Curina, A., de Wit, E., Bouwman, B.A.M., de Pretis, S., et al. Transcription of
571 Mammalian cis-Regulatory Elements Is Restrained by Actively Enforced Early
572 Termination. Molecular Cell *60*, 460–474.

573 Babbitt, C.C., Fedrigo, O., Pfefferle, A.D., Boyle, A.P., Horvath, J.E., Furey, T.S., and
574 Wray, G.A. (2010). Both Noncoding and Protein-Coding RNAs Contribute to Gene
575 Expression Evolution in the Primate Brain. Genome Biology and Evolution *2*, 67–79.

576 Bailey, T.L. (2011). DREME: motif discovery in transcription factor ChIP-seq data.
577 Bioinformatics *27*, 1653–1659.

578 Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li,
579 W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and
580 searching. Nucleic Acids Res *37*, W202-208.

581 Ballester, B., Medina-Rivera, A., Schmidt, D., Gonzàlez-Porta, M., Carlucci, M.,
582 Chen, X., Chessman, K., Faure, A.J., Funnell, A.P., Goncalves, A., et al. (2014).
583 Multi-species, multi-transcription factor binding highlights conserved control of tissue-
584 specific biological pathways. eLife *3*, e02626.

585 Barnabas, S., Hai, T., and Andrisani, O.M. (1997). The Hepatitis B Virus X Protein
586 Enhances the DNA Binding Potential and Transcription Efficacy of bZip Transcription
587 Factors. Journal of Biological Chemistry *272*, 20684–20690.

588 Bejerano, G., Lowe, C.B., Ahituv, N., King, B., Siepel, A., Salama, S.R., Rubin, E.M.,
589 James Kent, W., and Haussler, D. (2006). A distal enhancer and an ultraconserved
590 exon are derived from a novel retroposon. Nature *441*, 87–90.

591 Bourque, G., Leong, B., Vega, V.B., Chen, X., Lee, Y.L., Srinivasan, K.G., Chew, J.-
592 L., Ruan, Y., Wei, C.-L., Ng, H.H., et al. (2008). Evolution of the mammalian
593 transcription factor binding repertoire via transposable elements. Genome Research.

594 Boyd, J.L., Skove, S.L., Rouanet, J.P., Pilaz, L.-J., Bepler, T., Gordân, R., Wray,
595 G.A., and Silver, D.L. Human-Chimpanzee Differences in a FZD8 Enhancer Alter
596 Cell-Cycle Dynamics in the Developing Neocortex. Current Biology *25*, 772–779.

597 Boyle, A.P., Araya, C.L., Brdlik, C., Cayting, P., Cheng, C., Cheng, Y., Gardner, K.,
598 Hillier, L.W., Janette, J., Jiang, L., et al. (2014). Comparative analysis of regulatory
599 information and circuits across distant species. Nature *512*, 453–456.

600 Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csardi, G., Harrigan, P., Weier,
601 M., Liechti, A., Aximu-Petri, A., Kircher, M., et al. (2011). The evolution of gene
602 expression levels in mammalian organs. Nature *478*, 343–348.

603 Breschi, A., Djebali, S., Gillis, J., Pervouchine, D.D., Dobin, A., Davis, C.A.,
604 Gingeras, T.R., and Guigó, R. (2016). Gene-specific patterns of expression variation
605 across organs and species. Genome Biology *17*, 1–13.

606 Britten, R.J., and Davidson, E.H. (1969). Gene Regulation for Higher Cells: A
607 Theory. Science *165*, 349–357.

608 Brosius, J., Gould, S.J. (1992). On “genomenclature": a comprehensive (and
609 respectful) taxonomy for pseudogenes and other “junk DNA.”. Proceedings of the
610 National Academy of Sciences USA *89*, 10706–10710.

611 Brown, C.D., Johnson, D.S., and Sidow, A. (2007). Functional Architecture and
612 Evolution of Transcriptional Elements That Drive Gene Coexpression. Science *317*,
613 1557.

614 Cain, C.E., Blekhman, R., Marioni, J.C., and Gilad, Y. (2011). Gene Expression
615 Differences Among Primates Are Associated With Changes in a Histone Epigenetic
616 Modification. Genetics *187*, 1225–1234.

617 Chuong, E.B., Rumi, M.A.K., Soares, M.J., and Baker, J.C. (2013). Endogenous
618 retroviruses function as species-specific enhancer elements in the placenta. Nat
619 Genet *45*, 325–329.

620 Chuong, E.B., Elde, N.C., and Feschotte, C. (2016). Regulatory evolution of innate
621 immunity through co-option of endogenous retroviruses. Science *351*, 1083–1087.

622 Cohen, C.J., Lock, W.M., and Mager, D.L. (2009). Endogenous retroviral LTRs as
623 promoters for human genes: A critical assessment. Gene *448*, 105–114.

624 Coolon, J.D., McManus, C.J., Stevenson, K.R., Graveley, B.R., and Wittkopp, P.J.
625 (2014). Tempo and mode of regulatory evolution in Drosophila. Genome Research.

626 Cooper, G.M., and Brown, C.D. (2008). Qualifying the relationship between
627 sequence conservation and molecular function. Genome Research *18*, 201–205.

628 Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A., and Lis, J.T. (2014).
629 Analysis of nascent RNA identifies a unified architecture of initiation regions at
630 mammalian promoters and enhancers. Nat Genet *46*, 1311–1320.

631 Cotney, J., Leng, J., Yin, J., Reilly, S.K., DeMare, L.E., Emera, D., Ayoub, A.E.,
632 Rakic, P., and Noonan, J.P. The Evolution of Lineage-Specific Regulatory Activities
633 in the Human Embryonic Limb. Cell *154*, 185–196.

634 Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine,
635 E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone
636 H3K27ac separates active from poised enhancers and predicts developmental state.
637 Proceedings of the National Academy of Sciences *107*, 21931–21936.

638 Cusanovich, D.A., Pavlovic, B., Pritchard, J.K., and Gilad, Y. (2014). The Functional
639 Consequences of Variation in Transcription Factor Binding. PLoS Genet *10*,
640 e1004226.

641 Davidson, E., and Britten, R. (1979). Regulation of gene expression: possible role of
642 repetitive sequences. Science *204*, 1052–1059.

643 Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.B., Gaffney, D.J., et al. (2012).
644 DNaseI sensitivity QTLs are a major determinant of human expression variation.
645 Nature, *482*, 390–394.

646 DeLaForest, A., Nagaoka, M., Si-Tayeb, K., Noto, F.K., Konopka, G., Battle, M.A.,
647 and Duncan, S.A. (2011). HNF4A is essential for specification of hepatic progenitors
648 from human pluripotent stem cells. Development *138*, 4143.

649 De Souza, F.S.J., Franchini, L.F., and Rubinstein, M. (2013). Exaptation of
650 Transposable Elements into Novel Cis-Regulatory Elements: Is the Evidence Always
651 Strong? Molecular Biology and Evolution *30*(6), 1239-1251.

652 Dong, X., Wang, X., Zhang, F., and Tian, W. (2016). Genome-Wide Identification of
653 Regulatory Sequences Undergoing Accelerated Evolution in the Human Genome.
654 Mol Biol Evol *33*, 2565–2575.

655 Du, J., Leung, A., Trac, C., Lee, M., Parks, B.W., Lusis, A.J., Natarajan, R., and
656 Schones, D.E. (2016). Chromatin variation associated with liver metabolism is
657 mediated by transposable elements. Epigenetics Chromatin *9*, 28.

658 Eden, E., Lipson, D., Yogev, S., and Yakhini, Z. (2007). Discovering Motifs in
659 Ranked Lists of DNA Sequences. PLoS Comput Biol *3*, e39.

660 Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool
661 for discovery and visualization of enriched GO terms in ranked gene lists. BMC
662 Bioinformatics *10*, 1–7.

663 Emera, D., Yin, J., Reilly, S.K., Gockley, J., and Noonan, J.P. (2016). Origin and
664 evolution of developmental enhancers in the mammalian neocortex. Proceedings of
665 the National Academy of Sciences *113*, E2617–E2626.

666 ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements
667 in the human genome. Nature *489*, 57–74.

668 Fazio, G., Palmi, C., Rolink, A., Biondi, A., Cazzaniga, G. (2008). PAX5/TEL acts as
669 a transcriptional repressor causing down-modulation of CD19, enhances migration to
670 CXCL12, and confers survival advantage in pre-BI cells. Cancer Research *68*(1),
671 181–189.

672 FANTOM Consortium and the RIKEN PMI and CLST (DGT) (2014). A promoter-level
673 mammalian expression atlas. Nature *507*, 462–470.

674 Gittelman, R.M., Hun, E., Ay, F., Madeoy, J., Pennacchio, L., Noble, W.S., Hawkins,
675 R.D., and Akey, J.M. (2015). Comprehensive identification and analysis of human
676 accelerated regulatory DNA. Genome Res *25*, 1245–1255.

677 Gross, L. (2005). Selection on a Neural Gene Regulator Sheds Light on Human
678 Evolution. PLoS Biol *3*, e417.

679 Grueber, C.E., Wallis, G.P., and Jamieson, I.G. (2014). Episodic Positive Selection
680 in the Evolution of Avian Toll-Like Receptor Innate Immunity Genes. PLoS ONE *9*,
681 e89632.

682 Guo, C., Ludvik, A.E., Arlotto, M.E., Hayes, M.G., Armstrong, L.L., Scholtens, D.M.,
683 Brown, C.D., Newgard, C.B., Becker, T.C., Layden, B.T., et al. (2015). Coordinated
684 regulatory variation associated with gestational hyperglycaemia regulates expression
685 of the novel hexokinase HKDC1. Nat Commun *6*.

686 Hah, N., Murakami, S., Nagari, A., Danko, C.G., and Kraus, W.L. (2013). Enhancer
687 transcripts mark active estrogen receptor binding sites. Genome Research *23*,
688 1210–1223.

689 Hancks, D.C., and Kazazian Jr., H.H. (2010). SVA retrotransposons: Evolution and
690 genetic instability. Seminars in Cancer Biology *20*, 234–245.

691 Ho, M.C.W., Johnsen, H., Goetz, S.E., Schiller, B.J., Bae, E., Tran, D.A., Shur, A.S.,
692 Allen, J.M., Rau, C., Bender, W., et al. (2009). Functional Evolution of cis-Regulatory
693 Modules at a Homeotic Gene in Drosophila. PLoS Genet *5*, e1000709.

694 Holloway, A.K., Bruneau, B.G., Sukonnik, T., Rubenstein, J.L., and Pollard, K.S.
695 (2016). Accelerated Evolution of Enhancer Hotspots in the Mammal Ancestor.
696 Molecular Biology and Evolution *33*, 1008–1018.

697 Innocenti, F., Cooper, G.M., Stanaway, I.B., Gamazon, E.R., Smith, J.D., Mirkov, S.,
698 Ramirez, J., Liu, W., Lin, Y.S., Moloney, C., et al. (2011). Identification, Replication,
699 and Functional Fine-Mapping of Expression Quantitative Trait Loci in Primary Human
700 Liver Tissue. PLoS Genet *7*, e1002078.

701 Jain, D., Baldi, S., Zabel, A., Straub, T., and Becker, P.B. (2015). Active promoters
702 give rise to false positive “Phantom Peaks” in ChIP-seq experiments. Nucleic Acids
703 Res *43*, 6959–6968.

704 Jansa, A.S., Lundrigan, L.B., and Tucker, K.P. (2003). Tests for Positive Selection
705 on Immune and Reproductive Genes in Closely Related Species of the Murine
706 Genus Mus. Journal of Molecular Evolution *56*, 294–307.

707 Jordan, I.K., Rogozin, I.B., Glazko, G.V., and Koonin, E.V. (2003). Origin of a
708 substantial fraction of human regulatory sequences from transposable elements.
709 Trends in Genetics *19*, 68–72.

710 Kaneko, S., Bonasio, R., Saldaña-Meyer, R., Yoshida, T., Son, J., Nishino, K.,
711 Umezawa, A., and Reinberg, D. Interactions between JARID2 and Noncoding RNAs
712 Regulate PRC2 Recruitment to Chromatin. Molecular Cell *53*, 290–300.

713 Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G.,
714 Yandell, M., and Feschotte, C. (2013). Transposable Elements Are Major
715 Contributors to the Origin, Diversification, and Regulation of Vertebrate Long
716 Noncoding RNAs. PLoS Genet *9*, e1003470.

717 King, M., and Wilson, A. (1975). Evolution at two levels in humans and chimpanzees.
718 Science *188*, 107–116.

719 Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H.,
720 and Bourque, G. (2010). Transposable elements have rewired the core regulatory
721 network of human embryonic stem cells. Nat Genet *42*, 631–634.

722 Landeen, E.L., Muirhead, C.A., Wright, L., Meiklejohn, C.D., and Presgraves, D.C.
723 (2016). Sex Chromosome-wide Transcriptional Suppression and Compensatory Cis-
724 Regulatory Evolution Mediate Gene Expression in the Drosophila Male Germline.
725 PLoS Biol *14*, e1002499.

726 Lazzaro, B.P., and Schneider, D.S. (2014). The Genetics of Immunity. G3:
727 Genes|Genomes|Genetics *4*, 943–945.

728 Lecomte, V., Maugnier, E., Euthine, V., Durand, C., Freyssenet, D., Nemoz, G.,
729 Rome, S., Vidal, H. and Lefai, E. (2010). A new role for sterol regulatory element
730 binding protein 1 transcription factors in the regulation of muscle mass and muscle
731 cell differentiation. Molecular and Cell Biology *30*(5), 1182–1189.

732 Lesch, B.J., Silber, S.J., McCarrey, J.R., and Page, D.C. (2016). Parallel evolution of
733 male germline epigenetic poising and somatic development in animals. Nat Genet
734 *48*, 888–894.

735 Lewis, J.J., van der Burg, K.R.L., Mazo-Vargas, A., and Reed, R.D. ChIP-Seq-
736 Annotated Heliconius erato Genome Highlights Patterns of cis-Regulatory Evolution
737 in Lepidoptera. Cell Reports *16*, 2855–2863.

738 Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with
739 BWA-MEM. arXiv:1303.3997v1

740 Li, W., Notani, D., and Rosenfeld, M.G. (2016). Enhancers as non-coding RNA
741 transcription units: recent insights and future perspectives. Nat Rev Genet *17*, 207–
742 223.

743 Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general
744 purpose program for assigning sequence reads to genomic features. Bioinformatics
745 *30*, 923–930.

746 Loisel, D.A., Rockman, M.V., Wray, G.A., Altmann, J., and Alberts, S.C. (2006).
747 Ancient polymorphism and functional variation in the primate MHC-DQA1 5′ cis-
748 regulatory region. Proceedings of the National Academy of Sciences *103*, 16331–
749 16336.

750 Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change
751 and dispersion for RNA-seq data with DESeq2. Genome Biology *15*, 1–21.

752 Lynch, V.J., Leclerc, R.D., May, G., and Wagner, G.P. (2011). Transposon-mediated
753 rewiring of gene regulatory networks contributed to the evolution of pregnancy in
754 mammals. Nat Genet *43*, 1154–1159.

755 Lynch, V.J., Nnamani, M.C., Kapusta, A., Brayer, K., Plaza, S.L., Mazur, E.C.,
756 Emera, D., Sheikh, S.Z., Grützner, F., Bauersachs, S., et al. (2015) Ancient
757 Transposable Elements Transformed the Uterine Regulatory Landscape and
758 Transcriptome during the Evolution of Mammalian Pregnancy. Cell Reports *10*, 551–
759 561.

760 Markljung, E., Jiang, L., Jaffe, J.D., Mikkelsen, T.S., Wallerman, O., Larhammar, M.,
761 Zhang, X., Wang, L., Saenz-Vash, V., Gnirke, A., et al. (2009). ZBED6, a Novel
762 Transcription Factor Derived from a Domesticated DNA Transposon Regulates IGF2
763 Expression and Muscle Growth. PLoS Biol *7*, e1000256.

764 Marnetto, D., Molineris, I., Grassi, E., and Provero, P. (2014). Genome-wide
765 Identification and Characterization of Fixed Human-Specific Regulatory Regions.
766 American Journal of Human Genetics *95*, 39–48.

767 Martin, A., and Reed, R.D. (2014). Wnt signaling underlies evolution and
768 development of the butterfly wing pattern symmetry systems. Developmental Biology
769 *395*, 367–378.

770 Martin, A., Papa, R., Nadeau, N.J., Hill, R.I., Counterman, B.A., Halder, G., Jiggins,
771 C.D., Kronforst, M.R., Long, A.D., McMillan, W.O., et al. (2012). Diversification of
772 complex butterfly wing patterns by repeated regulatory evolution of a Wnt ligand.
773 Proceedings of the National Academy of Sciences *109*, 12632–12637.

774 Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., et al. (2012).
775 Systematic Localization of Common Disease-Associated Variation in Regulatory
776 DNA. Science, *337*, 1190–1195.

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

780 McClintock, B. (1950). The Origin and Behavior of Mutable Loci in Maize.
781 Proceedings of the National Academy of Sciences of the United States of America
782 *36*, 344–355.

783 McClintock, B. (1984). The significance of responses of the genome to challenge.
784 Science *226*, 792–801.

785 Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S.,
786 Gnirke, A., Callan, C.G., Kinney, J.B., et al. (2012). Systematic dissection and
787 optimization of inducible enhancers in human cells using a massively parallel
788 reporter assay. Nat Biotech *30*, 271–277.

789 Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N.E., Ahfeldt, T., Sachs, K.V.,
790 Li, X., Li, H., Kuperwasser, N., Ruda, V.M., et al. (2010). From noncoding variant to
791 phenotype via SORT1 at the 1p13 cholesterol locus. Nature, *466* (7307), 714–719.

792 Patwardhan, R.P., Hiatt, J.B., Witten, D.M., Kim, M.J., Smith, R.P., May, D., Lee, C.,
793 Andrie, J.M., Lee, S.-I., Cooper, G.M., et al. (2012). Massively parallel functional
794 dissection of mammalian enhancers in vivo. Nat Biotech *30*, 265–270.

795 Obbard, D.J., Welch, J.J., Kim, K.-W., and Jiggins, F.M. (2009). Quantifying Adaptive
796 Evolution in the Drosophila Immune System. PLoS Genet *5*, e1000698.

797 Ori, A., Atzmony, D., Haviv, I., and Shaul, Y. (1994). An NF1 Motif Plays a Central
798 Role in Hepatitis B Virus Enhancer. Virology *204*, 600–608.

799 Perelman, P., Johnson, W.E., Roos, C., Seuánez, H.N., Horvath, J.E., Moreira,
800 M.A.M., Kessing, B., Pontius, J., Roelke, M., Rumpler, Y., et al. (2011). A Molecular
801 Phylogeny of Living Primates. PLoS Genet *7*, e1001342.

802 Pickrell, J.K., Gaffney, D.J., Gilad, Y., and Pritchard, J.K. (2011). False positive
803 peaks in ChIP-seq and other sequencing-based functional assays caused by
804 unannotated high copy number regions. Bioinformatics *27*, 2144–2146.

805 Pollard, K.S., Salama, S.R., King, B., Kern, A.D., Dreszer, T., Katzman, S., Siepel,
806 A., Pedersen, J.S., Bejerano, G., Baertsch, R., et al. (2006). Forces Shaping the
807 Fastest Evolving Regions in the Human Genome. PLoS Genet *2*, e168.

808 Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of
809 nonneutral substitution rates on mammalian phylogenies. Genome Research *20*,
810 110–121.

811 Prabhakar, S., Visel, A., Akiyama, J.A., Shoukry, M., Lewis, K.D., Holt, A., Plajzer-
812 Frick, I., Morrison, H., FitzPatrick, D.R., Afzal, V., et al. (2008). Human-Specific Gain
813 of Function in a Developmental Enhancer. Science *321*, 1346–1350.

814 Prescott, S.L., Srinivasan, R., Marchetto, M.C., Grishina, I., Narvaiza, I., Selleri, L.,
815 Gage, F.H., Swigut, T., and Wysocka, J. Enhancer Divergence and cis-Regulatory
816 Evolution in the Human and chimpanzee Neural Crest. Cell *163*, 68–83.

817 Rayan, N.A., Del Rosario, R.C.H., and Prabhakar, S. (2016). Massive contribution of
818 transposable elements to mammalian regulatory sequences. Semin Cell Dev Biol *57*,
819 51–56.

820 Reilly, S.K., Yin, J., Ayoub, A.E., Emera, D., Leng, J., Cotney, J., Sarro, R., Rakic,
821 P., and Noonan, J.P. (2015). Evolutionary changes in promoter and enhancer activity
822 during human corticogenesis. Science *347*, 1155–1159.

823 del Rosario, R.C.H., Rayan, N.A., and Prabhakar, S. (2014). Noncoding origins of
824 anthropoid traits and a new null model of transposon functionalization. Genome
825 Research *24*, 1469–1484.

826 Sackton, T.B., Lazzaro, B.P., Schlenke, T.A., Evans, J.D., Hultmark, D., and Clark,
827 A.G. (2007). Dynamic evolution of the innate immune system in Drosophila. Nat
828 Genet *39*, 1461–1468.

829 Salazar-Jaramillo, L., Paspati, A., van de Zande, L., Vermeulen, C.J., Schwander, T.,
830 and Wertheim, B. (2014). Evolution of a cellular immune response in Drosophila: a
831 phenotypic and genomic comparative analysis. Genome Biol Evol *6*, 273–289.

832 Sasaki, T., Nishihara, H., Hirakawa, M., Fujimura, K., Tanaka, M., Kokubo, N.,
833 Kimura-Yoshida, C., Matsuo, I., Sumiyama, K., Saitou, N., et al. (2008). Possible
834 involvement of SINEs in mammalian-specific brain formation. Proceedings of the
835 National Academy of Sciences *105*, 4220–4225.

836 Schadt, E.E. (2009). Molecular networks as sensors and drivers of common human
837 diseases. Nature *461*, 218–223.

838 Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A.,
839 Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S., et al. (2010). Five-
840 Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor
841 Binding. Science *328*, 1036.

842 Schmidt, D., Schwalie, P.C., Wilson, M.D., Ballester, B., Gonçalves, Â., Kutter, C.,
843 Brown, G.D., Marshall, A., Flicek, P., and Odom, D.T. (2012). Waves of
844 Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in
845 Multiple Mammalian Lineages. Cell *148*, 335–348.

846 Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L.,
847 Yakhini, Z., Weinberger, A., and Segal, E. (2012). Inferring gene regulatory logic
848 from high-throughput measurements of thousands of systematically designed
849 promoters. Nat Biotech *30*, 521–530.

850 Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K.,
851 Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily
852 conserved elements in vertebrate, insect, worm, and yeast genomes. Genome
853 Research *15*, 1034–1050.

854 Sironi, M., Cagliani, R., Forni, D., and Clerici, M. (2015). Evolutionary insights into
855 host-pathogen interactions from mammalian sequence data. Nat Rev Genet *16*,
856 224–236.

857 Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O.,
858 Awedh, M.H., Baldock, R., Barbiera, G., et al. (2015). The BioMart community portal:
859 an innovative alternative to large, centralized data repositories. Nucleic Acids
860 Research *43*, W589–W598.

861 Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., Snyder, M.P., and Wang,
862 T. (2014). Widespread contribution of transposable elements to the innovation of
863 gene regulatory networks. Genome Res *24*, 1963–1976.

864 Takata, A., Iwayama, Y., Fukuo, Y., Ikeda, M., Okochi, T., Maekawa, M., Toyota, T.,
865 Yamada, K., Hattori, E., Ohnishi, T., et al. A Population-Specific Uncommon Variant
866 in GRIN3A Associated with Schizophrenia. Biological Psychiatry *73*, 532–539.

867 Tsankov, A.M., Thompson, D.A., Socha, A., Regev, A., and Rando, O.J. (2010). The
868 Role of Nucleosome Positioning in the Evolution of Gene Regulation. PLoS Biol *8*,
869 e1000414.

870 Vallender, E.J., and Lahn, B.T. (2004). Positive selection on the human genome.
871 Human Molecular Genetics *13*, R245–R254.

872 Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park, T.J.,
873 Deaville, R., Erichsen, J.T., Jasinska, A.J., et al. Enhancer Evolution across 20
874 Mammalian Species. Cell *160*, 554–566.

875 Wallerman, O., Motallebipour, M., Enroth, A., Patra, K., Bysani, M.S.R., Komorowski,
876 J., and Wadelius, C. (2009). Molecular interactions between HNF4a, FOXA2 and
877 GABP identified at regulatory DNA elements through ChIP-sequencing. Nucleid
878 Acids Research *37*, 7498–7508.

879 Wang, T., Zeng, J., Lowe, C.B., Sellers, R.G., Salama, S.R., Yang, M., Burgess,
880 S.M., Brachmann, R.K., and Haussler, D. (2007). Species-specific endogenous
881 retroviruses shape the transcriptional network of the human tumor suppressor
882 protein p53. Proceedings of the National Academy of Sciences *104*, 18613–18618.

883 Warner, L.R., Babbitt, C.C., Primus, A.E., Severson, T.F., Haygood, R., and Wray,
884 G.A. (2009). Functional consequences of genetic variation in primates on tyrosine
885 hydroxylase (TH) expression in vitro. Brain Research *1288*, 1–8.

886 Wertheim, B. (2015). Genomic basis of evolutionary change: evolving immunity.
887 Frontiers in Genetics *6*, 222.

888 Wray, G.A. (2007). The evolutionary significance of cis-regulatory mutations. Nat
889 Rev Genet *8*, 206–216.

890 Xie, M., Hong, C., Zhang, B., Lowdon, R.F., Xing, X., Li, D., Zhou, X., Lee, H.J.,
891 Maire, C.L., Ligon, K.L., et al. (2013). DNA hypomethylation within specific
892 transposable element families associates with tissue-specific enhancer landscape.
893 Nat Genet *45*, 836–841.

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

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

Young, R.S., Kumar, Y., Bickmore, W.A., and Taylor, M.S. (2016). Bidirectional transcription marks accessible chromatin and is not specific to enhancers. bioRxiv.

Zak, D.E., Tam, V.C., and Aderem, A. (2014). Systems-Level Analysis of Innate Immunity. Annu. Rev. Immunol. *32*, 547–577.

Zhang, L., and Reed, R.D. (2016). Genome editing in butterflies reveals that spalt promotes and Distal-less represses eyespot colour patterns. Nat Commun *7*.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biology *9*, 1–9.

Zhou, X., Cain, C.E., Myrthil, M., Lewellen, N., Michelini, K., Davenport, E.R., Stephens, M., Pritchard, J.K., and Gilad, Y. (2014). Epigenetic modifications are associated with inter-species gene expression variation in primates. Genome Biology *15*, 1–19.

915 ## Materials and Methods

916

917 **Tissue sampling**

918 We obtained liver tissue samples for three to four individuals belonging to each of
919 the studied primate species (three bushbabies, four chimpanzees, three humans,
920 three marmosets, three mouse lemurs, and three rhesus macaques; Table S1).
921 Samples for chimpanzees, marmoset and rhesus macaque were obtained from
922 Texas Biomedical Research Institute (San Antonio, TX); bushbaby and mouse lemur
923 livers were obtained from the Lemur Center of Duke University (Durham, NC).
924 Tissue samples were collected and flash-frozen immediately. With the exception of
925 the bushbaby, samples for all of the species included both males and females (Table
926 S1). Age was comparable across species (young adults) and all individuals died of
927 causes unrelated to liver disease.

928

929 **RNA-seq sample processing**

930 We processed samples from all species in random batches of four in order to
931 minimize batch effects. For each sample, 25 mg of frozen liver tissue was used to
932 extract total RNA and genomic DNA, using QIAGEN AllPrep DNA/RNA/miRNA
933 Universal Kit. Quality of total RNA was assessed computing the RNA Integrity
934 Number (RIN) using Agilent Bioanalyzer. All RNA samples had a RIN > 8. We used
935 4µg aliquots of total RNA to produce barcoded RNA sequencing libraries using the
936 Illumina TruSeq Stranded mRNA kit. The quality of generated libraries was assessed
937 using Agilent Bioanalyzer High Sensitivity DNA Kit and Kapa metrics. Libraries were
938 pooled in two different pools based on barcode compatibility, and each pool was
939 sequenced in two Illumina HiSeq2500 lanes, producing on an average of 42.1 million
940 single end (SE) 100-bp reads per sample.

941

942 **ChIP-seq sample processing**

943 We processed samples in six randomly assigned groups in order to minimize batch
944 effects. For each sample, we cut 90 mg of frozen liver tissue into 1 mm^3 pieces,
945 washed the cut tissue samples with cold phosphate-buffered saline (PBS), and fixed
946 with 1% formaldehyde for 5 minutes at room temperature. We prepared nuclei of
947 each washed sample using the Covaris truChIP Tissue Chromatin Shearing Kit.

948 Chromatin was then sheared for 16 minutes using a Covaris S220 Focused-
949 ultrasonicator. We quantified shearing efficiency and chromatin concentration using
950 Agilent Bioanalyzer High Sensitivity DNA Kit.
951 From each specimen, we kept aside a 0.5 µg aliquot of sheared chromatin to
952 be used as input. We used two 5 µg aliquots of chromatin per sample to perform
953 immunoprecipitation (IP) with antibodies directed at H3K27ac (ab4729) and
954 H3K4me1 (ab8895) respectively. We performed each IP using 5 µg of antibody with
955 an overnight incubation at 4°C as specified by the Magna ChIP A/G Chromatin
956 Immunoprecipitation Kit protocol. After elution and protein-DNA crosslink reversal,
957 we extracted DNA using Zymo Research ChIP DNA Clean & Concentrator kit, and
958 quantified extracted DNA using Agilent High Sensitivity kit and Qubit 2.0.
959 We used 5 to 15 ng of input and immunoprecipitated DNA to generate
960 sequencing libraries using the NEBNext Ultra ChIPseq library kit, following protocols
961 specified by the manufacturer. We assessed the quality of each constructed library
962 using Agilent Bioanalyzer High Sensitivity DNA Kit and Kapa metrics. Libraries were
963 multiplexed, pooled and sequenced on a total of 16 Illumina HiSeq2500 lanes,
964 producing on an average of 40.6 million SE 100-bp reads per sample.
965
966 **Sequence QC: ChIP-seq and RNA-seq**
967 We performed the standard quality control (QC) measures on both ChIP-seq and
968 RNA-seq fastq files using FastQC v0.11.3 (Andrews, 2010). We then trimmed
969 sequencing adapters and low quality base calls using TrimGalore! v0.4.1 with the
970 following parameters: -stringency 5 -length 50 -q 20
971 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/).
972
973 **RNA-seq alignment and gene expression quantification**
974 We aligned all sequences that passed QC to the reference genomes from the
975 Ensembl database (bushbaby: otoGar3; chimp: CHIMP2.1.4; humans: GRCh38;
976 rhesus macaque: Mmul1; marmoset: C_jacchus3.2.1; mouse lemur: micMur1) using
977 STAR v2.5 (Dobin et al., 2013) in 2-pass mode with the following parameters: --
978 quantMode TranscriptomeSAM --outFilterMultimapNmax 10 --
979 outFilterMismatchNmax 10 --outFilterMismatchNoverLmax 0.3 --alignIntronMin 21 --
980 alignIntronMax 0 --alignMatesGapMax 0 --alignSJoverhangMin 5 --runThreadN 12 --
981 twopassMode Basic --twopass1readsN 60000000 --sjdbOverhang 100. We filtered

982 bam files based on alignment quality (q = 10) and sorted using Samtools v0.1.19 sort
983 function (Li, 2009). We used the latest annotations for each species obtained from
984 Ensembl to build reference indexes for the STAR alignment:
985 Homo_sapiens.GRCh38.82.chr.gtf; Pan_troglodytes.CHIMP2.1.4.82.chr.gtf;
986 Macaca_mulatta.MMUL_1.82.chr.gtf; Callithrix_jacchus.C_jacchus3.2.1.82.chr.gtf;
987 Otolemur_garnettii.OtoGar3.82.gtf; Microcebus_murinus.micMur1.82.gtf (Aken et al.,
988 2016). We used FeatureCounts (Liao et al., 2014) to count reads mapping to each
989 protein coding gene/lincRNA/pseudogene, according to Ensembl annotations for the
990 six studied species. Read counts were then normalized based on feature (gene)
991 length.
992
993 **Differential gene expression analysis**
994 We analyzed differential gene expression levels for each species using read counts,
995 normalized by feature length with the DESeq2 software (Love et al., 2014), and with
996 the following model:
997 *design = ~condition*
998
999 where condition indicates the species or the group of species (e.g. apes).
1000 We used a set of 10,243 genes annotated as orthologs in the six species
1001 according to Ensembl (BioMart v. 0.9; Smedley et al., 2015; Table S3) and used 5%
1002 False Discovery Rate (FDR) as our multiple-testing-corrected significance threshold.
1003 The overall analysis included three main comparisons of our interest: 1) pairwise
1004 comparisons between human and each of the other five species; 2) comparisons for
1005 human-specific differential expression of orthologous genes (human versus other
1006 five primates grouped together); 3) comparisons for ape-specific differential
1007 expression of orthologous genes (human + chimpanzee versus other four primates).
1008 4) comparisons for Catarrhini-specific differential expression
1009 (human+chimpanzee+rhesus macaque versus other primates). Finally comparison
1010 between Strepsirrhini (human, chimpanzee, rhesus macaque and marmoset) and
1011 Haplorrhini (mouse lemur and bushbaby) was computed.
1012
1013 **ChIP-seq QC and alignment**
1014 We applied standard QC measures to ChIP-seq data as described above for RNA-
1015 seq data processing. We aligned the sequences that passed QC to the reference

1016 genomes from the Ensembl database (bushbaby: otoGar3; chimpanzee:
1017 CHIMP2.1.4; humans: GRCh38; rhesus macaque: Mmul1; marmoset:
1018 C_jacchus3.2.1; mouse lemur: micMur1) using Burrows Wheeler Alignment tool
1019 (BWA), with the MEM algorithm (Li, 2013). We sorted the filtered bam files using
1020 Samtools v0.1.19 (Li, 2009).
1021
1022 **ChIP-seq peak calling and QC**
1023 We called peaks for each individual using MACS2 (Zhang et al., 2008), at 1% FDR,
1024 and with parameters recommended for histone modifications (Liu, 2014): --m 30 40 -
1025 -ext size 147 -B. Input was used as a control in all differential histone modification
1026 analyses. We performed QC on peaks called for each specimen using metrics
1027 recommended by ENCODE (Landt et al., 2012): Fraction of Reads in Peaks (FRiP),
1028 Normalize Strand Correlation coefficient (NSC) and Relative Strand Correlation
1029 coefficient (RSC) and ENCODE quality score.
1030 In order to compute FRiP, we used Bedtools (v2.25.0; Quinlan and Hall, 2010)
1031 to intersect bed files containing all coordinates of called peaks (narrowPeak output of
1032 MACS2) with the original sorted bam file of the specific ChiP-seq sample. Then, we
1033 used a publicly available perl script to count the reads mapping in the intersection
1034 regions(https://github.com/mel-astar/mel-ngs/blob/master/mel-chipseq/chipseq-
1035 metrics/getCnt.pl). As recommended by the ENCODE consortium, we selected a
1036 threshold of 1% as acceptable FRiP values. We computed the two strand correlation
1037 metrics (NSC, RSC) using Phantompeakqualtools (Landt et al., 2012). For H3K27ac,
1038 NSC ≥ 1.05 and RSC ≥ 0.8 were used as threshold for retaining samples. For
1039 H3K4me1, that tends to produce broader peaks, we used NSC ≥ 1.05 and RSC ≥ 0.5
1040 (Table S1).
1041 Samples that did not pass at least two of the three main QC metrics (FRiP,
1042 NSC, RSC) were excluded for any downstream analysis. We then called human
1043 consensus peaks for H3K27ac and H3K4me1 using MACS2 and the above
1044 described parameters with the 1% FDR threshold. All human samples passing QC
1045 were considered as replicates of each other for the consensus peak calling. These
1046 human consensus H3K27ac and H3K4me1 profiles were used to perform all of the
1047 below described human-centric downstream analyses. Peaks called for the other
1048 species were only used for the above mentioned QC purposes but were not utilized
1049 for any of the downstream analyses, with the exception of the chimpanzee

1050 consensus peaks (see below). In order to assess how our data compares to known
1051 liver related regulatory regions, we overlapped our set of human consensus peaks to
1052 the set of H3K27ac and H3K4me1 peaks generated for HepG2 cells from the
1053 ENCODE consortium and to the set of permissive enhancers generated by the
1054 FANTOM5 consortium.
1055
1056 **Comparison to previous findings using human liver ChIP-seq data**
1057 We compared our human H3K27ac ChIP-seq data, with a set of published human
1058 liver H3K27ac peaks (Villar et al., 2015). We used the window function of Bedtools to
1059 quantify the number of H3K27ac peaks in the replication dataset that either overlap
1060 with, or are found within 1 kb up- and downstream from each of our human H3K27ac
1061 peaks. With this approach, we quantified and assessed overlaps between discovery
1062 and replication datasets.
1063 Next, we characterized sets of replicated and unreplicated peaks between the
1064 two datasets. Using the procedures previously described for the comparison of
1065 conserved and recently evolved CREs, we annotated several genomic features for
1066 both replicated and unreplicated peaks. Specifically, we included information
1067 regarding: 1) the average distance from the closest TSS; 2) the class of the
1068 associated gene (e.g., protein coding, lincRNA); 3) the number of cell types with an
1069 overlapping ENCODE DHS site and; 4) the number of ENCODE TFBS overlapping
1070 the peak. Moreover, a logistic regression on the q-values of replicated and not
1071 replicated peaks was performed.
1072
1073 **Parallelized reporter assay**
1074 We obtained a list of 334 putative 1-kb long CREs overlapping liver eQTLs from
1075 Brown and collaborators (Brown et al., 2013). This data included both enhancers
1076 (distance from TSS > 1Kb) and promoters (distance from TSS < 1kb). 122 CREs out
1077 of these 334 CREs overlapped our human ChIP-seq peaks (53 enhancers and 69
1078 promoters; Table S6). Within each of the loci defined by the investigated liver eQTLs,
1079 we predicted a 1-kb CRE. These predicted CREs were amplified in individual PCRs
1080 performed on 120 pooled Yoruban HapMap DNA samples. PCR products from each
1081 reaction therefore represent a complex mixture of haplotypes. We inserted barcodes
1082 (hereafter, tags) consisting of a 160-bp oligo, including a randomized 20-bp unique
1083 barcode for each construct, into luciferase reporter vectors (pGL4.23 and pGL4.10),

1084 immediately downstream of the luciferase gene, after linearizing the vector with the
1085 XbaI restriction enzyme.

1086 We pooled and cloned DNAs from each putative CRE into uniquely barcoded
1087 luciferase reporter vectors (pGL4.23 were used for enhancers and pGL4.10 for
1088 promoters), using the Gibson Assembly Kit (New England BioLabs). The CREs were
1089 specifically inserted upstream of the luciferase gene, after linearizing the vector with
1090 the restriction enzymes KpnI and XhoI. We then transfected the complex pool of
1091 CRE reporters into HepG2 cells in two replicates. 24 hours after transfection, we
1092 extracted total RNA, purified poly-A RNA, and produced cDNA that was used to
1093 amplify the tag, with the QIAGEN One Step RT-PCR Kit with primers that included
1094 Illumina adapters for sequencing. Tag libraries were pooled and sequenced on a
1095 single Illumina HiSeq2500 lane, producing single end (SE) 50-bp reads. We
1096 amplified the tags from the vector before the transfection and sequenced them in the
1097 same pool with the tag-RNA libraries as a control for tag read counts.

1098 In parallel, reporter tags were unambiguously associated with each specific
1099 CRE by sequence based sub-assembly. Briefly, we cut the luciferase gene from the
1100 vector by inverse PCR and then re-ligated the vector using the T4 Polynucleotide
1101 Kinase (PNK) + T4 ligase kit from NEB. In this way CREs and tags were flanking
1102 each other and CRE-tag complexes. The CRE-tag complexes were then PCR
1103 amplified using a reverse primer that included Illumina adapter for sequencing. Next,
1104 the CRE-tag PCR product was digested for 5 minutes at 55°C using Nexetera Tn5
1105 Transposase (TDE1) in order to produce fragments of variable length (from ca. 150
1106 bp to the entire length of the construct). When cutting the fragments, TDE1 also
1107 inserts an Illumina compatible adapter in proximity of the cutting site. We performed
1108 a PCR to enrich the libraries using the TDE1 inserted adapter as forward primer and
1109 the previously included Illumina adapter as reverse primer.

1110 We pooled the two libraries (one for pGL4.10 and one for pGL4.23 constructs)
1111 and sequenced them on an Illumina MiSeq, producing paired-end (PE) reads (250 +
1112 50 bp). After performing QC with FastQC v0.11.3, we aligned the sub-assembly
1113 sequences to the human genome (GRCh37/hg19) using BWA mem and the bam
1114 files were sorted and indexed with Samtools v0.1.19. Finally, we produced a matrix
1115 listing all of the CRE-tag associations. Tags associated with more than one CRE
1116 were discarded and not used for further analyses. After attributing each tag to its
1117 uniquely associated CRE, we used sequence based tag counts (HiSeq reads),

1118 normalized by sequencing depth, to quantify the gene expression level driven by
1119 each CRE, and therefore its functionality as enhancer/promoter.
1120 For each CRE, we used a count-based generalized linear model to quantify
1121 differential expression between RNA (after transfection) and DNA (before
1122 transfection), assuming a Poisson error function:
1123
1124 *model= count~condition*
1125
1126 where condition indicates that the read count comes either from RNA (replicates 1
1127 and 2) or DNA-control.
1128 In presence of a significant p-value, the model indicates a significant
1129 difference between the expression of the tags in the RNA samples compared to their
1130 DNA control. The effect size estimate was then used to infer whether the RNA
1131 samples were upregulated, hence showing significantly higher level of expression of
1132 the tags compared to their DNA controls, and therefore indicating that the CRE is a
1133 functional regulatory element.
1134
1135 **Detection of orthologous regions for human peaks in each primate**
1136 We mapped orthologous sequences using all identified consensus ChIP-seq peak
1137 regions in both H3K27ac and H3K4me1 experiments. We used the Ensembl multiple
1138 sequence alignment (MSA) reference database with the following specifications: 39
1139 Eutherian mammals; method_link_type: "EPO_LOW_COVERAGE";
1140 species_set_name: "mammals" (Herrero et al., 2015). For orthologous sequence
1141 analysis, 500 bp up- and downstream regions were considered to be a part of the
1142 identified consensus peaks in all six species. We queried all regions directly from the
1143 reference database using the REST API (Yates et al., 2015).
1144 All orthologous sequences retained gaps generated by MSA. In cases of
1145 incomplete chromosome assembly (e.g. mouse lemur), composite sequence
1146 representations containing parts of multiple scaffolds are used as a reference as
1147 provided by the Ensembl database. As a result, we independently queried each peak
1148 region as well as regions covering each peak +/- 500 bp. All orthologous sequences
1149 pulled from the references for downstream analysis contained only directly aligned
1150 sequences. All regions with no orthologous regions represented in the MSA
1151 reference were excluded from further analyses. All query results in .json format and

1152 extracted sequences formatted for the MSA alignment as well as genomic position
1153 information are provided in the repository mentioned in the final section.
1154
1155 **Correlation between human and marmoset read counts within orthologous**
1156 **regions**
1157 We assessed human and marmoset (i.e. the species with the smallest number of
1158 peaks called; Supplemental File S1) normalized read counts at the 47,673
1159 orthologous CREs, after splitting them into two groups: 1) regions with overlapping
1160 peaks present in both marmoset and human, and 2) regions with a peak present only
1161 in human. Spearman's correlation (ρ) between human and marmoset normalized
1162 read depths was then computed for each the two groups.
1163
1164 **Differential histone modification analysis**
1165 Using the above described procedure, for both H3K27ac and H3K4me1, we
1166 produced a single matrix including the human peaks having an ortholog in each of
1167 the studied species and the associated read count for the specific histone mark and
1168 for the input in all of the six species. The normalized read counts were used for
1169 differential ChIP-seq analysis with DESeq2, performing an interaction analysis using
1170 the Wald statistic between the histone marks read counts and their associated input
1171 values:
1172
1173 *design = ~assay + condition + assay:condition*
1174
1175 where the assay indicates either histone marks data or input data, and condition
1176 indicates instead the species or the group of species (e.g. apes, Catarrhini,
1177 Haplorrhini).
1178 Differential histone mark analysis included the same species × species and
1179 group × group comparisons described for RNA-seq. We used 10% FDR as our
1180 multiple testing corrected significance threshold. Further, different FDRs (5%, 10%,
1181 20%, 30%, 40%) were tested to assess the robustness of our approach.
1182 We analyzed differential histone modifications for the two marks
1183 independently. However, in order to quantify the fraction of differentially marked cis-
1184 regulatory elements (CREs) in all of the above described pairwise comparisons, we
1185 used bedtools to identify the CREs predicted by H3K27ac (i.e. H3K27ac peaks) that

1186 would overlap those predicted by H3K4me1 for at least a 25% of their length. The list
1187 of unique CREs was used to estimate the fraction of differentially bound CREs for
1188 each of the above mentioned human-centric pairwise comparisons. We defined
1189 CREs as evolutionarily conserved if they did not show significant differential histone
1190 mark in any of the above mentioned pairwise comparisons. Otherwise, the CRE
1191 were defined as recently evolved.
1192
1193 **Sequence conservation versus functional conservation analysis**
1194 We estimated per-nucleotide pairwise divergence for all five species in comparison
1195 to humans using MSA aligned sequences of orthologous regions for consensus
1196 peaks +/- 500 bps. All gaps in human were excluded from analysis. Regions not
1197 included in the set of six-way orthologous CREs were also pruned. Finally, we
1198 removed outliers - with respect to the distribution of the genetic distances in the
1199 given pairwise comparison - using the R package *outliers* (Komsta, 2006). We
1200 intersected the set of 47,673 orthologous liver CREs with the UCSC
1201 phastConsElements30wayPlacental track (Siepel et al., 2005), to assess whether
1202 genomic regions characterized by conserved nucleotide sequence (i.e. phastCons
1203 elements) are significantly more associated to CREs detected as evolutionarily
1204 conserved in primates.
1205
1206 **Analysis of features associated with evolutionary conservation of CREs**
1207 The following features were associated to each of predicted human CREs: nearest
1208 gene, distance from TSS, functional categories of genes, CRE category and histone
1209 mark (Table S2). Any human CREs without orthologous regions in other five species
1210 have been excluded from our analyses. To assess the correlation between the
1211 degree of conservation of a CRE and the number of cell types where the CRE is
1212 functional, we obtained publicly available data for DNase hypersensitivity sites (DHS)
1213 for over 200 cell types (ENCODE Project Consortium, 2012). For each CRE
1214 overlapping one or more DHS regions, we annotated the number of cell types where
1215 the specific CRE is putatively functional.
1216 We estimated the correlation between the degree of conservation of CREs
1217 and the number of TFBSs by comparing our human consensus peaks with
1218 previously published HepG2 TF-binding profiles (ENCODE Project Consortium,
1219 2012). Further, we quantified the proportion of putative primate CREs overlapping

1220 known transcribed enhancers (eRNAs) by using 43,011 known permissive
1221 transcribed enhancers (FANTOM5 Consortium, 2014). Similarly, for liver-specificity
1222 of human CREs analysis, we used coordinates of published liver eQTLs (Innocenti et
1223 al., 2011). We then selected all genes within 10 kb distance from evolutionarily
1224 conserved CREs for gene set enrichment analysis using GOrilla software (Eden et
1225 al., 2007; Eden et al., 2009). All genes found within 10 kb of any of the 47,673
1226 orthologous CREs are used as a background for the enrichment test.
1227
1228 **Known and *de novo* motif analysis**
1229 Genomic coordinates of orthologous regions were used to extract target sequences
1230 from the Ensembl references without MSA alignment gaps. All regions containing
1231 consensus peaks identified as human- and apes-specific and primate-conserved
1232 were used for the motif discovery and enrichment analysis. MEME-chip was used for
1233 known motif discovery and enrichment analysis using the Jaspar database (Bailey et
1234 al., 2009). We used DREME (Bailey, 2011). De novo motif identification was
1235 performed with AME (McLeay et al., 2010). Jaspar and Hocomoco (v10) databases
1236 were used as references to estimate similarities to known motifs. All motif discovery
1237 and enrichment analysis used default settings and parameters provided by the
1238 developers except for the maximum *de novo* motif discovery threshold (changed
1239 from 1 to 1000 for maximum threshold). Shuffled input sequences were used to
1240 estimate the background distribution of motifs.
1241
1242 **Overlap of transposable elements (TEs) with primate CREs**
1243 We used Bedtools to overlap the RepeatMasker track for GRCh38 to the set of
1244 unique human CREs that would overlap a TE for at least 25% of their length. TE
1245 enrichment analysis was performed using the *TEAnalysis* pipeline with TE-
1246 analysis_Shuffle_bed v. 2.0, setting 1000 replicates
1247 (https://github.com/4ureliek/TEanalysis; Kapusta et al., 2013). To test the regulatory
1248 effect of enriched young TEs on primate gene expression, we performed a
1249 differential gene expression analysis between Strepsirrhini (human, chimp, rhesus
1250 macaque, marmoset) and Haplorrhini (bushbaby, mouse lemur) and quantified the
1251 number of CREs associated to differential expressed genes that overlapped a TE
1252 younger than the Haplorrhini-Strepsirrhini divergence.
1253

1254 **Analysis of SINE-VNTR-*Alu*s (SVA) TEs enriched in human CREs**

1255 We intersected all known SVAs annotated in the human genome with all human
1256 consensus peaks from our study. We used the same 25% overlap threshold as
1257 described above and considered all human consensus peaks regardless of presence
1258 of orthologous regions in the other species. The two SVA lists (overlapping and not
1259 overlapping the CREs, respectively) were annotated for: the average distance from
1260 the closest TSS, functional categories of associated gene, the average number of
1261 cell types with available DHS data, the average number of TFBS overlapping the
1262 SVAs, and finally the average number of TFBS within 10 kb up- and downstream of
1263 the SVAs. We used AME (McLeay et al., 2010) to look for motifs enriched in the
1264 exapted SVAs, using the not exapted SVAs as a control.

1265

1266 **Luciferase reporter assay validation of *GRIN3A* and *JARID2***

1267 To test for species- or clade-specific regulatory activity, we compared activity of two
1268 predicted functional CREs with the empty pGL4.23 vector as a negative control. For
1269 *GRIN3A* we PCR amplified the CRE (Table S6), and cloned the fragment into
1270 pGL4.23 using the NEB Gibson Assembly Kit. The *JARID2* CRE was synthesized by
1271 GenScript and cloned into the same pGL4.23 vector. Cells were grown in DMEM
1272 high glucose (Gibco #11965084) supplemented with 10% fetal bovine serum (FBS)
1273 (GE Healthcare Life Sciences #SH3091003) containing antibiotic and antimycotic
1274 (Gibco #15240062) in a humidified incubator with 5% CO_2 at 37°C. HepG2 cells were
1275 seeded in 48-well CellBIND surface plates (Costar #3338) with 1.5 x 10^5 cells per
1276 well 24 h prior to transfection. Transfection complexes were formed using 800 ng of
1277 each construct with 1 µL of TransIT-LT1 transfection reagent (Mirus #MIR2304) and
1278 Opti-MEM (Gibso #31985070) in a total volume of 27 µL, incubated for 20 min and
1279 then added to cells. After transfection, cells were incubated for 24 h and were lysed
1280 in passive lysis buffer. To read firefly luciferase activity, 100 µL of LARII were added
1281 to 20 µL of cell lysate (from the dual-luciferase reporter assay system from Promega
1282 #E1910). We read Luminescence for 2 seconds per well on a 96-well compatible
1283 plate luminometer (ThermoFisher Luminoskan Ascent). The constructs were tested
1284 using three vector preparations in three to four technical transfection replicates (9 to
1285 12 measurements for for each construct). We normalized for transfection replicates
1286 effect using a linear model:

1287 *lm(log10(luciferase) ~ replicate + element*.

1288

1289 **Validation of the gene regulatory functionality of TE families**

1290 HepG2 cells were cultured in DMEM + GlutaMAX (Gibco) supplemented with 10%
1291 Fetal Bovine Serum (Gibco) and Normocin (InvivoGen). Transposable element
1292 constructs were built by synthesizing (GenScript) the Dfam (Hubley et al., 2016)
1293 consensus sequence for each element and cloning into the pGL3 Basic vector
1294 (Promega) with an added minimal promoter (pGL3 Basic[minP]). pGL3 BASIC[minP]
1295 with no insert was used as the negative expression control. pRL null (Promega) was
1296 the renilla control for transfection efficiency. TAP2 cloned into the pGL3 Basic[minP]
1297 was the positive control. Confluent HepG2 cells in opaque 96 well plates in 90ml of
1298 Opti-MEM (Gibco) were transfected according to the Lipofectamine p3000 protocol
1299 (Invitrogen) with 100 ng of the luciferase containing plasmid, 1 ng of pRL null, 0.3 ml
1300 of Lipofectamine 3000, and 0.2ml of p3000 reagent in 10 ml of Opti-MEM per well.
1301 The cells incubated in the transfection mixture for 24h hours then the media was
1302 then changed to the regular FBS containing media for an additional 24 hours. Dual
1303 Luciferase Reporter Assays (Promega) were started by incubating the cells for 15
1304 mins in 20 ml of 1x passive lysis buffer. Luciferase and renilla expression were then
1305 measured using the Glomax multi+ detection system (Promega). Luciferase
1306 expression values of the transposable elements and TAP2 were standardized by the
1307 renilla expression values and background expression values as determined by
1308 pGL3-Basic expression. Enriched motifs were found by analyzing the Dfam (Hubley
1309 et al., 2016) consensus sequences of the TEs found to have a regulatory ability
1310 significantly different from the pGL3 Basic[minP] empty vector using the MEME
1311 Suite. TomTom (Gupta et al., 2007) was used to match binding site motifs in the
1312 Jaspar database to the enriched motifs found in our data.

1313

1314 **Additional notes on analyses used throughout the project**

1315 All statistical analyses (DESeq2 analysis, Fisher's exact tests, logistic regressions,
1316 Spearman's correlations, Wilcoxon tests, General Linear Models, binomial test, and
1317 quantiles calculations) were performed using R v3.3.1. Figures were made with the
1318 package ggplot2 (Wickham, 2009) in R v3.3.1. Bedtools v2.25.0 (Quinlan et al.,
1319 2010) was used for overlap and closest feature/window analyses. All relevant scripts
1320 and pipelines are available online (https://github.com/ypar/cre_evo_primates.git). All

1321 supplementary data are also available online
1322 (https://github.com/ypar/cre_evo_primates_data.git).
1323
1324

1325 ## Supplementary references
1326

1327 Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez
1328 Banet, J., Billis, K., García Girón, C., Hourlier, T., et al. (2016). The Ensembl gene
1329 annotation system. Database 2016.

1330 Andrews S., (2010). FastQC: a quality control tool for high throughput sequence
1331 data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

1332 Bailey, T.L. (2011). DREME: motif discovery in transcription factor ChIP-seq data.
1333 Bioinformatics *27*, 1653–1659.

1334 Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li,
1335 W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and
1336 searching. Nucleic Acids Research *37*, W202-208.

1337 Brown, C.D., Mangravite, L., and Engelhardt B. (2013) Integrative Modeling of
1338 eQTLs and Cis-Regulatory Elements Suggests Mechanisms Underlying Cell Type
1339 Specificity of eQTLs. Plos Genetics http://dx.doi.org/10.1371/journal.pgen.1003649

1340 Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P.,
1341 Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner.
1342 Bioinformatics *29*(1), 15–21.

1343 Eden, E., Lipson, D., Yogev, S., and Yakhini, Z. (2007). Discovering Motifs in
1344 Ranked Lists of DNA Sequences. PLoS Comput Biol *3*, e39.

1345 Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool
1346 for discovery and visualization of enriched GO terms in ranked gene lists. BMC
1347 Bioinformatics *10*, 1–7.

1348 FANTOM Consortium and the RIKEN PMI and CLST (DGT) (2014). A promoter-level
1349 mammalian expression atlas. Nature *507*, 462–470.

1350 Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella,
1351 A.J., Searle, S.M.J., Amode, R., Brent, S., et al. (2016). Ensembl comparative
1352 genomics resources. Database 2016.

1353 Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H.,
1354 Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC Genome
1355 Browser Database: update 2006. Nucleic Acids Research *34*, D590–D598.

1356 Hubley, R., Finn, R.D., Clements, J., Eddy, S.R., Jones, T.A., Bao, W., Smit, A.F.A.,
1357 and Wheeler, T.J. (2016). The Dfam database of repetitive DNA families. Nucleid
1358 Acids Research *44*, D81-D89.

1359 Gupta, S, Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007).
1360 Quantifying similarity between motifs. Genome Biology *8*, R24.

1361 Innocenti, F., Cooper, G.M., Stanaway, I.B., Gamazon, E.R., Smith, J.D., Mirkov, S.,
1362 Ramirez, J., Liu, W., Lin, Y.S., Moloney, C., et al. (2011). Identification, Replication,
1363 and Functional Fine-Mapping of Expression Quantitative Trait Loci in Primary Human
1364 Liver Tissue. PLoS Genet *7*, e1002078.

1365 Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G.,
1366 Yandell, M., and Feschotte, C. (2013). Transposable Elements Are Major
1367 Contributors to the Origin, Diversification, and Regulation of Vertebrate Long
1368 Noncoding RNAs. PLoS Genet *9*, e1003470.

1369 Komsta, L. (2006). Processing data for outliers. R News, 6(2), 10-13.

1370 Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S.,
1371 Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., et al. 2012. ChIP-seq guidelines
1372 and practices of the ENCODE and modENCODE consortia. Genome Research
1373 *22*(9): 1813–1831.

1374 Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general
1375 purpose program for assigning sequence reads to genomic features. Bioinformatics
1376 *30*, 923–930.

1377 Liu, T. (2014). https://github.com/taoliu/MACS/wiki/Call-differential-binding-events.

1378 Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with
1379 BWA-MEM. arXiv:1303.3997v1

1380 Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change
1381 and dispersion for RNA-seq data with DESeq2. Genome Biology *15*, 1–21.

1382 McLeay, R.C., and Bailey, T.L. (2010). Motif Enrichment Analysis: a unified
1383 framework and an evaluation on ChIP data. BMC Bioinformatics *11*, 1–11.

1384 Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S.,
1385 Gnirke, A., Callan, C.G., Kinney, J.B., et al. (2012). Systematic dissection and
1386 optimization of inducible enhancers in human cells using a massively parallel
1387 reporter assay. Nat Biotech *30*, 271–277.

1388 Patwardhan, R.P., Hiatt, J.B., Witten, D.M., Kim, M.J., Smith, R.P., May, D., Lee, C.,
1389 Andrie, J.M., Lee, S.-I., Cooper, G.M., et al. (2012). Massively parallel functional
1390 dissection of mammalian enhancers in vivo. Nat Biotech *30*, 265–270.

1391 Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for
1392 comparing genomic features. Bioinformatics *26*, 841–842.

1393 R Core Team (2016). R: A language and environment for statistical computing. R.
1394 Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-
1395 project.org/.

1396 Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L.,
1397 Yakhini, Z., Weinberger, A., and Segal, E. (2012). Inferring gene regulatory logic

from high-throughput measurements of thousands of systematically designed promoters. Nat Biotech *30*, 521–530.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research *15*, 1034–1050.

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

Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park, T.J., Deaville, R., Erichsen, J.T., Jasinska, A.J., et al. Enhancer Evolution across 20 Mammalian Species. Cell *160*, 554–566.

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

Yates, A., Beal, K.,Keenan, S., McLaren, W., Pignatelli, M., Ritchie, G.R.S., Ruffler, M., Tylor, K., Vullo, A. and Flicek, P. (2015). The Ensembl REST API: Ensembl Data for any language. Bioinformatics *31*(1): 143–145.

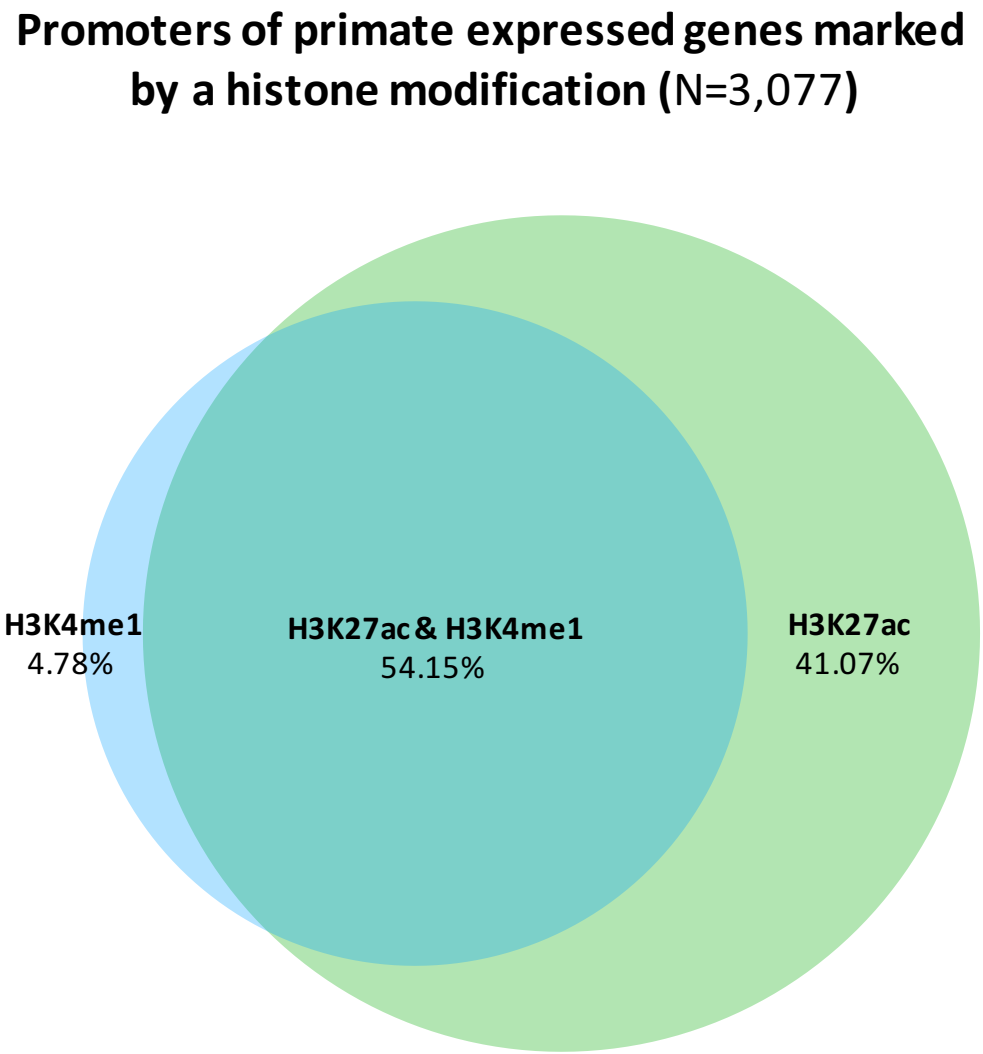
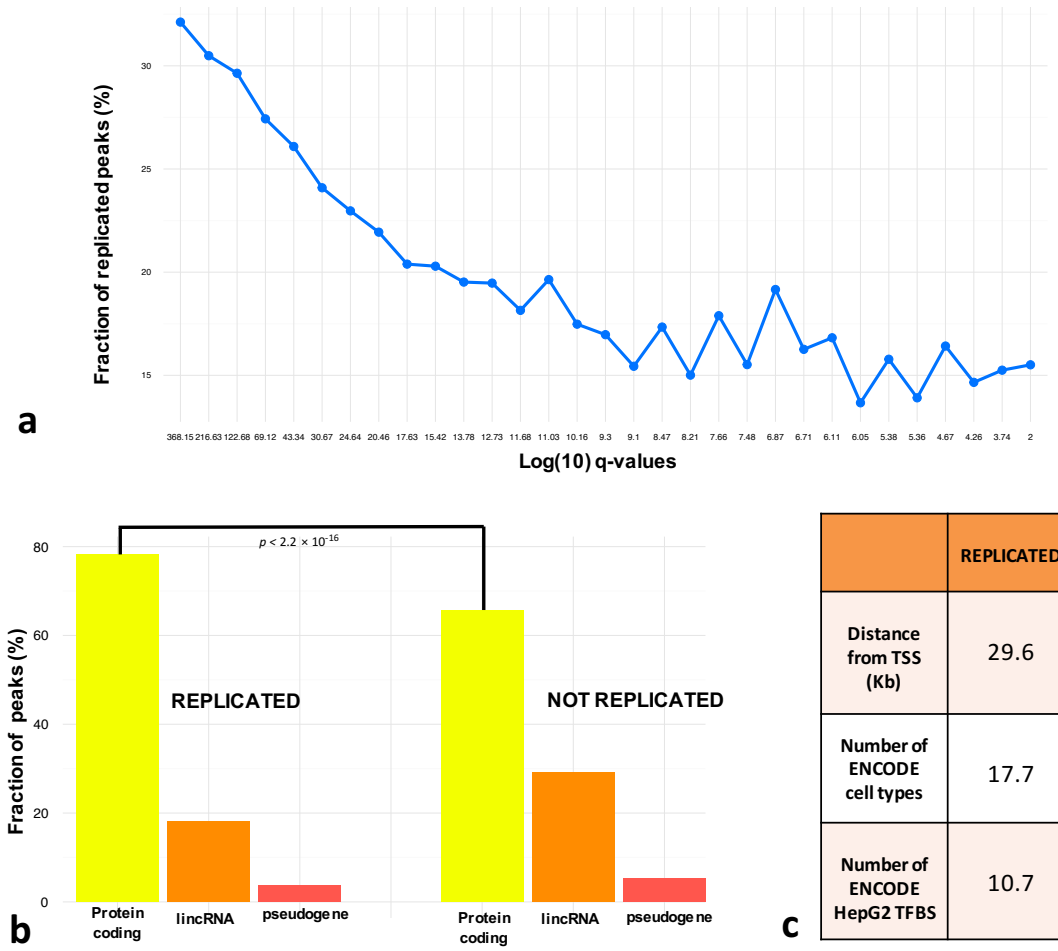


Figure S1: Promoters of genes that are expressed in primates are significantly histone modified. Venn diagram showing the distribution of histone marks on the promoters of genes that are expressed in the primate liver.



	REPLICATED	NOT REPLICATED
Distance from TSS (Kb)	29.6	90
Number of ENCODE cell types	17.7	9.3
Number of ENCODE HepG2 TFBS	10.7	3.3

Figure S2: Peaks bearing signatures of robust and broad regulatory activities are largely reproducible across studies. We compared our human H3K27ac data to a recent study focused on liver CREs in mammals (Villar et al., 2015). Overall, slightly less than 40% of the peaks identified by Villar et al. (2015) overlapped one of our H3K27ac regions. Further, 69.3% of the ENCODE HepG2 H3K27ac peaks were replicated in the Villar et al. (2015) dataset. We thus investigated possible features associated with ChIP-seq peak reproducibility.
(A) Peak discovery significance (q-value) is significantly correlated with cross-dataset reproducibility (logistic regression $p < 2.2\times10^{-16}$). (B) Replicated peaks are significantly more likely than non-replicated peaks to be associated with protein coding genes rather than with lincRNAs or pseudogenes (Fisher's exact test $p < 2.2\times10^{-16}$). (C) Replicated peaks are: 1) systematically closer to the nearest TSS (29.6 kb for replicated peaks, 90.8 kb for not replicated; Wilcoxon rank-sum test $p < 2.2\times10^{-16}$); 2) overlap chromatin accessible regions in significantly higher numbers of ENCODE cell types (an average of 17.7 cell types for replicated peaks and 9.3 cell types for unreplicated peaks; logistic regression $p < 2.2\times10^{-16}$); 3) contain a significantly higher number of transcription factor binding sites (TFBSs) per peak region as identified by ENCODE in HepG2 cells (10.7 TFBSs in the replicated peaks, 3.6 in the unreplicated; logistic regression $p < 2.2\times10^{-16}$).

Figure S3: Correlation of the normalized ChIP-seq read depths between human and marmoset. Human consensus peaks with orthologous in all the six species were split into two groups for this analysis: 1) regions with overlapping peaks present in both marmoset and human, and 2) regions with a peak present only in human. While human and marmoset normalized read counts were more highly correlated with each other in group 1 (Spearman's ρ = 0.67; $p < 2.2\times10^{-16}$), we found a nearly as strong correlation in group 2 (Spearman's ρ = 0.57; $p < 2.2\times10^{-16}$). These findings are consistent with the results of our differential histone modification state analyses, which demonstrated that only a small fraction of the 47,673 orthologous CREs (5.97%, FDR < 10%) are differentially modified, despite the fact that we had a much smaller total number of peak calls in the marmoset samples.

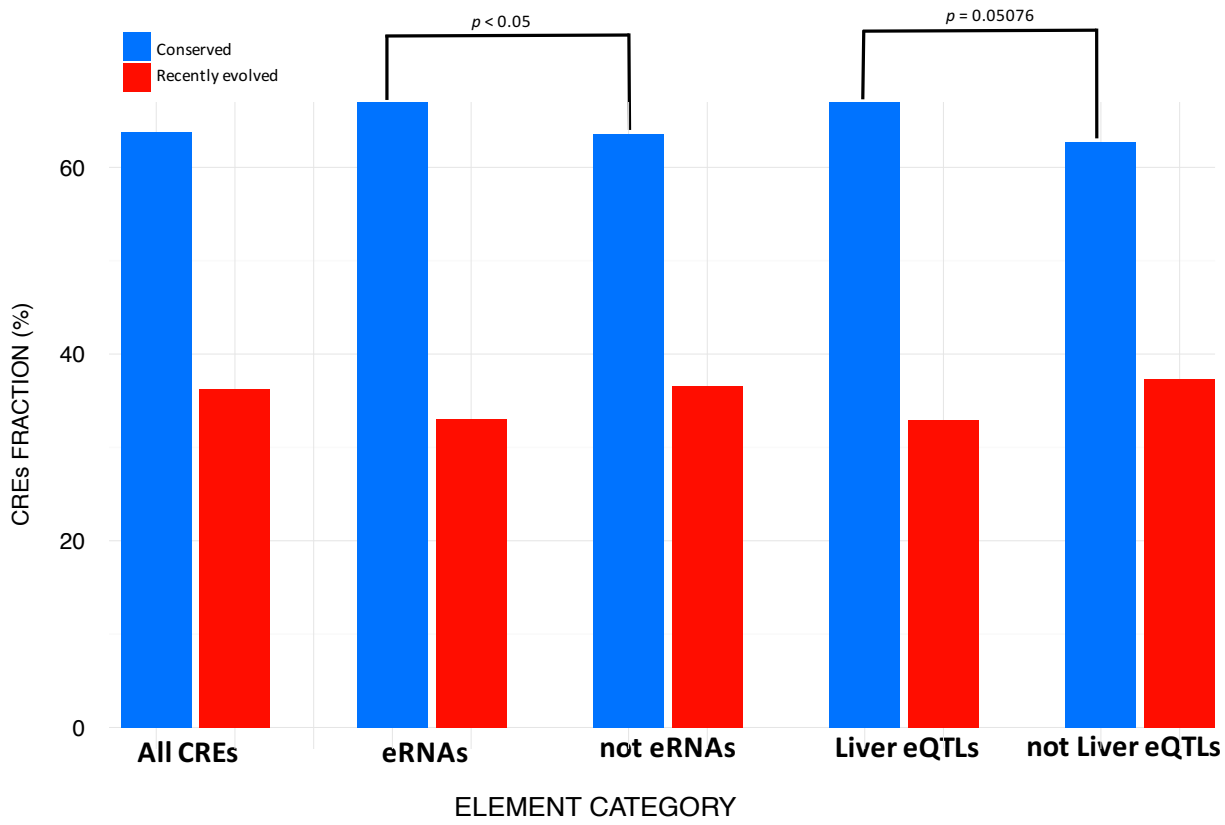


Figure S4: Enhancers with signature of transcription are conserved. Fraction of conserved and recently evolved CREs with signature of transcription (eRNA) based on FANTOM5 Consortium data. Fraction of conserved and recently evolved CREs with no signature of transcription (eRNA) based on FANTOM5 Consortium data, Fraction of conserved and recently evolved CREs overlapping and not overlapping liver eQTLs. Specifically, predicted human CREs overlapped 500 eQTLs detected in a recent study on human liver (Innocenti et al., 2011). We tested whether enhancers and promoters overlapping liver eQTLs would lean toward being more conserved or more labile than a random liver CRE, and we found that neither of these two conditions are satisfied (Fisher's Exact Test p = 0.05076), as we show that liver eQTLs behave as “average” liver regulatory elements.

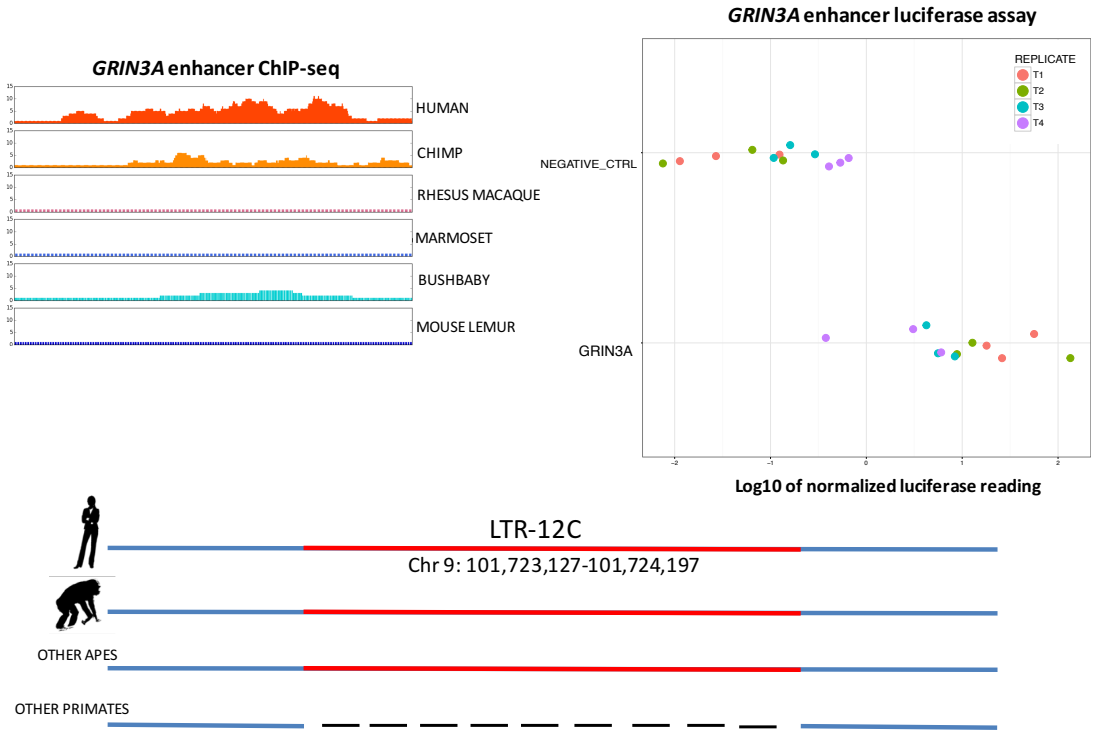


Fig S5: **Functional analysis on *GRIN3A locus*.** ChIP-seq read depth distributions and luciferase assays reporter activity for the CRE associated to *GRIN3A*.

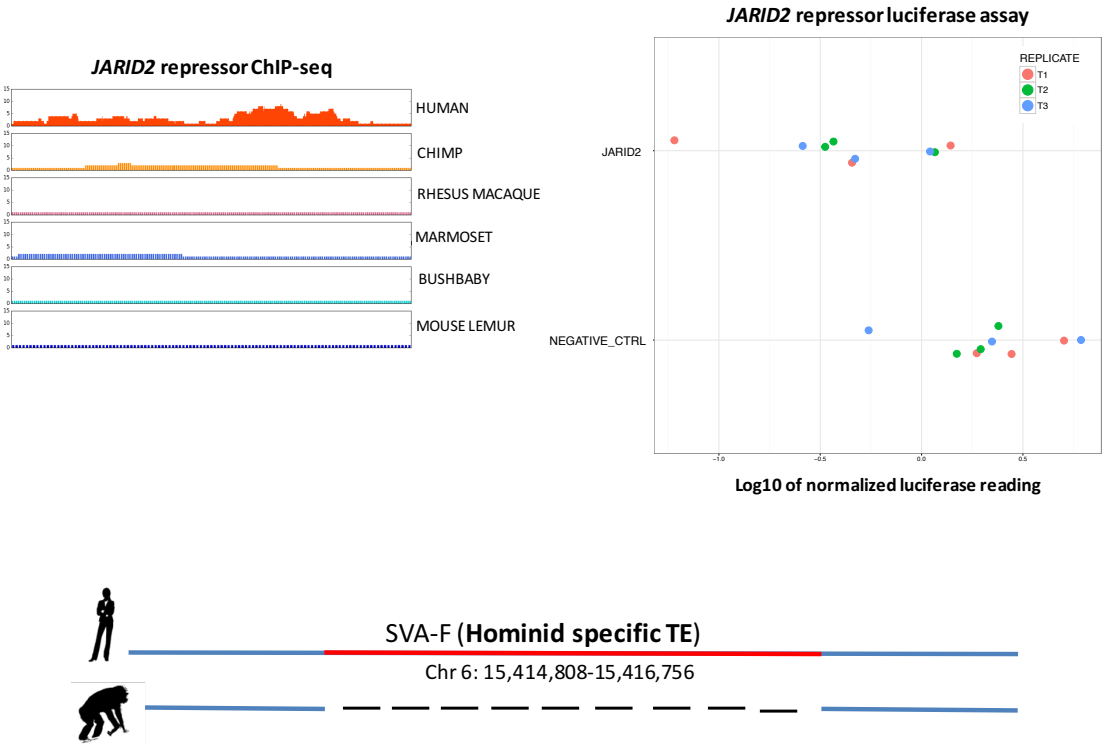


Fig S6: **Functional analysis on *JARID2 locus.*** ChIP-seq read depth distributions and luciferase assays reporter activity for the CRE associated to *JARID2*.

Figure S7: **Lineage specificity of enriched TEs.** Word-cloud representing the lineage specificity of the enriched TE families.