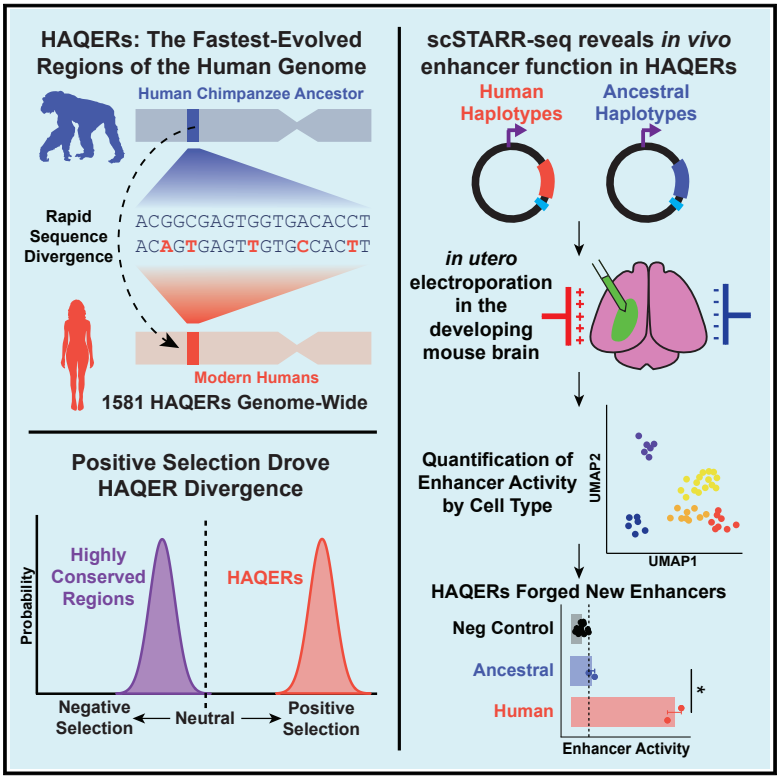
Cell

Adaptive sequence divergence forged new neurodevelopmental enhancers in humans

Graphical abstract



Highlights

- HAQERs are human genomic regions highly divergent from the human-chimpanzee ancestor
- HAQERs evolved under elevated mutation rates and positive selection
- HAQERs are enriched for bivalent chromatin and disease-linked variation
- HAQER divergence forged hominin-unique enhancers in the developing cerebral cortex

Authors

Riley J. Mangan, Fernando C. Alsina, Federica Mosti, ..., Timothy E. Reddy, Debra L. Silver, Craig B. Lowe

Correspondence

craig.lowe@duke.edu

In brief

Most comparative genomics studies focus on conserved regions. However, in this study, Mangan et al. identify the fastest-evolved regions across the entire human genome and provide insights into which genomic regions underlie human-specific disease risks and adaptations.

Mangan et al., 2022, Cell *185*, 4587–4603
November 23, 2022
https://doi.org/10.1016/j.cell.2022.10.016

CellPress

Cell

CellPress
OPEN ACCESS

Article

Adaptive sequence divergence forged new neurodevelopmental enhancers in humans

Riley J. Mangan,[1] Fernando C. Alsina,[1,6] Federica Mosti,[1,6] Jesús Emiliano Sotelo-Fonseca,[1] Daniel A. Snellings,[1,7] Eric H. Au,[1,8] Juliana Carvalho,[1] Laya Sathyan,[1] Graham D. Johnson,[2,3] Timothy E. Reddy,[2,3] Debra L. Silver,[1,4,5] and Craig B. Lowe[1,2,9,*]

[1]Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, NC 27710, USA
[2]Center for Genomic and Computational Biology, Duke University, Durham, NC 27705, USA
[3]Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC 27710, USA
[4]Duke Institute for Brain Sciences and Duke Regeneration Center, Duke University Medical Center, Durham, NC 27710, USA
[5]Departments of Cell Biology and Neurobiology, Duke University Medical Center, Durham, NC 27710, USA
[6]These authors contributed equally
[7]Present address: Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, USA
[8]Present address: Mammoth Biosciences, Inc., Brisbane, CA 94005, USA
[9]Lead contact
*Correspondence: craig.lowe@duke.edu
https://doi.org/10.1016/j.cell.2022.10.016

SUMMARY

Searches for the genetic underpinnings of uniquely human traits have focused on human-specific divergence in conserved genomic regions, which reflects adaptive modifications of existing functional elements. However, the study of conserved regions excludes functional elements that descended from previously neutral regions. Here, we demonstrate that the fastest-evolved regions of the human genome, which we term "human ancestor quickly evolved regions" (HAQERs), rapidly diverged in an episodic burst of directional positive selection prior to the human-Neanderthal split, before transitioning to constraint within hominins. HAQERs are enriched for bivalent chromatin states, particularly in gastrointestinal and neurodevelopmental tissues, and genetic variants linked to neurodevelopmental disease. We developed a multiplex, single-cell *in vivo* enhancer assay to discover that rapid sequence divergence in HAQERs generated hominin-unique enhancers in the developing cerebral cortex. We propose that a lack of pleiotropic constraints and elevated mutation rates poised HAQERs for rapid adaptation and subsequent susceptibility to disease.

INTRODUCTION

Humans can be distinguished from our recent great ape ancestors by many unique phenotypes, including bipedal locomotion,[1] craniofacial morphology,[2] and our remarkable cognitive capabilities.[3,4] Intertwined with adaptations like these are human-specific disease susceptibilities, including knee osteoarthritis[5] and schizophrenia.[6] Notwithstanding that both researchers and the public have a long-standing interest in understanding the genetic basis of human uniqueness, we have struggled to partition the millions of mutations separating humans from their great ape ancestors into those that are neutrally evolving and those that are significantly contributing to human-specific traits.

Initial systematic searches for the genetic basis of human traits focused on protein-coding regions to enrich for genetic changes with phenotypic effects.[7,8] More recent studies have identified several human-specific gene duplications that have been implicated in the expansion of the human neocortex.[9–11] However, humans and chimpanzees harbor few differences in amino acid sequences, and it has long been hypothesized that the mutations responsible for human-specific phenotypes lie primarily in non-protein-coding regulatory regions.[12–14]

A second generation of screens began with the insight that cross-species conservation could be utilized to enrich for functionally significant mutations in the non-protein-coding genome. This allowed screens to expand from the 1% of the genome that is protein coding to the 5% of the genome that includes highly conserved regulatory elements. Genomic regions from these screens are termed human accelerated regions (HARs).[15] These screens identified HARs based on acceleration in the rate of nucleotide substitutions, positing that an increase in the rate of molecular evolution from prior constraint reflects a change in the mode of selection. Over the past 15 years, additional studies have expanded the set of HARs with the addition of more genome assemblies, specific tissues of interest, and alternative statistical methods.[16–19]

Many HARs act as developmental enhancers, demonstrating the feasibility of expanding beyond protein-coding regions to identify modifications to regulatory elements.[20,21] One example

is a distal enhancer of the neurodevelopmental gene *FZD8*, where human-specific sequence changes increased enhancer activity in mouse embryonic brain, which was sufficient to accelerate neural precursor cell-cycle dynamics and increase brain size.[22] As would be expected for genomic regions with important roles in neurodevelopment, mutations in HARs have been associated with schizophrenia and autism spectrum disorder.[6,23]

Preconditioning studies of human-specific traits on highly conserved regions restrict analyses to 5% of the genome. However, a growing body of evidence suggests that much more of the genome is functional.[24,25] We propose that the combination of consortium efforts to catalog human genetic variation[26] and recent advances in high-throughput functional genomic technologies[27–30] provides an avenue for identifying functionally significant regulatory innovations across the entire genome through the integration of comparative, population, and functional genomics.

The remaining 95% of the genome is likely to include two types of evolutionarily significant genomic regions not targeted in past studies: functional elements recurrently modified on independent lineages and functional elements unique to humans. Many distinctive characteristics of human anatomy, such as brain size, limb proportions, and craniofacial morphology, are not static in non-human species but rather are dynamic across the panoply of vertebrate life. Therefore, we expect many genetic determinants of these dynamic traits to be fast-evolving in both humans and non-human species and thus to exhibit function without the stringent condition of past constraint. Furthermore, regions with cross-species conservation, which have evolved under purifying selection, will not contain recently evolved functional elements that are held under constraint only in humans. Both of these classes of regulatory innovations will be discovered in the underexplored non-conserved genome.

In this work, we integrate comparative genomics with genetic variation data from human populations to demonstrate that the fastest-evolved regions of the human genome, which we term "human ancestor quickly evolved regions" (HAQERs), diverged rapidly through the combination of elevated mutation rates and positive selection. While HAQERs diverged rapidly from the human-chimpanzee ancestor, they are highly similar among extant and archaic hominins. HAQERs are enriched in bivalent domains that are associated with spatiotemporally restricted developmentally or environmentally responsive regulatory elements. We developed *in vivo* single-cell self-transcribing active regulatory region sequencing (scSTARR-seq) as a multiplex, single-cell enhancer assay in the developing mouse cerebral cortex to demonstrate that rapid HAQER divergence forged functional elements that are exclusive to hominins. HAQERs are also enriched for disease-linked variation, suggesting an active role in shaping human-specific susceptibilities to disease.

RESULTS

Acceleration and velocity are associated with signatures of positive selection

Historically, it has been thought that higher rates of divergence in genomic regions are primarily associated with variation in the local mutation rate,[31] as opposed to selection. This is based on the notion that the vast majority of genetic differences between humans and great apes are selectively neutral and that positive selection is rare.[32]

Acceleration has been employed as a metric to mitigate the influence of local mutation rates, as a change in the rate of divergence is proposed to reflect a change in selective pressure. Even though this approach has been fruitfully applied to identify HARs as exciting candidates for further study in the highly conserved genome, this strategy limits the scope of investigation to regions with an initial velocity near zero, excluding regions that have accelerated from neutrality to rapid divergence. Therefore, we sought to generalize acceleration to identify targets of positive selection in the remaining 95% of the genome.

We define the acceleration, **a**, of a genomic region as the difference between the current velocity of divergence, **v**, and the initial velocity of divergence, $\mathbf{v}_0$, as in $\mathbf{a} \propto \Delta\mathbf{v} = \mathbf{v} - \mathbf{v}_0$. We define $\mathbf{v}_0$ as the divergence rate on the branch from the human-gorilla ancestor to the human-chimpanzee ancestor, and **v** as the divergence rate on the branch from the human-chimpanzee ancestor to extant humans (Figure 1A). Both $\mathbf{v}_0$ and **v** are measured in units of genetic distance per base pair per million years, where distance is counted as the sum of substitutions, insertions, and deletions in a 500-bp window. Unlike previous work, we do not place a threshold filter on $\mathbf{v}_0$, allowing us to calculate **a** genome-wide from a syntenic alignment of the human, chimpanzee, gorilla, and orangutan reference genomes (STAR Methods).

To understand if acceleration is predictive of selective pressure, we analyzed the frequencies of derived alleles in African populations[26] and inferred the direction and magnitude of selection acting on variants in genomic regions binned by acceleration values (STAR Methods). Under negative selection, a derived allele is more deleterious than the ancestral allele and thus unlikely to spread to high frequencies. Under positive selection, a derived allele is beneficial and more likely to be found at higher frequencies (Figure S1A). We implemented a statistical model to infer the mean selection parameter from derived allele frequency spectra (dAFS) using Markov chain Monte Carlo (MCMC)[34] and corrected for ascertainment bias present when regions are identified based on divergence[35] (Figure S1; STAR Methods). We report that highly positive acceleration is associated with positive selection coefficients (Figure 1B) and may be an informative identifier of adaptive innovation genome-wide.

However, the most dramatically accelerated regions will still preferentially include regions with modified function and past constraint (low $\mathbf{v}_0$, high **v**) at the expense of recurrently modified functional elements (high $\mathbf{v}_0$, high **v**) and recently functional elements from neutrally evolving sequence (moderate $\mathbf{v}_0$, high **v**). This motivated us to examine the relationship between the current velocity of a genomic region and selection, using the dAFS of variants in genomic regions binned by velocity. We observed a stronger relationship between velocity and selection than between acceleration and selection (Figures 1B and 1C).

We also observed a robust relationship between **a** and **v** (Figure S2A) but not between $\mathbf{v}_0$ and **v** (Figure S2B). This observation suggests that unlike at the megabase scale,[36] regional differences in the divergence rate at small scales are unlikely to reflect an intrinsic variation in mutation rates that is stable across

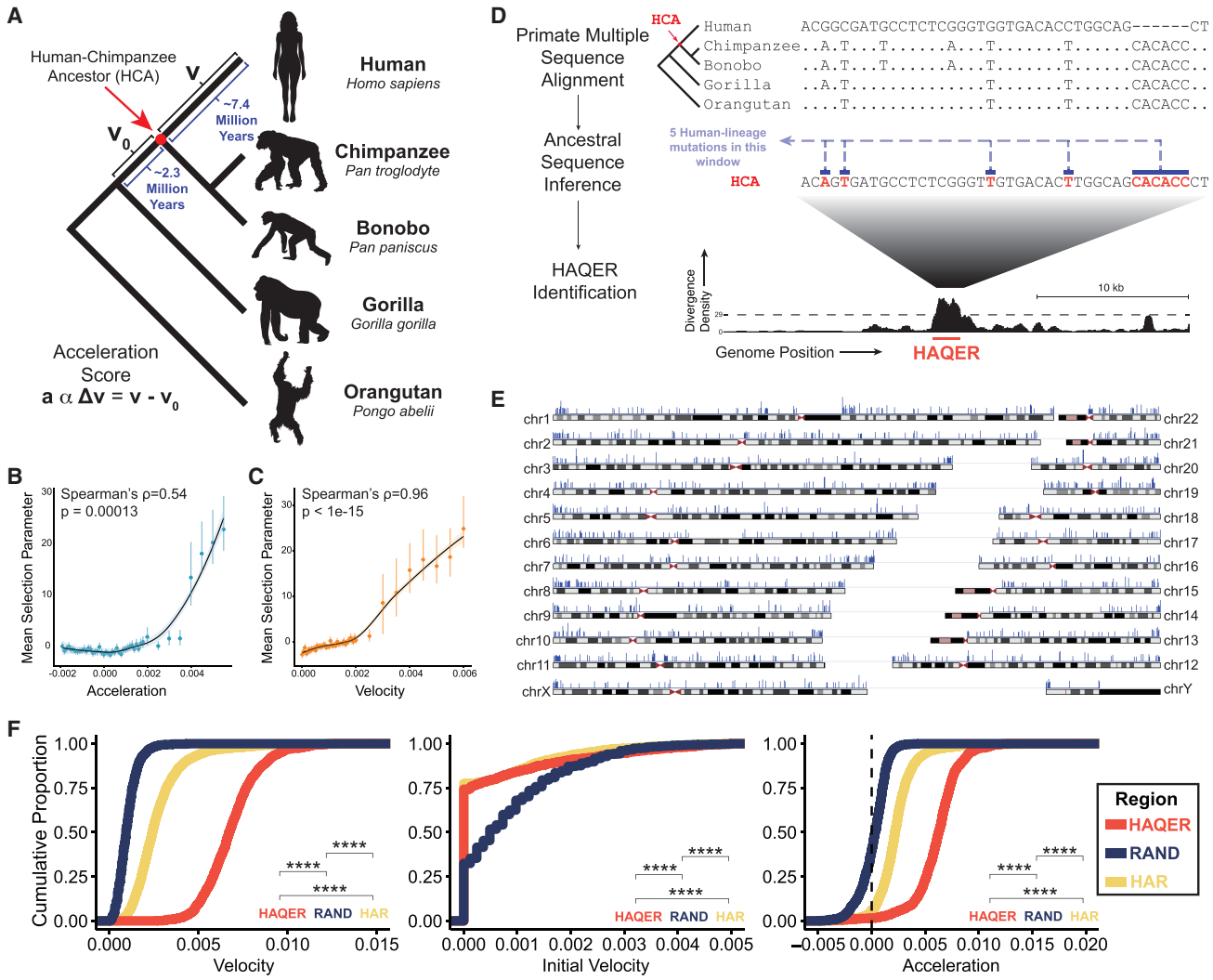


Figure 1. HAQERs, the fastest-evolved regions of the human genome

(A) We display the values of velocity (**v**), initial velocity ($\mathbf{v}_0$), and acceleration (**a**) in the phylogenetic context of recent human evolution.

(B and C) Mean selection parameter estimates for 500-bp genomic regions binned by either acceleration (B) or velocity (C). Error bars display the 95% highest density credible interval. Both acceleration and velocity correlate with signatures of selection in human populations.

(D) HAQERs (human ancestor quickly evolved regions) are identified as regions containing at least 29 mutations in a 500-bp window ($p < 10^{-6}$) that separate the inferred human-chimpanzee ancestor sequence from the human genome. We count insertions and deletions as one mutation regardless of their length.

(E) Locations in the human genome of the 1,581 HAQERs (blue markers). Marker amplitude reflects the maximum divergence density observed in each region. HAQERs are distributed across all human chromosomes and enriched near chromosome ends.

(F) Cumulative distribution of velocity, initial velocity, and acceleration observed across HAQERs, human accelerated regions (HARs), and random neutral proxy regions (RAND). Regions are filtered to a minimum element size of 50 bp. (Bonferroni-adjusted Wilcoxon; **** $p < 0.0001$).

See also Figures S1 and S2.

phylogenetic branches. Since we found that velocity and acceleration covary across the genome, we sought to disentangle the individual relationship between each of the two metrics and selection. When controlling for the other metric, we saw a strong relationship between velocity and selection but not between acceleration and selection (Figure S2C). These results indicate that rapid acceleration is associated with selection primarily in that it correlates with rapid velocity, a strong indicator of selection.

The fastest-evolved regions of the human genome

Encouraged by these findings, we implemented a computational screen to identify the most rapidly evolved regions in the human lineage (Figure 1A). Using a syntenic genome-wide multiple alignment of great apes (human, chimpanzee, bonobo, gorilla, and orangutan), we inferred the probability of each nucleotide state in the human-chimpanzee ancestor at each alignment position (STAR Methods). In order to more conservatively estimate genetic differences, we only considered a site divergent between

the human-chimpanzee ancestor and humans when the ancestral inference estimated a base change with a probability of 80% or more (STAR Methods). We use this conservative method to define ''divergence density'' as the genetic distance between the human-chimpanzee ancestor and the human genome for every 500-bp window. If mutations were uniformly distributed across the genome at the rate observed in the fastest evolving 10-Mb genomic region, it would be unlikely to observe 29 or more mutations in a 500-bp window ($p < 10^{-6}$, Bonferroni-corrected binomial; STAR Methods). Thus, we define HAQERs as genomic regions with a divergence density of at least 29 evolutionary operations separating the human-chimpanzee ancestor and the human genome (Figure 1D). We identified 1,581 HAQERs with an average length of 892 bp, which collectively include ~1.41 Mb of the human genome (Figure 1E).

As we ascertained HAQERs based on their rapid divergence, it follows that they exhibit higher velocities than either HARs or randomly selected neutral proxy regions (RAND) (Figures 1F and S2D; STAR Methods). HAQERs exhibit significantly lower initial velocity than RAND, even though HAQERs were not directly ascertained based on conservation. HAQERs are also significantly more accelerated than HARs or RAND, reflecting the combination of their slightly lower initial velocity and their dramatic velocity.

HAQERs and HARs are largely independent genomic regions, with only six out of 2,733 expanded HARs[23] overlapping HAQERs. One notable overlap is HAQER0035, which corresponds to HAR1, part of a well-studied RNA gene expressed in neurodevelopment[33] (Figure S6C). HAQERs are also largely distinct from the fastest-evolved regions in chimpanzees and gorillas (Figure S2E). Thus, we have expanded beyond the highly conserved genome to identify over one thousand previously uncharacterized regions that represent the most rapidly evolved regions in the human genome.

Sequence evolution in HAQERs was driven by both elevated mutation rates and directional positive selection prior to the Neanderthal split

As rapid sequence divergence in a genomic region can be generated either by variation in the local mutation rate or by positive selection, we sought to determine the relative influence of these forces in HAQER evolution, using recently available high-coverage human population sequencing data.[26] We first partitioned variants from 501 unrelated African individuals (STAR Methods) to subsets overlapping HAQERs, HARs, RAND, ultraconserved elements (UCEs), ENCODE candidate *cis*-regulatory elements (cCREs),[25] or missense variants (MISSENSE).

We calculated the density of polymorphic sites and divergent sites between modern humans and the inferred human-chimpanzee ancestor in these regions (Figure S3A; STAR Methods). UCEs—regions that have undergone minimal sequence divergence during the last 100 million years[45]—exhibit very limited divergence and polymorphism density, compared with RAND, whereas HAQERs exhibit significantly elevated densities of both polymorphic sites and divergent sites.

We observed the co-occurrence of HAQERs and genomic features associated with higher mutation rates, suggesting an underlying mechanism for the increased density of polymorphic sites in HAQERs. HAQERs are enriched for meiotic recombination double-stranded break hotspots (106 overlaps, 1.4× enrichment, $p < 10^{-3}$), and toward the ends of chromosomes (Figures 1E and S2F), both of which have been associated with elevated local mutation rates.[46–48] We also found that HAQERs are enriched for early replication timing[49] (Figure S3C), consistent with the enrichment for meiotic recombination double-stranded break hotspots. Meiotic double-stranded breaks and subtelomeric regions are also associated with higher recombination rates,[50] and we observed a slight, yet significant, elevation of recombination rates in HAQERs[51] (Figure S3B). GC-biased gene conversion has been previously explored as a possible contributor to the divergence observed in HARs.[15] We found that HAQERs demonstrate a slight enrichment for weak to strong divergent sites. However, this enrichment is significantly weaker than we observed in HARs (Figures S3D and S3E). These observations are consistent with the hypothesis that rapid HAQER divergence is driven by elevated mutation rates.

While many HAQERs appear to have elevated mutation rates, this does not rule out that these same elements harbor function and were positively selected. Indeed, we observed a significantly elevated proportion of fixed alleles relative to polymorphic alleles at sites that are divergent between the human-chimpanzee ancestor and the human genome, a statistic associated with positive selection (Figure S3F).[52] To further explore positive selection as a contributing force to HAQER evolution, we constructed dAFS for each set of genomic regions. Again, HAQERs show signatures of positive selection driven by an enrichment of high-frequency derived alleles and a depletion of intermediate frequency alleles relative to RAND and the other sets of genomic regions (Figures 2A and 2D).

To infer the magnitude of selective pressure across populations, we partitioned each dAFS into five component dAFS containing segregating variants from individuals in each of five populations (Gambian in Western Division—Mandinka; Mende in Sierra Leone; Esan in Nigeria; Yoruba in Nigeria; and Luhye in Webuye, Kenya). We evaluated the mean selection parameters acting on each population, using MCMC (Figures 2B and 2C; STAR Methods). For HAQERs, the 95% credible intervals for the mean selection parameter acting on segregating sites are within the range of 12.7–16.5, and they did not overlap intervals from any other variant set (Figure 2C). Roughly estimating the effective population size in humans at 10^4 individuals,[53] we estimated a mean selection coefficient for bases in HAQERs ranging from $s = 0.000635$ to $s = 0.000825$.

If HAQERs evolved under directional selection, we would expect variation between humans and chimpanzees to be much larger than the variation within humans for these regions. Alternatively, under diversifying selection, HAQER divergence between the human and chimpanzee reference genomes is instead the result of an increase in human variation without directionality. To investigate these alternatives, we analyzed the distribution of the Dunn index, a conservative metric of cluster separation,[54] among clusters of modern human, ancient hominin, and chimpanzee sequences for HAQERs, HARs, and RAND (Figures 2E and S3H–S3J). Dunn index values of less than 1 suggest overlapping clusters, whereas values greater than 1 suggest distinct, well-defined clusters.

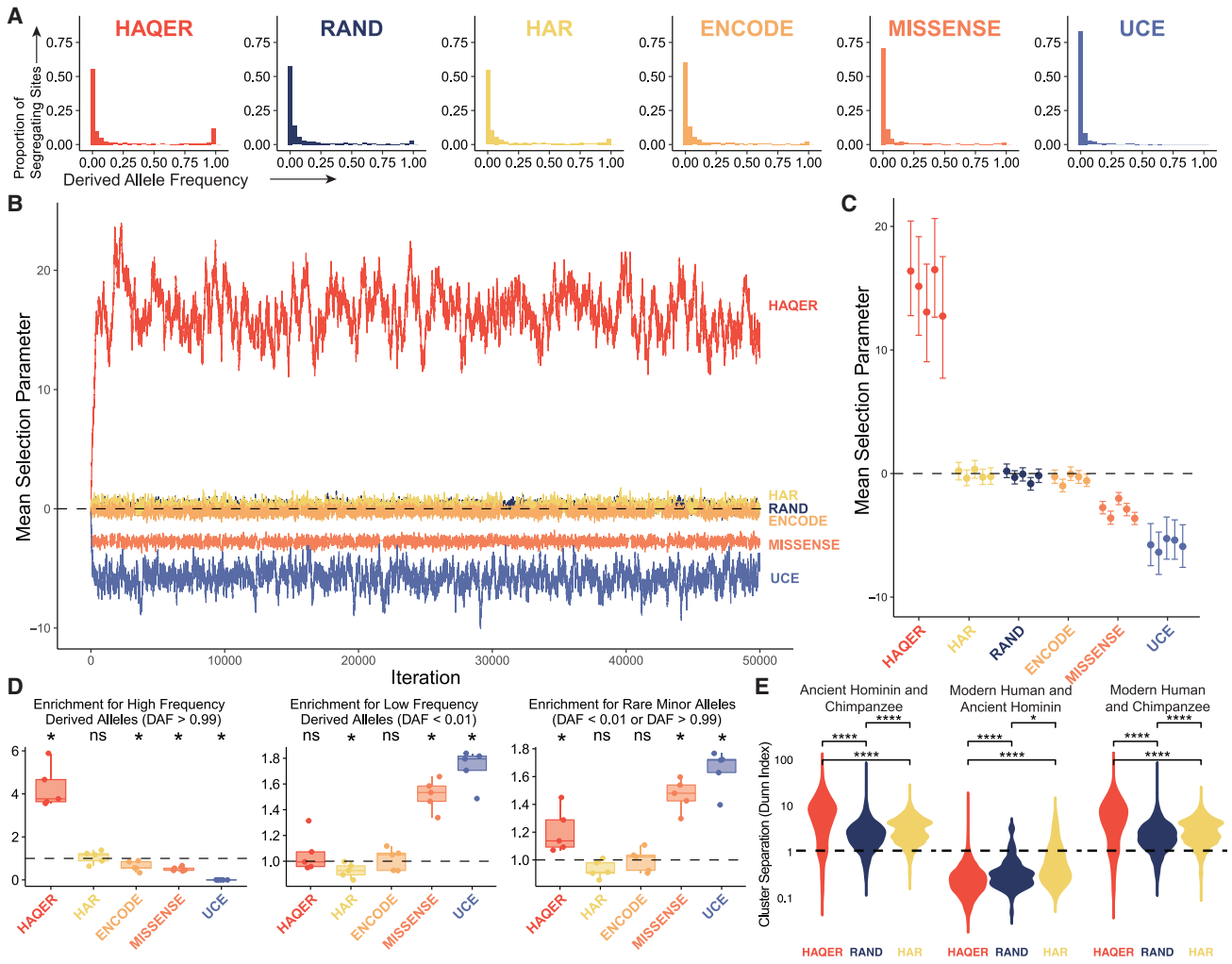


Figure 2. HAQER sequence divergence was driven by positive selection prior to the human-Neanderthal split

(A) Derived allele frequency spectra representing 501 individuals from African populations (1,002 alleles) for segregating sites within HAQERs, RAND, HARs, ENCODE candidate *cis*-regulatory elements (cCREs), missense variants (MISSENSE), or ultraconserved elements (UCEs).

(B) Representative MCMC trace for the mean selection parameter acting on segregating sites within each set of regions.

(C) Posterior mean and 95% highest density credible intervals describing the mean selection parameters for each set of regions inferred from segregating sites from five independent populations of unrelated African individuals.

(D) Enrichment for high derived allele frequency (DAF > 0.99, left), low frequency (DAF < 0.01, center), and rare minor allele (DAF < 0.01 or DAF > 0.99, right) segregating sites relative to RAND (* $p < 0.05$; Bonferroni-adjusted Mann-Whitney U). Each point represents the enrichment for one population of individuals partitioned from the set of all African individuals.

(E) Distribution of the cluster separation (measured as the Dunn index) between ancient hominins and chimpanzees (left), modern humans and ancient hominins (center), or modern humans and chimpanzees (right). Comparisons are presented between HAQERs, RAND, and HARs (Bonferroni-adjusted Mann-Whitney U; * $p < 0.05$; ** $p < 0.01$; and **** $p < 0.0001$).

See also Figure S3.

HAQERs demonstrate greater cluster separation relative to RAND when comparing either human or ancient hominin sequences with that of chimpanzees (Figure 2E). Significantly, most HAQERs have a Dunn index of less than 1 between humans and ancient hominins, suggesting that ancient hominin HAQER sequences largely fall within the range of human variability (Figure 2E). These results are consistent with rapid directional selection in humans after the split with chimpanzees, followed by a transition to constraint prior to the human-Neanderthal split.

While the dAFS model assumes infinite sites, we observed that sites in HAQERs with high derived allele frequencies exhibit an elevated proportion of transitions, which is characteristic of sites with back mutations to the ancestral state (Figure S3G; STAR Methods). If a derived allele is advantageous for many sites in HAQERs, back mutations to the ancestral state would be deleterious by comparison, and these sites would be unlikely to drift from high to intermediate derived allele frequency. Thus, the enrichment for high-frequency derived alleles observed in HAQERs may be magnified by the overabundance of fixed

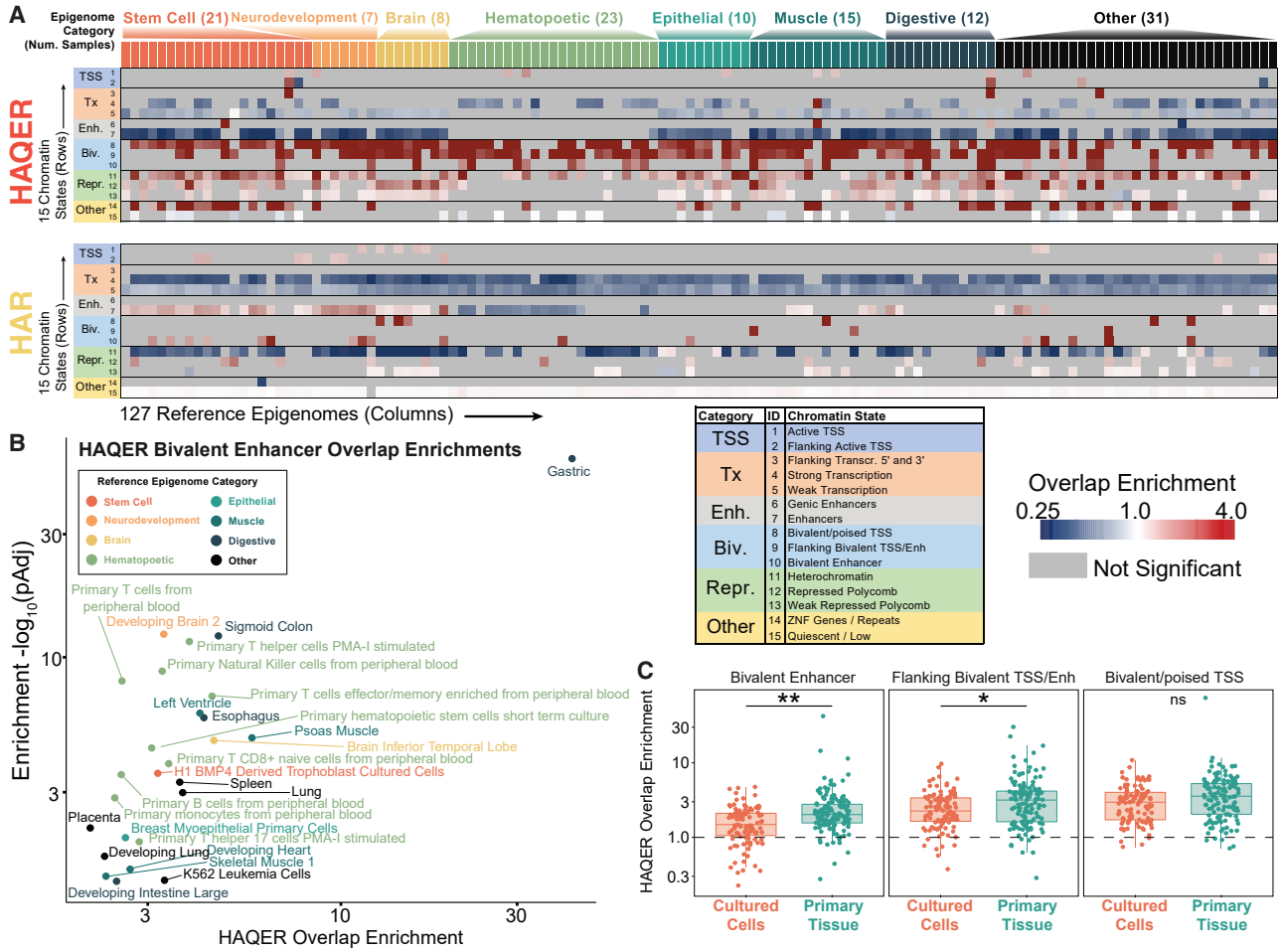


Figure 3. HAQERs are enriched in bivalent chromatin states

(A) Overlap enrichment/depletion matrix between HAQERs (top) or HARs (bottom) for 15 chromatin states (rows) from 127 reference epigenomes (columns). HAQERs are enriched for bivalent chromatin states but not for active enhancer and promoter states. An expanded matrix with individual sample annotations is presented in Figure S5A.

(B) Volcano plot displaying significant overlap enrichments for HAQERs and the bivalent enhancer chromatin state in various tissues.

(C) HAQER overlap enrichment for bivalent chromatin states compared between reference epigenomes derived from cultured cells and those derived from primary tissue (t test; * $p < 0.05$ and ** $p < 0.01$).

See also Figure S4.

differences among divergent sites (itself a signifier of positive selection), an elevated mutation rate back to the ancestral state, and the maintenance of the derived state by purifying selection in modern humans.

While mutation rate variation impacts allele frequency spectra,[55] our results do not suggest that elevated mutation rates in neutral regions are the exclusive cause of rapid divergence in HAQERs. First, the relative depletion of intermediate frequency alleles (presented in Figure 2D as an enrichment for rare alleles) and the overabundance of fixed divergent sites compared with polymorphic sites in HAQERs is not expected in selectively neutral regions (Figure S3F). Furthermore, greater HAQER sequence cluster separation relative to RAND between modern humans and chimpanzees suggests directional evolution rather than the expansion of intraspecies variability as the cause of elevated divergence.

HAQERs are enriched in chromatin states

The conclusion that HAQERs evolved through directional positive selection implies adaptive function in these regions. To test this hypothesis, we analyzed genome-wide patterns of enrichment and depletion in chromatin states across 127 reference epigenomes[37] (Figures 3A and S4A). Both HAQERs and HARs are significantly depleted in transcriptionally active chromatin states, consistent with past reports that most rapid evolution occurs outside of protein-coding regions[15] and the predicted significance of non-coding regulatory regions to evolution.[13]

Surprisingly, while HAQERs are not enriched for active enhancer or promoter states, they are strongly enriched for bivalent chromatin states (Figure 3A). Bivalent chromatin, which harbors both the polycomb repression mark H3K27me3 and the active promoter mark H3K4me3 and/or the active enhancer mark H3K4me1, is proposed to maintain expression of

developmentally and environmentally responsive genes at low levels through active, yet rapidly reversible, silencing that allows precise activation.[56,57]

HAQERs are significantly enriched for bivalent chromatin states in both developing and adult primary tissues (Figure S4C). Evolutionary changes to developmental gene regulatory programs can alter adult morphology including allometric relationships. One example is gut reduction and brain expansion on the human lineage, which have been linked by the expensive tissue hypothesis.[58] Consistent with these dramatic changes, we observed the most significant enrichments for the bivalent enhancer chromatin state in gastrointestinal and neurodevelopmental reference epigenomes (Figures 3B and S4E). As a glimpse of environmental response in adult tissues, we observed that two HAQERs transition from bivalent to active enhancer states in adult epithelial cells, following exposure to dexamethasone, an anti-inflammatory glucocorticoid[29] (enrichment $p < 0.01$; STAR Methods).

While many of the observed bivalent states may represent domains in which individual histones bear both active and repressive modifications simultaneously (true bivalency), the observation of bivalent states in bulk ChIP-seq data may be a consequence of differential states of activation and repression in distinct cell types within heterogeneous tissues (mixed cell bivalency). We observed stronger HAQER overlap enrichments for bivalent chromatin states in reference epigenomes derived from primary tissues than reference epigenomes derived from cultured cells, which represent a single-cell type (Figure 3C); however, even reference epigenomes derived from cultured cells exhibit significant enrichments for bivalent states, suggesting both mixed cell and true bivalency in HAQERs.

In either scenario, genomic regions in bivalent states are likely to demonstrate more restricted spatial and temporal patterns of activity than regions with uniform active regulatory states in heterogeneous tissues. In contrast to HAQERs, HARs are associated with active enhancer states (Figure S4B), which are thought to be associated with more broadly expressed genes.[56] Thus, enrichments for bivalency suggest that HAQERs encode gene regulatory elements with a high degree of specificity in development and environmental response.

HAQERs are enriched for recently evolved neurodevelopmental gene regulatory elements

If the adaptive divergence observed in HAQERs underlies the innovation of developmental gene regulatory functions, we would expect differences in the epigenomic profiles between humans and closely related species. While cross-species epigenomic profiles of developing tissue are not broadly available, the developing cerebral cortex, owing to its association with human cognition,[3] has been profiled across humans, rhesus macaques, and mice to identify putative enhancers and promoters in the human genome that were gained after the split between humans and rhesus macaques.[38] Despite not being significantly associated with active enhancer or promoter states in the developing brain overall (Figures 3A and S4A), HAQERs exhibit an enrichment for overlapping the subset of active enhancer or promoter chromatin states that were gained after the rhesus split (Figure S4D). While we observe enrichments between HAQERs and putatively gained gene regulatory activity identified across developmental stages and brain regions, HAQERs demonstrate the greatest enrichment for gained elements in the frontal lobe in late embryonic neurodevelopment (Figure S4D).

A multiplex, single-cell *in vivo* enhancer assay reveals hominin-specific neurodevelopmental enhancer activity in HAQERs

We identified the developing brain as a tissue of interest for *in vivo* analysis of HAQER function because HAQERs are enriched near genes associated with olfaction and cell recognition (Figure S2H), make 3D chromatin contacts with nervous system genes (Figure S2G), are enriched for neurodevelopmental regulatory elements gained after the human-rhesus split (Figure S4D), and are highly enriched for bivalent enhancers in the developing brain (Figure 3B). Notably, the brain has changed dramatically on the human lineage[3] and is associated with many human-specific disease susceptibilities.[23]

STARR-seq is a high-throughput sequencing-based assay in which the abundance of RNA transcripts containing a particular test sequence provides a quantitative measure of enhancer activity.[27] While STARR-seq has been effectively employed in cultured cell lines,[27,29,59] our results suggest that HAQERs function in spatiotemporally restricted contexts in highly heterogeneous tissues, such as in late embryonic neurodevelopment.[60] Thus, we performed *in vivo* scSTARR-seq to measure the enhancer activity of multiple test sequences simultaneously in developing brain tissue.

In this assay, we cloned DNA sequences into a STARR-seq vector[27] to form input libraries. We injected input libraries, along with a constitutive GFP transfection reporter plasmid, into embryonic mouse cerebral cortices via *in utero* electroporation (Figure 4A; STAR Methods). Following dissection 16–18 h later, we used fluorescence-activated cell sorting (FACS) to enrich for GFP+ cells for subsequent single-cell RNA sequencing. This approach allowed us to interrogate enhancer activity in electroporated cells as well as their immediate progeny.

To identify candidates for human-evolved neurodevelopmental regulatory elements, we identified 105 HAQERs that overlap one of three datasets: functional elements gained after the rhesus split;[38] open chromatin in the developing human brain;[37] or regions with differential chromatin accessibility between human and chimpanzee cerebral organoids, which recapitulate many features of early neurodevelopment[39,61] (Figure 4A). We were able to commercially synthesize 40 of these sequences, a requirement for the analysis of extinct and ancestral alleles. We conducted a pilot assay with only the human alleles and selected the 13 with the strongest signal for a full comparative analysis between the hominin (human, Neanderthal, and Denisovan) and non-hominin (chimpanzee and inferred human-chimpanzee ancestor) alleles (Data S1).

We performed two independent *in vivo* scSTARR-seq experiments with this injection library and recovered STARR-seq reporter reads, endogenous RNA reads, and transfection reporter reads simultaneously from 7,170 single cells (Figure 4B). As these two experiments were performed at temporally close

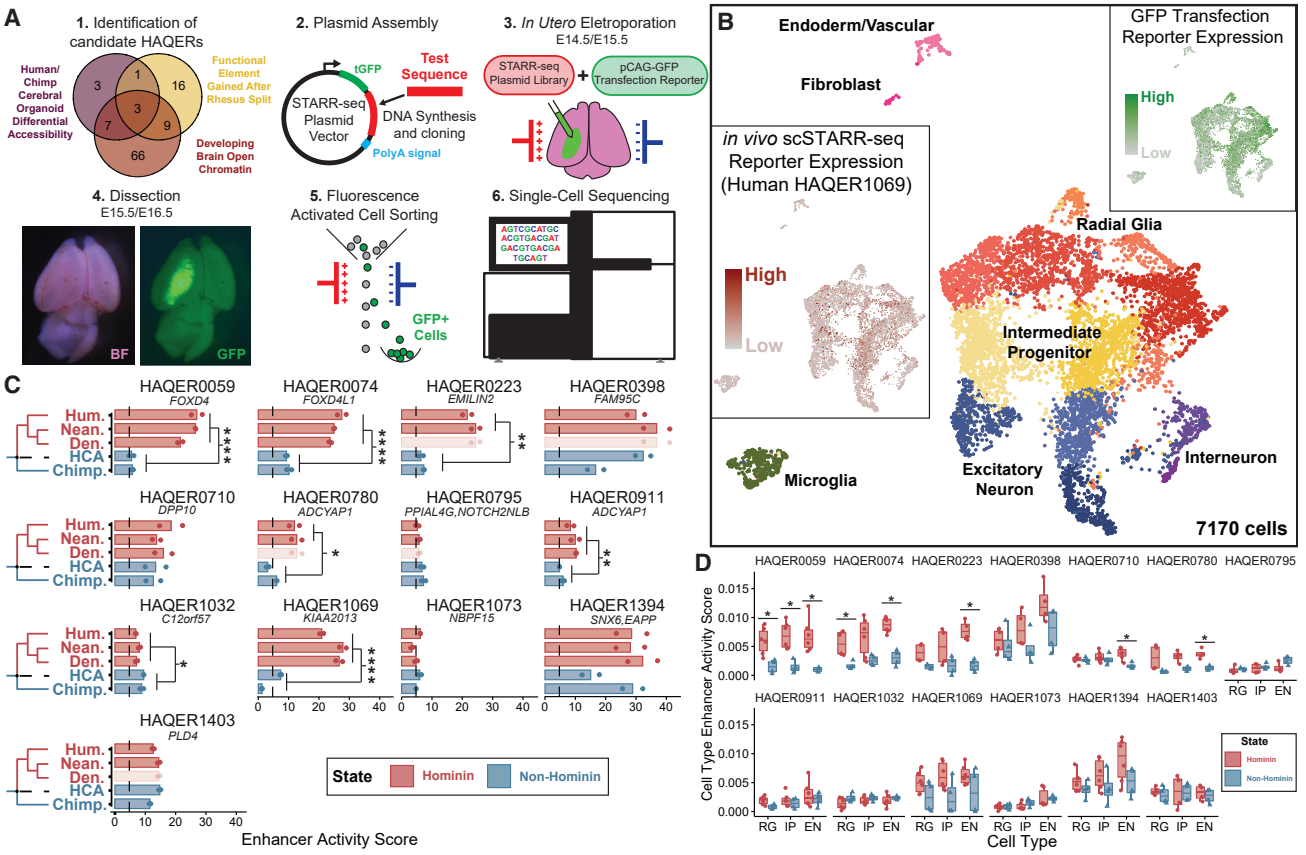


Figure 4. Rapid sequence divergence in HAQERs generated hominin-specific neurodevelopmental enhancers

(A) Experimental design. Candidate HAQERs were prioritized based on overlaps with epigenomic datasets, cloned into a STARR-seq vector, and electroporated into developing mouse brains along with a pCAG-GFP transfection reporter. Single-cell sequencing followed dissection and FACS enrichment of GFP+ cells.

(B) UMAP projection of 7,170 single cells from two scSTARR-seq experiments, labeled with metacluster identities. Inserts display cells colored by expression of the GFP transfection reporter and human HAQER0169.

(C) Enhancer activity score, defined as the input-normalized unique molecular identifier (UMI) count pooled across all cells per 1,000 reporter UMIs, for 13 HAQERs. Nearest gene name is displayed below each HAQER ID. We display significant differences in enhancer activity between hominin (human, Neanderthal, and Denisova) and non-hominin (chimpanzee, human-chimpanzee ancestor [HCA]) sequences (Bonferroni-adjusted t test, $p < 0.05$). Faded bars represent sequences where Neanderthal and Denisovan had the same sequence in the 500-bp genomic region, and these duplicate sequences are not included in the statistical analysis.

(D) Cell-type enhancer activity score or the input-normalized reporter UMI count normalized to the pCAG-GFP UMI count for each cell averaged across all cells in each metacluster (RG, radial glia; IP, intermediate progenitor; EN, excitatory neuron) (FDR-corrected t test, $p < 0.05$).

See also Figures S5 and S6.

developmental time points (injections at E14.5 and E15.5), we observed a strong correlation between enhancer activity scores in both experiments (Figures S5E and S5F).

As most rapid sequence divergence in HAQERs occurred prior to the human-Neanderthal split, we expected similar patterns of enhancer activity among hominin sequences and compared enhancer activity between hominin and non-hominin sequences. Critically, 6 of the 13 HAQERs demonstrated significantly greater enhancer activity in the hominin ortholog test sequences than in the non-hominin sequences (Figures 4C, S6A, S6B, and S6D). Additionally, HAQER1032 showed a small but statistically significant decrease in enhancer activity in hominin orthologous sequences (Figure 4C). Many of the non-hominin sequences exhibit similar enhancer activity to random sequence negative controls. The lack of observed functionality of non-hominin alleles suggests that these HAQERs represent hominin-specific functional elements forged from a previously neutrally evolving sequence, a class of elements excluded from previous comparative genomic screens reliant on functional constraint outside of humans.

We next sought to leverage the single-cell resolution of *in vivo* scSTARR-seq to determine the cell-type specificity of enhancer activity in developing tissue for HAQERs. We annotated cell types, utilizing developing brain cell atlases[60,62] to calculate an enhancer activity score specific to each cell type (Figures 4B, 4D, S5A, and S5B; STAR Methods). *In utero* electroporation preferentially targets ventricular progenitors. Thus, we observed the most GFP signal in radial glia and radial glial progeny, including intermediate progenitors and newborn excitatory neurons (Figures 4B and S5C). While we resolved clusters with inhibitory neuron, microglia,

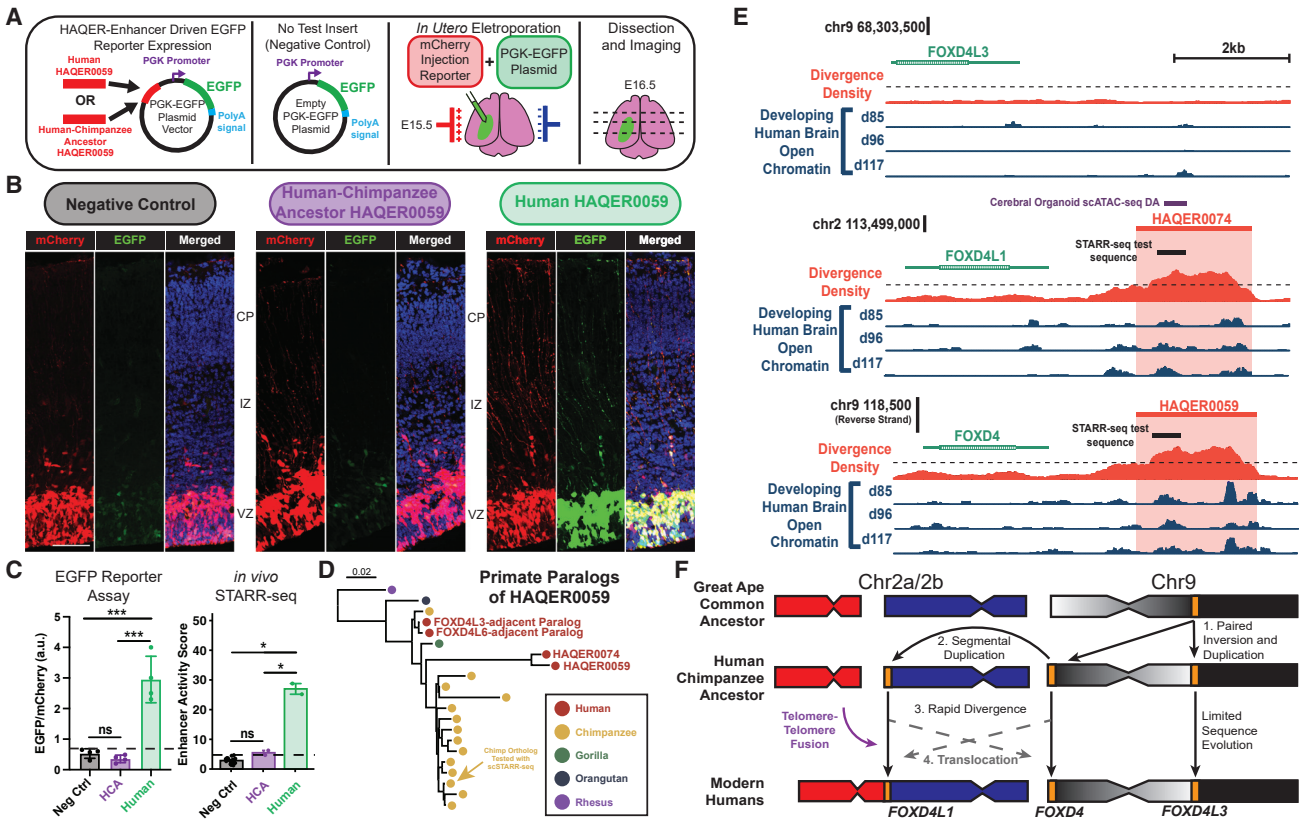


Figure 5. Rapid divergence of hominin-specific neurodevelopmental enhancers near *FOXD4* family genes followed multiple segmental duplications

(A) Experimental design. We cloned the human or inferred human-chimpanzee ancestral sequence of HAQER0059 into an PGK-EGFP reporter plasmid and delivered plasmids to the developing cerebral cortex via *in utero* electroporation at E15.5 alongside an mCherry injection reporter. We performed dissection, sectioning, and imaging 24 h later.

(B) Representative images of Hoechst-stained coronal sections imaged for the mCherry injection reporter and EGFP enhancer reporter. Scale bars, 100 μm.

(C) Left: quantification of PGK-EGFP reporter signal normalized to the mCherry injection reporter for HAQER0059. Right: corresponding *in vivo* STARR-seq results (* $p < 0.05$; *** $p < 0.001$; FDR t test. Dotted line, negative control mean + 3SD).

(D) Phylogeny of HAQER0059 homologs in humans and other great apes.

(E) Genomic context for the paralogous regions near the genes *FOXD4L3*, *FOXD4L1*, and *FOXD4*. We present genomic context for the region near *FOXD4*, which contains HAQER0059, on the reverse strand. The region near *FOXD4L3* does not have a nearby HAQER and shares synteny with the great ape ortholog.

(F) A model of recent *FOXD4* evolution. The great ape ortholog of the human gene *FOXD4L3* generated the paralog *FOXD4* in the chromosome 9 subtelomere through paired inversion and duplication. Subsequent duplication produced the paralog *FOXD4L1* at the fusion site between the ancestral chromosomes 2a/2b, which formed the modern human chromosome 2.

See also Figure S6.

and fibroblast cell-type identities, these clusters exhibited limited GFP expression as they were not targeted by electroporation (Figure S5C). Therefore, to control for differences in transfection efficiency when calculating cell-type-specific enhancer activity, enhancer activity scores were normalized to the amount of GFP observed in each cluster (STAR Methods). We observed that 5 of our 13 HAQER sequences demonstrated a significant increase in enhancer activity in hominin sequences in at least one cell type. While HAQER0911 and HAQER1032 exhibited significant hominin/non-hominin differences in bulk tissue, we did not observe a similar result at the metacluster level where we had less statistical power. Notably, HAQER0710 demonstrated hominin-specific enhancer activity in excitatory neurons, a result that was not visible in bulk tissue (Figures 4B and 4D). This result highlights the potential of single-cell technologies to uncover cell-type-specific gene regulatory function in complex tissues.

As an orthogonal confirmation to our discovery of human-specific brain enhancers, we introduced the human and the human-chimpanzee ancestor sequence of HAQER0059 into an additional plasmid to test for enhancer-driven EGFP expression (Figure 5A; STAR Methods). After *in utero* electroporation in the developing mouse brain, we observed robust expression of enhancer-driven EGFP for the human construct but not the ancestral ortholog of HAQER0059 (Figures 5B and 5C), validating our multiplex sequencing assay with an independent fluorescence-based methodology.

Segmental duplication of human-specific paralogs follows rapid divergence in HAQERs

We observed that many HAQERs are contained within recent segmental duplications. This is consistent with the prevalence of differential expression between paralogous genes created by human-specific segmental duplications.[61,63] Two hominin-specific enhancers that we identified, HAQER0059 and HAQER0074, are located near the paralogs *FOXD4* and *FOXD4L1*, respectively. *FOXD4* encodes a forkhead-family transcription factor that is necessary for neuronal differentiation[64,65] and implicated in psychiatric disorders.[66] The genome assemblies of mouse, gorilla, and orangutan contain one *FOXD4* paralog, corresponding to the location of the human gene *FOXD4L3* on chromosome 9. This suggests that one *FOXD4* paralog was present in the great ape common ancestor. The short arm of the modern human chromosome 9 is inverted relative to that of gorilla and orangutan. In humans, *FOXD4L3* is found near the inversion breakpoint, and an additional paralog, *FOXD4*, is found at the other end of the inversion in the chromosome 9 subtelomere, suggesting a paired inversion and duplication event following the split with gorilla (Figure 5F). Chimpanzees exhibit an additional paralog in the subtelomere of chromosome 2b, suggesting an additional segmental duplication. In humans, this paralog corresponds to *FOXD4L1* and is located at the site of the end-to-end fusion[67] of the ancestral chromosomes 2a and 2b that formed the modern human chromosome 2 (Figure 5F). Even though HAQER0059 and HAQER0074 are both highly divergent from the human-chimpanzee ancestor, they exhibit 97.6% identity in the 500-bp regions used as STARR-seq inserts[68] (Figures 5D and 5E). However, the orthologous region near *FOXD4L3* is not highly divergent from the ancestral sequence. While the similarity between HAQER0059 and HAQER0074 could be explained by convergent evolution, this would require over 100 parallel mutations. Thus, we propose that one of two paralogs rapidly diverged, and a subsequent event translocated the highly diverged paralog to the paralogous location on the other chromosome, resulting in the same highly diverged sequence on the ends of both chromosomes 2b and 9.

Additionally, we observed 26 HAQERs in the 1q21.1-2 region containing the NBPF gene cluster (Figure S6E), which contains several human-specific segmental duplications.[10,69,70] NBPF genes contain Olduvai domains, which have undergone the most dramatic copy-number increase of any protein-coding region in the human lineage.[70,71] Copy number of Olduvai domains is implicated in a dose-dependent manner with brain size, and deletions and duplications in this region are associated with microcephaly and macrocephaly, respectively.[72] These results are consistent with the hypothesis that adaptive increases of expression in *FOXD4* and *NBPF* were achieved through the paired action of *cis*-regulatory innovation and segmental duplication. We propose that the cooperation between independent molecular mechanisms may be a common method of rapid evolution.

HAQER evolution shapes human disease susceptibility

To investigate the relationship between HAQERs and disease, we calculated if segregating variants in HAQERs are linked to SNPs that have been associated with human diseases and disorders through genome-wide association studies (GWASs). For each variant in the GWAS catalog, we used population resequencing data to identify the additional segregating sites that are in linkage disequilibrium with the reported SNP. We performed this calculation for all GWAS SNPs associated with an annotated trait to get the set of all observed linked variations for a particular GWAS trait (Figure 6; STAR Methods). HAQERs are highly enriched for variation linked to GWAS traits like hypertension, neuroblastoma, and unipolar depression/schizophrenia/bipolar disorder (Figure 6B; STAR Methods).

Variants in HAQERs could be associated with disease risk due to pleiotropic effects, where selection for an advantageous mutation in the HAQER element is accompanied by a deleterious side effect in an independent trait. In single-locus pleiotropy, the DNA segment of the HAQER element has multiple functions, and the same variant that is selected for an advantageous change has an accompanying deleterious change, which is realized as a susceptibility to a disease. HAQERs do not show signs of locus-specific pleiotropy, as their single-locus pleiotropy scores are much lower than that of HARs and similar to that of RAND, suggesting that many HAQERs perform more specific functions (Figure S6H; STAR Methods). Alternatively, if HAQERs exhibit elevated haplotype lengths due to recent positive selection, we may expect HAQER disease enrichments to be the result of linkage disequilibrium-induced pleiotropy, where causal disease variants occur in elements distinct from HAQERs but in the same haplotype. However, we do not expect elevated haplotype lengths in HAQERs, as the divergence in HAQERs had largely subsided prior to the human-Neanderthal split (Figure 2E), and selection-associated haplotype length elevation dissipates via recombination on the timescale of tens of thousands of years.[73] Indeed, segregating sites in HAQERs occur on smaller haplotypes than in random regions (Figures S6F and S6G; STAR Methods), a reflection of their slightly elevated recombination frequency (Figure S3B). Thus, it is unlikely that HAQER disease enrichments are the result of linkage disequilibrium-induced pleiotropy. HAQER-associated disease susceptibility is not driven primarily by pleiotropic effects, as HAQERs do not exhibit significant pleiotropy either at their genomic position or through linkage disequilibrium.

We propose that HAQERs confer disease susceptibilities in humans as they are located in genomic regions with elevated mutation rates (Figure S3B). We expect subsequent mutations to commonly occur in HAQERs and these mutations often to be deleterious and associated with disorders. It is likely that these disease susceptibilities will be specific to humans since many HAQERs are only functional in humans, and more generally, HAQERs have largely distinct gene ontology enrichments from HAQER-like regions in other species (Figure S2H). Thus, although HAQER evolution was adaptive in the human lineage, the association with disease variants suggests rapid divergence generated human-specific disease susceptibilities as consequences.

DISCUSSION

While there has been substantial disagreement in whether highly divergent regions reflect the action of natural selection[74] or

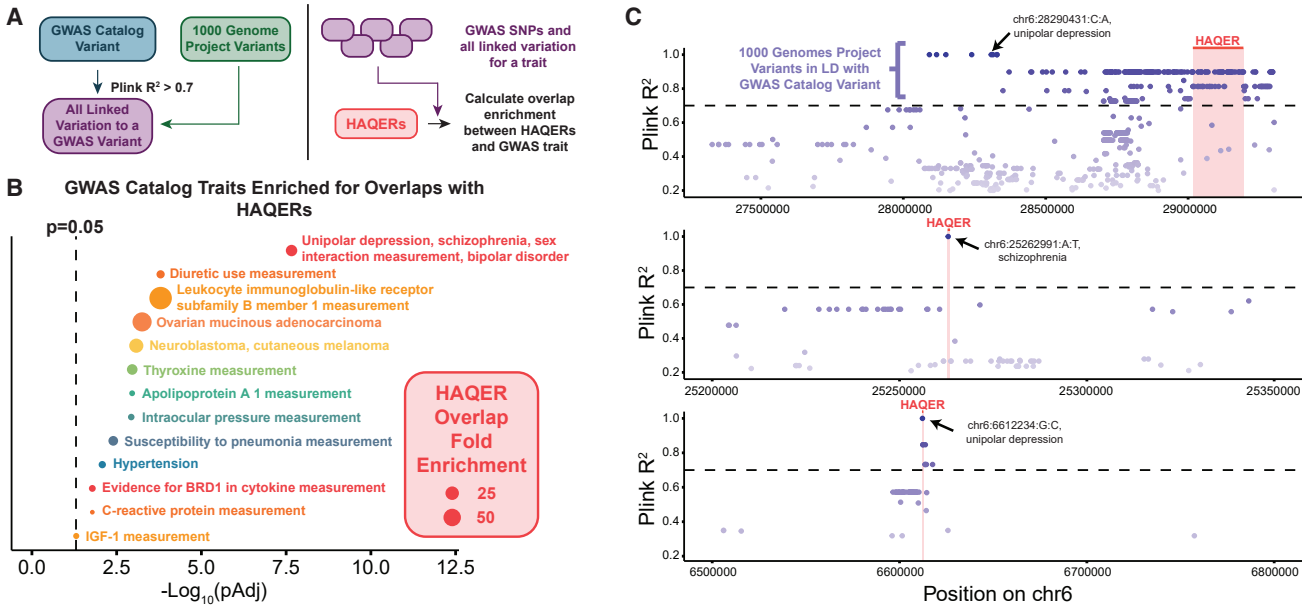


Figure 6. HAQERs are enriched near genetic variants linked to human disease

(A) Experimental design. For each GWAS catalog variant, we identified all linked variants ($R^2 > 0.7$). We calculated overlap enrichments between HAQERs and the set of all linked variations for all SNPs associated with a GWAS trait.

(B) FDR-corrected p values for overlap enrichments between HAQERs and GWAS catalog traits. The dotted line marks pAdj = 0.05.

(C) Representative overlaps between HAQERs (red) and 1000 Genomes variants (purple) linked to GWAS catalog variants (black labels).

See also Figure S6.

variation in the local mutation rate,[36] researchers have speculated that the careful integration of human population genetic data into comparative genomic efforts could effectively resolve the mutually confounding signatures of selection and mutation rate variation.[75]

Even though variation in local mutation rate and positive selection are often presented as mutually exclusive explanations for the generation of rapidly evolved regions, we have found evidence for both positive selection and elevated local mutation rates in HAQERs, suggesting that the combination of these two forces shaped the most divergent regions in the human genome.

Importantly, we have identified that the adaptive evolution of HAQERs produced functional consequences in humans and ancient hominins. HAQERs are strongly enriched in bivalent chromatin, particularly in the gastrointestinal tract, immune system, and developing brain. We developed a multiplex, single-cell enhancer assay to demonstrate that rapid sequence divergence in HAQERs forged hominin-specific gene regulatory elements.

HAQERs transitioned from rapid evolution following the human-chimpanzee ancestor to constraint among modern humans. Neanderthal and Denisovan HAQER sequences fall in the range of human variability for both sequence and function, suggesting that the rapid divergence of HAQERs largely predates this population split. While the recent accessibility of Neanderthal and Denisovan genomes has spurred substantial investigation into the differences between humans and these extinct hominins, many of the defining phenotypic transitions of the human lineage, including bipedalism and brain expansion, are shared among us. HAQERs, at the level of both sequence and function, separate us as humans from our great ape ancestors through rapid divergence and yet unite us as a species through modern constraint.

HAQERs and HARs show striking similarities in the anatomical specificity of their function. Both sets show enrichments for the brain and gastrointestinal tract. These consistent genomic enrichments parallel known anatomical changes on the human lineage of brain expansion and gut reduction. These two changes are proposed to have co-evolved to maintain a relatively constant basal metabolic rate.[58]

While HAQERs and HARs show similarities in the tissues they impact, we propose that these sets represent distinct classes of regulatory innovation during vertebrate evolution. HAQERs include *de novo* functional elements generated from neutral regions, whereas HARs represent modifications of existing functional elements. This view is consistent with differences we observe between HAQERs and HARs in selection parameters, chromatin states, and pleiotropic effects. In terms of selection, HAQERs may be a better fit for a unimodal model of selection where many bases are under positive selection as a regulatory element is forged from neutral sequence. By contrast, HARs are modifications of existing functional elements, and we expect their composition to be a mixture of bases under negative selection that maintain prior function and bases influenced by positive selection. Therefore, it is unsurprising that our selection model, which evaluates a selection parameter averaged across all sites, does not observe a substantial deviation from neutrality in HARs. In terms of chromatin states, HAQERs demonstrate strong and

consistent enrichments for bivalent chromatin states, which are associated with spatiotemporally restricted regulatory contexts, whereas HARs are associated with active enhancer states that function more broadly. Consistent with this functional specificity, we observe limited pleiotropic variation in HAQERs while HARs are substantially pleiotropic, as may be expected from modifying highly conserved active enhancers. This difference is consistent with newer and more specific functions in HAQERs, compared with older and more multifunctional regulatory elements modified in HARs. Importantly, the relative contributions of gene regulatory element gain, loss, and modification to vertebrate evolution and disease remain unknown. We propose that forging functional elements from previously non-functional regions is likely to play an outsized role in regulatory differences among species by circumventing pleiotropic constraints that reduce the evolvability of many highly conserved developmental enhancers.[76]

The observation of high mutation rates in positively selected HAQERs is explained by the non-uniformity of evolvability in vertebrate genomes. As an example, populations of marine stickleback fish have independently adapted to freshwater habitats by reducing their pelvises through the deletion of a developmental enhancer.[77] While more than one enhancer deletion can achieve pelvic reduction,[78] wild populations recurrently exhibit deletions of the same enhancer located in a region that is highly susceptible to double-stranded breaks.[79] Often, many possible mutations can produce the same adaptive phenotype. When similarly adaptive mutations occur at different rates, mutations with higher rates of occurrence will be used preferentially for adaptation. In fact, we observed elevated mutation rates in HAQERs and expect this pattern of elevated mutation rates in positively selected regions to be common throughout vertebrate life.

Some hypermutable regions utilized by adaptive evolution will retain their mutability in the derived state, such as regions prone to double-stranded breaks during meiosis, whereas other regions will not, including deletions at fragile sites.[79] We propose that positively selected regions that maintain hypermutability in the derived state will predispose organisms to disease susceptibility through subsequent deleterious mutations. Indeed, HAQERs are enriched for human genetic variants linked to diseases ranging from hypertension to neuropsychiatric disease. Thus, we anticipate a general correspondence between mutation rate, positive selection, and species-specific disease susceptibility across vertebrate evolutionary history.

Limitations of the study

First, as we conservatively limited our analysis to well-assembled syntenic regions to avoid the overestimation of divergence from paralog misalignment, we believe many highly divergent regions between great apes have yet to be found. Many genome assembly gaps are located near centromeres, telomeres, and highly paralogous regions, which are also regions enriched for HAQERs. The discovery of these regions will likely require the completion of telomere-to-telomere assemblies of great ape species to resolve syntenic relationships. Second, confident ancestral sequence reconstruction of the human-chimpanzee ancestor allele requires a minimal level of identity between great ape species. Therefore, HAQERs may be missed in alignable regions where different mutations at the same base position occurred in many independent lineages. Similarly, our current method will not detect rapid evolution in positions where humans and other great apes have all independently evolved to the same derived state. Third, we focused our *in vivo* functional analysis on the developing brain. While we propose that HAQERs impact many anatomical locations, future work will be required to uncover how HAQER-mediated regulatory innovation impacts target gene expression and phenotypic changes across diverse tissues and stages. Finally, several HAQERs of interest overlapped simple repeat sequences. We were unable to test these HAQERs for enhancer activity due to limitations in current methods of DNA synthesis, which is required to investigate haplotypes of extinct and ancestral species.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Human genetic variation preprocessing
  - Bayesian model design
  - Likelihood calculations
  - MCMC evaluation of selection parameters
  - Divergence-based ascertainment corrections
  - MCMC validation with synthetic datasets
  - Genome-wide multiple alignment
  - Divergence velocity and acceleration analysis
  - Ancestral state inference
  - HAQER Identification
  - Chromosome location
  - GREAT Ontology Analysis
  - Mutation rate and fixation estimation
  - Recombination and replication timing
  - Mutation spectrum analysis
  - Back mutation analysis
  - Great ape genome divergence analysis
  - Chromatin state enrichment analysis
  - Functional annotation of HAQERs
  - Gene synthesis and plasmid preparation
  - *In utero* electroporation
  - Immunofluorescence staining and image acquisition
  - Fluorescence activated cell sorting
  - scSTARR-seq reporter read targeted enrichment
  - scSTARR-seq sequencing and preprocessing
  - Enhancer activity quantification
  - Single-cell cluster identification and cell-type specific enhancer activity quantification
  - Enhancer paralog phylogenetic analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2022.10.016.

ACKNOWLEDGMENTS

We thank the Duke Human Vaccine Institute Research Flow Cytometry and Viral Genome Analysis core facilities. Shataakshi Dube and Scott Soderling generously provided the pCAG-GFP plasmid. We thank Douglas A. Marchuk, Christiana Fauci, Chelsea R. Shoben, Seth Weaver, Shae Simpson, and Yanting Luo for critical feedback. This research was supported by the Duke Whitehead Scholarship, National Human Genome Research Institute (R35HG011332), the Sigma Xi Grants in Aid of Research Program, North Carolina Biotechnology Center (2016-IDG-1013 and 2020-IIG-2109), and the Triangle Center for Evolutionary Medicine.

AUTHOR CONTRIBUTIONS

R.J.M. and C.B.L. conceived the study and wrote the paper. R.J.M., F.C.A., F.M., J.E.S.-F., D.A.S., E.H.A., J.C., L.S., and G.D.J. performed experiments and analyses. T.E.R., D.L.S., and C.B.L. supervised and funded research. All authors reviewed, edited, and approved the manuscript prior to submission.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 9, 2022
Revised: September 8, 2022
Accepted: October 14, 2022
Published: November 23, 2022

REFERENCES

1. Sockol, M.D., Raichlen, D.A., and Pontzer, H. (2007). Chimpanzee locomotor energetics and the origin of human bipedalism. Proc. Natl. Acad. Sci. USA *104*, 12265–12269. https://doi.org/10.1073/pnas.0703267104.
2. Vick, S.J., Waller, B.M., Parr, L.A., Smith Pasqualini, M.C., and Bard, K.A. (2007). A cross-species comparison of facial morphology and movement in humans and chimpanzees using the facial action coding system (FACS). J. Nonverbal Behav. *31*, 1–20. https://doi.org/10.1007/s10919-006-0017-z.
3. Geschwind, D.H., and Rakic, P. (2013). Cortical evolution: judge the brain by its cover. Neuron *80*, 633–647. https://doi.org/10.1016/j.neuron.2013.10.045.
4. Silver, D.L. (2016). Genomic divergence and brain evolution: how regulatory DNA influences development of the cerebral cortex. Bioessays *38*, 162–171. https://doi.org/10.1002/bies.201500108.
5. Richard, D., Liu, Z., Cao, J., Kiapour, A.M., Willen, J., Yarlagadda, S., Jagoda, E., Kolachalama, V.B., Sieker, J.T., Chang, G.H., et al. (2020). Evolutionary selection and constraint on human knee chondrocyte regulation impacts osteoarthritis risk. Cell *181*. 362.e28–381.e28. https://doi.org/10.1016/j.cell.2020.02.057.
6. Xu, K., Schadt, E.E., Pollard, K.S., Roussos, P., and Dudley, J.T. (2015). Genomic and network patterns of schizophrenia genetic variation in human evolutionary accelerated regions. Mol. Biol. Evol. *32*, 1148–1160. https://doi.org/10.1093/molbev/msv031.
7. Clark, A.G., Glanowski, S., Nielsen, R., Thomas, P.D., Kejariwal, A., Todd, M.A., Tanenbaum, D.M., Civello, D., Lu, F., Murphy, B., et al. (2003). Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science *302*, 1960–1963. https://doi.org/10.1126/science.1088821.
8. Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D., White, T.J., et al. (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. *3*, e170. https://doi.org/10.1371/journal.pbio.0030170.
9. Florio, M., Heide, M., Pinson, A., Brandl, H., Albert, M., Winkler, S., Wimberger, P., Huttner, W.B., and Hiller, M. (2018). Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex. eLife *7*, e32332. https://doi.org/10.7554/eLife.32332.
10. Fiddes, I.T., Lodewijk, G.A., Mooring, M., Bosworth, C.M., Ewing, A.D., Mantalas, G.L., Novak, A.M., van den Bout, A., Bishara, A., Rosenkrantz, J.L., et al. (2018). Human-specific NOTCH2NL genes affect Notch signaling and cortical neurogenesis. Cell *173*. 1356.e22–1369.e22. https://doi.org/10.1016/j.cell.2018.03.051.
11. Heide, M., Haffner, C., Murayama, A., Kurotaki, Y., Shinohara, H., Okano, H., Sasaki, E., and Huttner, W.B. (2020). Human-specific ARHGAP11B increases size and folding of primate neocortex in the fetal marmoset. Science *369*, 546–550. https://doi.org/10.1126/science.abb2401.
12. King, M.C., and Wilson, A.C. (1975). Evolution at two levels in humans and chimpanzees. Science *188*, 107–116. https://doi.org/10.1126/science.1090005.
13. Wray, G.A. (2007). The evolutionary significance of cis-regulatory mutations. Nat. Rev. Genet. *8*, 206–216. https://doi.org/10.1038/nrg2063.
14. Carroll, S.B. (2008). Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell *134*, 25–36. https://doi.org/10.1016/j.cell.2008.06.030.
15. Pollard, K.S., Salama, S.R., King, B., Kern, A.D., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J.S., Bejerano, G., Baertsch, R., et al. (2006). Forces shaping the fastest evolving regions in the human genome. PLoS Genet. *2*, e168. https://doi.org/10.1371/journal.pgen.0020168.
16. Bird, C.P., Stranger, B.E., Liu, M., Thomas, D.J., Ingle, C.E., Beazley, C., Miller, W., Hurles, M.E., and Dermitzakis, E.T. (2007). Fast-evolving noncoding sequences in the human genome. Genome Biol. *8*, R118. https://doi.org/10.1186/gb-2007-8-6-r118.
17. Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature *478*, 476–482. https://doi.org/10.1038/nature10530.
18. Prabhakar, S., Noonan, J.P., Pääbo, S., and Rubin, E.M. (2006). Accelerated evolution of conserved noncoding sequences in humans. Science *314*, 786. https://doi.org/10.1126/science.1130738.
19. Gittelman, R.M., Hun, E., Ay, F., Madeoy, J., Pennacchio, L., Noble, W.S., Hawkins, R.D., and Akey, J.M. (2015). Comprehensive identification and analysis of human accelerated regulatory DNA. Genome Res. *25*, 1245–1255. https://doi.org/10.1101/gr.192591.115.
20. Capra, J.A., Erwin, G.D., McKinsey, G., Rubenstein, J.L., and Pollard, K.S. (2013). Many human accelerated regions are developmental enhancers. Philos. Trans. R. Soc. Lond. B Biol. Sci. *368*, 20130025. https://doi.org/10.1098/rstb.2013.0025.
21. Girskis, K.M., Stergachis, A.B., DeGennaro, E.M., Doan, R.N., Qian, X., Johnson, M.B., Wang, P.P., Sejourne, G.M., Nagy, M.A., Pollina, E.A., et al. (2021). Rewiring of human neurodevelopmental gene regulatory programs by human accelerated regions. Neuron *109*. 3239.e7–3251.e7. https://doi.org/10.1016/j.neuron.2021.08.005.

22. Boyd, J.L., Skove, S.L., Rouanet, J.P., Pilaz, L.-J., Bepler, T., Gordân, R., Wray, G.A., and Silver, D.L. (2015). Human-chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex. Curr. Biol. *25*, 772–779. https://doi.org/10.1016/j.cub.2015.01.041.

23. Doan, R.N., Bae, B.-I., Cubelos, B., Chang, C., Hossain, A.A., Al-Saad, S., Mukaddes, N.M., Oner, O., Al-Saffar, M., Balkhy, S., et al. (2016). Mutations in human accelerated regions disrupt cognition and social behavior. Cell *167*. 341.e12–354.e12. https://doi.org/10.1016/j.cell.2016.08.071.

24. Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., and Kaul, R. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. https://doi.org/10.1038/nature11247.

25. ENCODE Project Consortium, Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710. https://doi.org/10.1038/s41586-020-2493-4.

26. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell *185*. 3426.e19–3440.e19. https://doi.org/10.1016/j.cell.2022.08.004.

27. Arnold, C.D., Gerlach, D., Stelzer, C., Boryń, Ł.M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science *339*, 1074–1077. https://doi.org/10.1126/science.1232542.

28. Shen, S.Q., Myers, C.A., Hughes, A.E.O., Byrne, L.C., Flannery, J.G., and Corbo, J.C. (2016). Massively parallel cis-regulatory analysis in the mammalian central nervous system. Genome Res. *26*, 238–255. https://doi.org/10.1101/gr.193789.115.

29. Johnson, G.D., Barrera, A., McDowell, I.C., D'Ippolito, A.M., Majoros, W.H., Vockley, C.M., Wang, X., Allen, A.S., and Reddy, T.E. (2018). Human genome-wide measurement of drug-responsive regulatory activity. Nat. Commun. *9*, 5317. https://doi.org/10.1038/s41467-018-07607-x.

30. Klein, J.C., Agarwal, V., Inoue, F., Keith, A., Martin, B., Kircher, M., Ahituv, N., and Shendure, J. (2020). A systematic evaluation of the design and context dependencies of massively parallel reporter assays. Nat. Methods *17*, 1083–1091. https://doi.org/10.1038/s41592-020-0965-y.

31. Ellegren, H., Smith, N.G., and Webster, M.T. (2003). Mutation rate variation in the mammalian genome. Curr. Opin. Genet. Dev. *13*, 562–568. https://doi.org/10.1016/j.gde.2003.10.008.

32. Kimura, M. (1983). The Neutral Theory of Molecular Evolution (Cambridge University Press).

33. Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., et al. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. Nature *443*, 167–172. https://doi.org/10.1038/nature05113.

34. Katzman, S., Kern, A.D., Bejerano, G., Fewell, G., Fulton, L., Wilson, R.K., Salama, S.R., and Haussler, D. (2007). Human genome ultraconserved elements are ultraselected. Science *317*, 915. https://doi.org/10.1126/science.1142430.

35. Kern, A.D. (2009). Correcting the site frequency spectrum for divergence-based ascertainment. PLoS One *4*, e5152. https://doi.org/10.1371/journal.pone.0005152.

36. Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. Nature *437*, 69–87. https://doi.org/10.1038/nature04072.

37. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330. https://doi.org/10.1038/nature14248.

38. Reilly, S.K., Yin, J., Ayoub, A.E., Emera, D., Leng, J., Cotney, J., Sarro, R., Rakic, P., and Noonan, J.P. (2015). Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. Science *347*, 1155–1159. https://doi.org/10.1126/science.1260943.

39. Kanton, S., Boyle, M.J., He, Z., Santel, M., Weigert, A., Sanchís-Calleja, F., Guijarro, P., Sidow, L., Fleck, J.S., Han, D., et al. (2019). Organoid single-cell genomic atlas uncovers human-specific features of brain development. Nature *574*, 418–422. https://doi.org/10.1038/s41586-019-1654-9.

40. Girirajan, S., Dennis, M.Y., Baker, C., Malig, M., Coe, B.P., Campbell, C.D., Mark, K., Vu, T.H., Alkan, C., Cheng, Z., et al. (2013). Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. Am. J. Hum. Genet. *92*, 221–237. https://doi.org/10.1016/j.ajhg.2012.12.016.

41. Paulitti, A., Andreuzzi, E., Bizzotto, D., Pellicani, R., Tarticchio, G., Marastoni, S., Pastrello, C., Jurisica, I., Ligresti, G., Bucciotti, F., et al. (2018). The ablation of the matricellular protein EMILIN2 causes defective vascularization due to impaired EGFR-dependent IL-8 production affecting tumor growth. Oncogene *37*, 3399–3414. https://doi.org/10.1038/s41388-017-0107-x.

42. Watanabe, J., Nakamachi, T., Matsuno, R., Hayashi, D., Nakamura, M., Kikuyama, S., Nakajo, S., and Shioda, S. (2007). Localization, characterization and function of pituitary adenylate cyclase-activating polypeptide during brain development. Peptides *28*, 1713–1719. https://doi.org/10.1016/j.peptides.2007.06.029.

43. Wang, Y.Q., Qian, Y.P., Yang, S., Shi, H., Liao, C.H., Zheng, H.K., Wang, J., Lin, A.A., Cavalli-Sforza, L.L., Underhill, P.A., et al. (2005). Accelerated evolution of the pituitary adenylate cyclase-activating polypeptide precursor gene During human origin. Genetics *170*, 801–806. https://doi.org/10.1534/genetics.105.040527.

44. Ressler, K.J., Mercer, K.B., Bradley, B., Jovanovic, T., Mahan, A., Kerley, K., Norrholm, S.D., Kilaru, V., Smith, A.K., Myers, A.J., et al. (2011). Post-traumatic stress disorder is associated with PACAP and the PAC1 receptor. Nature *470*, 492–497. https://doi.org/10.1038/nature09856.

45. Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved elements in the human genome. Science *304*, 1321–1325. https://doi.org/10.1126/science.1098119.

46. Brown, C.A., Murray, A.W., and Verstrepen, K.J. (2010). Rapid expansion and functional divergence of subtelomeric gene families in yeasts. Curr. Biol. *20*, 895–903. https://doi.org/10.1016/j.cub.2010.04.027.

47. Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G.V., and Camerini-Otero, R.D. (2014). DNA recombination. Recombination initiation maps of individual human genomes. Science *346*, 1256442. https://doi.org/10.1126/science.1256442.

48. Arbeithuber, B., Betancourt, A.J., Ebner, T., and Tiemann-Boege, I. (2015). Crossovers are associated with mutation and biased gene conversion at recombination hotspots. Proc. Natl. Acad. Sci. USA *112*, 2109–2114. https://doi.org/10.1073/pnas.1416622112.

49. Ding, Q., Edwards, M.M., Wang, N., Zhu, X., Bracci, A.N., Hulke, M.L., Hu, Y., Tong, Y., Hsiao, J., Charvet, C.J., et al. (2021). The genetic architecture of DNA replication timing in human pluripotent stem cells. Nat. Commun. *12*, 6746. https://doi.org/10.1038/s41467-021-27115-9.

50. Jensen-Seaman, M.I., Furey, T.S., Payseur, B.A., Lu, Y., Roskin, K.M., Chen, C.F., Thomas, M.A., Haussler, D., and Jacob, H.J. (2004). Comparative recombination rates in the rat, mouse, and human genomes. Genome Res. *14*, 528–538. https://doi.org/10.1101/gr.1970304.

51. Zhou, Y., Browning, B.L., and Browning, S.R. (2020). Population-specific recombination maps from segments of identity by descent. Am. J. Hum. Genet. *107*, 137–148. https://doi.org/10.1016/j.ajhg.2020.05.016.

52. Andolfatto, P. (2005). Adaptive evolution of non-coding DNA in *Drosophila*. Nature *437*, 1149–1152. https://doi.org/10.1038/nature04107.

53. Tenesa, A., Navarro, P., Hayes, B.J., Duffy, D.L., Clarke, G.M., Goddard, M.E., and Visscher, P.M. (2007). Recent human effective population size estimated from linkage disequilibrium. Genome Res. *17*, 520–526. https://doi.org/10.1101/gr.6023607.

54. Dunn, J.C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J. Cybern. *3*, 32–57. https://doi.org/10.1080/01969727308546046.

55. Harpak, A., Bhaskar, A., and Pritchard, J.K. (2016). Mutation rate variation is a primary determinant of the distribution of allele frequencies in humans. PLoS Genet. *12*, e1006489. https://doi.org/10.1371/journal.pgen.1006489.

56. Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315–326. https://doi.org/10.1016/j.cell.2006.02.041.

57. Voigt, P., Tee, W.W., and Reinberg, D. (2013). A double take on bivalent promoters. Genes Dev. *27*, 1318–1338. https://doi.org/10.1101/gad.219626.113.

58. Aiello, L.C., and Wheeler, P. (1995). The expensive-tissue hypothesis: the brain and the digestive system in human and primate evolution. Curr. Anthropol. *36*, 199–221.

59. Klein, J.C., Keith, A., Agarwal, V., Durham, T., and Shendure, J. (2018). Functional characterization of enhancer evolution in the primate lineage. Genome Biol. *19*, 99. https://doi.org/10.1186/s13059-018-1473-6.

60. Nowakowski, T.J., Bhaduri, A., Pollen, A.A., Alvarado, B., Mostajo-Radji, M.A., Di Lullo, E., Haeussler, M., Sandoval-Espinosa, C., Liu, S.J., Velmeshev, D., et al. (2017). Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. Science *358*, 1318–1323. https://doi.org/10.1126/science.aap8809.

61. Pollen, A.A., Bhaduri, A., Andrews, M.G., Nowakowski, T.J., Meyerson, O.S., Mostajo-Radji, M.A., Di Lullo, E., Alvarado, B., Bedolli, M., Dougherty, M.L., et al. (2019). Establishing cerebral organoids as models of human-specific brain evolution. Cell *176*. 743.e17–756.e17. https://doi.org/10.1016/j.cell.2019.01.017.

62. La Manno, G., Siletti, K., Furlan, A., Gyllborg, D., Vinsland, E., Mossi Albiach, A., Mattsson Langseth, C., Khven, I., Lederer, A.R., Dratva, L.M., et al. (2021). Molecular architecture of the developing mouse brain. Nature *596*, 92–96. https://doi.org/10.1038/s41586-021-03775-x.

63. Shew, C.J., Carmona-Mora, P., Soto, D.C., Mastoras, M., Roberts, E., Rosas, J., Jagannathan, D., Kaya, G., O'Geen, H., and Dennis, M.Y. (2021). Diverse molecular mechanisms contribute to differential expression of human duplicated genes. Mol. Biol. Evol. *38*, 3060–3077. https://doi.org/10.1093/molbev/msab131.

64. Neilson, K.M., Klein, S.L., Mhaske, P., Mood, K., Daar, I.O., and Moody, S.A. (2012). Specific domains of FoxD4/5 activate and repress neural transcription factor genes to control the progression of immature neural ectoderm to differentiating neural plate. Dev. Biol. *365*, 363–375. https://doi.org/10.1016/j.ydbio.2012.03.004.

65. Sherman, J.H., Karpinski, B.A., Fralish, M.S., Cappuzzo, J.M., Dhindsa, D.S., Thal, A.G., Moody, S.A., LaMantia, A.S., and Maynard, T.M. (2017). Foxd4 is essential for establishing neural cell fate and for neuronal differentiation. Genesis *55*, e23031. https://doi.org/10.1002/dvg.23031.

66. Minoretti, P., Arra, M., Emanuele, E., Olivieri, V., Aldeghi, A., Politi, P., Martinelli, V., Pesenti, S., and Falcone, C. (2007). A W148R mutation in the human FOXD4 gene segregating with dilated cardiomyopathy, obsessive-compulsive disorder, and suicidality. Int. J. Mol. Med. *19*, 369–372. https://doi.org/10.3892/ijmm.19.3.369.

67. Fan, Y., Newman, T., Linardopoulou, E., and Trask, B.J. (2002). Gene content and function of the ancestral chromosome fusion site in human chromosome 2q13-2q14.1 and paralogous regions. Genome Res. *12*, 1663–1672. https://doi.org/10.1101/gr.338402.

68. Kent, W.J. (2002). BLAT—the BLAST-like alignment tool. Genome Res. *12*, 656–664. https://doi.org/10.1101/gr.229202.

69. Suzuki, I.K., Gacquer, D., Van Heurck, R., Kumar, D., Wojno, M., Bilheu, A., Herpoel, A., Lambert, N., Cheron, J., Polleux, F., et al. (2018). Human-specific NOTCH2NL genes expand cortical neurogenesis through delta/Notch regulation. Cell *173*. 1370.e16–1384.e16. https://doi.org/10.1016/j.cell.2018.03.067.

70. Fiddes, I.T., Pollen, A.A., Davis, J.M., and Sikela, J.M. (2019). Paired involvement of human-specific Olduvai domains and NOTCH2NL genes in human brain evolution. Hum. Genet. *138*, 715–721. https://doi.org/10.1007/s00439-019-02018-4.

71. O'Bleness, M.S., Dickens, C.M., Dumas, L.J., Kehrer-Sawatzki, H., Wyckoff, G.J., and Sikela, J.M. (2012). Evolutionary history and genome organization of DUF1220 protein domains. G3 (Bethesda) *2*, 977–986. https://doi.org/10.1534/g3.112.003061.

72. Dumas, L.J., O'Bleness, M.S., Davis, J.M., Dickens, C.M., Anderson, N., Keeney, J.G., Jackson, J., Sikela, M., Raznahan, A., Giedd, J., et al. (2012). DUF1220-domain copy number implicated in human brain-size pathology and evolution. Am. J. Hum. Genet. *91*, 444–454. https://doi.org/10.1016/j.ajhg.2012.07.016.

73. Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D., and Lander, E.S. (2006). Positive natural selection in the human lineage. Science *312*, 1614–1620. https://doi.org/10.1126/science.1124309.

74. Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K.-D., and Wray, G.A. (2007). Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nat. Genet. *39*, 1140–1144. https://doi.org/10.1038/ng2104.

75. Taylor, M.S., Massingham, T., Hayashizaki, Y., Carninci, P., Goldman, N., and Semple, C.A.M. (2008). Rapidly evolving human promoter regions. Nat. Genet. *40*, 1262–1263. author reply 1263–1264. https://doi.org/10.1038/ng1108-1262.

76. Fuqua, T., Jordan, J., van Breugel, M.E., Halavatyi, A., Tischer, C., Polidoro, P., Abe, N., Tsai, A., Mann, R.S., Stern, D.L., and Crocker, J. (2020). Dense and pleiotropic regulatory information in a developmental enhancer. Nature *587*, 235–239. https://doi.org/10.1038/s41586-020-2816-5.

77. Chan, Y.F., Marks, M.E., Jones, F.C., Villarreal, G., Shapiro, M.D., Brady, S.D., Southwick, A.M., Absher, D.M., Grimwood, J., Schmutz, J., et al. (2010). Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science *327*, 302–305. https://doi.org/10.1126/science.1182213.

78. Thompson, A.C., Capellini, T.D., Guenther, C.A., Chan, Y.F., Infante, C.R., Menke, D.B., and Kingsley, D.M. (2018). A novel enhancer near the Pitx1 gene influences development and evolution of pelvic appendages in vertebrates. eLife *7*, e38555. https://doi.org/10.7554/eLife.38555.

79. Xie, K.T., Wang, G., Thompson, A.C., Wucherpfennig, J.I., Reimchen, T.E., MacColl, A.D.C., Schluter, D., Bell, M.A., Vasquez, K.M., and Kingsley, D.M. (2019). DNA fragility in the parallel evolution of pelvic reduction in stickleback fish. Science *363*, 81–84. https://doi.org/10.1126/science.aan1425.

80. Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C., et al. (2012). A high-coverage genome sequence from an archaic denisovan individual. Science *338*, 222–226. https://doi.org/10.1126/science.1224344.

81. Wei, X., Xiang, Y., Peters, D.T., Marius, C., Sun, T., Shan, R., Ou, J., Lin, X., Yue, F., Li, W., et al. (2022). HiCAR is a robust and sensitive method to analyze open-chromatin-associated genome organization. Mol. Cell *82*. 1225.e6–1238.e6. https://doi.org/10.1016/j.molcel.2022.01.023.

82. Prado-Martinez, J., Sudmant, P.H., Kidd, J.M., Li, H., Kelley, J.L., Lorente-Galdos, B., Veeramah, K.R., Woerner, A.E., O'Connor, T.D., Santpere, G., et al. (2013). Great ape genetic diversity and population history. Nature *499*, 471–475. https://doi.org/10.1038/nature12228.

83. Navarro Gonzalez, J., Zweig, A.S., Speir, M.L., Schmelter, D., Rosenbloom, K.R., Raney, B.J., Powell, C.C., Nassar, L.R., Maulding, N.D., Lee, C.M., et al. (2021). The UCSC Genome Browser database: 2021 update. Nucleic Acids Res. *49*, D1046–D1057. https://doi.org/10.1093/nar/gkaa1070.

84. Prüfer, K., de Filippo, C., Grote, S., Mafessoni, F., Korlević, P., Hajdinjak, M., Vernot, B., Skov, L., Hsieh, P., Peyrégne, S., et al. (2017). A high-coverage Neandertal genome from Vindija Cave in Croatia. Science *358*, 655–658. https://doi.org/10.1126/science.aao1887.

85. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinform. Oxf. Engl. *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

86. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinform. Oxf. Engl. *23*, 2947–2948. https://doi.org/10.1093/bioinformatics/btm404.

87. 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. https://doi.org/10.1038/nbt.1630.

88. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. Nat. Methods *9*, 676–682. https://doi.org/10.1038/nmeth.2019.

89. Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. (2003). Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc. Natl. Acad. Sci. USA *100*, 11484–11489. https://doi.org/10.1073/pnas.1932072100.

90. Harris, R.S. (2007). Improved pairwise alignment of genomic DNA. PhD thesis (The Pennsylvania State University).

91. Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F.A., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. *14*, 708–715. https://doi.org/10.1101/gr.1933104.

92. Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113. https://doi.org/10.1186/1471-2105-5-113.

93. Shank, S.D., Weaver, S., and Kosakovsky Pond, S.L. (2018). phylotree.js - a JavaScript library for application development and interactive data visualization in phylogenetics. BMC Bioinformatics *19*, 276. https://doi.org/10.1186/s12859-018-2283-2.

94. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

95. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing (Austria). https://www.R-project.org/.

96. Hubisz, M.J., Pollard, K.S., and Siepel, A. (2011). PHAST and RPHAST: phylogenetic analysis with space/time models. Brief. Bioinform. *12*, 41–51. https://doi.org/10.1093/bib/bbq072.

97. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*. 3573.e29–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

98. Cingolani, P., Platts, A., Wang, le L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80–92. https://doi.org/10.4161/fly.19695.

99. Wright, S. (1984). Evolution and the Genetics of Populations, Volume 2: Theory of Gene Frequencies (University of Chicago Press).

100. Karolchik, D., Hinrichs, A.S., and Kent, W.J. (2012). The UCSC Genome Browser. Curr. Protoc. Bioinforma *40*, 1.4.1–1.4.33. https://doi.org/10.1002/0471250953.bi0104s40.

101. Felsenstein, J. (1997). An alternating least squares approach to inferring phylogenies from pairwise distances. Syst. Biol. *46*, 101–111. https://doi.org/10.1093/sysbio/46.1.101.

102. Fitch, W.M., and Margoliash, E. (1967). Construction of phylogenetic trees. Science *155*, 279–284. https://doi.org/10.1126/science.155.3760.279.

103. Schrago, C.G., and Voloch, C.M. (2013). The precision of the hominid timescale estimated by relaxed clock methods. J. Evol. Biol. *26*, 746–755. https://doi.org/10.1111/jeb.12076.

104. Yang, Z., Kumar, S., and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. Genetics *141*, 1641–1650.

105. Jukes, T.H., and Cantor, C.R. (1969). Chapter 24 - Evolution of protein molecules. In Mammalian Protein Metabolism, H.N. Munro, ed. (Academic Press), pp. 21–132. https://doi.org/10.1016/B978-1-4832-3211-9.50009-7.

106. Kimura, M. (1968). Evolutionary rate at the molecular level. Nature *217*, 624–626. https://doi.org/10.1038/217624a0.

107. Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43–49. https://doi.org/10.1038/nature12886.

108. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.

109. Lowe, C.B., Kellis, M., Siepel, A., Raney, B.J., Clamp, M., Salama, S.R., Kingsley, D.M., Lindblad-Toh, K., and Haussler, D. (2011). Three periods of regulatory innovation during vertebrate evolution. Science *333*, 1019–1024. https://doi.org/10.1126/science.1202702.

110. 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. https://doi.org/10.1093/bioinformatics/btp352.

111. Vockley, C.M., D'Ippolito, A.M., McDowell, I.C., Majoros, W.H., Safi, A., Song, L., Crawford, G.E., and Reddy, T.E. (2016). Direct GR binding sites potentiate clusters of TF binding across the human genome. Cell *166*. 1269.e19–1281.e19. https://doi.org/10.1016/j.cell.2016.07.049.

112. Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., and Smith, H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods *6*, 343–345. https://doi.org/10.1038/nmeth.1318.

113. Saito, T., and Nakatsuji, N. (2001). Efficient gene transfer into the embryonic mouse brain using in vivo electroporation. Dev. Biol. *240*, 237–246. https://doi.org/10.1006/dbio.2001.0439.

114. Gasperini, M., Hill, A.J., McFaline-Figueroa, J.L., Martin, B., Kim, S., Zhang, M.D., Jackson, D., Leith, A., Schreiber, J., Noble, W.S., et al. (2019). A genome-wide framework for mapping gene regulation via cellular genetic screens. Cell *176*. 377.e19–390.e19. https://doi.org/10.1016/j.cell.2018.11.029.

115. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420. https://doi.org/10.1038/nbt.4096.

116. Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. *37*, 38–44. https://doi.org/10.1038/nbt.4314.

117. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*, D1005–D1012. https://doi.org/10.1093/nar/gky1120.

118. Jordan, D.M., Verbanck, M., and Do, R. (2019). HOPS: a quantitative score reveals pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases. Genome Biol. *20*, 222. https://doi.org/10.1186/s13059-019-1844-7.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
Stbl3 Chemically Competent E. Coli	Invitrogen	Cat# C737303
Chemicals, peptides, and recombinant proteins		
Agencourt AMPure Beads	Beckman	Cat# A63881
DNase-I	New England Biolabs	Cat# M0303S
EcoRI	New England Biolabs	Cat# R3101S
Fast Green FCF	Sigma-Aldrich	Cat# F7252
FBS	ThermoFisher	Cat# 10438026
Hoechst 33342	Invitrogen	Cat# H1399
NEG-50	Richard-Allan Scientific	Epredia 6502
Phusion DNA Polymerase	New England Biolabs	Cat# M0530S
SphI	New England Biolabs	Cat# R3182S
Trypsin-EDTA	Sigma-Aldrich	Cat# 59428C
Vectashield	Vector Laboratories	H-1000-10
ZymoPURE II Plasmid Maxiprep Kit	Zymo Research	Cat# D4203
Critical commercial assays		
LIVE/DEAD Fixable Near-IR Dead Cell Stain Kit	Invitrogen	Cat# L10119
NovaSeq 6000 S-Prime Reagents	Illumina	Cat# 20040719
Chromium Next GEM Single Cell 3' Reagent Kit v3.1	10x Genomics	https://www.10xgenomics.com/support/single-cell-gene-expression/documentation/steps/library-prep/chromium-single-cell-3-reagent-kits-user-guide-v-3-1-chemistry
NEBuilder HiFi DNA Assembly Master Mix	New England Biolabs	Cat# E2621L
NEBNext Ultra II FS DNA Library Prep Kit	New England Biolabs	Cat# E6177
Deposited data		
1000 Genomes Project genomes	Byrska-Bishop et al.[26]	https://www.internationalgenome.org/data-portal/data-collection/30x-grch38
Altai Neanderthal genome	Meyer et al.[80]	https://www.eva.mpg.de/genetics/genome-projects/neandertal/
Combined Human Accelerated Region locations	Doan et al.[23]	Table S1 of Doan et al.[23]
Denisovan genome	Meyer et al.[80]	https://www.eva.mpg.de/denisova/index.html
ENCODE cCRE locations and ChromHmm Datasets	Moore et al.[25]	https://www.encodeproject.org/
GWAS Catalog Variants	GWAS Catalog	https://www.ebi.ac.uk/gwas/
HiCAR H1 and GM12878	Wei et al.[81]	GEO: GSE162819
Human ChromHmm Roadmap Epigenomics Data	Kundaje et al.[37]	http://www.roadmapepigenomics.org/
Human gained enhancer locations	Reilly et al.[38]	GEO: GSE63648
Individual chimpanzee genomes	Prado-Martinez et al.[82]	https://www.ncbi.nlm.nih.gov/sra?term=SRP018689
knownGene	Navarro Gonzalez et al.[83]	https://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/
Meiotic Recombination DSB hotspots	Pratto et al.[47]	GEO: GSE59836
Raw and processed sequencing reads	This study	GEO: GSE212159
Recombination frequency maps	Zhou et al.[51]	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130507_omni_recombination_rates
Reference genomes: *hg38, panTro6, panPan2, gorGor5, gorGor6, ponAbe3*	UCSC Genome Browser	https://hgdownload.soe.ucsc.edu/downloads.html

(*Continued on next page*)

Continued		
REAGENT or RESOURCE	SOURCE	IDENTIFIER
Replication timing datasets	Ding et al.[49]	https://www.thekorenlab.org/data
Ultraconserved Element locations	Bejerano et al.[45]	https://hgwdev.gi.ucsc.edu/
Vindija Cave Neanderthal genome	Prüfer et al.[84]	https://www.eva.mpg.de/neandertal/draft-neandertal-genome/data.html
Experimental models: Organisms/strains		
Mouse: C57BL/6J (B6) (WT)	The Jackson Laboratory	JAX: 000664
Oligonucleotides		
Synthetic STARR-seq Insert Sequences	This study	Data S1
Targeted Enrichment Primers	This study	Data S1
Recombinant DNA		
hSTARR-seq ORI vector	Addgene	RRID: Addgene 99296
PGK-EGFP	Addgene	RRID: Addgene 169744
Software and algorithms		
bcl2fastq2 Conversion Software v2.20	Illumina	https://support.illumina.com/downloads/bcl2fastq-conversion-software-v2-20.html
BWA 0.7.17	Li and Durbin[85]	https://github.com/lh3/bwa
CellRanger v6.0	10x Genomics	https://support.10xgenomics.com/
ClustalW2	Larkin et al.,[86]	https://www.ebi.ac.uk/Tools/phylogeny/simplephylogeny/
gonomics	Vertebrate Genetics Laboratory	https://github.com/vertgenlab/gonomics
GraphPad Prism	GraphPad	https://www.graphpad.com/
GREAT version 4.0.4	McLean et al.[87]	http://great.stanford.edu/public/html/
ImageJ	Schindelin et al.[88]	https://imagej.net/software/fiji/
kentUtils	Kent et al.[89]	https://github.com/ENCODE-DCC/kentUtils
lastz	Harris[90]	https://github.com/lastz/lastz
multiz	Blanchette et al.[91]	https://bio.tools/multiz
muscle	Edgar[92]	https://www.ebi.ac.uk/Tools/msa/muscle/
phylotree	Shank et al.[93]	https://phylotree.hyphy.org/
Plink	Purcell et al.[94]	https://zzz.bwh.harvard.edu/plink/
R version 4.0.5	R Foundation for Statistical Computing[95]	https://www.r-project.org/
RPHAST	Hubisz et al.[96]	https://github.com/CshlSiepelLab/RPHAST
Seurat v4.0	Hao et al.[97]	https://satijalab.org/seurat/
SNPeff	Cingolani et al.[98]	http://pcingola.github.io/SnpEff/

RESOURCE AVAILABILITY

Lead Contact

Further information and request for resources and reagents should be directed to and will be fulfilled by the lead contact, Craig B. Lowe craig.lowe@duke.edu.

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All software written for this manuscript was implemented as a part of Gonomics, an ongoing effort to develop an open-source genomics platform in the Go programming language (golang). Gonomics can be accessed at https://github.com/vertgenlab/gonomics.
- Raw and analyzed datasets, including browser tracks, sequencing files, multiple alignments, and variant sets used in selection analysis, have been made freely available on our lab website at https://www.vertgenlab.org/. Raw and analyzed datasets have

also been deposited at GEO and are publicly available as of the date of publication at the accession number listed in the key resources table.

- Any additional information required to reanalyze the data reported in this work paper is available from the Lead Contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

- Wild type B6 mouse embryos at stages E14.5 and E15.5 were used for *in utero* electroporations for both scSTARR-seq and GFP enhancer reporter assays as described in method details. Embryos were assigned to experimental or control injection plasmids sequentially by position in the uterine horn. We did not restrict our scSTARR-seq or GFP enhancer reporter assays to embryos of only one sex; our data includes both developing males and females.
- All experiments were performed in agreement with the guidelines from the Division of Laboratory Animal Resources from Duke University School of Medicine and the Institutional Animal Care and Use Committee of Duke University.

METHOD DETAILS

Human genetic variation preprocessing

To analyze the role of selection in shaping the fast-evolved regions of the human genome, we accessed haplotype-phased high-coverage genotype data from 2,504 human samples gathered by the 1000 Genomes Project[26] from the url: https://www.internationalgenome.org/data-portal/data-collection/30x-grch38.

This genotype data underwent a series of transformations to prepare it for use in our selection analysis. First, we used *gonomics: vcfFilter* to retain only autosomal biallelic substitution variants in unrelated individuals. To reduce the impact of population bottlenecks introduced by human migration events, we only considered individuals from African populations (Gambian in Western Division – Mandinka, Mende in Sierra Leone, Esan in Nigeria, Yoruba in Nigeria, or Luhye in Webuye, Kenya). We implemented *gonomics: vcfAncestorAnnotation* to determine the ancestral allele for each variant using a pairwise alignment between the human reference sequence and the inferred Human-Chimpanzee ancestor sequence (see ancestral state inference below). We retained variants where the ancestral and derived states could be clearly determined because one of the two alleles present in the extant human population matched the allele present in the inferred ancestral sequence. We removed polymorphic sites where neither allele matched the inferred ancestral sequence. We retained a total of 29,739,731 bi-allelic sites with genotype calls in 501 individuals (a total of 1002 alleles per site) after filtering and annotation.

We created subsets of these variants that overlap regions of interest for our comparative analyses. These regions of interest include six sets: HAQERs, HARs,[33] Ultraconserved Elements[45] (UCEs), missense variants, ENCODE candidate *cis*-regulatory elements,[25] a random neutral proxy (RAND), which includes regions of the genome that do not overlap exons in known genes[83] or ENCODE cCREs pseudorandomly selected from all ungapped bases in the hg38 assembly (*gonomics: simulateBed; kentUtils: featureBits*). We generated a set of missense variants from the 501 individuals from the 1000 Genomes Project using *SnpEff*.[98] We then subsampled these variant sets to contain a maximum of 1000 segregating sites for ease of computability in subsequent analyses using *gonomics:vcfFilter -subset* and *gonomics:sampleVcf*. To limit the impact of linkage disequilibrium on the shape of the derived allele frequency spectrum, we retained variants that had a minimum of 10,000 bases from any other variant in the sample set using *gonomics:proximityBlockVcf*. We generated derived allele frequency spectra from variant data with *gonomics: vcfAfs*.

For each population, we measured the proportion of three categories of derived allele frequencies (DAF): high frequency derived alleles (DAF > 0.99), low frequency derived alleles (DAF < 0.01), and rare minor alleles (DAF < 0.01 or DAF > 0.99). We then calculated the enrichment of each category as the proportion of alleles observed in each category relative to the proportion observed in our random neutral proxy set (RAND). Enrichments for a category of allele frequencies for a set of regions were calculated by a Bonferroni-adjusted Mann-Whitney *U* test compared to RAND (n = 5, corresponding to the five African populations).

Bayesian model design

To infer the direction and magnitude of selective pressure acting on the HAQERs, we implemented a hierarchical Bayesian model based on a statistical framework developed to infer the selective pressure acting in highly conserved genomic regions, using allele frequency data from human populations.[34]

We abstracted all filtered variant calls (see Human genetic variation preprocessing) for all base positions within the HAQER, or other set of genomic regions, into a set of segregating sites. We define a segregating site as a tuple, S_k, containing the quantities n_k, the number of individual alleles with a genotype call for that segregating site, and i_k, the number of individuals with the derived allele at the k^{th} segregating site.

$$S_k = \{i_k, n_k\}$$

i_k/n_k therefore represents the derived allele frequency, or the proportion of individual sequences with the derived allele at that segregating site. Furthermore, we define $\boldsymbol{S}$ as a set of segregating sites, referred to henceforth as a derived allele frequency spectrum.

$$\boldsymbol{S} = (S_1, S_2, S_3, \ldots S_n)$$

We assume that each segregating site in an allele frequency spectrum $\boldsymbol{S}$ is associated with its own selection parameter α, which is two times the product of a selection coefficient, s, and the haploid effective population size, N_e.

$$\alpha = 2N_e s$$

Therefore, the set of selection parameters corresponding to each of n segregating sites in a derived allele frequency spectrum $\boldsymbol{S}$ is represented by the vector quantity $\boldsymbol{\alpha}$.

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \ldots \alpha_n)$$

We assume that each α_k in $\boldsymbol{\alpha}$ is independently selected from a normal distribution with mean μ and standard deviation σ, where the probability that an individual value α is selected follows the function $f(\alpha|\mu, \sigma)$. Therefore, μ represents the mean selection parameter of a set of variants. Regions under neutral selection should exhibit $\mu \approx 0$ with $\mu < 0$ and $\mu > 0$ indicating negative and positive selection, respectively.

We also define the quantity $\boldsymbol{\Theta}$ to represent the following set of parameters.

$$\boldsymbol{\Theta} = \{\alpha, \mu, \sigma\}$$

Using Bayes' rule, we can represent the posterior distribution of a particular parameter set given an observed allele frequency spectrum $P(\boldsymbol{\Theta}|\boldsymbol{S})$ with the following equation.

$$P(\boldsymbol{\Theta}|\boldsymbol{S}) = \frac{P(\boldsymbol{S}|\boldsymbol{\alpha})f(\boldsymbol{\alpha}|\mu, \sigma)g(\mu)h(\sigma)}{\int\int\int P(\boldsymbol{S}|\boldsymbol{\alpha})f(\boldsymbol{\alpha}|\mu, \sigma)g(\mu)h(\sigma)d\alpha d\mu d\sigma}$$

Here $P(\boldsymbol{S}|\alpha)$ represents the likelihood function of a derived allele frequency spectrum $\boldsymbol{S}$ for a given $\boldsymbol{\alpha}$, $f(\boldsymbol{\alpha}|\mu, \sigma)$ is a normally distributed prior, while $g(\mu)$ and $h(\sigma)$ are hyperpriors. $h(\sigma)$ is a gamma-distributed hyperprior on σ, Gamma(2, 10), and $g(\mu)$ is a normally-distributed hyperprior on μ, Normal(0, 3). This model is therefore a hierarchical Bayesian model as $f(\boldsymbol{\alpha}|\mu, \sigma)$, the prior distribution for the parameter set $\boldsymbol{\alpha}$, is governed by the hyperparameters μ and σ.

Likelihood calculations

In the Wright-Fisher model, the stationarity distribution of derived allele frequencies p can be described as a function of α_k, the selection parameter for a particular segregating site, with the following equation:[99]

$$\varphi(p|\alpha_k) = \frac{1 - e^{-\alpha_k(1-p)}}{1 - e^{-\alpha_k}} \frac{2}{p(1-p)}$$

When a finite number of alleles, n_k, are sampled from a population, we do not know the true derived allele frequency, but for a particular segregating site, a density function, F, can be defined as the product of the stationarity density and the binomial density of observing a segregating site at a particular discrete allele frequency i_k/n_k integrated over all possible derived allele frequencies p[34]:

$$F(i_k|n_k, \alpha_k) = \int_0^1 \frac{n_k!}{i_k!(n_k - i_k)!} p^{i_k}(1-p)^{n_k - i_k} \varphi(p|\alpha_k) dp$$

The probability of observing a particular derived allele frequency i_k can then be expressed as follows:

$$P(i_k|n_k, \alpha_k) = \frac{F(i_k|n_k, \alpha_k)}{\sum_{j=1}^{n_k - 1} F(j|n_k, \alpha_k)}$$

Thus, the likelihood of observing a derived allele frequency spectrum, $\boldsymbol{S}$, for a given set of selection parameters, $\boldsymbol{\alpha}$, can be represented as the product of the allele frequency probability for each segregating site.

$$P(\boldsymbol{S}|\boldsymbol{\alpha}) = \prod_k P(i_k|n_k, \alpha_k)$$

MCMC evaluation of selection parameters

We evaluated the posterior distribution $P(\boldsymbol{\Theta}|\boldsymbol{S})$ with *gonomics: selectionMcmc*, which implements the Metropolis-Hastings algorithm, a method of Markov Chain Monte Carlo (MCMC) sampling.

The Metropolis-Hastings algorithm begins with an initial set of parameters, Θ, and draws a new set of parameters, Θ', based on the current parameter set. To draw this new parameter set, a new value for σ denoted σ' is first selected as a random value from a Normal distribution, Normal(σ, *sigmaStep*), where σ is the value from the previous iteration and *sigmaStep* is a constant that may be changed for optimal parameter space exploration (we use 0.01). This makes it possible for a proposed σ' to be less than zero, which is outside the support for $h(\sigma)$ and will be evaluated to have a zero probability of acceptance. Next, a new value of μ (μ') is drawn from a normal distribution, Normal(μ, *muStep*), where *muStep* is a second tuning parameter that controls parameter space exploration, which we set to 0.5. We tuned *sigmaStep* and *muStep* to arrive at an acceptance probability near 0.5. We generated a proposal for $\boldsymbol{\alpha}$ ($\boldsymbol{\alpha}'$) by drawing each α'_k from a Normal(μ', σ').

Due to symmetry in the proposal functions, where proposing μ' and σ' when at the current values of μ and σ would be equal to proposing μ and σ when at current values of μ' and σ', we are able to reduce the acceptance probability for the candidate parameter set Θ' to:

$$P(\text{accept}) = \min\left\{1, \frac{P(\boldsymbol{S}|\boldsymbol{\alpha}')}{P(\boldsymbol{S}|\boldsymbol{\alpha})} * \frac{g(\mu')h(\sigma')}{g(\mu)h(\sigma)}\right\}$$

If a new parameter set Θ' is accepted, it serves as the initial parameter set in the following iteration. Over many iterations, the random walk of the Θ parameter set forms a Markov Chain whose stationarity distribution represents the posterior distribution for its parameters.

We implemented *gonomics: mcmcTraceStats* to calculate the mean and 95% highest density credible interval for each chain, discarding the first 5,000 iterations as burn-in for variant sets overlapping regions of evolutionary interest.

Divergence-based ascertainment corrections

HAQERs, and other sets of genomic regions in our analyses, were defined based on the level of divergence between the human reference assembly and other species. This creates a systematic bias where regions in the reference assembly with low divergence are enriched for segregating sites with a low derived allele frequency. Similarly, regions in the reference assembly with a high divergence are enriched for segregating sites with a high derived allele frequency. This is because segregating sites with low derived allele frequencies are likely to appear non-divergent when sampling a single human allele (the reference assembly) and segregating sites with high derived allele frequencies are more likely to appear divergent when sampling a single human allele. This issue has been extensively explored by Kern,[35] who describes a mathematical framework for correcting this ascertainment bias. Utilizing this framework, we introduce a corrected version of the likelihood function that is conditioned on the divergence-based ascertainment, *Asc*, of a set of variants:

$$P(\mathbf{S}|\mathbf{Asc}, \alpha) = \prod_k P(i_k|n_k, \text{Asc}_k, \alpha_k)$$

To calculate this corrected likelihood function, we use a special case of the Kern correction where only one human allele (the allele from the reference genome) has been used for ascertainment. We represent the probability that a segregating site $S_k = \{i_k, n_k\}$ is identified as divergent between two genomes as:

$$P(\text{Asc}_k|i_k, n_k, \alpha_k) = \frac{i_k}{n_k}$$

Here, n_k represents the number of individuals with a genotype call for the segregating site k, including the reference genome as an additional observation of an allele. Conversely, the probability of ascertaining a segregating site in the ancestral state of a function of its allele frequency is:

$$P(\text{Asc}_k|i_k, n_k, \alpha_k) = \frac{n_k - i_k}{n_k}$$

Using Bayes' Theorem, we can then represent the corrected allele frequency probability expression as:

$$P(i_k|n_k, \text{Asc}_k, \alpha_k) = \frac{P(i_k|n_k, \alpha_k)P(\text{Asc}_k|i_k, n_k, \alpha_k)}{P(\text{Asc}_k|n_k, \alpha_k)}$$

In this equation, the denominator represents a constant normalization factor:

$$P(Asc_k|n_k, \alpha_k) = \sum_{j=1}^{n_k - 1} P(j|\alpha_k)P(Asc_k|j, n_k, \alpha_k)$$

We applied this correction to each segregating site in region sets generated through divergence-based criteria (*i.e.*, HAQERs, HARs, and UCEs) by using the options *-divergenceAscertainment* and *-includeRef* in the program *gonomics: selectionMcmc*.

MCMC validation with synthetic datasets

In order to validate our MCMC selection model, we evaluated the ability of our model to recover known selection parameters used to generate synthetic data. To this end, we designed and implemented *gonomics: simulateVcf* to generate synthetic allele frequency spectra based on a particular selection parameter, α.

To generate an allele frequency spectrum, $\boldsymbol{S}$, we generated individual segregating sites S_k with the parameters $\{i_k, n_k\}$. To simulate segregating sites for a particular selection parameter, α, our program first generated Beta-distributed random variates $p \in (0, 1)$ from a distribution with the parameters 5000 * Beta($\alpha = 0.001$, $\beta = 0.5$). We selected these parameters so that the resulting distribution $B(p)$ could serve as a bounding function for the allele frequency stationarity distribution $\varphi(p|\alpha)$ when α is between -10 and 10. In symbolic terms:

$$[B(p) \geq \varphi(p|\alpha)] \forall p \in (0,1), \alpha \in [-10, 10]$$

With this function in hand, we could then perform bounded rejection sampling to recover random variates from the stationarity distribution $\varphi(p|\alpha)$ by accepting variates from $B(p)$ with the following probability:

$$P_{accept} = \frac{\varphi(p|\alpha)}{B(p)}$$

This provides us with a method for generating synthetic derived allele frequencies for a set of segregating sites in a large population that are evolving under the given value of α.

To test our method we need segregating sites to be represented as finite samples from this population in the form (i_k, n_k). To that end, we simulate n_k draws from a binomial distribution with a success probability of p_k. The number of successes becomes i_k. If i_k were equal to 0 or n_k (representing the cases where a site that is segregating in the population is not detected as segregating in the sample), this result was discarded and the process repeated with a new p_k.

We generated 10 independent synthetic datasets for five values of the selection parameter α (i.e. -4, -2, ~0, 2, 4) for a total of 50 synthetic derived allele frequency spectra. As the stationarity distribution is undefined at $\alpha = 0$, we used $\alpha = 0.01$ to represent near neutral selection (~0). Representative spectra are displayed in Figure S1A.

To estimate selection parameters from synthetic data, we performed MCMC sampling on each dataset for 50,000 iterations starting from near neutral initial parameters Figure S1B. The mean and 95% credible intervals from the inferred posterior distributions for the mean selection parameter are displayed in Figure S1C, calculated after discarding the first 5,000 iterations as burn-in. The true value of the selection parameter used to generate each dataset is displayed as a vertical dashed line.

We implemented the program *gonomics: simulateDivergentWindowsVcf* to verify our ability to correct for divergence-based ascertainment biases in synthetic derived allele frequency data sets, using our special case of the Kern correction.[35] For each replicate experiment, we generated 1000 sets of variants, each containing 100 simulated segregating sites generated from a stationarity distribution with a fixed selection parameter α. The number of divergent sites generated in each set was then calculated, and we returned the top 1% or bottom 1% of sets ordered by the number of divergent sites. We generated 10 replicates of upper and lower divergence variant sets for three values of the selection parameter α: strong positive selection ($\alpha = 5$), strong negative selection ($\alpha = -5$), and neutral expectation ($\alpha = 0.01$) We then used *gonomics: selectionMcmc* with and without the divergence-based ascertainment bias correction to assess its impact on our estimation of the mean selection parameter.

Genome-wide multiple alignment

We generated a genome-wide alignment to identify the fastest-evolved regions in the human, chimpanzee, and gorilla genome using the following assemblies: Human (*Homo sapiens*, hg38), Chimpanzee (*Pan troglodytes*; panTro6), Bonobo (*Pan paniscus*; panPan2), Gorilla (*Gorilla gorilla*, gorGor5), and Orangutan (*Pongo abelii*, ponAbe3).

We downloaded each reference assembly from the UCSC Genome Browser[100] and generated local pairwise alignments with LASTZ.[90] We used the human-chimp.v2 scoring matrix with parameters (O=600 E=150 T=2 M=254 K=4500 L=4500 Y=15000).[100] We then chained the local alignments together using *kentUtils: axtChain*.[89]

We took several additional steps to prevent and remove misalignments during chaining. First, to prevent the generation of chained alignments bridging assembly gaps, we chained alignments in each gapless regions of the genome independently and only considered gapless regions greater than 1 Mb of the human genome and greater than 20 kb for each of the query genomes. This filtering allowed us to ensure a large genomic context to better separate orthologs from paralogs. We also generated a custom scoring matrix (O=20 E=5):

	A	C	G	T
A	3	-11	-8	-12
C	-11	3	-11	-8
G	-8	-11	3	-11
T	-12	-8	-11	3

and gap penalty function:

tableSize	5				
smallSize	11				
Position	1	2	3	11	111
qGap	12	19	24	43	420
tGap	12	19	24	43	420
bothGap	25	40	50	90	700

for the *axtChain* program to more conservatively chain local alignments by preventing the chaining of alignments spanning large gaps in the target or query. We filtered the chains to have a minimum score of 50,000 and used *kentUtils: chainNet* to generate the final pairwise alignments for each alignable position of the human genome.

We used MultiZ[91] to generate the multi-species genome-wide alignment and converted the output into an aligned FASTA file (*gonomics: mafToFa*). Subsections of this alignment were displayed using *gonomics: multFaVisualizer*.

Divergence velocity and acceleration analysis

To analyze the velocity and acceleration of genomic regions on the human branch we reduced our genome-wide alignment to four species: Human (*Homo sapiens*, hg38), Chimpanzee (*Pan troglodytes*; panTro6), Gorilla (*Gorilla gorilla*, gorGor5), and Orangutan (*Pongo abelii*, ponAbe3) using *gonomics: faFilter* and estimated the branch lengths in 500bp windows with *gonomics: multiFaAcceleration*.

This method estimates the branch lengths for a phylogenetic tree as the set of branch lengths that minimizes the error term Q, which represents the squared difference between the pairwise distances between the sequence of two species, D, and the patristic distance separating these two species on the tree, d, while constraining branch lengths to be non-negative.[101,102] We measured pairwise distances in terms of the number of differences separating two sequences, which includes both substitutions, insertions, and deletions, where each insertion or deletion counts as one difference regardless of length. As all species needed to be present in the alignment for us to calculate the branch lengths for a given region, we implemented *gonomics: mafToBed* to generate a BED file of all such regions.

$$Q = \sum_{i \in S} \sum_{j \in S} (D_{ij} - d_{ij})^2$$

Two branch lengths from this tree are used in the subsequent calculations: b_0, which represents the distance between the human-gorilla ancestor and the human-chimpanzee ancestor, and b_1, which represents the distance between the human-chimpanzee ancestor and the extant human genome assembly. We then defined the quantity $\mathbf{v}$ as the velocity score, or the rate of divergence over the branch b_1 measured in units of mutations per site per million years of evolution. With 500 base pair windows and 7.4 million years of evolution between the human-chimpanzee ancestor and extant humans,[103] $\mathbf{v}$ can be calculated as follows:

$$\mathbf{v} = \frac{b_1}{500\text{bp} \cdot 7.4\text{My}}$$

Similarly, we define the initial velocity score $\mathbf{v}_0$, or the rate of divergence over the branch b_0 in units of differences per site per 1 million years of evolution, as follows:

$$\mathbf{v}_0 = \frac{b_0}{500\text{bp} \cdot 2.3\text{My}}$$

Finally, we define the quantity $\mathbf{a}$, the acceleration score, as the change in velocity between branches b_0 and b_1:

$$\mathbf{a} \propto \Delta\mathbf{v} = \mathbf{v} - \mathbf{v}_0$$

The genome-wide average velocity score is $9.18 \cdot 10^{-4}$ differences per site per 1 million years of evolution. Given an estimate of 7.4 million years of evolution between humans and the HCA, or a total of 14.8 million years of independent evolution separating extant humans from extant chimpanzees, our model would estimate an average sequence divergence of 1.36% in alignable regions between humans and chimpanzees, which is consistent with past estimates.[36] The genome-wide average acceleration score is 3.12×10^{-5}.

To calculate the relationships between initial velocity, velocity, and acceleration we first pseudorandomly sampled 2.9 million 500-bp genomic windows (*gonomics: bedFilter -subset*). We then partitioned these genomic windows into subsets

corresponding to particular ranges of velocity and acceleration scores (*gonomics: bedFilter -minNameFloat/maxNameFloat*, *gonomics: intervalOverlap*).

For each acceleration and velocity subset, we identified the biallelic SNPs that were segregating in these windows and calculated mean selection parameters associated with the given range of velocity or acceleration, as described in *MCMC Evaluation of Selection Parameters*. MCMC chains were run for 10,000 iteration with the first 1,000 iterations discarded as burn-in. As these regions were identified on the basis of divergence, we applied the Kern correction for divergence-based ascertainment bias for all chains.

Along with calculating initial velocity, velocity, and acceleration for 500-bp windows, we also calculate these scores for diverse sets of genomic regions where the length of genomic segments is variable (*gonomics: branchLengthsMultiFaBed*). We use a length of 50bp as a minimum to prevent large fluctuations in these scores seen in very small elements and use the more general equation with the genomic length, *l*, is a variable:

$$\mathbf{v} = \frac{b_1}{l \cdot 7.2My}$$

$$\mathbf{v}_0 = \frac{b_0}{l \cdot 2.3My}$$

Ancestral state inference

We implemented *gonomics: primateRecon* to estimate the ancestral allele states using a maximum likelihood framework[104] from our alignment of the human, chimpanzee, bonobo, gorilla, and orangutan genomes. We used this program to estimate both the human-chimpanzee ancestor and the human-gorilla ancestor.

We first estimated the neutral rate of evolution based on four-fold degenerate sites in codons using the *knownGenes* track on the UCSC Genome Browser as our gene set. We used *PHAST: msa_view* to extract four-fold degenerate codon sites and estimated branch lengths for a fixed-topology tree using a Jukes-Cantor model of evolution[105] by maximum likelihood[96] (*PHAST: phyloFit*).

A base was determined to be present in the ancestral node if a base is present in at least two species on two independent lineages connected to the ancestral node. For alignment columns where an ancestral base was determined to be present, we first reconstructed the probabilities of A, C, G, and T in the ancestral node using the tree inferred from four-fold degenerate sites.[104] We then used one of two methods to assign a single base to the ancestor from these four probabilities. These distinct methods of ancestral state inference reflect the specific experimental use cases for the resulting inferred sequences. In the first method, we bias the reconstruction towards an extant species base by mandating that the sum of probabilities for the three other bases must be greater than or equal to 0.8 for the most likely base to be assigned as the ancestral state. This method produced a conservative estimation of divergent sites between modern and ancestral species and was used in the ascertainment of HAQERs, *chimp*-AQERs, and *gorilla*-AQERs. We used our second method of ancestral state inference for annotating the ancestral allele for segregating sites among modern humans. In this method, we first implemented *gonomics: vcfToFa* to construct a FASTA format sequence of the human reference genome where the reference allele at each segregating site is replaced with the alternate allele from a VCF format file. We then appended this sequence to our multiple alignment and treated both the reference and alternate human sequence with equal weight. we then calculated the four base probabilities for the human-chimpanzee ancestor and accepted the most likely allele as the ancestral state if its probability was greater than or equal to 99%. For uncertain positions, we assigned an N to the ancestral state to ensure that only high confidence SNPs were retained for subsequent analysis of derived allele frequencies.

HAQER Identification

We identified Human Ancestor Quickly Evolved Regions (HAQERs), or regions of the human genome with an increased density of differences when compared to the human-biased estimate of the human-chimpanzee ancestor.

We calculated the number of evolutionary operations (including substitutions, insertions, and deletions) that would be needed to convert our reconstruction of the human-chimpanzee ancestor's genome into the human reference genome (hg38), for every 500 bp sliding window (*gonomics: faFindFast*). We used our reconstruction of the human-chimpanzee ancestor that conservatively uses the identity of the human base when the statistical model is uncertain (when the most likely base has a probability of less than 0.8). This results in us having high confidence in the changes on the human lineage that we do identify, which are likely to be a lower-bound on the total number of evolutionary operations that occurred in each window.

To assign statistical significance to HAQERs, we first constructed a null model by scanning the genome with a 10 Mb window to calculate the number of mutations in the fastest evolving 10 Mb section of the human genome since the split with chimpanzee. This rate of high-confidence genomic changes is 0.0126899 evolutionary operations per site, which we use as the rate of divergence μ in our null model.[45,106] With this rate of divergence for our null model, we are able to calculate uncorrected p-values associated with observing *N* changes within a 500 bp window with the R command *pbinom*($N - 1$, 500, μ, lower.tail = *FALSE, log.p = FALSE*).

When $N = 29$, our false discovery rate is $1.52096 \cdot 10^{-7}$. We merged all overlapping 500 bp windows containing at least 29 evolutionary operations separating the human-chimpanzee reconstruction and the human reference genome (*gonomics: bedFilter, bedMerge*). This resulted in our final set of 1581 HAQERs.

We performed nearly identical procedures to identify the corresponding fastest-evolved regions in the chimpanzee genome (*chimp*-AQERs) and gorilla genome (*gorilla*-AQERs) using biased ancestor estimates for each of these species. We identified 2497 *chimp*-AQERs and 2885 *gorilla*-AQERs. We report overlap enrichments between HAQERs, *chimp*-AQERs, and *gorilla*-AQERs using *gonomics:overlapEnrichments*.

To generate ideograms for the visualization of the genomic locations of HAQERs, we converted a BED file of HAQER coordinates into a text file compatible with the UCSC Genome Graphs visualization tool such that the amplitude of each region is proportional to its maximum divergence density (*gonomics: formatIdeogram*). To visualize divergence density between the reconstructed human-chimpanzee ancestor and the human reference genome, we converted the BED file listing divergences for each 500 bp window into a wiggle (WIG) format track for the UCSC Genome Browser (*gonomics: faFindFast, bedScoreToWig*). We then converted this WIG track into a binary wiggle (bigWig) format track for final visualization on the browser (*kentUtils: wigToBigWig*).

Chromosome location

We generated sets of BED files containing genomic elements that are pseudorandomly generated and uniformly distributed in the human, chimpanzee, and gorilla genomes to quantify the enrichment of HAQERs near chromosome ends (*gonomics: simulateBed*). As *chimp*-AQERs and *gorilla*-AQERs were identified on hg38 coordinates, we used *kentUtils:liftOver* to project these regions onto coordinates for *panTro6* and *gorGor6*, respectively, to measure distance from chromosome ends in the correct syntenic context for each species. We then calculated the distance to chromosome ends for both HAQERs and pseudorandom regions and compared the mean distance from the chromosome end (*t*-test) and proportion of elements within 5 megabases of the chromosome end (Chi-squared).

GREAT Ontology Analysis

We used the Genomic Regions Enrichment of Annotations Tool[87] (GREAT) to identify enriched Gene Ontology (GO) Biological Process gene sets nearby HAQERs, *chimp*-AQERs, and *gorilla*-AQERs lifted to the human reference genome *hg38*. We report significant enrichments in terms of Bonferroni-adjusted Binomial p values using the whole genome as the background region.

We also used GREAT to identify GO Biological Processes enriched near 3D chromatin contact sites of HAQERs. To this end, we accessed chromatin contact sites identified in H1 hESC and GM12878 cell lines by HiCAR.[81] We then identified the set of all genomic regions forming distant 3D chromatin contacts with HAQERs in each cell type *gonomics:intervalContacts*. We report significant enrichments in terms of Bonferroni-adjusted hypergeometric p-values using the set of all chromatin contact sites for each cell line as background regions.

Mutation rate and fixation estimation

We first generated a list of all divergent positions between hg38 and the inferred human-chimpanzee ancestor (*gonomics: multiFaToVcf*). We then generated a set of all divergent sites that overlap specified genomic regions, including HAQERs, HARs, RAND, ENCODE, and UCE *gonomics: intervalOverlap*. Next, we calculated the divergent sites per base as the number of divergent sites divided by the total length in base pairs of the input set of genomic regions. Similarly, polymorphic sites per base were calculated as the number of variants from the 501 African individual subset of the 1000 Genomes Project data (see Human genetic variation preprocessing) that overlapped each set of genomic regions divided by the length in base pairs of that set.

We also intersected the set of divergent positions with the set of all polymorphic sites identified in the 501 African individual subset of the 1000 Genomes Project data (*gonomics: intervalOverlap*). Positions found in both sets were labeled as polymorphic divergent sites, and divergent sites not found in the 1000 Genomes Data were labeled as fixed divergent sites. We then determined the sets of fixed and polymorphic divergent sites overlapping each set of genomic regions (*gonomics: intervalOverlap*). We plotted the proportion of divergent sites that are polymorphic as the number of polymorphic sites divided by the sum of polymorphic and fixed divergent sites and assigned significance via the Chi-squared test of independence against the observed ratio of polymorphic divergent sites in RAND. The 2x2 contingency table for this analysis for a set of regions, X, had the dimensions {X, RAND} and {Fixed, Polymorphic}.

Recombination and replication timing

To measure recombination frequencies for regions of interest, we intersected a genome-wide recombination map estimated from all Yoruba individuals in the 1000 Genomes Project[51] with HAQERs, HARs, and RAND. We also accessed a BED format file of meiotic double stranded break hotspots[47] (GEO: GSE59836) and intersected this dataset with HAQERs. Overlap enrichments were quantified with *gonomics:overlapEnrichments* and intersecting HAQERs were identified with *gonomics:intervalOverlap*. We also accessed a dataset of replication timing in 300 iPSC lines[49] and generated a dataset representing the average replication timing for each genomic region across these 300 iPSC lines. This dataset was then lifted to hg38 coordinates (*kentUtils: liftOver*) and intersected with HAQERs, HARs, and RAND *gonomics: intervalOverlap*.

Mutation spectrum analysis

For each genomic region of interest, we gathered the set of all divergent positions between hg38 and the inferred human-chimpanzee ancestor and partitioned this set into six classes of mutations ($A \rightarrow G/T \rightarrow C$) ($G \rightarrow A/C \rightarrow T$) ($A \rightarrow T/T \rightarrow A$) ($G \rightarrow C/C \rightarrow G$) ($A \rightarrow C/T \rightarrow G$) ($C \rightarrow A/G \rightarrow T$) (*gonomics: divergenceSpectrum*). We then calculated the proportion of HCA divergent sites that are weak to strong mutations for each genomic element ($A \rightarrow G/T \rightarrow C$) or ($A \rightarrow C/T \rightarrow G$). We constructed a matrix of six values for each genomic region, with each value relating to the proportion of overlapping HCA divergent sites in each mutation class for principal component analysis in the R programming language.

Back mutation analysis

We hypothesized that while most low frequency derived alleles (DAF<0.1) will represent nearly exclusively forward mutations (in which the ancestor allele mutates to a derived variant), high frequency derived alleles (DAF>0.9) will represent a mixture of forward mutations and back mutations (in which a since diverged derived allele mutates back to the ancestral state). Forward mutations occur with unequal probabilities of transitions and transversions. We estimated the proportion of transitions in forward mutations t_f to be equal to the proportion of transitions across all segregating sites in RAND ($t_f = t_{seg} = 0.685$; *gonomics: vcfInfo*). Back mutations will also occur with unequal probabilities of transitions and transversions. For the ancestral allele to be phased at a segregating site, a back mutation must revert to the ancestral allele state. We define t_{div} to be equal to the proportion of transitions across all divergent sites in RAND ($t_{div} = 0.668$). This model allows t_{seg} to differ from t_{div}, as would be the case if transition/transversion biases change over evolutionary time. However, in our analysis of the human lineage, t_f and t_{div} are similar. There are two scenarios in which a back mutation can occur. First, the inverse transition of a divergent transition will occur at a rate proportional to $t_f \cdot t_{div}$. Second, the inverse transversion of a divergent transversion will occur at a rate proportional to $(1 - t_f)(1 - t_{div})/2$, as there are two possible reverse transversions for a divergent site.

Thus, the expected proportion of transitions in back mutations will be equal to:

$$t_b = \frac{t_f \cdot t_{div}}{\frac{(1-t_f)(1-t_{div})}{2} + (t_f \cdot t_{div})}$$

Based on our estimates of t_f and t_{div}, we estimate $t_b = 0.897$. In other words, segregating sites that are back mutations should exhibit a quantifiable elevation in the proportion of transitions.

We use the following mixture model to estimate the relative proportion of forward and back mutations in a set of segregating sites with a proportion of transitions x:

$$x = t_f f + t_b(1 - f)$$

Here f represents the proportion of forward mutations and $(1 - f)$ represents the proportion of back mutations. We measured that segregating sites in HAQERs with $DAF > 0.9$ exhibit a proportion of transitions $x \approx 0.75$. This figure implies that approximately 30% of segregating sites in HAQERs at $DAF > 0.9$ are back mutations.

Great ape genome divergence analysis

We constructed a 30-way whole-genome multiple alignment[91] to analyze patterns of divergence and constraint in HAQERs. This alignment included five reference genomes: *hg38*, *panTro6*, *panPan2*, *ponAbe3*, and *gorGor5*. In addition to these reference genomes, we generated reference-based haploid assemblies from individuals within a species to survey intraspecific variability. To this end, we aligned the short-read sequencing data from individuals to the corresponding reference assembly and calculated the consensus allele for each position (*gonomics: samConsensus*).

We generated consensus sequences for three high coverage sequencing data sets from archaic hominins: a 30x coverage Denisovan genome from Denisova Cave in the Altai Mountains,[80] a 52x coverage Neanderthal genome also from the Denisova Cave in the Altai Mountains,[107] and a 30x coverage Neanderthal genome from the Vindija Cave in Croatia.[84] In addition to these archaic genomes, we also included the consensus sequences from 10 diverse, unrelated human individuals accessed from the 1000 Genomes Project (*HG00096*, *HG01112*, *HG03052*, *NA18525*, *NA20502*, *HG00419*, *HG01879*, *HG01500*, *HG03742*, *NA18939*).[108] Additionally, we included sequencing data from the following 12 chimpanzee individuals,[82] comprised of three individuals from each chimpanzee subspecies: *Pan troglodytes verus* (*SRX243499*, *SRX243488*, *SRX243446*), *Pan troglodytes schweinfurthii* (*SRX237583*, *SRX237539*, *SRX237526*), *Pan troglodytes ellioti* (*SRX243519*, *SRX243518*, *SRX24351*), and *Pan troglodytes troglodytes* (*SRX243489*, *SRX243492*, *SRX243496*).

We used the Dunn Index[54] to quantify interspecies divergence in the context of intraspecies variability. We calculated the Dunn Index for each region in a set of regions as the ratio of the minimum intercluster sequence distance to the maximum intracluster distance (*gonomics: dunnIndex*). We restricted our Dunn Index analysis to regions with at least 5 segregating sites and where every individual had aligned sequence to the region.

Chromatin state enrichment analysis

To analyze chromatin state enrichments and depletions we used the ChromHMM classification of 127 epigenomes, which was produced as part of the Roadmap Epigenomics consortium.[37] We calculated the overlap enrichment and depletion between two sets of

genomic elements (set 1 and set 2) in using our previously described statistical framework[109] (*gonomics: overlapEnrichments*). We define the search space for this method as the area in the genome in which elements of set 1 and set 2 can be found, which includes all ungapped genomic regions greater than 1mb in length. If an individual genomic element from set 2 of length *L* were randomly distributed in the search space, the probability that it overlaps an element in set 1 can be expressed as the number of positions an element of size *L* can be placed in the search space that overlap an element of set 1 divided by the total number of positions in the search space in which an element of length *L* can be placed. The probability of observing k overlaps out of n trials, where n is equal to the number of elements in set 2, thus follows the Poisson binomial distribution. When the number of trials is large, the Poisson binomial distribution can be approximated with a normal distribution with the following mean, μ, and variance, σ^2:

$$\mu = \sum_{i=1}^{n} p_i$$

$$\sigma^2 = \sum_{i=1}^{n} (1 - p_i) p_i$$

We report the enrichment between two sets of elements as the ratio between the observed number of overlaps and μ, the expected value of overlaps. We calculate Bonferroni-adjusted p values for enrichment and depletion with the following formulas:

$$p_{enrichment} = \min\left(1, 2C \cdot \sum_{i=k}^{n} \text{Normal}(\mu, \sigma)\right)$$

$$p_{depletion} = \min\left(1, 2C \cdot \sum_{i=0}^{k} \text{Normal}(\mu, \sigma)\right)$$

We used 2*C*, where *C* is the number of comparisons, as the Bonferroni adjustment, as we tested for both enrichment and depletion for each pair of genomic elements. Significance was assigned for enrichment and depletion at $p < 0.05$.

To investigate the relationship between HAQER bivalent chromatin enrichments and environmental response, we accessed 17 ChromHMM datasets from untreated human A549 cells or A549 cells at various timepoints following dexamethasone (dex) treatment. These datasets were accessed from the ENCODE consortium website at the following accession numbers: ENCFF107YWL, ENCFF662GGJ, ENCFF161LGJ, ENCFF524GBP, ENCFF877NZN, ENCFF246IPY, ENCFF146UIL, ENCFF113TCU, ENCFF324PWA, ENCFF255QUQ, ENCFF052NXZ, ENCFF646AJN, ENCFF108TED, ENCFF910RII, ENCFF845TIM, ENCFF513UFQ, ENCFF418WHV. From here, we classified 410 dex-responsive bivalent enhancers as genomic regions that were in the *EnhBiv* state in untreated cells and in the *EnhA1* or *EnhA2* state in any post-treatment dataset. 2 HAQERs, HAQER0547 and HAQER0919, overlapped a dex-responsive bivalent enhancer (expected overlap: 0.32. $p < 0.01$, *gonomics:overlapEnrichments*).

Functional annotation of HAQERs

To identify enrichments between HAQERs and gene regulatory elements gained after the rhesus split, we accessed 15-state chromHmm data from the Roadmap Epigenomics consortium[37] for the active enhancer and active promoter states (*7 Enh* and *1 TssA*) from the developing brain reference epigenomes E081 and E082. Next, we concatenated and merged BED format files with *gonomics: bedMerge* to produce BED files representing all regions identified as either active enhancers or active promoters in either fetal brain reference epigenome. We then used *gonomics: intervalOverlap* to identify promoter and enhancer regions that overlapped gene regulatory elements gained after the rhesus split.[38] We used *gonomics: overlapEnrichments* to calculate the enrichment between HAQERs and these recently-evolved regulatory elements.

To identify overlap between HAQERs and open chromatin, human fetal brain DHS-seq data was obtained from the Roadmap Epigenomics Consortium data[37] from the following three individuals: GSM595920, GSM595922, and GSM595926. BAM format alignment files aligned to hg19 were disassembled to FASTQ format sequencing files using *samtools:bam2fq*[110] and aligned to hg38 with BWA MEM.[85] To visualize DNase hypersensitivity sequencing (DHS-seq) data on the UCSC genome browser, we developed *gonomics: samToWig* to convert SAM/BAM format alignment files into WIG graphing track format. WIG files were then converted to binary bigWig files with *kentUtils: wigToBigWig*. We developed *gonomics: bedValueWig* to generate a score for each region in an input BED format file corresponding to the highest value of an input WIG file in the coordinate range of the queried region. Regions with at least 10 reads overlapping a single position were considered for further analysis as possible regions of open chromatin. Overlaps between HAQERs and other genomic regions, including functional elements gained after the rhesus split[38] and differential Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) peaks from human and chimpanzee cerebral organoids[39] were determined using *gonomics: intervalOverlap*.

Gene synthesis and plasmid preparation

Test sequences for single-cell Self-transcribing active regulatory region sequencing (scSTARR-seq) were synthesized and cloned into the STARR-seq screening vector[111] using a commercial service (Twist Bioscience). We added an 8 base pair unique barcode

to the 3′ end of each insert to differentiate between closely related orthologous sequences with short read 3′ RNA sequencing. Plasmids were transformed into One Shot Stbl3 chemically competent E. coli (Thermo Fisher), selected for ampicillin resistance, and amplified in Luria broth (Invitrogen) with 100 μg/mL ampicillin. Endotoxin-free plasmids were then purified with the ZymoPURE II Plasmid Maxiprep kit per manufacturer's instructions (Zymo Research). As necessary, purified plasmids were then precipitated with 3M Na-Acetate pH 5.2, and 100% ethanol for 2 hours to achieve desirable concentrations. For in vivo STARR-seq, an equimolar solution containing each STARR-seq plasmid was prepared at a total plasmid concentration of 3 μg/mL. This pooled STARR-seq solution was then mixed with a pCAG-GFP injection reporter plasmid, which represented 1/6 of the total plasmid content of the final injection solution. Our input STARR-seq library included plasmids with a total of 77 distinct inserts: 60 corresponding to the orthologs of 13 HAQER sequences, 7 sequences used only in the analysis in Figures S5E and S5F, and 10 pseudorandom sequences which served as negative controls (STAR Methods). For the PGK-EGFP enhancer reporter assay, we amplified HAQER inserts from the STARR-seq plasmid vector via polymerase chain reaction and introduced these inserts to a PGK-EGFP plasmid vector (Addgene #169744) via Gibson assembly cloning,[112] which we confirmed with Sanger sequencing.

In utero electroporation

In utero electroporation was performed as previously reported.[113] Briefly, E14.5 or E15.5 wild type B6 pregnant females were anesthetized with isoflurane. Uterine horns were exposed by making an incision in the abdomen. Each embryo was injected with 1-1.5 μl of plasmid solution (containing 0.01% fast green and 1-2 μg/μl of plasmids) and electroporated using the following parameters: five 50 ms-pulses at 50V (E14.5) or 60V (E15.5) with 950 ms pulse-interval, using platinum-plated BTX Tweezertrodes. Uterine horns were then repositioned into the abdominal cavity and the muscle and skin incisions were sutured. Dams were then placed on a heating pad for recovery and monitored.

Immunofluorescence staining and image acquisition

Brains were fixed overnight in 4% PFA-PBS at 4°C, rinsed in PBS, and submerged in 30% sucrose-PBS until sinking (24 hours). Brains were frozen in NEG-50 medium (Richard-Allan Scientific) and cryostat sections (20 μm) were prepared and stored at -80°C until use. Sections were washed 3 times 10 minutes with PBS and incubated 1 μg/ml Hoechst 33342 (Invitrogen) for 30 minutes at room temperature. Sections were then mounted using Vectashield (Vector Laboratories) as mounting media. Images were acquired with a Zeiss Axio Observer Z.1 microscope coupled with an apotome2. Image measurements and quantifications were blindly performed using Fiji.[88] Statistical significance was assigned by 2-way ANOVA in GraphPad Prism. We analyzed anatomically comparable regions from sections from 2 embryos from 2 IUEs (n=4) per injection construct.

Fluorescence activated cell sorting

Electroporated brains were harvested after approximately 18 hours and dissected in ice-cold sterile PBS. Meninges were removed and GFP+ portions of the cortices were incubated at 37°C for 10 minutes in 0.25% trypsin-EDTA supplemented with 0.1% DNAse I (New England Biolabs cat# M0303S). Following incubation, the trypsin solution was removed and replaced with ice-cold 10% FBS/HBSS/Propidium iodide (Invitrogen) supplemented with 0.01% DNAse I. A single cell suspension was then generated by trituration with a fire-polished glass pipette and filtered with a 30 μm cell strainer. Cells were then stained with the LIVE/DEAD Near-IR Dead Cell Stain per manufacturer's instructions (Thermo Fisher). Following staining, viable GFP+ cells were bulk sorted using a FACS Aria II cytometer (BD Biosciences).

scSTARR-seq reporter read targeted enrichment

In order to enrich reporter read sequences from cDNA generated from endogenous mouse mRNA and STARR-seq reporter RNA, we performed a three-step PCR reaction based on a 10x targeted enrichment protocol developed by Gasperini et al.[114] We began with approximately 10-13 ng of unfragmented scRNA-seq cDNA and performed qPCR-monitored 50 μl Phusion PCR (annealing temp 62°C, 1.5μl DMSO) with the following primers:

F1: tGFPOuter 5- ATGGCTAGCAAAGGAGAAGAACTCT -3

R1: R1-PCR1 5- ACACTCTTTCCCTACACGACG -3

Following 1x Agencourt AMPure XP bead cleanup (Beckman Coulter), 2 μl of cleaned product was amplified in a subsequent 50μl Phusion reaction (12 cycles) with the following primers:

F2: tGFPInner 5-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTTGTTGAATTAGATTGATCT -3

R2: RP5 5-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACG -3

Following 1x AMPure cleanup, 2μl of cleaned product was used in a third 50 μl Phusion reaction (12 cycles), with the following primers:

F3: 5- CAAGCAGAAGACGGCATACGAGATIIIIIIIIIIGTCTCGTGGGCTCGG -3 (standard NEXTERA P7 indexing primer)

R3: Same as R2.

Following this reaction, final libraries were cleaned once more with 1X AMPure and quantified using the Bioanalyzer.

scSTARR-seq sequencing and preprocessing

Up to 10,000 GFP+ cells were captured per lane of a 10X Chromium device and single cell libraries were prepared using protocols from the Chromium Next GEM Single Cell 3′ Reagent Kits v3.1 (Rev D) User Guide (10X Genomics, Inc.). Final libraries were quantified

using the Bioanalyzer (Agilent) according to manufacturer's protocols. Prior to enzymatic fragmentation, an aliquot of cDNA was separated for targeted enrichment. Final 10X libraries were sequenced using the NovaSeq 6000 S-Prime reagents (Illumina; R1 28, I1 10, I2 10, R2 90). Reporter-targeted enrichment libraries were sequenced independently on the Illumina NextSeq platform. Fastq files from both libraries were recovered with *bcl2fastq* (v2.20.0.422, Illumina). For targeted enrichment libraries, we used the following bases mask: Y28n*,n*,I8n*,Y75n*. Output fastq files from both the unenriched and reporter-targeted enrichment libraries across multiple lanes were concatenated together for downstream analysis.

For input normalization, we used the NEBNext Ultra II FS DNA Library Prep Kit (New England Biolabs) to sequence our STARR-seq injection plasmid library and sequenced the resulting library on the Illumina iSeq 100 platform. Fastq reads from the input library were aligned using the Burrows-Wheeler Aligner (BWA) to a custom STARR-seq reference genome including the sequence of the mouse reference mm10 with additional FASTA records containing the sequences of each STARR-seq reporter construct and the pCAG-GFP injection reporter sequence. For an input STARR-seq library with n constructs, the input normalization factor C_s for an individual STARR-seq construct, s, was then calculated as the ratio of the expected number of reads from s if all n constructs were present at equimolar concentration in the input library to the observed number of reads from s: $C_s = E_s/O_s$ where:

$$E_s = \left(\sum_{i=1}^{n} O_i\right) \Big/ n$$

Constructs with an input normalization factor greater than 5 (indicating a greater than 5-fold depletion that equimolar expectation in the input library) were excluded from all subsequent analysis.

GFP+ cells were pooled from all embryos in each experiment to control for batch effects associated with anatomical differences in electroporation and dissection. We sequenced 3494 single cells from the first library, which was comprised of GFP+ cells pooled from 6 embryos from a singlemouse injected at E14.5. The second library, which was composed of GFP+ cells from 9 embryos injected at E15.5, yielded 3676 single-cell transcriptomes.

Enhancer activity quantification

To score enhancer activity from scSTARR-seq data, we implemented *gonomics: fastqFilter -collapseUmi* to remove unique molecular identifier (UMI) duplicates from our 10x libraries. We then used *gonomics: fastqFormat -singleCell* to parse the cell barcode and UMI from R1 into the read name for the R2 fastq. We then used the BWA to align reads to our custom STARR-seq reference genome described above. The enhancer activity score for each construct was then calculated as the input-normalized UMI count per 1000 total reporter UMI counts. To determine the basal level of transcription from the STARR-seq plasmid vector, we synthesized 10 plasmids with inserts of 500bp of pseudorandom DNA sequences generated using *gonomics: randSeq*. To guard against spurious enhancer activity present in pseudorandom sequences, we used six of the ten negative control constructs with the lowest enhancer activity scores to determine a limit of detection for enhancer activity, which we defined as the three standard deviations above the average enhancer activity score from these six pseudorandom sequences. We attempted to generate STARR-seq orthologous (human, Neanderthal, Denisova, chimpanzee, HCA) test sequences for each region of interest. However, some ortholog pairs exhibited the same sequence for the 500bp region of interest. Duplicate constructs were not included in statistical analysis, but are still displayed as faded bars in Figure 4C.

Single-cell cluster identification and cell-type specific enhancer activity quantification

Count matrices were produced using CellRanger v6.0 (10x Genomics) with the custom reference genome described above. Subsequent analysis for cluster identification was performed in Seurat v4.0.[97] For each library, cells were removed which contained 200 or fewer genes or more than 5,000 genes. Each library was independently normalized and 2,000 highly variable features were identified for each library. Cells across independent libraries were integrated for joint analysis via canonical correlation analysis.[115] Variation in gene expression based on cell-cycle related genes was regressed from cluster analysis in dataset scaling using an annotated set of G2M and S phase related genes provided in Seurat. k-nearest neighbors (k=20) were calculated in the space of significant principal components (in this case, 30 principal components) and clustering was performed with the Louvain-Jaccard method. Visualizations were generated in uniform manifold approximation and projection (UMAP) space.[116] We identified the top 10 positive markers for each cluster and manually assigned cluster identities based on marker gene expression in two previously published neurodevelopmental single-cell atlases in mouse[62] and human.[60] Multiple clusters corresponding to the same cell type (ex. Excitatory Neuron I-IV) were pooled as metaclusters for subsequent analysis.

To perform cell-type specific enhancer activity quantification, reads from each library aligned to the custom STARR-seq reference genome described above were sorted by cell barcode using *gonomics: mergeSort -singleCellBx* and input-normalized count matrices were generated with *gonomics: scCount*. Input-normalized count matrices were then partitioned by metacluster using cluster identities determined for each cell barcode by Seurat. Cells with fewer than 4 pCAG-GFP UMIs were discarded. The input-normalized reporter UMI counts for each cell were then further normalized to the pCAG-GFP UMI count for that cell. The cell type enhancer activity score was then calculated as the average transfection-normalized, input-normalized UMI count per cell in each cluster.

Enhancer paralog phylogenetic analysis

We began with the hg38 human reference sequence for HAQER0059 and gathered the sequences for all paralogs in the human (hg38), chimpanzee (panTro6), gorilla (gorGor5), orangutan (ponAbe3), and rhesus (rheMac10) assemblies as identified with BLAT.[68] From here, we used *gonomics: faFormat -revComp* for all reverse strand sequences before aligning all forward-strand paralogous sequences with *muscle*.[92] We then constructed a Newick-format phylogenetic tree from this alignment with ClustalW2.[86] Finally, we visualized phylogenies with phylotree.[93]

GWAS catalog trait enrichment analysis

We first sampled the set of all GWAS Catalog variants that report an association in European populations to obtain a record for each SNP.[117] We retained only those variants that also appeared as segregating among individuals in the GBR subpopulation in the 1000 Genomes Project variant set.[26] To generate a comprehensive list of possible causal variants, we used plink --r2[94] to identify all other 1000 Genomes Project variants in linkage disequilibrium (Plink $R^2 > 0.7$) with each GWAS Catalog variant.

We then merged the set of all GWAS variants and linked variation for each mapped trait from the Experimental Factor Ontology from the GWAS Catalog association table and calculated overlap enrichment between this merged set of variants and HAQERs *gonomics: overlapEnrichments*. We report significant enrichments for mapped traits with an FDR-adjusted $p < 0.05$.

We also calculated the distributions of the number of all possible causal variants (including a GWAS variant and all linked variation (Plink $R^2 > 0.7$)) and median distance of each linked variant to its corresponding GWAS variant for all possible causal variants overlapping HAQERs, HARs, or RAND.

The observed disease enrichments are unlikely to be influenced solely by haplotype structure or density of linked variation around GWAS variants, as these features were similar between RAND and HAQERs (Figures S6F and S6G).

Horizontal pleiotropy score quantification

We accessed a dataset of 1,183,386 human genetic variants annotated with LD-corrected horizontal number of traits pleiotropy scores (P_n^{LD} generated by Jordan et al.[118]). We intersected these variant sets with HAQERs, RAND, and HARs and compared the distribution of P^{LD} scores in variants overlapping each set of genomic regions.

Briefly, Jordan et al. leveraged PheWAS relationships between genetic variants and human traits to calculate P_n^{LD} as the expected value of the number of statistically independent traits for which a given variant is associated in a set of 100 traits. This approach starts with a matrix Z^{raw} of Z-scores associating each genetic variant to a human trait. Many clinical traits exhibit covariance as a result of either partially redundant or ambiguous terminology (i.e. Alzheimer's Disease and Dementia) or vertical pleiotropy (i.e. a causal relationship between traits, such as between hypertension and heart disease). Thus, P_n^{LD} corrects for covariance between traits by applying the following Mahalanobis whitening transformation to Z^{raw}:

$$Z = \Sigma^{-\frac{1}{2}} Z^{raw}$$

where Σ is the covariance matrix of Z^{raw}. The result of this transformation is that the covariance matrix of the resulting matrix Z will be equal to the identity matrix, indicating no covariance between traits. P_n for a variant n is then calculated as the scaled number of whitened traits significantly associated with the variant n:

$$P_n = \frac{100}{l} \sum_{i=1}^{1} H(z_i - 2)$$

Where $H(z_i\text{-}2)$ is the Heaviside step function, which is equal to 1 when $|z_i| > 2$ and 0 otherwise. The term $100/l$ scales the number of significantly associated whitened traits by the number of traits l and the constant 100 so that the resulting term represents the expected value of the number of significantly associated whitened traits in a dataset of 100 traits. Finally, this value is corrected for linkage disequilibrium with the following transformation:

$$P_n^{LD} = P_n - \beta_n x$$

where x is the LD score of the variant position and β_n is the regression coefficient for LD on P_n.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical parameters were reported either in individual figures or corresponding figure legends. Statistical details of experiments can be found in method details. All statistical analyses were performed in R or Go.

ADDITIONAL RESOURCES

The raw data and analyzed results are available at our website: https://vertgenlab.org/.

CellPress
OPEN ACCESS

Supplemental figures

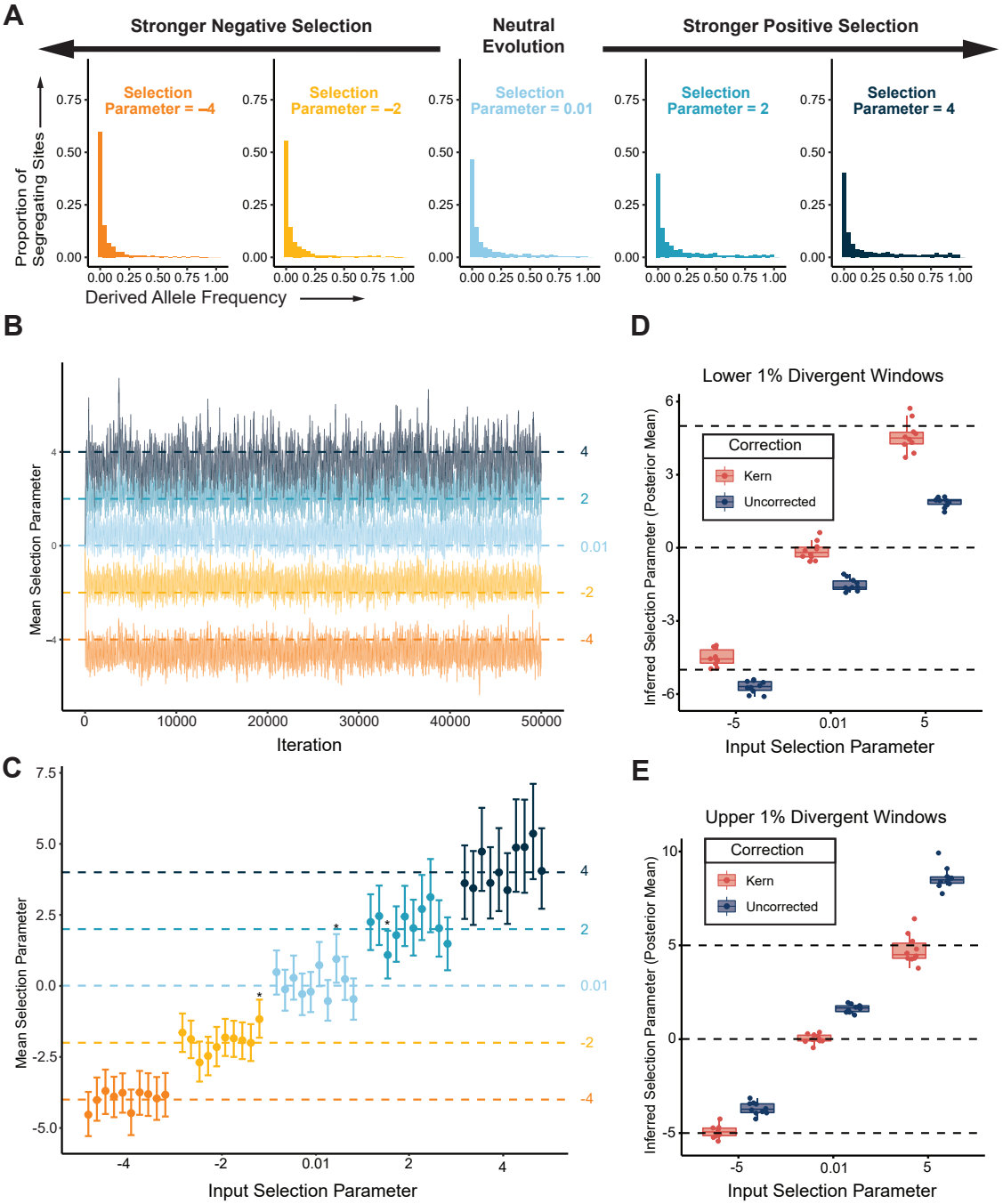


Figure S1. A model to detect the strength and direction of selection from derived allele frequency spectra, validated with synthetic allele frequency data, related to Figure 1

(A) Synthetic allele frequency spectra ordered by value of the selection parameter used to generate each spectrum.

(B) Representative Markov chain Monte Carlo (MCMC) traces for the mean selection parameter acting on variants from synthetic derived allele frequency spectra. Traces are colored by the value of the input selection parameter used to simulate each spectra.

(legend continued on next page)

(C–E) Mean and 95% highest density credible intervals for the posterior distribution of the mean selection parameter for 50 synthetic variant sets, 10 for each input selection parameter (C). Posterior estimates are displayed from 50,000 iterations with the first 5,000 iterations discarded as burn-in. 47/50 (94%) of traces contained the true input selection parameter in the 95% credible interval. Traces where the input selection parameter was not contained in the credible interval are marked with an asterisk. Distribution of MCMC trace means for the mean selection parameter with and without the Kern correction for divergence-based ascertainment bias are displayed for the least divergent 1% (D) or most divergent 1% (E) of variant sets. Mean selection parameters are estimated as the posterior mean of 10,000 iterations with the first 1,000 iterations discarded as burn-in.

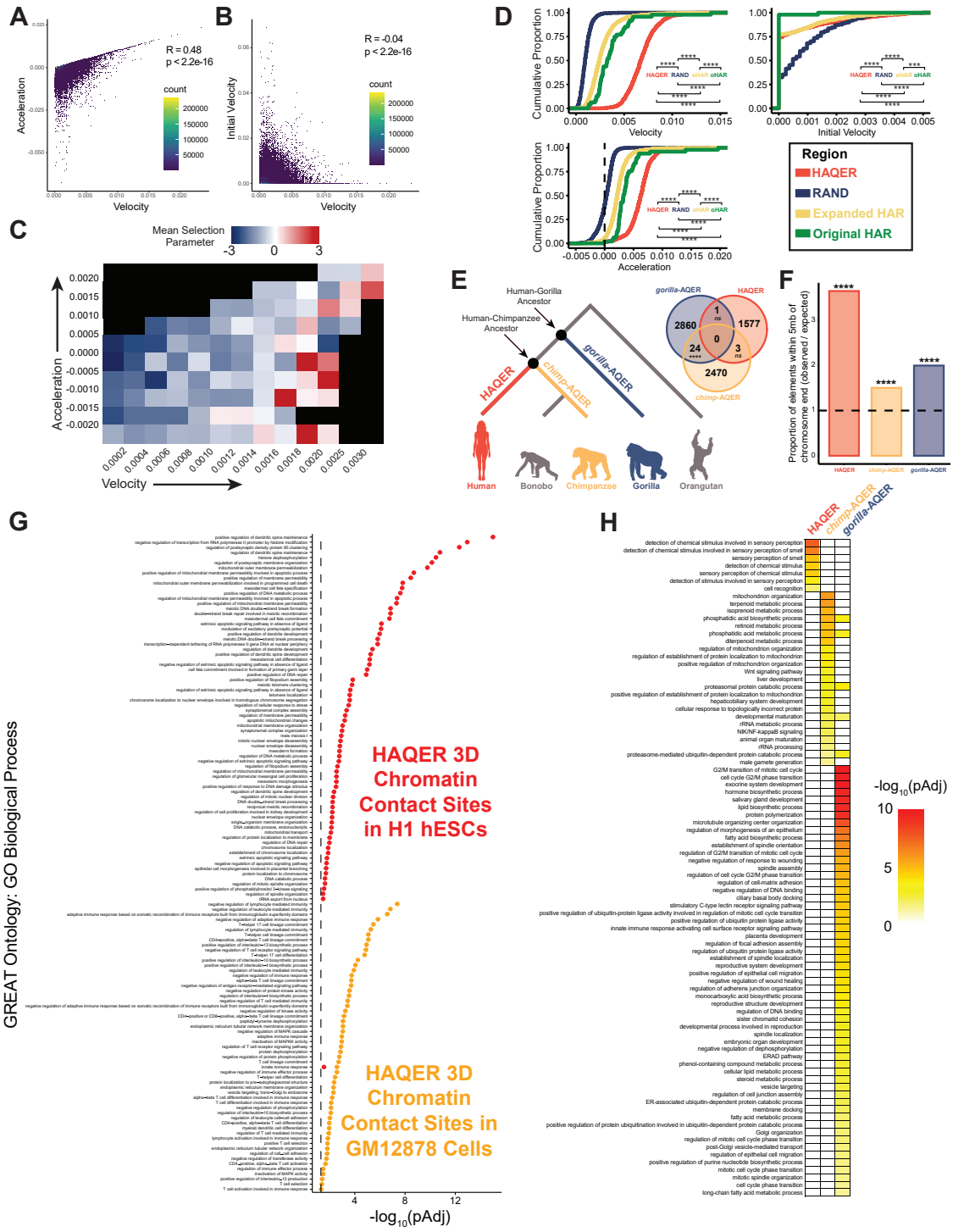


Figure S2. Relationship between phylogenetic scores and selection; gene ontology of rapidly evolved regions across species, related to Figure 1

(A and B) Pearson correlations between velocity and acceleration (A) or initial velocity (B) for 2,902,532 500-bp genomic regions.

(C) Heatmap of mean selection parameters for sets of variants overlapping all genomic regions binned by velocity and acceleration score. Mean selection parameters are estimated as the posterior mean of 10,000 iterations with the first 1,000 iterations discarded as burn-in.

(D) Cumulative proportions of velocity, initial velocity, and acceleration for variants overlapping regions of interest. Here, HARs are split into the original HARs (oHARs[33]] and the expanded HARs (eHARs[23]) (Bonferroni-adjusted Wilcoxon; *** $p < 0.001$; **** $p < 0.0001$).

(legend continued on next page)

(E) Phylogenetic context of HAQERs, *chimp*-AQERs, and *gorilla*-AQERs. Venn diagram displays overlaps between rapidly evolved primate regions (**** $p < 0.001$ for overlap enrichment between these sets of genomic regions).
(F) Observed over expected proportion of elements within 5 mb of chromosome ends. Expected proportion was estimated from randomly distributed regions in each genome.
(G) GREAT ontology enrichments for HAQER chromatin contact sites in H1 hESCs (red) and GM12878 (yellow). One ontology term, *innate immune response*, was significant in both cell types.
(H) GREAT ontology enrichments for HAQERs, *chimp*-AQERs, and *gorilla*-AQERs.

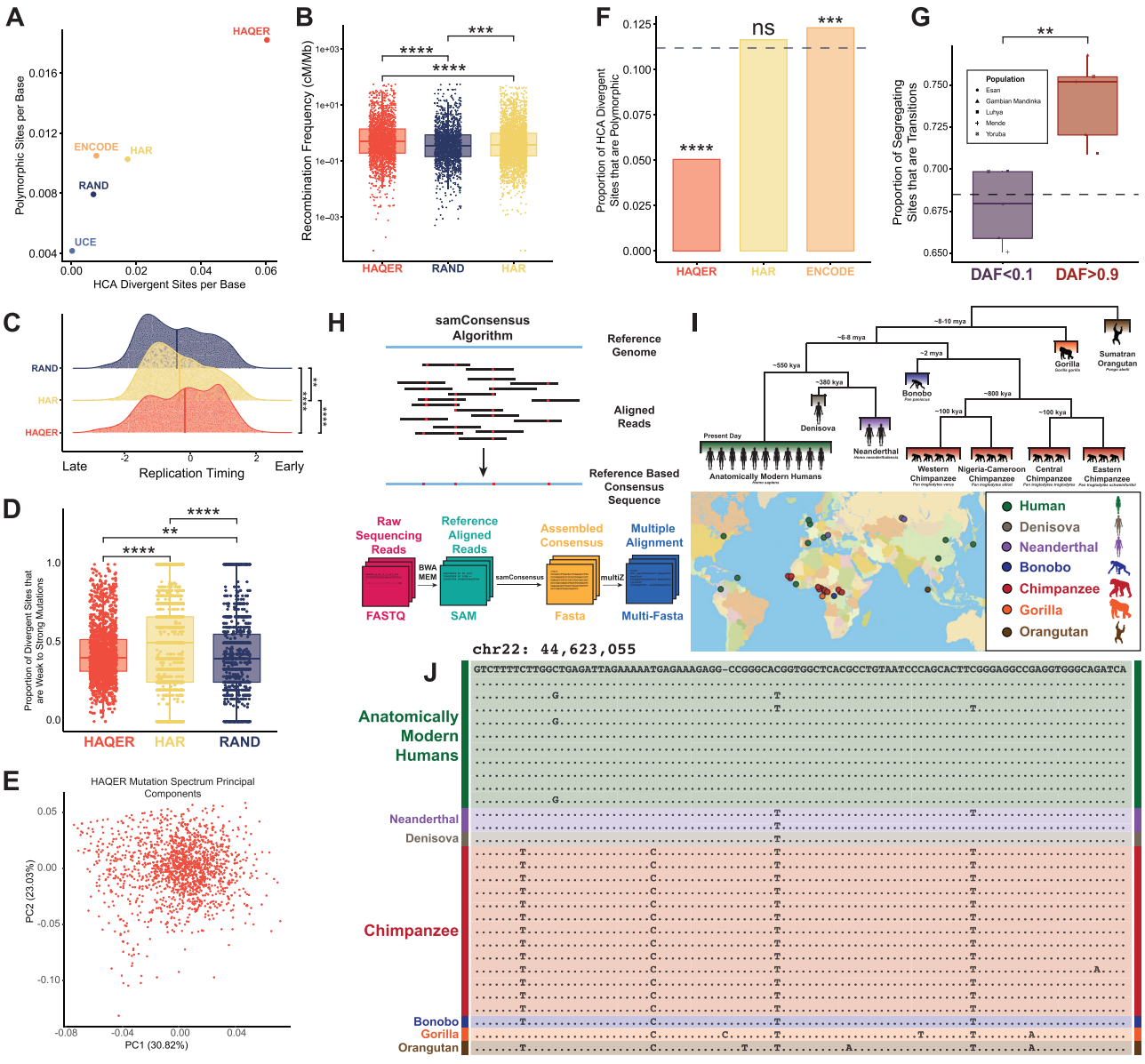


Figure S3. Mutation rate, fixation, and spectra in HAQER evolution; a 30-way alignment for population structure analysis in HAQERs, related to Figure 2

(A) Density of sites divergent between the modern human sequence and the inferred human-chimpanzee ancestor (HCA) sequence against the density of polymorphic sites observed in 501 unrelated African individuals for human ancestor quickly evolved regions (HAQERs), pseudo-randomly selected neutral proxy regions (RAND), human accelerated regions (HARs), ENCODE candidate *cis*-regulatory elements (ENCODE), and ultraconserved elements (UCEs).

(B) Distribution of recombination frequency in HAQER, RAND, and HARs (Bonferroni-adjusted Wilcoxon).

(C) Distribution of replication timing in HAQERs, HARs, and RAND (Bonferroni-adjusted Wilcoxon).

(D) Proportion of HCA divergent sites that are weak to strong mutations (A to G/T to C or A to C/T to G) in HAQERs, HARs, and RAND.

(E) Principal component analysis of the mutation spectrum for all HAQERs. HAQERs do not demonstrate distinct mutation spectrum subtypes.

(F) Proportion of HCA divergent sites that are polymorphic (observed as segregating in the 501 unrelated African individuals from A) as opposed to fixed (not observed as segregating). The dotted line represents the proportion of polymorphic variants in RAND (1,069 polymorphic sites; 8,487 fixed sites) (chi-square test).

(G) Proportion of segregating sites that are transitions in segregating sites from five African populations binned by low derived allele frequency (DAF < 0.1) and high derived allele frequency (DAF > 0.9) (Wilcoxon).

(H) The *samConsensus* program generates consensus sequences from individuals calling substitutions (represented as red positions) from sequencing reads aligned to a reference genome and editing the corresponding positions of the reference assembly.

(I) Phylogenetic history and geographic distribution of 30 great ape genomes used to construct the multiple alignment.

(J) A representative 100-bp block of sequence sampled from the 30-way whole-genome alignment (** $p < 0.01$; *** $p < 0.001$; and **** $p < 0.0001$).

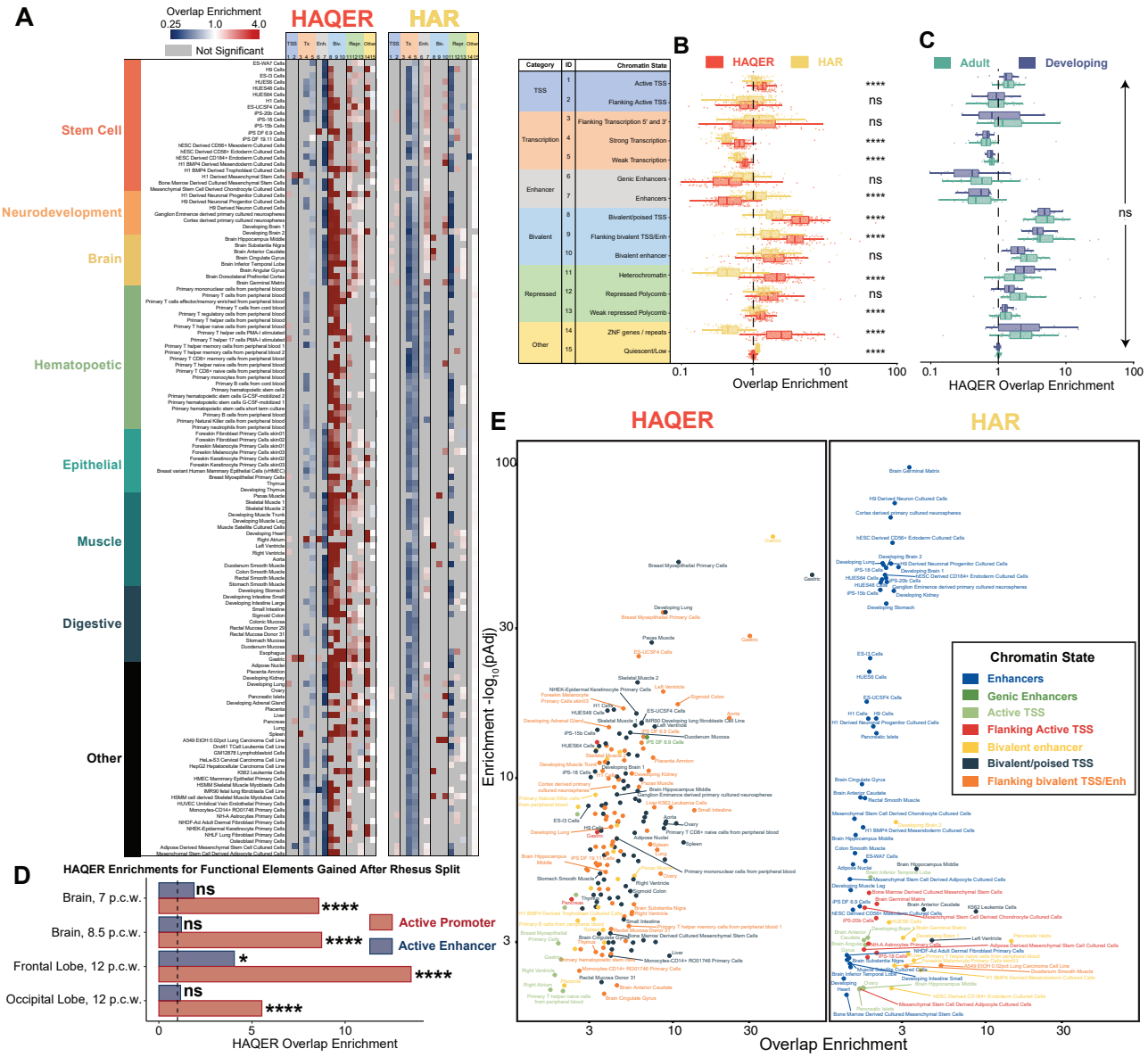


Figure S4. HAQER chromatin state enrichment analysis, related to Figure 3

(A) Overlap enrichment/depletion matrix between HAQERs (left) or HARs (right) and chromatin states from 127 reference epigenomes with sample level annotation. Both sets are depleted from transcribed regions; this reflects the ascertainment bias that some HAR studies specifically excluded coding regions.[18]

(B) Comparisons between HAQER and HAR overlap enrichment scores for each chromatin state across 127 reference epigenomes, quantified by Bonferroni-adjusted t test (**** $p < 0.0001$).

(C) Comparison of HAQER overlap enrichment between adult and developing tissue-derived reference epigenomes.

(D) HAQERs are enriched for overlaps with active promoters and active enhancers gained after rhesus split, defined as increased H3K27ac or H3K4me2 ChIP-seq signal in the developing human brain relative to mouse and rhesus macaque[38] (Bonferroni-adjusted overlap enrichment; **** $p < 0.0001$ and * $p < 0.05$).

(E) Volcano plot of significant overlap enrichments for HAQERs (left) or HARs (right) for 7 gene regulatory chromatin states.

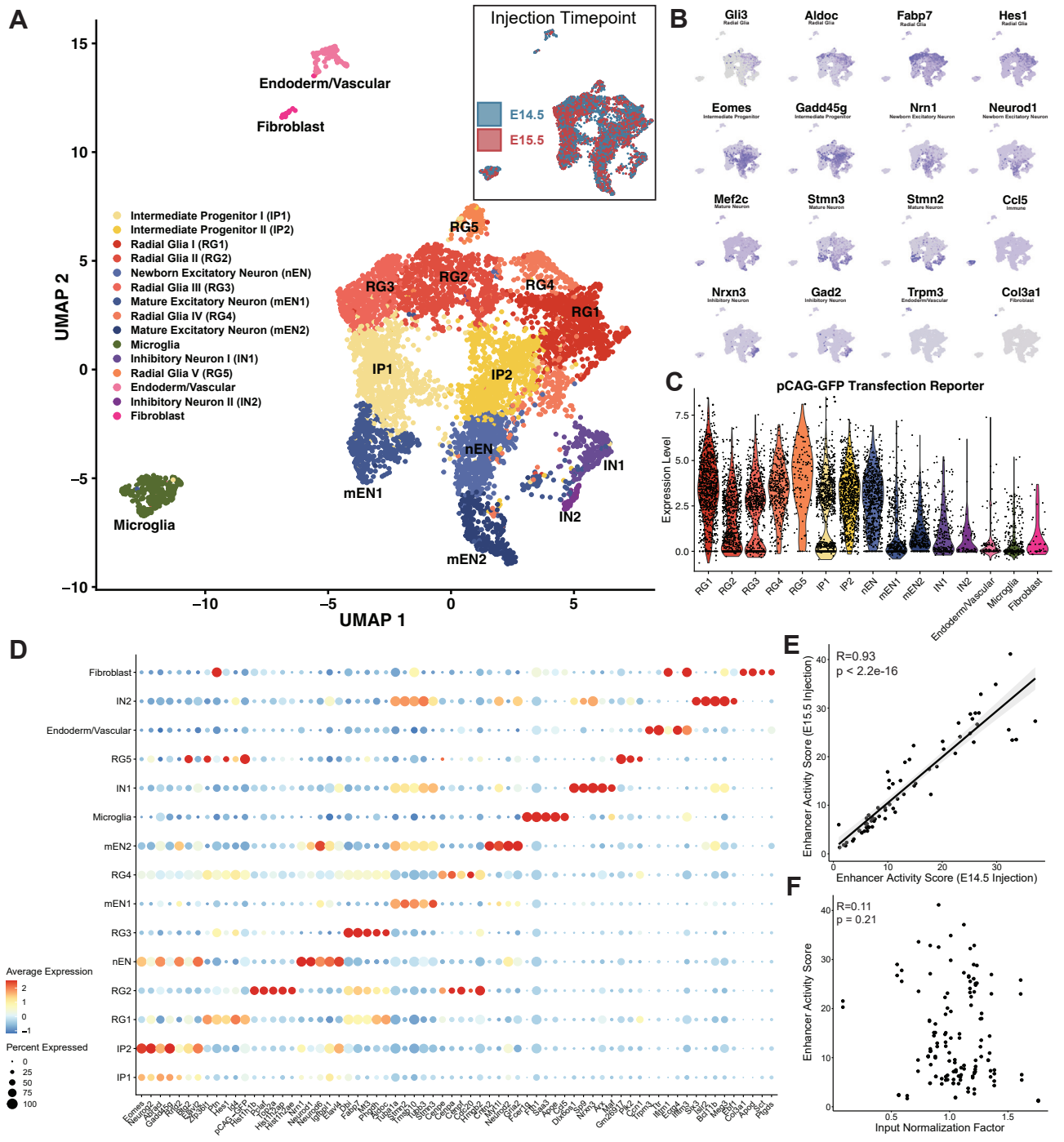


Figure S5. scSTARR-seq cluster analysis, marker gene expression, and impact of injection time point and input normalization, related to Figure 4

(A) UMAP representation of 7,170 cells from scSTARR-seq in the developing mouse brain. Insert UMAP labels cells from two independent STARR-seq experiments, performed at E14.5 and E15.5.

(B) Gene expression for canonical markers of each cell identity.

(C) pCAG-GFP transfection reporter expression in each cell cluster. As radial glia and their progeny were targeted, limited GFP expression is observed in the inhibitory neuron, microglia, fibroblast, and vascular clusters.

(legend continued on next page)

(D) Dotplot representation of the top five marker genes by cluster. The pCAG-GFP transfection reporter, which was electroporated preferentially into radial glia, was identified as a radial glial marker.
(E) Pearson correlation between enhancer activity score measurements for each test sequence at the E14.5 and E15.5 injection time points.
(F) Pearson correlation between the input normalization factor and enhancer activity score estimates for all test sequences with input normalization factors below 5.

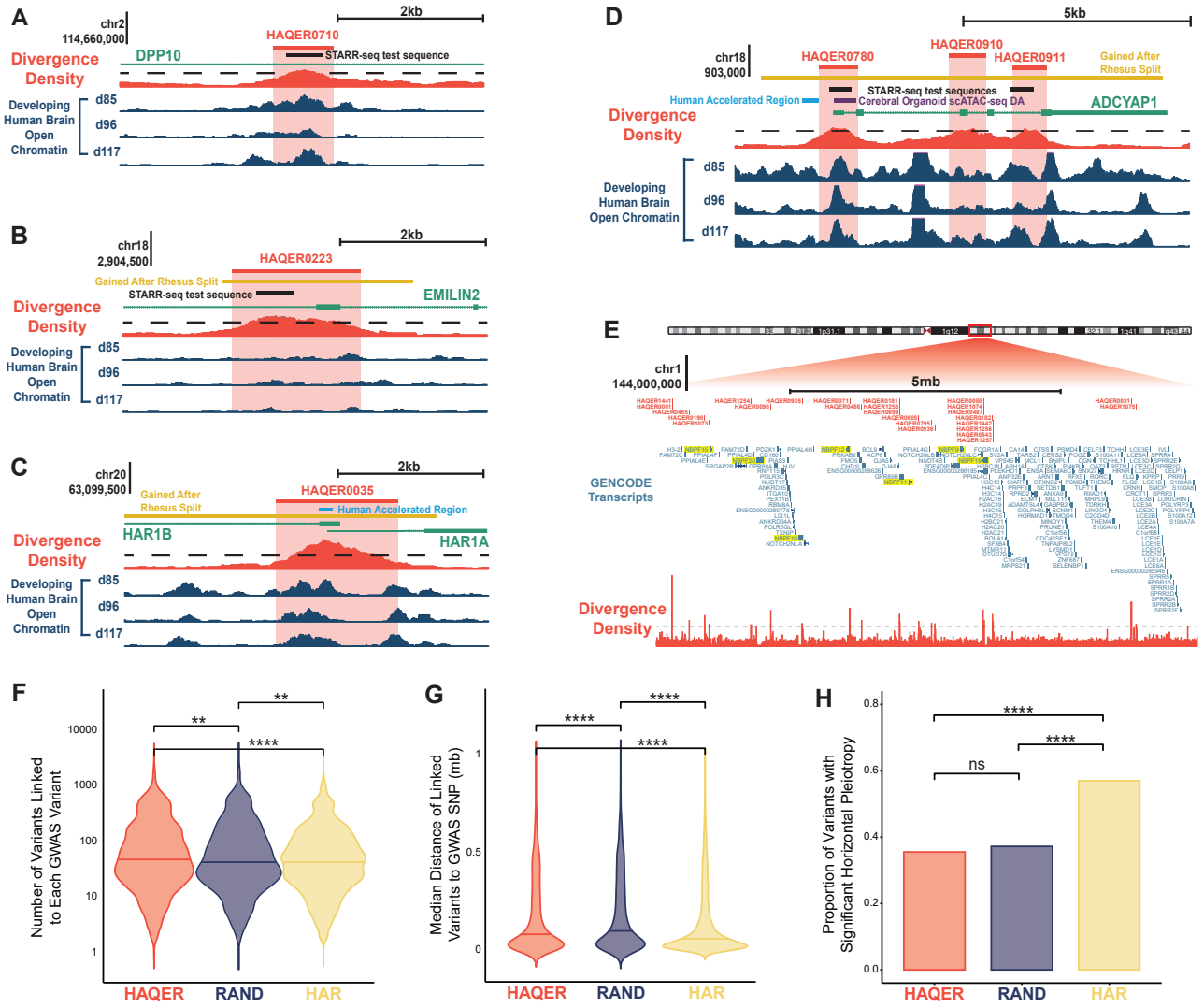


Figure S6. Genomic context of HAQERs with hominin-specific neurodevelopmental regulatory innovation; features of disease-linked variation overlapping HAQERs, related to Figures 4, 5, and 6

Each panel displays the GENCODE gene track, open chromatin in the developing human brain,[37] and divergence density in the genomic context of HAQERs with neurodevelopmental function. The dashed line corresponds to a divergence density of 29 mutations per 500 bases, the statistical threshold for HAQER identification. Also labeled are the positions of STARR-seq test sequences used in this study, human accelerated regions,[23] functional elements gained after the rhesus split,[38] and differentially accessible (DA) scATAC-seq sites in cerebral organoids.[39]

(A–C) The genomic context is shown for HAQER0710, a hominin-specific enhancer in the locus of *DPP10*, an autism spectrum disorder-related gene[40] (A); HAQER0223, a hominin-specific enhancer in the locus of *EMILIN2*[41] (B); and HAQER0035, which corresponds to HAR1, a previously described rapidly evolving region[33] (C).

(D) The genomic context for the gene *ADCYAP1*, which harbors three HAQERs. HAQER0780 and HAQER0911 were both observed as hominin-specific neurodevelopmental enhancers. *ADCYAP1* is associated with neurodevelopment,[42] human evolution,[43] and psychiatric disease.[44]

(E) Genome browser snapshot of a 10-Mb region of hg38 chr1. Members of the NBPF gene family are highlighted in yellow.

(F) Distributions of the number of variants in significant linkage disequilibrium (Plink $R^2 > 0.7$) for each disease-linked variant overlapping HAQERs, RAND, and HARs (Bonferroni-adjusted Wilcoxon; **** $p < 0.0001$ and ** $p < 0.01$).

(G) Distributions of the median distance of significantly linked variants (Plink $R^2 > 0.7$) to corresponding GWAS variants for disease-linked variants overlapping HAQERs, RAND, and HARs (Bonferroni-adjusted Wilcoxon; **** $p < 0.0001$).

(H) Proportion of variants overlapping HAQERs, RAND, and HARs with significant ($p < 0.05$) HOPS (horizontal pleiotropy scores) (chi-square test; **** $p < 0.0001$).