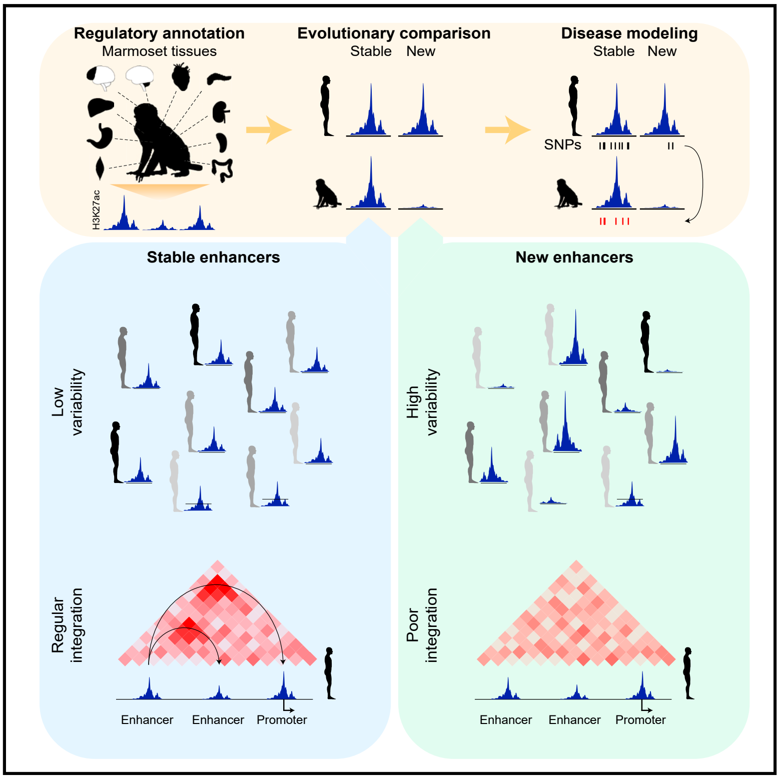
Cell Reports

Recently Evolved Enhancers Emerge with High Interindividual Variability and Less Frequently Associate with Disease

Graphical Abstract



Highlights

- Enhancer annotation in marmoset tissues for disease model suitability analysis
- Newly evolved enhancers are highly variable between individuals
- New enhancers are poorly integrated in the transcriptional machinery
- Disease-associated enhancers are more often conserved in marmoset

Authors

Bas Castelijns, Mirna L. Baak,
Geert Geeven, ..., Ivanela Kondova,
Wouter de Laat, Menno P. Creyghton

Correspondence

m.creyghton@erasmusmc.nl

In Brief

Modeling diseases in non-human species is complicated as many enhancers that regulate expression are species specific. Castelijns et al. demonstrate that species-specific enhancers are highly variable and poorly integrated in the regulatory network, suggesting their biological impact is low. Instead, disease-associated enhancers are typically conserved and can therefore be modeled.

Castelijns et al., 2020, Cell Reports *31*, 107799
June 23, 2020 © 2020
https://doi.org/10.1016/j.celrep.2020.107799

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

Recently Evolved Enhancers Emerge with High Interindividual Variability and Less Frequently Associate with Disease

Bas Castelijns,[1,5] Mirna L. Baak,[1,5] Geert Geeven,[1,5] Marit W. Vermunt,[1] Caroline R.M. Wiggers,[1,2] Ilia S. Timpanaro,[1] Ivanela Kondova,[3] Wouter de Laat,[1] and Menno P. Creyghton[1,4,6,*]
[1]Hubrecht Institute-KNAW & University Medical Center Utrecht, Utrecht, the Netherlands
[2]Department of Pediatric Hematology, University Medical Center Utrecht, Utrecht, the Netherlands
[3]Biomedical Primate Research Center, Lange Kleiweg 161, 2288GJ Rijswijk, the Netherlands
[4]Erasmus University Medical Center, Department of Developmental Biology, Wytemaweg 80, 3015 CN Rotterdam, the Netherlands
[5]These authors contributed equally
[6]Lead Contact
*Correspondence: m.creyghton@erasmusmc.nl

https://doi.org/10.1016/j.celrep.2020.107799

SUMMARY

Mutations in non-coding regulatory DNA such as enhancers underlie a wide variety of diseases including developmental disorders and cancer. As enhancers rapidly evolve, understanding their function and configuration in non-human disease models can have important clinical applications. Here, we analyze enhancer configurations in tissues isolated from the common marmoset, a widely used primate model for human disease. Integrating these data with human and mouse data, we find that enhancers containing trait-associated variants are preferentially conserved. In contrast, most human-specific enhancers are highly variable between individuals, with a subset failing to contact promoters. These are located further away from genes and more often reside in inactive B-compartments. Our data show that enhancers typically emerge as instable elements with minimal biological impact prior to their integration in a transcriptional program. Furthermore, our data provide insight into which trait variations in enhancers can be faithfully modeled using the common marmoset.

INTRODUCTION

The gene expression programs that dictate cellular behavior are controlled by non-coding gene regulatory elements (GREs) such as promoters and enhancers (Schoenfelder and Fraser, 2019). Alterations in GREs lead to a panoply of diseases, including developmental disorders and cancer (Rickels and Shilatifard, 2018). Furthermore, many disease-associated variants reside in enhancer sequences suggesting that complex diseases may be associated with combinations of common variants that modulate enhancer activity (Maurano et al., 2012). As such, the contributions of several common enhancer variants to disease susceptibility have now been resolved biologically (Soldner et al., 2016). During evolution, coding sequences remain relatively stable, while enhancer elements that regulate the expression of genes in an often cell-type-specific manner show rapid evolutionary turnover (Villar et al., 2014, 2015). As this rapid enhancer turnover is a hallmark of mammalian evolution, it is unclear to what extent diseases that are driven in part by enhancer variants or mutations can be correctly recapitulated in non-human model systems. Therefore, elucidation of their configuration in model systems is required to be able to determine whether disease states can be faithfully captured.

A promising animal model that is gaining attention to study human disease is the common marmoset (*Callithrix jacchus*), a new-world primate originating from the northeast of Brazil (Carrion and Patterson, 2012; Cyranoski, 2014). Due to its relatively small size, short gestation period, and early sexual maturation, it is increasingly being proposed as a more suitable model compared to mice or larger primates to study a variety of diseases such as age-related diseases, immunological disorders, and most prominently research into neuropsychiatric and neurodegenerative disorders (Carrion and Patterson, 2012; Cyranoski, 2014). To facilitate these efforts high-quality genome assemblies as well as transcriptome data have recently been generated (Shimogori et al., 2018; Worley et al., 2014). Nevertheless, insight into the gene regulatory network underlying these gene expression programs is still lacking. As many disease-associated variants are located in non-coding regulatory elements, it is thus crucial to understand the differences and similarities in gene regulation with humans to fully exploit the common marmoset as non-human primate model for biomedical studies.

Here, we annotate putative promoters and enhancers in distinct marmoset tissues and compare these to human and mouse data. We provide evidence that repurposed enhancers are in part misclassified interindividual variabilities. Following this, we find that most enhancers that are new in the human lineage emerge with high interindividual variability and often fail to contact genes, providing insight into their attenuated effect on gene expression (Berthelot et al., 2018). Consequently, enhancers containing DNA variants that link to trait variation or disease susceptibility tend to be more often conserved between species. Thus, our analyses provide new insights into the evolution and function of gene regulatory networks and provide a framework for modeling disease variant containing regulatory elements in the common marmoset.

RESULTS

Annotation of Regulatory DNA in Marmoset Tissue

To analyze gene regulation in marmosets, we annotated GREs in eight different marmoset tissues (colon, heart, kidney, liver, pancreas, skeletal muscle, spleen, stomach) and two brain regions (cerebellum and prefrontal cortex (Castelijns et al., 2020)) (Figure 1A). We used chromatin immunoprecipitation followed by sequencing (ChIP-seq) for histone 3 lysine 27 acetylation (H3K27ac), a robust mark with a good balance between sensitivity and specificity (both ~70%) to identify active promoters and enhancers (Arnold et al., 2013; Bonn et al., 2012; Nord et al., 2013; Vermunt et al., 2014; Villar et al., 2015). H3K27ac serves as a footprint for activity of its acetyltransferase, *p300/CBP*, which is sufficient to convey enhancer activity upon a genomic region (Hilton et al., 2015) and responsible for the acetylation of several histone residues (Calo and Wysocka, 2013) as well as the polymerase complex itself thus affecting transcriptional output (Boija et al., 2017; Schröder et al., 2013). In addition, we performed ChIP-seq for histone 3 lysine 4 tri-methylation (H3K4me3), which is specifically found on active transcriptional start sites (TSSs) (Guenther et al., 2007), to distinguish between putative active promoters and putative active enhancers. Data were within quality standards (Landt et al., 2012) and reproducible between biological replicates (H3K27ac: average $\rho = 0.86$; H3K4me3: average $\rho = 0.96$, Figures 1B and 1C; Table S1) and data published previously for marmoset liver (Villar et al., 2015) (Figures S1A and S1B).

In total, we annotated 60,824 H3K27ac-enriched and 21,136 H3K4me3-enriched regions in the marmoset genome across all marmoset tissues combined (Table S2; Figure S1C). Most (55%) of the H3K4me3-enriched GREs overlap an annotated marmoset TSS (Figure S1D). Moreover, as the human genome is more extensively annotated, mapping of all human TSSs on the marmoset genome showed that ~75% of the H3K4me3-enriched regions that were annotated in marmoset overlay a marmoset and/or human TSS (Figure S1D). Of the remaining H3K4me3-enriched regions, the majority (67%) are located in non-coding DNA (Figure S1E), of which most (84%) are actively transcribed, as determined using RNA sequencing (RNA-seq) data derived from different marmoset tissues (Figure S1F) (Cortez et al., 2014). These could therefore represent marmoset-specific splice variants, novel genes, or actively transcribed GREs in the marmoset genome. Mapping of these regions to the human genome revealed that these regions are enriched for species-specific DNA sequences not found in humans (4.7-fold enrichment, $p < 2.2e{-}16$, Fisher's exact test).

Of all 60,824 H3K27ac-enriched regions in the marmoset genome, 42,128 did not co-localize with H3K4me3 enrichment, thus representing putative active enhancers (Bonn et al., 2012; Nord et al., 2013; Vermunt et al., 2014; Villar et al., 2015), although additional analyses are required for each putative active enhancer or promoter separately to confirm its activity within the genome. For the purpose of simplicity throughout the manuscript, H3K27ac-enriched regions will be referred to as active GREs, or when specified, as active promoters and enhancers. Most enhancers were found active in only a single tissue, consistent with their known tissue specificity (Figures 1D and S1G). Functional analysis of genes in proximity of active enhancers, as defined by GREAT (McLean et al., 2010), reflected the tissues in which they were identified (Figure S1H).

A Proportion of Repurposed GREs Are Misclassified Interindividual Variabilities

To analyze to what extent human regulatory elements can be faithfully modeled in the common marmoset, we compared our data to human datasets of matching tissues (Kundaje et al., 2015; Shen et al., 2012; Villar et al., 2015). Using the same analysis as was done for marmoset samples, we identified 72,900 H3K27ac and 22,426 H3K4me3-enriched regions in human samples. Consistent with the evolutionary distance to their common ancestors, 39.7% of human GREs could not be mapped to the mouse genome, while only 21.4% of human GREs could not be mapped to the marmoset genome (Figure 1E). We combined all marmoset and human GREs by selecting GREs that could be mapped on both species' genomes (39,153 regions) using reciprocal liftover (see STAR Methods) (Figure 1F). Of these shared regions, 89.8% were enriched in the same tissue between the species (Figure 1G). For example, GREs linked to *KLF5* were enriched in human colon and stomach, a tissue-specific pattern that was conserved in marmoset tissues (Figure S2A). Consistent with previous data (Vermunt et al., 2016; Vierstra et al., 2014), we found a fraction of H3K27ac-enriched regions (10.2%) that were repurposed between tissues and species (Figure 1G), with no major influence of confounder variables observed (Figure S2B). Repurposed GREs are elements that lose activity in one tissue while gaining it in another between two species. For instance, an enhancer in the *TMEM175* gene, a risk factor in Parkinson disease (Jinn et al., 2017) (Figure 1H), switches its broad H3K27ac signal specifically to the human brain while being enriched in several other tissues in marmoset. Closer analysis of these regions revealed that most repurposing occurs between tissue-specific GREs (Figure S2C). To assess whether this repurposing was specific between human and marmoset, we compared our data to H3K27ac enrichment in matching mouse tissues (Shen et al., 2012). We observed that human-marmoset repurposed GREs are mainly specific to the primate lineage since these regions were often either depleted from H3K27ac signal in mouse tissues (53%) or their sequence could not be mapped onto the mouse genome (13.1%) ($p < 2.2e{-}16$, Figures S2D and S2E). However, a proportion of human-marmoset repurposed elements (5.8%) showed conservation of H3K27ac enrichment between human and mouse as well as between marmoset and mouse; e.g., the mouse enhancer was enriched in both tissues in mouse (called "ambiguous,"

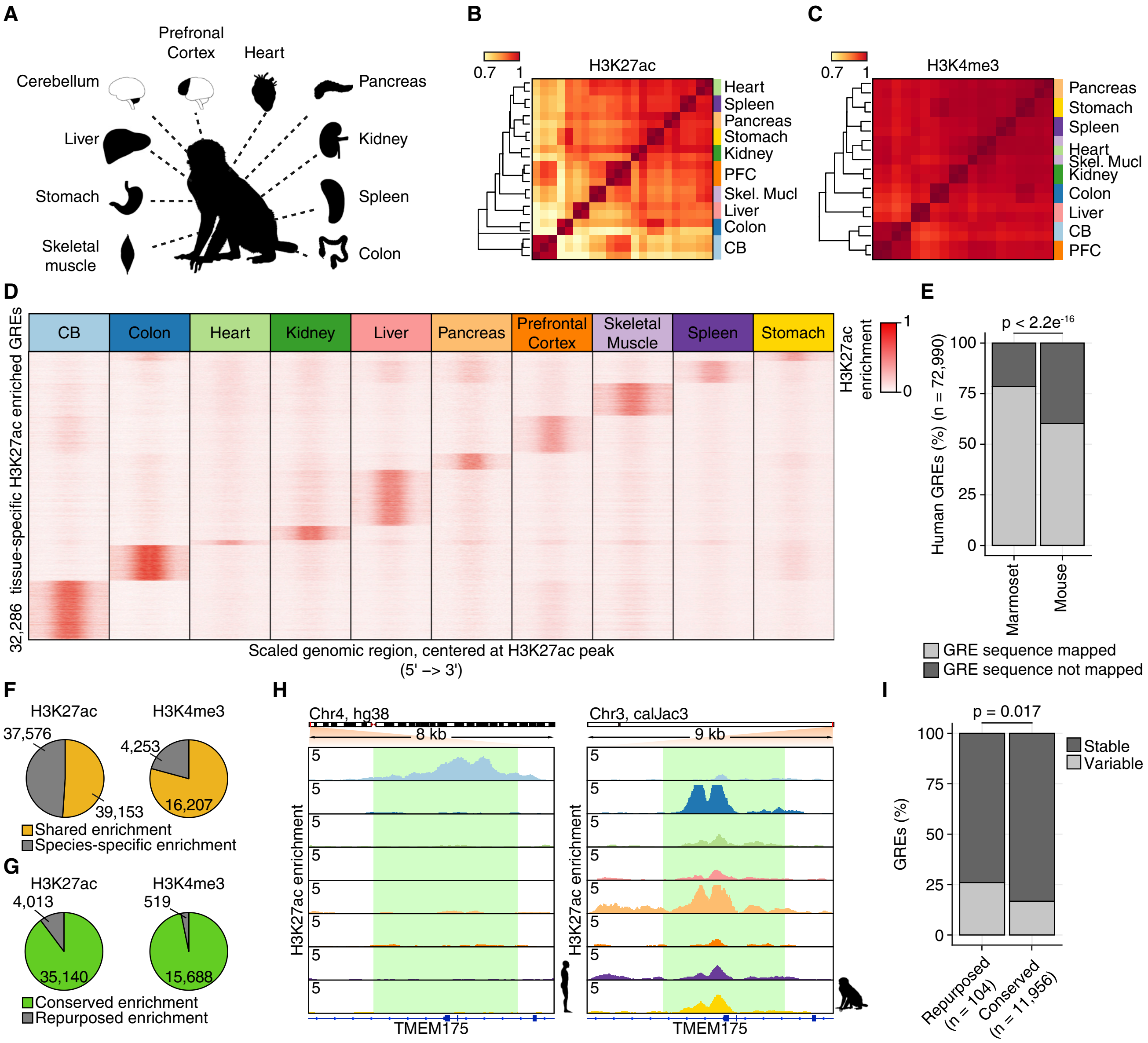


Figure 1. Conservation of Regulatory Elements across the Primate Lineage

(A) Schematic representation of the common marmoset and all the tissues sampled.

(B) Hierarchical clustering of all sampled tissues based on H3K27ac-enriched regions (n = 60,824). Correlation map colors indicate Pearson's correlation coefficients between samples. Side bars are color coded to indicate different tissues. Pearson distances are represented by the tree on the left. (CB, cerebellum; PFC, prefrontal cortex; Skel. Mucl, skeletal muscle).

(C) Hierarchical clustering as in (B) based on H3K4me3-enriched regions (n = 21,136).

(D) Heatmap depicting the H3K27ac enrichment in different tissues across 5′ to 3′ scaled tissue-specific GREs, centered around the H3K27ac peak. Heatmap colors indicate RPKM normalized H3K27ac enrichment.

(E) Bar plot depicting the percentage of human GRE sequences that could be mapped to marmoset and mouse genomes. Dissimilarity between the ratio was calculated with a Fisher's exact test.

(F) Pie charts showing the degree of shared activity between marmoset and human for both H3K27ac and H3K4me3-enriched regions.

(G) Pie charts showing the degree of repurposed activity between human and marmoset for both H3K27ac and H3K4me3-enriched regions.

(H) ChIP-seq tracks of RPM normalized H3K27ac enrichment in different tissues from both human and marmoset samples across an 8 kb region within the *TMEM175* gene. The repurposed GRE is highlighted in green.

(I) Bar plot depicting the percentage of GREs that have variable H3K27ac enrichment between different human LCL samples for both conserved as well as repurposed regions. Dissimilarity between the ratio was calculated with a Fisher's exact test.

See also Figures S1 and S2.

Figure S2D). For example, an enhancer in the *FLNB* gene was annotated as repurposed from cerebellum and liver to heart tissue between human and marmoset (Figure S2F). Instead, analysis of the mouse data showed that this enhancer was enriched in all of these tissues, suggesting that some repurposing events may rather represent species-specific losses of H3K27ac enrichment.

Surprisingly, we found that elements that were repurposed between human and marmoset and enriched for H3K27ac in any mouse tissue were also more likely to be repurposed to a different tissue in mouse ($p < 2.2e–16$, Figures S2D and S2E). For example, a GRE within the *GLT8D2* gene is repurposed between human heart and marmoset liver while being enriched in different tissues in mouse (Figure S2G). A similar pattern in H3K27ac enrichment was observed for the promoter of *GLT8D2*. As repurposing the same elements multiple times across species seemed unlikely, we wondered whether these elements were truly repurposed or merely shared between tissues but variable between individuals within a species. To analyze this, we compared our GREs to data generated in lymphoblastoid cell lines (LCLs), analyzing enhancer variability across 19 different human individuals using ANOVA (see STAR Methods; Figure S2H) as shown previously (Kasowski et al., 2013). By assessing variation between samples from different human individuals compared to the variation of replicate samples from the same individuals, we found that GREs that were repurposed were more likely to be variable between LCL samples derived from different individuals ($p = 0.017$, Figure 1I), suggesting that a proportion of repurposed elements may be misclassified as such due to interindividual variability.

Recently Evolved Regulatory DNA Is Predominantly Variable between Individuals

To further analyze whether there was a relationship between recent evolutionary changes at regulatory DNA and individual variation, we reanalyzed H3K27ac data from ten human, seven chimpanzee, and seven rhesus macaque LCLs (Zhou et al., 2014), using the same analysis as the marmoset tissues. This allowed us to analyze individual variation and evolution in the same cell type in a more quantitative manner, while also providing a more precise annotation of when these regulatory elements emerged as compared to directly contrasting human and marmoset. We therefore generated a non-redundant list of H3K27ac-enriched regions based on samples from all three primate species using reciprocal liftover. We annotated 54,793 regions that could be mapped on the genomes of all three primate species (see STAR Methods, Figure 2A), of which 40,924 were classified as putative enhancers based on not overlaying a known TSS. Using DESeq, 3,347 gains and 1,666 losses of H3K27ac enrichment were identified at enhancers, which were specific to the human lineage (Figures 2B and S3A), with expression analysis of nearby genes showing the typical correlation of modest expression change with enhancer change (Figures 2C and S3B).

We sub-classified three distinct classes of putative enhancers. Enhancers that increased H3K27ac enrichment in humans compared to the other two primate species (‘‘gains’’), elements that were new in humans based on absence of H3K27ac enrichment in the other two primates using a stringent background model (‘‘new’’), and elements that were stable and not differentially enriched across primates (‘‘stable’’). For each region, we determined its variability using ANOVA in the panel of 19 human LCL lines as described above (Kasowski et al., 2013). Surprisingly more than half of the elements that recently evolved were classified as variable between human individuals ($p < 2.2e–16$, Figure 2D). Increased variability was observed for both enhancers classified as gains (44.8%) and those classified as new (65.3%). In contrast, elements that were classified as stable across primates were less frequently (25.2%) variable between humans ($p < 2.2e–16$, Figure 2D). The increase in variability for recently evolved enhancers was not due to major differences in H3K27ac enrichment at recently evolved elements, as the relationship between variability and H3K27ac enrichment for gains and new enhancers was not significantly different from stable enhancers or enhancers in general ($p = 0.2$, Figure S3C). We also found no indication that line to line variability was a factor in our analysis as flagging enhancers that were variable between 5 independent LCL lines from the same individual (Ozgyin et al., 2019) did not affect our observations (Figure S3D). In addition, we found no indication that variability of enhancers that were evolutionary new was linked to differences in ancestry between the individuals analyzed (see STAR Methods; Figures S3E–S3H). For example, a putative enhancer that recently evolved near *ING1* was highly variable across individuals, independent of their ancestry (Figures 2E and 2F).

We and others recently established a link between recently evolved enhancers and cell-type specificity suggesting that cell-type specificity and variability may be linked (Fish et al., 2017; Vermunt et al., 2016). By comparing H3K27ac enrichment at putative enhancers in lymphoblastoid cells with H3K27ac datasets generated in 7 unrelated cell types (Figures 3A and 3B) (Ernst et al., 2011), we confirmed that recently evolved enhancers were more often cell type specific (Figures 3C–3E). In addition, we found that cell-type-specific enhancers were also more often variable in LCLs between individuals (28% increase, $p < 2.2e–16$, Figure 3F). To test whether there was a direct dependence between all three categorical characteristics; variability, cell-type specificity, and enhancer type (gain/new/stable/all), we used LLM3D (Geeven et al., 2011), a method to assess the interdependences of the different variables of these enhancers. We found that evolutionarily new and variable enhancers differ significantly in their rate of cell-type specificity ($p = 2.0e–218$). This demonstrates that the degree of variability between individuals and cell-type specificity are two independent characteristics of recently evolved regulatory elements. We conclude that high interindividual variability is a property of the majority of recently evolved enhancers.

Recently Evolved Enhancers Less Frequently Engage in Functional Contacts with Gene Promoters

As the gain or loss of an enhancer can be buffered by the presence of other enhancers that regulate the same gene (Berthelot et al., 2018; Hong et al., 2008; Perry et al., 2010; Vermunt et al., 2016), we next assessed whether enhancer variability was also linked to the number of other enhancers engaging their target gene. As gene proximity is an imperfect measure to couple enhancers to genes, we increased the specificity of our analysis

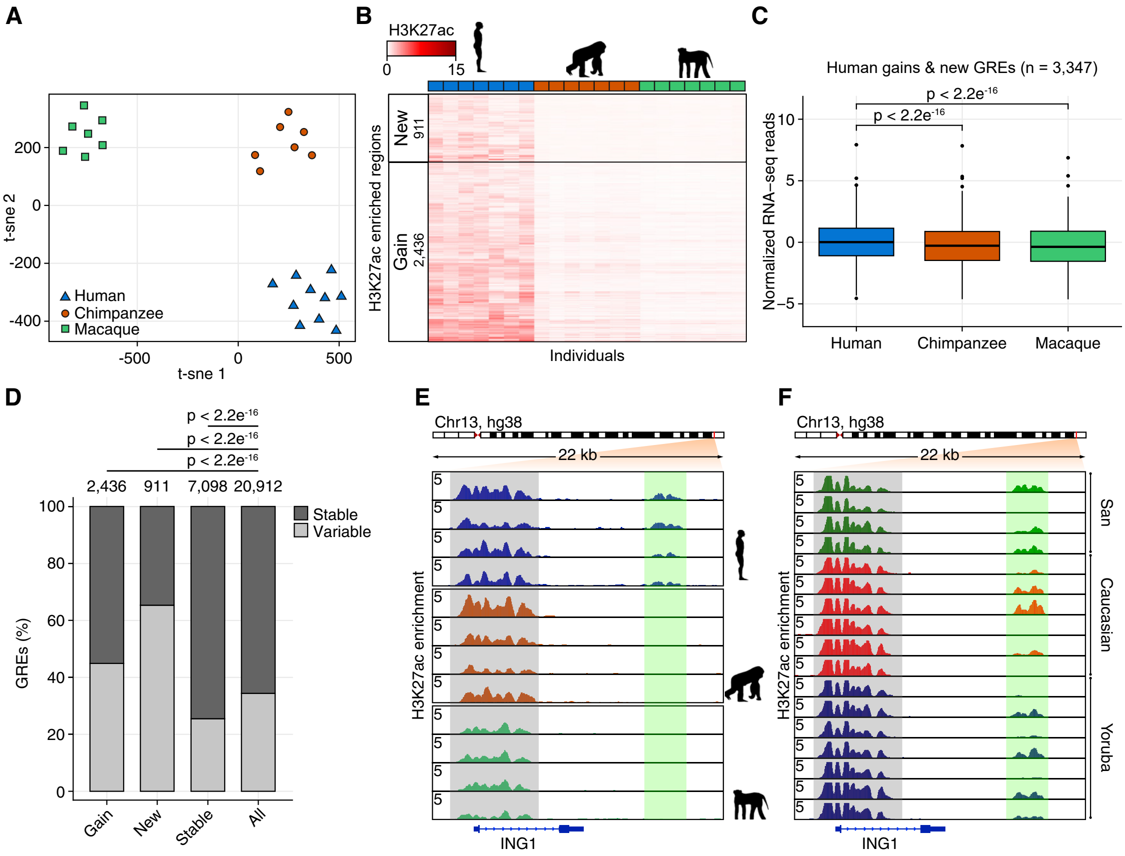


Figure 2. Variability of Enhancer Is Linked to Evolutionary Novelty

(A) t-Stochastic Neighbor Embedding (t-SNE) analysis of 10 human, 7 chimpanzee, and 7 macaque samples, based on a non-redundant list of H3K27ac-enriched regions mappable on all three species (n = 54,793).

(B) Heatmap of 911 evolutionary new enhancers and 2,436 enhancer gains in human compared to chimpanzee and macaque. Read counts were normalized for library size and length of the region. Individual samples per species are aligned at the x axis; H3K27ac-enriched regions are depicted on the y axis. Heatmap color represents normalized H3K27ac enrichment.

(C) Boxplot depicting gene expression in human, chimpanzee and macaque for genes linked to evolutionary enhancer gains. Values are log2 zero-mean normalized RNA-seq counts. Enhancers were linked based on proximity. Dissimilarity between the distributions was calculated using a Student's t test.

(D) Bar plot depicting the percentage of GREs that have variable H3K27ac enrichment in different human LCL samples for different categories of enhancers as indicated. Dissimilarities between the ratios was calculated using a Fisher's exact test.

(E) ChIP-seq tracks of RPM normalized H3K27ac enrichment in LCL cells from different primate species as indicated across a 22 kb region containing the *ING1* gene. A human-specific GRE is highlighted in green. A conserved GRE is highlighted in gray.

(F) ChIP-seq tracks showing the same region as in (E) for multiple human LCL samples from distinct individuals and distinct ancestries. The same human-specific GRE as in (E) is highlighted in green.

See also Figure S3.

by integration of HiC data from an LCL using a new method to determine distal contacts at kilobase resolution (Geeven et al., 2018; Rao et al., 2014). We analyzed 422,144 reciprocal contacts between 155,084 viewpoints (GREs, TSSs, and CTCF sites) and their anchors, including 84,855 reciprocal contacts with the H3K27ac-enriched putative enhancer elements. Using direct interactions over proximity-based analysis, we found that enhancers that are evolutionarily new and variable between individuals are not more likely to contact gene promoters with multiple GREs suggesting variability and buffering by alternative enhancers are not directly linked (p = 0.31, Figure S4A).

Unexpectedly, we found that new enhancers were less frequently in contact with any other anchor (p = 2.8e–6, Figure 4A). This modest drop in contact frequency was specific for enhancers that were classified as new and not for existing enhancers that were classified as gains (p = 0.047). As both classes of enhancers are more often variable, the specificity for new enhancers to be less frequently engaged is unlikely the result of increased variability at new enhancers. Analyzing this reduction in contact frequency for new enhancers further, we found a much stronger decrease in contact frequency when assessing contacts with TSSs (p < 2.2e–16, Figure 4B). A similar reduction in TSS contacts was observed for new enhancers when assessing only those enhancers that were significantly enriched for H3K27ac in the GM12878 cell line, in which the HiC data were generated in (Figure S4B). This suggests that the reduction in

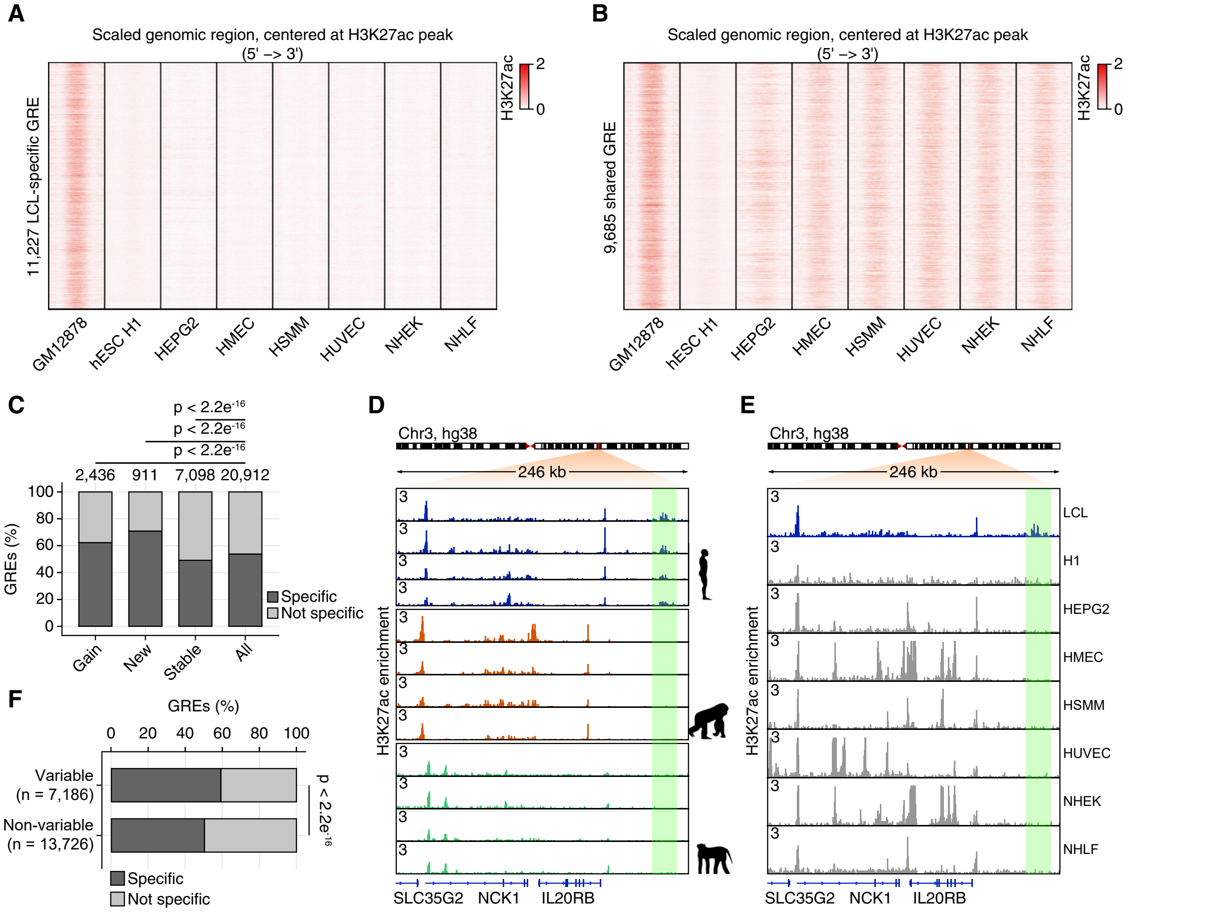


Figure 3. Cell-Type Specificity of Novel GREs

(A) Heatmap showing RPKM normalized H3K27ac enrichment in different cell types on all GREs defined as LCL cell type specific (n = 11,227). Regions are scaled 5′ to 3′ and centered around the H3K27ac peak. Heatmap colors indicate RPKM normalized H3K27ac enrichment.

(B) Heatmap as in (A) but for all non-LCL cell-type-specific GREs (n = 9,685).

(C) Bar plot depicting the percentage of GREs, classified by their evolutionary status as indicated on the axis, that are LCL cell type specific or shared with one of the other cell types. Dissimilarities between the ratios were calculated with a Fisher's exact test.

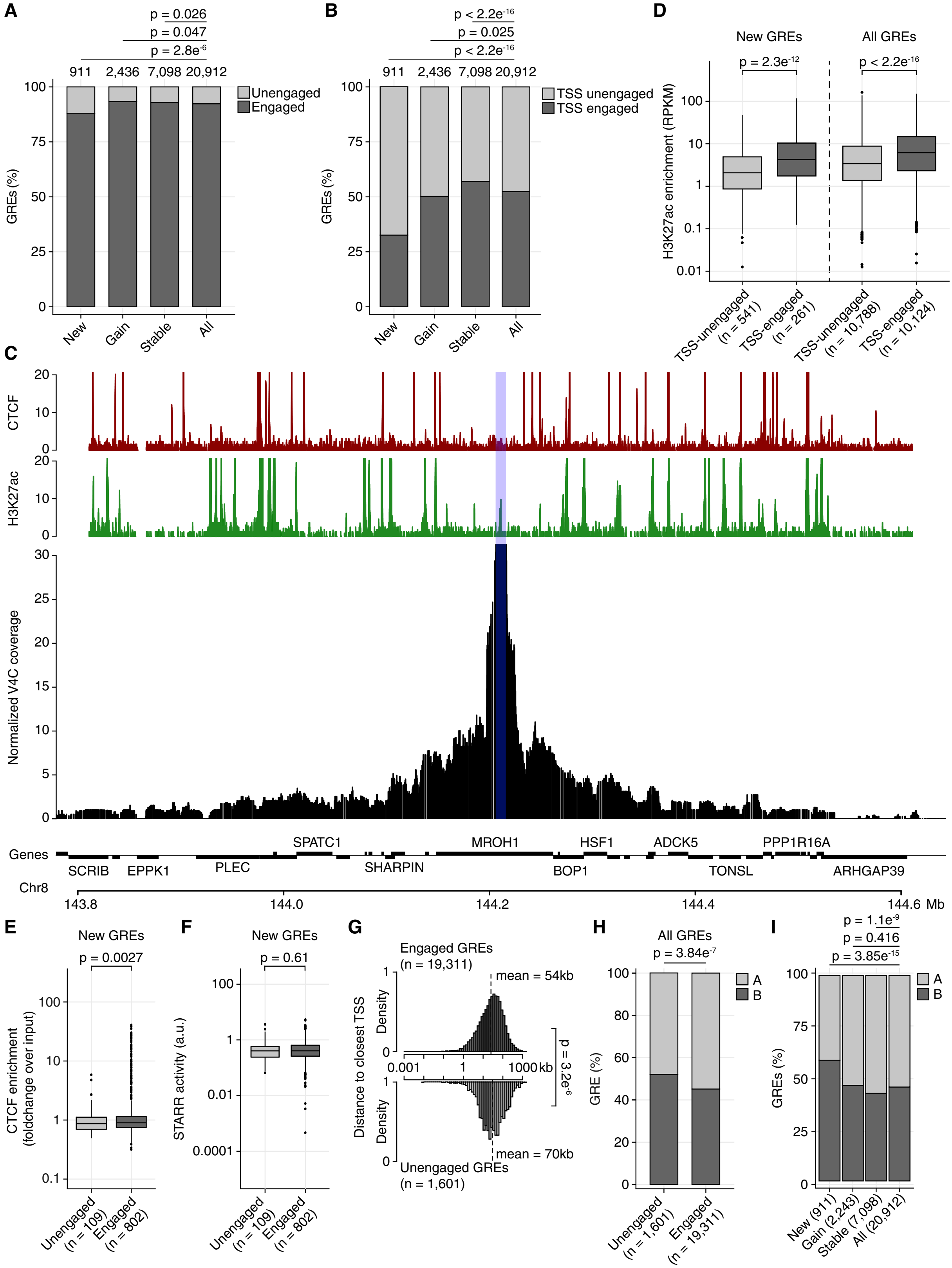
(D) ChIP-seq tracks of RPM normalized H3K27ac enrichment on LCL samples of different primates as shown, across a 246 kb genomic regions containing several genes and a human-specific GRE (highlighted in green).

(E) ChIP-seq tracks as in (D) for the different human cell types. The same human-specific GRE is highlighted in green.

(F) Bar plot depicting the percentage of cell-type-specific GREs for regions that are variable between human individuals and those that are not variable. Dissimilarity between the ratios was calculated with a Fisher's exact test.

promoter contacts at new enhancers is not due to the HiC analysis being limited to a single cell line. The opposite effect (increased TSS contact frequency) was observed for enhancers that were classified as stable (p < 2.2e–16, Figure 4B). Putative enhancers that failed to contact other regions were classified as unengaged enhancers and those that failed to contact promoters as TSS unengaged enhancers. For instance, a putative enhancer in the *MROH1* gene is new in humans and is not engaging any of its surrounding elements above the detection limit despite a variety of active GREs being present in the region (Figures 4C and S4C).

Further analysis of these unengaged enhancers revealed that they showed attenuated H3K27ac enrichment compared to enhancers that do contact other GREs (Figure S4D). This was not a specific property of enhancers that recently evolved as enrichment of H3K27ac was also lower for unengaged enhancers regardless of their evolutionary status (p < 2.2e–16, Figure S4D). While reduced H3K27ac enrichment could be the result of detection issues at these sites, it could also be the consequence of not engaging a promoter, as promoters are typically highly enriched for this modification. We therefore analyzed enhancers that did engage in detectable contacts with other GREs but not with promoters (TSS unengaged). We found that TSS unengaged enhancers that were able to contact other GREs were also less enriched for H3K27ac compared to enhancers that did contact promoters (p < 2.2e–16, Figure 4D) regardless of their evolutionary status. This argues against detection issues at new and unengaged elements and for a lack of promoter engagement as a basis for the reduction in H3K27ac enrichment seen at these elements. Moreover, we observed only a minor depletion of



(legend on next page)

CTCF binding at unengaged enhancers (Figure 4E), again suggesting that the lack of engagement and lower H3K27ac enrichment is not due to an inability to detect signal at these regions. Finally, further analysis of enhancer activity of new enhancer regions in humans using mass parallel reporter assays (Wang et al., 2018) revealed no difference in intrinsic enhancer activity between engaged and unengaged enhancers (Figures 4F and S4E). Thus, the lower H3K27ac enrichment at unengaged enhancers can at least in part be attributed to their failure to interact with an H3K27ac-enriched promoter and not due to a lack of signal or intrinsic activity.

Interestingly, analyzing the genomic positions of unengaged enhancers revealed that they are typically located further away from promoters compared to regular enhancers (p = 3.2e–6, Figure 4G). In addition, they are more often found in inactive B compartments, as identified in GM12878 cells (Rao et al., 2014), that are associated with closed and repressive chromatin (Rowley and Corces, 2018) (Figures 4H, S4F, and S4G). Enrichment in B-compartments was also observed for new enhancers, regardless of whether they were specific to the GM12878 cell line (Figures 4I and S4H). As expected, gene expression of proximal genes was not impacted by the emergence of new enhancers that are unengaged (Figures S4I and S4J). Therefore, we propose that recently evolved enhancers more often emerge further away from TSSs and are less likely to engage in productive contacts. Thus, even though these regions are intrinsically active they are less likely to impact gene expression.

Enhancers Containing Trait-Associated Variants Are More Often Conserved

As enhancers that recently evolved within a species are less frequently integrated in a regulatory network and may therefore be more often biologically irrelevant, we asked what that would mean for the utility of marmosets as non-human model organism for human disease. Using the tissue datasets described above, we assessed which human GREs contained known variants (single nucleotide polymorphisms [SNPs]) that are coupled to known phenotypic traits including various diseases (see STAR Methods; Figure 5A) and assessed their conservation in both marmoset and mouse tissues. We observed that H3K27ac enrichment of trait variant containing GREs was more likely to be conserved between species across evolution compared to GREs without trait variants (p < 2.2e–16, Figure 5B). In contrast, enhancers that are new in humans are also depleted for trait variants (Figure 5C). For example, the *BSN* gene, a regulator of neurotransmission, contains several brain-specific GREs that are conserved between human, marmoset, and mouse (Figure 5D). These contain several SNPs associated with neurodegenerative diseases such as Parkinson disease and Alzheimer disease (Leslie et al., 2014).

In agreement with trait variants occurring preferentially at conserved enhancers, we also observed that these variant containing enhancers were less frequently variable within the human lineage compared to other GREs (Figure 5E). This may be counterintuitive as variation is expected in trait variant containing GREs (Kasowski et al., 2013). However, the strong interindividual variation that characterizes recently evolved enhancers may not reflect the subtle effects of common trait variation (Tam et al., 2019). Furthermore, common variants in GREs are expected to alter GREs that are of impact on their target gene, which would disfavor their appearance in enhancers that recently evolved. In agreement with trait variant containing enhancers being more often conserved and functional, we observed that these were less often unengaged enhancers and more frequently contacted a gene promoter (p = 1.4e–15, Figure 5F). Moreover, trait variant containing enhancers were intrinsically more active in reporter assays than GREs without trait variants (p < 2.2e–16, Figure 5G). Overall these results demonstrate that, while enhancers that recently evolved are often species specific, variable and unengaged, trait variants typically occur in regulatory DNA that is conserved and likely functional.

DISCUSSION

While rapid turnover of regulatory DNA has put non-coding elements in the spotlight of evolutionary research and has raised questions about the validity of other species as model organisms, the impact of these novel elements on phenotypical

Figure 4. New Enhancers Are More Frequently Unengaged

(A) Bar plot depicting the percentage of enhancers that have a reciprocal HiC interaction with any anchor for the different evolutionary categories as indicated on the axis. Dissimilarities between the ratios were calculated using a Fisher's exact test.

(B) Bar plot as in (A) showing the percentage of enhancers reciprocally contacting a TSS.

(C) Virtual 4C profile using an enhancer that is evolutionary new and unengaged, located within the *MROH1* gene, as a viewpoint (highlighted in blue). Normalized ChIP-seq enrichment for H3K27ac (green) and CTCF (red) in GM12878 cells are shown in the top two tracks.

(D) Boxplots depicting normalized H3K27ac enrichment on GREs that are either contacting a TSS (TSS-engaged) or that are TSS-unengaged. Data are shown for both evolutionary new and for all enhancers. Dissimilarity between the distributions was calculated using a Student's t test.

(E) Boxplots depicting CTCF enrichment (fold change over input as defined by ENCODE) for enhancers that are evolutionary new in humans and are either unengaged or engaged as indicated on the axis. Dissimilarity between the distributions was calculated using a Student's t test.

(F) Boxplots depicting the maximum STARR activity of enhancers that are evolutionary new in humans and either unengaged or engaged as indicated on the axis. Dissimilarity between the distributions was calculated using a Student's t test.

(G) Histograms depicting the absolute distance of human-specific elements to their closest TSS for both engaged (upper) and unengaged (lower) GREs. Dissimilarity between the distributions was calculated using a Student's t test.

(H) Bar plot depicting the percentage of both engaged and unengaged GREs that are located in either the A- or B-compartments as defined in the GM12878 cell line. Dissimilarity between the ratios was calculated with a Fisher's exact test.

(I) Bar plot as in (H) for the different evolutionary enhancer categories as indicated on the x axis. Dissimilarities between the ratios were calculated with a Fisher's exact test.

See also Figure S4.

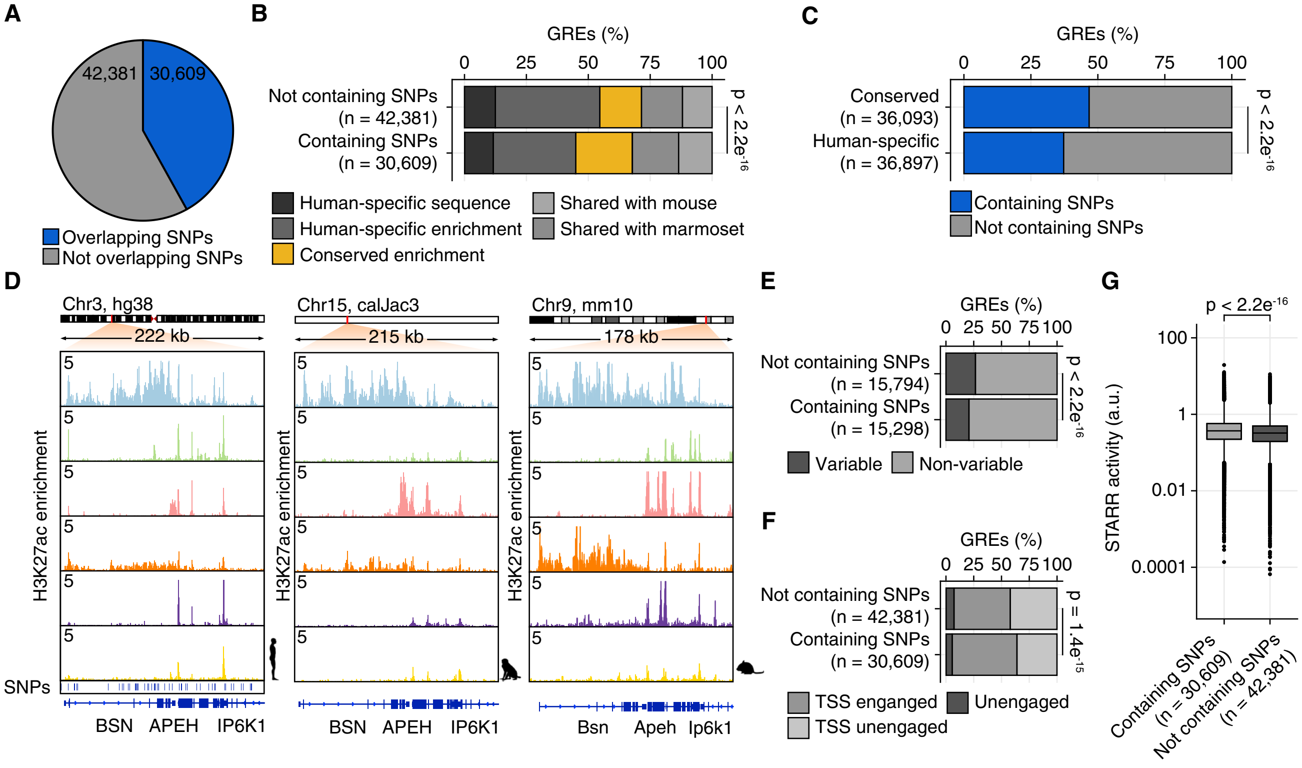


Figure 5. DNA Trait Variant Containing GREs Are Conserved Across Evolution

(A) Pie chart depicting the number of human GREs that contain a known trait variants (SNP).

(B) Bar plots depicting the percentage of GREs with conserved activity in marmoset and mouse. For both SNP containing GREs and those that do not. Dissimilarity between the ratio of human-specific activity was calculated with a Fisher's exact test.

(C) Bar plot depicting the percentage of GRE that contain a known SNP for both evolutionary conserved as well as human-specific GREs. Dissimilarity between the ratios was calculated with a Fisher's exact test.

(D) ChIP-seq tracks of RPM normalized H3K27ac enrichment in different tissues of human, marmoset, and mouse as indicated, across a 222 kb region containing the *BSN* gene. Known SNPs are depicted as blue bars underneath the human panel.

(E) Bar plot depicting the percentage of variable GREs, for GREs that are both SNP containing and those that are not as indicated on the axis. Dissimilarity between the ratios was calculated with a Fisher's exact test.

(F) Bar plot depicting the percentage of GREs that are unengaged or engaged with a promoter or other anchor. Data are shown for elements containing SNPs or those that do not. Dissimilarity in TSS engagement was calculated with a Fisher's exact test.

(G) Boxplots depicting the maximum STARR-activity of both SNP containing enhancers and those that do not as indicated on the axis. Dissimilarity between the distributions was calculated using a Student's t test.

evolution may be more modest than initially anticipated. This is reflected in a general paucity in identifying evolutionary changes in non-coding DNA with large phenotypical consequences. Several mechanisms may contribute to this. For instance, regulatory buffering may occur when multiple non-coding elements are modulating expression control over the same gene (Berthelot et al., 2018; Hong et al., 2008; Osterwalder et al., 2018; Perry et al., 2010), allowing for non-detrimental changes to occur at elements without major gene expression disturbances. In addition, the loss of an enhancer may be offset against the gain of a different enhancer, which is known as compensation (Vermunt et al., 2016). Furthermore, our current work shows that a subset of enhancers emerges without an apparent integration in the transcriptional network providing additional mechanisms that can in part explain the paucity of finding enhancers that affected phenotypical evolution. In our data, this is also reflected by an altered genomic distribution of unengaged enhancers, which are found to be located further away from genes and more often in inactive B-compartments. These unengaged enhancers may be "evolutionarily poised" and serve as a breeding pool from which future evolutionary novelty may be selected but may be dispensable at the current evolutionary stage. Nevertheless, these effects are only seen for small subsets of enhancers and do not fully explain the inability to link phenotypical change to regulatory innovations.

Instead, our data show a surprisingly high variability of H3K27ac enrichment on enhancers that are evolutionarily new within the human population. This high interindividual variability affects over 50% of enhancers that changed H3K27ac enrichment during recent human evolution. This could suggest that over half of the enhancers that are affected by evolution have not yet stabilized within the population and are thus unlikely to support key phenotypical species changes. Combining variable enhancers with compensated enhancers, unengaged enhancers and buffered enhancers suggest that many enhancers that are evolutionary new may lack strong functional impact. This complicates our understanding of species-specific gene expression control in relationship to phenotypical change. Supporting this notion, we find that common trait variants more often reside in conserved enhancers, which makes sense if they are to affect function. Thus, our data suggest that there is an inverse relationship between the observed evolutionary flexibility at enhancers and their biological impact. In addition, we provide new insight into how enhancers evolve, and where their lack of impact originates from. Finally, as conserved enhancers are more relevant to understanding the control of a human biological process in a marmoset model, our data demonstrate that the species specificity of enhancers as affecting the suitability of other species as disease model organisms is likely overestimated.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESEARCH AVAILABILITY
  - ○ Lead Contact
  - ○ Materials Availability
  - ○ Data and Code Availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Common marmoset tissue collection
- METHOD DETAILS
  - ○ Chromatin immunoprecipitation and sequencing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ ChIP-Seq enrichment analysis
  - ○ Comparison with public marmoset ChIP-Seq data
  - ○ Identification of repurposed elements
  - ○ Evolutionary classification of GREs in LCLs
  - ○ Classification of evolutionary new elements
  - ○ Analysis of variability and human ancestry
  - ○ Gene expression analysis
  - ○ Determination of interindividual variability
  - ○ Analysis of line to line variability
  - ○ t-SNE analysis and hierarchical clustering
  - ○ Cell type specificity of human gains
  - ○ Enhancer gene associations using HiC data
  - ○ Enhancer compartmentalization and activity
  - ○ Gene ontology analysis and SNP enrichment
  - ○ Analysis of confounder variable importance
  - ○ Analysis of interdependence

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.celrep.2020.107799.

ACKNOWLEDGMENTS

M.P.C., I.S.T., B.C., and M.L.B. were supported by the Royal Academy of Sciences in the Netherlands (KNAW), The Erasmus University Medical Center, and by funding acquired by the Friends of the Hubrecht Institute.

AUTHOR CONTRIBUTIONS

B.C., M.L.B., and M.P.C. conceived and designed the study. B.C. performed the experiments. M.L.B., B.C., M.W.V., G.G., W.d.L., and M.P.C. designed the analysis. M.L.B., B.C., and G.G. performed the analysis and were supervised by W.d.L. and M.P.C. I.S.T., I.K., and C.R.M.W. collected data. M.W.V. contributed data analysis tools. M.L.B., B.C., M.W.V., C.R.M.W., I.S.T., and M.P.C. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 28, 2020
Revised: April 23, 2020
Accepted: June 1, 2020
Published: June 23, 2020

REFERENCES

Arnold, C.D., Gerlach, D., Stelzer, C., Boryn, L.M., Rath, M., Stark, A., Boryń, Ł.M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science *339*, 1074–1077.

Berthelot, C., Villar, D., Horvath, J.E., Odom, D.T., and Flicek, P. (2018). Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. Nat. Ecol. Evol. *2*, 152–163.

Boija, A., Mahat, D.B., Zare, A., Holmqvist, P.H., Philip, P., Meyers, D.J., Cole, P.A., Lis, J.T., Stenberg, P., and Mannervik, M. (2017). CBP Regulates Recruitment and Release of Promoter-Proximal RNA Polymerase II. Mol. Cell *68*, 491–503.

Bonn, S., Zinzen, R.P., Girardot, C., Gustafson, E.H., Perez-Gonzalez, A., Delhomme, N., Ghavi-Helm, Y., Wilczyński, B., Riddell, A., and Furlong, E.E. (2012). Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. Nat. Genet. *44*, 148–156.

Cain, C.E., Blekhman, R., Marioni, J.C., and Gilad, Y. (2011). Gene expression differences among primates are associated with changes in a histone epigenetic modification. Genetics *187*, 1225–1234.

Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? Mol. Cell *49*, 825–837.

Carrion, R., Jr., and Patterson, J.L. (2012). An animal model that reflects human disease: the common marmoset (Callithrix jacchus). Curr. Opin. Virol. *2*, 357–362.

Castelijns, B., Baak, M.L., Timpanaro, I.S., Wiggers, C.R.M., Vermunt, M.W., Shang, P., Kondova, I., Geeven, G., Bianchi, V., de Laat, W., et al. (2020). Hominin-specific regulatory elements selectively emerged in oligodendrocytes and are disrupted in autism patients. Nat. Commun. *11*, 301.

Cortez, D., Marin, R., Toledo-Flores, D., Froidevaux, L., Liechti, A., Waters, P.D., Grützner, F., and Kaessmann, H. (2014). Origins and functional evolution of Y chromosomes across mammals. Nature *508*, 488–493.

Cyranoski, D. (2014). Marmosets are stars of Japan's ambitious brain project. Nature *514*, 151–152.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Fish, A., Chen, L., and Capra, J.A. (2017). Gene Regulatory Enhancers with Evolutionarily Conserved Activity Are More Pleiotropic than Those with Species-Specific Activity. Genome Biol. Evol. *9*, 2615–2625.

Geeven, G., Macgillavry, H.D., Eggers, R., Sassen, M.M., Verhaagen, J., Smit, A.B., de Gunst, M.C.M., and van Kesteren, R.E. (2011). LLM3D: a log-linear modeling-based method to predict functional gene regulatory interactions from genome-wide expression data. Nucleic Acids Res. *39*, 5313–5327.

Geeven, G., Teunissen, H., de Laat, W., and de Wit, E. (2018). peakC: a flexible, non-parametric peak calling package for 4C and Capture-C data. Nucleic Acids Res. *46*, e91.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77–88.

Hilton, I.B., D'Ippolito, A.M., Vockley, C.M., Thakore, P.I., Crawford, G.E., Reddy, T.E., and Gersbach, C.A. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat. Biotechnol. *33*, 510–517.

Hong, J.W., Hendrix, D.A., and Levine, M.S. (2008). Shadow enhancers as a source of evolutionary novelty. Science *321*, 1314.

Jinn, S., Drolet, R.E., Cramer, P.E., Wong, A.H.-K., Toolan, D.M., Gretzula, C.A., Voleti, B., Vassileva, G., Disa, J., Tadin-Strapps, M., and Stone, D.J. (2017). TMEM175 deficiency impairs lysosomal and mitochondrial function and increases α-synuclein aggregation. Proc. Natl. Acad. Sci. USA *114*, 2389–2394.

Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J.B., Kundaje, A., Liu, Y., Boyle, A.P., Zhang, Q.C., Zakharia, F., Spacek, D.V., et al. (2013). Extensive Variation in Chromatin States Across Humans. Science *342*, 750–752.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al.; Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Kursa, M.B., and Rudnicki, W.R. (2010). Feature selection with the boruta package. J. Stat. Softw. *36*, 1–13.

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

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Leslie, R., O'Donnell, C.J., and Johnson, A.D. (2014). GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. Bioinformatics *30*, i185–i194.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. *28*, 495–501.

McVicker, G., van de Geijn, B., Degner, J.F., Cain, C.E., Banovich, N.E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y., and Pritchard, J.K. (2013). Identification of genetic variants that affect histone modifications in human cells. Science *342*, 747–749.

Nord, A.S., Blow, M.J., Attanasio, C., Akiyama, J.A., Holt, A., Hosseini, R., Phouanenavong, S., Plajzer-Frick, I., Shoukry, M., Afzal, V., et al. (2013). Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. Cell *155*, 1521–1531.

Osterwalder, M., Barozzi, I., Tissières, V., Fukuda-Yuzawa, Y., Mannion, B.J., Afzal, S.Y., Lee, E.A., Zhu, Y., Plajzer-Frick, I., Pickle, C.S., et al. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. Nature *554*, 239–243.

Ozgyin, L., Horvath, A., Hevessy, Z., and Balint, B.L. (2019). Extensive epigenetic and transcriptomic variability between genetically identical human B-lymphoblastoid cells with implications in pharmacogenomics research. Sci. Rep. *9*, 4889.

Perry, M.W., Boettiger, A.N., Bothma, J.P., and Levine, M. (2010). Shadow enhancers foster robustness of Drosophila gastrulation. Curr. Biol. *20*, 1562–1567.

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

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

Rickels, R., and Shilatifard, A. (2018). Enhancer Logic and Mechanics in Development and Disease. Trends Cell Biol. *28*, 608–630.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26.

Rowley, M.J., and Corces, V.G. (2018). Organizational principles of 3D genome architecture. Nat. Rev. Genet. *19*, 789–800.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109–113.

Schoenfelder, S., and Fraser, P. (2019). Long-range enhancer-promoter contacts in gene expression control. Nat. Rev. Genet. *20*, 437–455.

Schröder, S., Herker, E., Itzen, F., He, D., Thomas, S., Gilchrist, D.A., Kaehlcke, K., Cho, S., Pollard, K.S., Capra, J.A., et al. (2013). Acetylation of RNA polymerase II regulates growth-factor-induced gene transcription in mammalian cells. Mol. Cell *52*, 314–324.

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V., and Ren, B. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116–120.

Shimogori, T., Abe, A., Go, Y., Hashikawa, T., Kishi, N., Kikuchi, S.S., Kita, Y., Niimi, K., Nishibe, H., Okuno, M., et al. (2018). Digital gene atlas of neonate common marmoset brain. Neurosci. Res. *128*, 1–13.

Sirugo, G., Williams, S.M., and Tishkoff, S.A. (2019). The Missing Diversity in Human Genetic Studies. Cell *177*, 26–31.

Soldner, F., Stelzer, Y., Shivalila, C.S., Abraham, B.J., Latourelle, J.C., Barrasa, M.I., Goldmann, J., Myers, R.H., Young, R.A., and Jaenisch, R. (2016). Parkinson-associated risk variant in distal enhancer of α-synuclein modulates target gene expression. Nature *533*, 95–99.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nat. Rev. Genet. *20*, 467–484.

Van Der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

Vermunt, M.W.W., Reinink, P., Korving, J., de Bruijn, E., Creyghton, P.M.M., Basak, O., Geeven, G., Toonen, P.W.W., Lansu, N., Meunier, C., et al.; Netherlands Brain Bank (2014). Large-scale identification of coregulated enhancer networks in the adult human brain. Cell Rep. *9*, 767–779.

Vermunt, M.W., Tan, S.C., Castelijns, B., Geeven, G., Reinink, P., de Bruijn, E., Kondova, I., Persengiev, S., Bontrop, R., Cuppen, E., et al.; Netherlands Brain Bank (2016). Epigenomic annotation of gene regulatory alterations during evolution of the primate brain. Nat. Neurosci. *19*, 494–503.

Vierstra, J., Rynes, E., Sandstrom, R., Zhang, M., Canfield, T., Hansen, R.S., Stehling-Sun, S., Sabo, P.J., Byron, R., Humbert, R., et al. (2014). Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science *346*, 1007–1012.

Villar, D., Flicek, P., and Odom, D.T. (2014). Evolution of transcription factor binding in metazoans - mechanisms and functional implications. Nat. Rev. Genet. *15*, 221–233.

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. (2015). Enhancer evolution across 20 mammalian species. Cell *160*, 554–566.

Wang, X., He, L., Goggin, S.M., Saadat, A., Wang, L., Sinnott-Armstrong, N., Claussnitzer, M., and Kellis, M. (2018). High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. Nat. Commun. *9*, 5380.

Worley, K.C., Warren, W.C., Rogers, J., Locke, D., Muzny, D.M., Mardis, E.R., Weinstock, G.M., Tardif, S.D., Aagaard, K.M., Archidiacono, N., et al.; Marmoset Genome Sequencing and Analysis Consortium (2014). The common marmoset genome provides insight into primate biology and evolution. Nat. Genet. *46*, 850–857.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

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 Biol. *15*, 547.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
H3K27ac antibody	Abcam	#ab4729
H3K4me3 antibody	Abcam	#ab8580
Biological Samples		
Common Marmoset tissues	Biomedical Primate Research Centre	https://www.bprc.nl
Critical Commercial Assays		
TruSeq ChIP Sample Prep Kit	Illumina	IP-202-1012
TruSeq Stranded mRNA Library Prep	Illumina	#20020594
Dynabeads Protein G	Life Technologies	#10004D
Deposited Data		
Raw and processed data	This paper	GSE141563
Marmoset gene expression data	Cortez et al., 2014	GSE50747
Human tissues H3K27ac ChIP-seq	Roadmap Epigenomics	GSE16256
Mouse tissues H3K27ac ChIP-seq	Shen et al., 2012	GSE29184
Human Liver H3K27ac ChIP-seq	Villar et al., 2015	E-MTAB- 2633
Human LCL H3K27ac ChIP-seq	McVicker et al., 2013	GSE47991
Primate LCL H3K27ac ChIP-seq	Zhou et al., 2014	GSE60269
Human LCL H3K27ac ChIP-seq	Kasowski et al., 2013	GSE50893
Human LCL H3K27ac ChIP-seq	Ozgyin et al., 2019	GSE121926
Primate LCL RNA-seq	Cain et al., 2011	GSE24111
Human Cell lines H3K27ac ChIP-seq	Ernst et al., 2011	GSE26386
Human LCL HiC	Rao et al., 2014	GSE63525
Human LCL CTCF ChIP-seq	ENCODE	GSE29611
Software and Algorithms		
Bowtie version 1.1.2	Langmead and Salzberg, 2012	http://bowtie-bio.sourceforge.net
Samtools version 1.3.1	Li et al., 2009	http://samtools.sourceforge.net/
MACS2 version 2.1.1	Zhang et al., 2008	https://pypi.org/project/MACS2/
Bedtools version 2.26.0	Quinlan and Hall, 2010	https://bedtools.readthedocs.io/en/latest/
Fastx-toolkit	Hannon Lab	http://hannonlab.cshl.edu/fastx_toolkit/
R version 3.6.1	R Development Core Team, 2012	https://cran.r-project.org
IGV version 2.3.40	Robinson et al., 2011	http://software.broadinstitute.org/software/igv
T-SNE R package	Van Der Maaten and Hinton, 2008	https://cran.r-project.org/web/packages/tsne/
DESeq2 R package	Love et al., 2014	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
Boruta R package	Kursa and Rudnicki, 2010	https://cran.r-project.org/web/packages/Boruta/index.html
GREAT version 3.0.0	McLean et al., 2010	http://bejerano.standford.edu/great/public/html/
PeakC	Geeven et al., 2018	https://github.com/deWitLab/peakC
LLM3D	Geeven et al., 2011	https://academic.oup.com/nar/article/39/13/5313/2409399

RESEARCH AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Menno P. Creyghton (m.creyghton@erasmusmc.nl).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

The ChIP-Seq data generated for this study is available at Gene Expression Omnibus with accession code GSE141563. Public datasets analyzed during this study are available in the NCBI Gene Expression Omnibus and available under accessions GSE5074 (Marmoset gene expression) (Cortez et al., 2014), GSE16256 and E-MTAB-2633 (Human tissues H3K27ac ChIP-Seq), GSE29184 (Mouse tissues H3K27ac ChIP-Seq), GSE47991 (H3K27ac in human LCL from Yoruba descent) (Kasowski et al., 2013), GSE60269 (H3K27ac in chimpanzee and rhesus macaque LCL) (Zhou et al., 2014), GSE50893 (H3K27ac in human LCL from Caucasian, San, Yoruba and Asian descent) (Kasowski et al., 2013), GSE121926 (H3K27ac in five human LCL from the same individual), GSE24111 (RNA-Seq in human, chimpanzee and rhesus macaque LCL)(Cain et al., 2011), GSE26386 (H3K27ac for 7 non-LCL cell types human)(Ernst et al., 2011), GSE63525 (HiC data in human LCL GM12878)(Rao et al., 2014) and GSE29611 (CTCF in human LCL GM12878). This study did not generate any unique code.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Common marmoset tissue collection

Common marmoset (*Callithrix jacchus*), tissue samples were collected for three specimens (*cj1 (5 years, female), cj2 (5 years, female), cj3 (8 months, male)*) at the Biomedical Primate Research Centre (BPRC) in Rijswijk, the Netherlands (https://www.bprc.nl) and represent rest material involving no animal experimentation for the purpose of this work as determined by the Animal Experimental Committee (DEC) (Table S1). Samples were frozen down as fast as possible after death and stored at −80°C. Ten different tissue samples, two brain regions and eight organs, were collected from cj1 and cj2. These samples include cerebellum, prefrontal cortex, colon, heart, kidney, liver, pancreas, skeletal muscle, spleen and stomach. Two brain region, cerebellum and prefrontal cortex were collected from cj3.

METHOD DETAILS

Chromatin immunoprecipitation and sequencing

Chromatin immunoprecipitation (ChIP) was performed as previously described (Castelijns et al., 2020). In short, 60mg of tissue was used per sample and homogenized in 1mL Dulbecco's Modified Eagle Medium (DMEM) with 0.2% Bovine Serum Albumine (BSA) in a glass douncer (Kontes Glass Co.). Cells were crosslinked at Room Temperature (RT) in 10ml fixation buffer (freshly made; 1% formaldehyde, 0.5mM Ethylenediaminetetraacetic acid (EDTA), 0.05mM ethylene glycol-bis(β-aminoethyl ether)-N,N,N′,N′-tetraacetic acid (EGTA), 10mM NaCl, 5mM HEPES-KOH, pH 7.5) while rotating for 10'. Next, samples were washed twice with PBS, pelleted for 5′ at 2095xg and 4°C, resuspended in 10ml lysis buffer (50mM HEPES-KOH pH 7.5, 140mM NaCl, 1mM EDTA, 10% glycerol, 0.5% Igepal, 0.25% Triton X-100) and lysed for 10' at RT while rotating. Samples were pelleted for 5′ at 2095xg at 4°C and resuspended in 10ml wash buffer (200mM NaCl, 1mM EDTA, 0.5mM EGTA, 10mM Tris-HCl pH 8.0) and incubated for 10' at RT while rotating. Cells were pelleted for 5′ at 2095xg at 4°C and resuspended in 150μl sonication buffer (1mM EDTA, 0.5mM EGTA, 10mM Tris-HCl pH 8.0, 100mM NaCl, 0.1% Na-Deoxycholate, 0.5% N-lauroyl sarcosine) and sonicated using the Covaris S series (12 cycles of 60 s: intensity 3, duty cycle 20%, 200 cycles/burst) in two microtubes (Covaris 520045) per sample. After sonication, the microtubes per sample were pooled and sonication buffer and Triton X-100 (final concentration 1%) was added to a total volume of 550μl. Immunoprecipitation with antibody (H3K27ac: ab4729 abcam, H3K4me3: ab8580 abcam) coated Protein G Dynabeads (Invitrogen 10003D) was performed overnight at 4°C. The following day beads were washed 4 times with RIPA (50mM HEPES-KOH pH 7.5, 1mM EDTA, 0.7% DOC, 1% NP40 and 0.5M LiCl) and once with 50mM NaCl in TE. DNA elution from the beads was done in elution buffer (50mM Tris pH 8.0, 10mM EDTA, 1% Sodium dodecyl sulfate (SDS) overnight at 65°C. Samples were centrifuged shortly to remove the beads and the supernatant was 1:1 diluted with TE, followed by a 2h incubation with RNase (final concentration 0.2μg/μl) at 37°C and a 2h incubation with proteinase K (final concentration 0.2μg/μl) at 55°C. Finally, the DNA was extracted using phenol/chloroform and MaXtract High Density gel tubes (QIAGEN) followed by ethanol purification. Sequencing libraries were prepared according to the Illumina Truseq DNA library protocol and samples were sequenced at the MIT BioMicro Center (https://openwetware.org/wiki/BioMicroCenter) or at the Utrecht DNA Sequencing facility (http://useq.nl) using the Illumina HiSeq 2000 or NextSeq 500 genome sequencer.

QUANTIFICATION AND STATISTICAL ANALYSIS

ChIP-Seq enrichment analysis

To ensure comparable mapping between the samples, all reads were trimmed to a length of 36 bp using the Fastx-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Reads where mapped to the most recent genome available (human; hg38, chimpanzee; panTro5, rhesus macaque; rheMac8, marmoset; calJac3, mouse; mm10) using Bowtie version 1.1.2 (Langmead and Salzberg, 2012), allowing 1 mismatch and discarding reads that could not be uniquely mapped. Samtools version 1.3.1 (Li et al., 2009) was used to discard duplicate reads. Enriched regions were called with MACS2 version 2.1.1 (Zhang et al., 2008) using a p value $< 10^{-5}$, extsize 300 and local lambda 100,000. Regions were extended to a minimum of 2,000 bp in length, (+/− 1,000 bp from peak center) consistent with typical ChIP-Seq resolution (Vermunt et al., 2014; Villar et al., 2015). Lists of GREs per tissue per species were obtained by merging identified regions of replicate samples, regions with a minimal 1 bp overlap were stitched together. This cutoff was chosen to ensure that enriched regions with multiple summits were analyzed as one. All enriched regions were classified as enhancers or promoters based on the hg38 RefSeq list. Regions overlapping a 1,000 bp window around a known human transcriptional start site (TSS) were considered promoters, others were annotated as enhancers. Distance to the closest TSS was determined for each enhancer using the RefSeq gene list of hg38 with Bedtools version 2.26.0 (Quinlan and Hall, 2010).

Comparison with public marmoset ChIP-Seq data

Marmoset H3K27ac samples were compared to publicly available liver H3K27ac ChIP-Seq data (Villar et al., 2015). Liver data from human, rhesus macaque, vervet monkey and marmoset was obtained and analyzed as described above. Integration of public marmoset data was done by merging enriched regions from public liver samples with our non-redundant H3K27ac marmoset enriched regions. Integration of the new liver data with the publicly available datasets was done by the generation of a non-redundant, cross-species list of H3K27ac-enriched liver regions. Therefore, H3K27ac regions from liver samples of different primate species were reciprocally mapped onto the human genomes using the UCSC liftOver tool (-minMatch = 0.1) as described previously (Castelijns et al., 2020). Reciprocal mapping had to yield unique regions and all regions that changed more than 50% in size were excluded. To ensure equal mappability, > 90% of the bases within a regulatory element had to be properly annotated in all reference genomes (< 10% overlap with UCSC Table Brower's gap locations lists). This cut-off was chosen as unknown bases generally occur in stretches rather than single nucleotides scattered across the genome. Moreover, to account for repetitive or duplicated genomic regions, which are susceptible to poor annotation in lower-quality genomes, enrichment scores were not allowed to change significantly in the target genome when allowing reads to map to multiple locations. These repetitive regions were defined per species by mapping reads from every sample to unique locations (bowtie: –best –strata –m 1) as well as to multiple locations (bowtie: –best –strata –M 1). Genomic regions that were enriched using the multimap settings but not with unique mapping are potential repetitive elements that are not annotated at similar depth across all the genomes.

Identification of repurposed elements

A non-redundant list of regions mappable on the human and marmoset genome was created using reciprocal liftover as described above. For every region its activity across the different tissues was determined based on peak calling. Repurposed regions were defined as being active in different tissues between human and marmoset based on significant enrichment calls using MACS2. These where then further refined by only considering regions where H3K27ac enrichment differed at least 2-fold. Similar repurposing analysis was used to compare human and marmoset regions with mouse data. The use of presence / absence calls omits the need to control for batch effects which is not feasible when the batches align with the different species. No strong contribution of confounders was detected (Figure S2B).

Evolutionary classification of GREs in LCLs

Non-redundant (NR) lists per species were created using LCL data from a single source (Zhou et al., 2014) and overlapping regions were merged as described above. Only regions called at least twice in a single species were kept for analysis. To compare enriched regions between species, the NR-list of rhesus macaque and chimpanzee were reciprocally mapped on the human genome as described above. Next, enriched regions of rhesus macaque and chimpanzee were combined with human enriched regions, merging all overlapping regions and discarding all regions that overlapped ENCODE blacklisted regions (Landt et al., 2012), resulting in a total list of 54,793 H3K27ac enriched regions that could be assessed across all three species. The number of reads per region was counted using Bedtools version 2.26.0 (Quinlan and Hall, 2010) and normalized for library size and length of the region (RPKM).

To determine differentially enriched (DE) regions between species, seven samples per species were used in a DESeq2 (Love et al., 2014) analysis. Regions with a log2FC > 1 and FDR < 0.01 were considered as DE between species. Human gained elements were defined as regions that were DE between both human and chimpanzee, as well as human and rhesus macaque. Losses were defined as significantly lower in human versus chimpanzee and human versus rhesus macaque. One explanation for this is that to be able to assign differentially enrichment in human, target elements require a time period of evolutionary stability (between chimpanzee and rhesus macaque). Stable elements were determined as not differentially enriched between all the three species.

Classification of evolutionary new elements

To define regions as new in human, GREs were compared to the genome-wide background enrichment of H3K27ac in each sample of rhesus macaque and chimpanzee. To calculate the background enrichment, reads were counted using a sliding window of 3,000 bp with 500 bp steps across the genome as shown previously (Vermunt et al., 2016). Read count values for GREs were normalized to a 3,000 bp size and compared with the background windows. Human gain elements were defined as evolutionary new when the enrichment was below the 90th percentile of the background windows in all seven samples of chimpanzee and rhesus.

Analysis of variability and human ancestry

As we defined ‘‘human-specific’’ changes in GREs using samples from Yoruba ancestry only, we verified that variation in ancestry was not causing miss-classification of GREs as human-specific (Sirugo et al., 2019). Therefore, 4 samples from individuals from San, Caucasian and Yoruba descent were selected based on high FRIP scores and exclusion of direct family members. A NR-list was created based on enriched regions identified using MACS2 in at least 2 technical replicates per individual and 3 out of 4 individuals of a single ancestry ($n = 39,495$). As a control, an additional set was generated in a similar manner but composed of 3 independent groups of different individuals of Yoruba descent ($n = 31,360$). For both sets the elements were selected that were previously classified as human gains. To investigate the effect of ancestry on the classification of human gains; the number of human gains present in one ancestry lineage (Yoruba, San, or Caucasian) was plotted on the x axis = 1. The number of human gain enhancers present in two groups was plotted on x axis = 2 and number of enhancers present in all groups was on x axis = 3 (Figure S3E). This analysis was done for all possible combinations and a regression line was plotted using the ggplot R package. Comparison of the slopes of the two linear regression lines was performed with ANOVA. Venn diagrams of the overlap between analyzed groups were created using the venn package in R. The same analysis was performed for all enhancers without intersection with human gains. Close to 25% of human-specific gains that were based on the Yoruba samples were not found enriched in individuals from both San and Caucasian descent (Figure S3E). However, when comparing the human gains to two control sets containing different Yoruba individuals, we observed a comparable reduction in the number of GREs classified as human-specific (Figure S3E). Indicating that these regions are sample rather than ancestry-specific. Similar results were obtained when analyzing all human GREs, including those that were not gained in the human lineage (Figure S3F) suggesting that this effect was not related to recent evolution. Ancestry specific gains and losses were defined using DE-Seq ($\log2FC > 1$ and $FDR < 0.01$) between a single ancestry and both other ancestries using the NR -list generated above ($n = 39,495$ of which 28,113 putative enhancers). DE analysis was performed on read counts of one technical replicate per individual, with four individuals per ancestry. Next, ancestry-specific regions were overlaid with human-specific gains and evolutionary novel elements using bedtools. Differences in overlap frequencies were calculated using a Fisher exact test. While some examples of evolutionary miss-classifications were found (Figure S3G), i.e., 6 San-specific gains, their numbers were negligible and could be explained by random overlap ($p = 0.474$, Figure S3H). Thus, human ancestry plays no measurable role in the classification of human-specific regulatory DNA.

Gene expression analysis

Gene expression in LCL cells was determined using the exon read count tables derived from RNA-Seq data generated from 3 humans, chimpanzees and rhesus macaque samples (Cain et al., 2011). Expression was calculated as the sum of reads per exon divided by gene length and the total number of reads followed by log2 zero-mean transformation. Only genes with orthologous exons present in all three species were used for analysis. Significant differences in expression between human, chimpanzee and rhesus macaque gene sets were determined using a Student’s t test. Gene expression in common marmoset tissues was assessed using a previously published list of normalized FPKM values for which the female tissue samples were plotted.

Determination of interindividual variability

Variation between human individuals was defined using H3K27ac ChIP-Seq data generated from human LCL samples covering individuals ($n = 19$) of Caucasian, San, Yoruba and Asian descent (Kasowski et al., 2013), which were analyzed as described above. Aligned reads that were not paired were discarded. A read count table across all individuals and input data was generated for the above-mentioned NR-list, created on human, chimpanzee and rhesus macaque ($n = 54,793$). Variability of enhancers was calculated as described previously (Kasowski et al., 2013), using only those regions that could also be identified in at least 2 individuals by MACS2. In short, read counts of the ChIP-Seq samples and input samples, were normalized for library size and region length (RPKM) followed by asinh-transformation and quantile normalization using the R package preprocessCore. Fold change over input was calculated as signal/input followed by calculating the F-value (inter-replicate variation to inter-individual variation) using the ANOVA R package. Obtained F-values were log10 transformed and elements with a log10 F-value larger than 1 were considered variable. To refine the variable set, a pairwise differentially enrichment analysis of the 19 individuals was performed using DESeq2. A region with $\log2FC > 3$ and q-value $< 1e^{-5}$ was defined as DE. Only GREs called DE between any pairwise comparison and the log10 F-value > 1 was annotated as variable.

Analysis of line to line variability

Line to line variability was determined using H3K27ac data generated in 5 independent cell lines that were obtained from a single individual (Ozgyin et al., 2019). H3K27ac data was obtainted and mapped to the human genome (hg38) as described above.

H3K27ac enrichment for the here identified GREs was counted and RPKM normalized. Only GREs that were significantly enriched for H3K27ac using MACS2 in at least 2 samples were selected and further processed using an asinh-transformation and quantile normalization. Variability was then calculated per GRE on the normalized readcount using ANOVA as described above.

t-SNE analysis and hierarchical clustering

Read counts were normalized for library size and length of the region. Pearson correlations between the samples were calculated and t-Distributed Stochastic Neighbor Embedding (t-SNE) multiple scaling coordinates were defined with the t-SNE R package (Van Der Maaten and Hinton, 2008) on the distance matrix. For hierarchical clustering, Pearson correlations between samples were calculated and samples were clustered based on Pearson distance with average linkage. Heatmaps were generated using the heatmap.2 function from the gplots R package.

Cell type specificity of human gains

To assess cell type specificity for human-specific enhancers, seven H3K27ac ChIP-Seq datasets of unrelated cell types were analyzed; including H1, HepG2, HMEC, HSMM, Huvec, NHEK, and NHLF (Ernst et al., 2011). Enriched regions were identified using MACS2 as described above. NR-lists of all cell types were compared to the LCL enriched regions defined using the primate samples. Regions not found in any of the other cell types were classified as cell-type-specific for LCL. Difference in the ratio of cell-type-specific regions and regions active in multiple cell types within the human gains was performed using a Fisher's exact test.

Enhancer gene associations using HiC data

A high-resolution map of 3D contacts was obtained from a GM12878 HiC dataset (Rao et al., 2014), using the here identified H3K27ac enriched GREs, known TSS and ENCODE CTCF-sites as viewpoints (Sanyal et al., 2012). Loops were defined from every viewpoint's virtual 4C profile using PeakC (Geeven et al., 2018), with a maximum distance of 1 Mb between viewpoint and anchor. Viewpoints and anchors where then resized to +-5 kb from the center and only those loops retained that showed reciprocal interactions. Viewpoints were considered to interact with a TSS when their resized anchors overlapped a known promoter region. To analyze enhancer redundancy, the number of enhancers looping to each gene were counted from the HiC data. Enhancers that looped toward a TSS that was also contacting other enhancers, were considered redundant. Enhancers were considered as unengaged when they had no reciprocal contacts. Differences in contact frequencies between sets of enhancers with distinct evolutionary properties were calculated using a Fisher's exact test.

Enhancer compartmentalization and activity

To determine whether GREs emerged in distinct chromatin domains, all enhancers were overlapped with A and B compartments of the GM12878 cell line, as defined previously (Rao et al., 2014). Enhancers that were located on a boundary were assigned to the domain with the largest overlap. Differences in the distribution between A and B compartments were calculated using a Fisher's exact test. To determine whether evolutionary novel enhancers harbored intrinsic activity, we calculated the activity of each element by determining the maximum reporter-activity within each region as defined using a mass parallel reporter assay (Wang et al., 2018). Differences in enhancer activity between enhancer groups was calculated using a Student's t test.

Gene ontology analysis and SNP enrichment

Gene ontology analysis was done using the Genomic Regions Enrichment of Annotations Tool (GREAT version 3.0.0, http://bejerano.stanford.edu/great/public/html) (McLean et al., 2010) with basal plus extension setting. Multiple genes are can therefore by assigned to the supplied enriched regions. For the marmoset tissue-specific H3K27ac enriched regions sets, regions that were specifically enriched in a single tissue type were analyzed (based on MACS2 peak calling). SNP regions were obtained using the annotated SNPs from the NIH GRASP database (https://grasp.nhlbi.nih.gov/Overview.aspx). SNPs with a reported p value < $1e^{-6}$ were extracted and these were extended to a +-2.5 kb, merging any resulting overlapping regions. These were then overlapped with the here identified regulatory elements to obtain both SNP containing GREs and those that do not.

Analysis of confounder variable importance

To test the contribution of confounding variables across different tissue and species samples in the ChIP-Seq datasets, we analyzed the importance of various potential confounder variables including: species, tissue, gender, sequencing depth, lab of origin, GC-content, conservation score and fraction of reads in peaks on the observed read counts using the Boruta package in R (Kursa and Rudnicki, 2010) (Figure S2B). Sequence conservation was determined by calculating the average PhastCon score (20 mammals, UCSC) per GRE. The Boruta algorithm defines the importance of confounder variables on the observed read count by using an iterative random forest classification. Every iteration, the observed variables are randomly shuffled to create shadow variables and the importance of the original variables on the observed read count is compared to these shadows. Variables that perform significantly better are confirmed confounders (green boxplot Figure S2B), while variables that perform significantly worse are considered as having no confounding effect on the observed read counts. (red boxplots Figure S2B).

Analysis of interdependence

The interdependence between categorical variables (evolutionary timing, variability and cell-type specificity) of the identified evolutionary novel enhancers, was calculated using LLM3D as described previously (Geeven et al., 2011). LLM3D fits a number of log-linear models to 3D contingency tables of enhancer counts and selects the model that best fits the observed enhancer characteristics. These models imply different (in)dependence relationships between the variables, with the null hypothesis assuming complete independence between the variables. Model selection assigned a model in which no (conditional) independence between any pair of the three variables was implied, meaning all levels of every attribute have different probabilities of occurring jointly. The significant p value indicates that enhancers that are new and variable differ in their rate of cell type specificity.