Massively parallel dissection of human accelerated regions in human and chimpanzee neural progenitors

Hane Ryu[1,2,3], Fumitaka Inoue[1,2], Sean Whalen[4], Alex Williams[4], Martin Kircher[5,6], Beth Martin[5], Beatriz Alvarado[7], Md. Abul Hassan Samee[2,4], Kathleen Keough[3,4], Sean Thomas[4], Arnold Kriegstein[7,9], Jay Shendure[5,8], Alex Pollen[2,7,9], Nadav Ahituv[1,2]*, Katherine S. Pollard[2,4,10,11]*

[1]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA

[2]Institute for Human Genetics, University of California San Francisco, San Francisco, CA, USA

[3]Pharmaceutical Sciences and Pharmacogenomics Graduate Program, University of California San Francisco, San Francisco, CA, USA

[4]Gladstone Institutes, San Francisco, CA 94158, USA

[5]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA

[6]Berlin Institute of Health, Berlin, Germany

[7]Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, University of California, San Francisco, CA 94143

[8]Howard Hughes Medical Institute, Seattle, Washington 98195, USA

[9]Department of Neurology, University of California, San Francisco, San Francisco, CA 94158, USA

[10]Department of Epidemiology and Biostatistics and Institute for Computational Health Sciences, University of California San Francisco, San Francisco, CA, USA

[11]Chan-Zuckerberg Biohub, San Francisco, CA, USA

*Correspondence: nadav.ahituv@ucsf.edu (N.A.), katherine.pollard@gladstone.ucsf.edu (K.S.P.)

SUMMARY

How mutations in gene regulatory elements lead to evolutionary changes remains largely unknown. Human accelerated regions (HARs) are ideal for exploring this question, because they are associated with human-specific traits and contain multiple human-specific variants at sites conserved across mammals, suggesting that they alter or compensate to preserve function. We performed massively parallel reporter assays on all human and chimpanzee HAR sequences in human and chimpanzee iPSC-derived neural progenitors at two differentiation stages. Forty-three percent (306/714) of HARs function as neuronal enhancers, with two-thirds (204/306) showing consistent changes in activity between human and chimpanzee sequences. These changes were almost all sequence dependent and not affected by cell species or differentiation stage. We tested all evolutionary intermediates between human and chimpanzee sequences of seven HARs, finding variants that interact both positively and negatively. This study shows that variants acquired during human evolution interact to buffer and amplify changes to enhancer function.

INTRODUCTION

Human accelerated regions (HARs) are highly conserved sequences that acquired many nucleotide substitutions in humans since we diverged from our common ancestor with chimpanzees and, more recently, from archaic hominins (reviewed in (Franchini and Pollard, 2017; Hubisz and Pollard, 2014)). This genetic signature suggests that HARs are important and that their functions changed during human evolution, perhaps altering traits that distinguish us from chimpanzees and other animals, such as morphological differences, our unique diet, reproductive challenges and cognitive skills. Rather than changing function, some human-specific substitutions in HARs could be the result of compensatory evolution to maintain ancestral functions, making them ideal for exploring how nearby regulatory variants interact over evolutionary time scales. Furthermore, it has been hypothesized that HARs and other uniquely human genomic regions could be responsible for our high rates of psychiatric disorders, such as schizophrenia and autism spectrum disorder (ASD), which might be maladaptive by-products of the same changes in the human brain that enabled our unique linguistic and cognitive skills (Burns, 2004; Crow, 1997). Indeed, HARs are enriched in disease-associated loci and nearby genes expressed during embryonic development, especially neurodevelopment (Babbitt et al., 2011; Capra et al., 2013; Kamm et al., 2013; Pollard et al., 2006; Prabhakar et al., 2006). HARs are also enriched for *de novo* copy number variants and biallelic mutations in individuals with ASD (Doan et al., 2016).Thus, HARs are exciting candidates for understanding human-specific traits, including our unique susceptibilities to disease, and for elucidating general principles of gene regulatory evolution.

The majority of HARs (96%) reside in noncoding regions. Amongst these noncoding HARs, at least 30% are predicted to be developmental enhancers, the majority of which are thought to be involved in brain development (Capra et al., 2013). Fifty-one prioritized HARs have been analyzed for their regulatory activity via mouse enhancer assays, with ~70% (36/51) found to be functional enhancers. Around a third (9/28) of those, where the human and non-human primate sequence were both tested, show differential enhancer activity (Franchini and Pollard, 2017). These differentially active HARs drive human-specific expression patterns during development of the central nervous system (Boyd et al., 2015; Capra et al., 2013; Kamm et al., 2013) and/or limbs (Capra et al., 2013; Prabhakar et al., 2008). HARE5 (ANC516), for example, is thought to accelerate the cell cycle in neural progenitor cells and increase brain size in transgenic mice (Boyd et al., 2015). These findings indicate that sequence changes in HARs during human evolution had the potential to alter developmental gene regulation and phenotypes.

To date, HARs have been functionally characterized on a ‘one-by-one’ basis with low-throughput techniques, primarily in a specific developmental time point using transgenic mice. Massively parallel reporter assays (MPRAs) provide a high-throughput method that can assay for enhancer function *en masse* (Inoue and Ahituv, 2015). MPRAs were recently used to analyze the effect of ASD-associated biallelic mutations on the human sequences of 279 HARs in primary mouse neurospheres, showing that 29% of the mutations alter enhancer activity (19% decrease and 10% increase) in an episomal context (Doan et al., 2016). Lentivirus-based MPRA (lentiMPRA) enables the testing of regulatory sequences in hard to transfect cells such as neurons with genomic integration, which is more reproducible than episomal MPRA (Inoue et al., 2017). Induced pluripotent stem cell (iPSC) differentiation provides a system in which to apply lentiMPRA in primate development, which has been inaccessible to investigation. By differentiating human and chimpanzee iPSCs into neuronal cell types, it is possible to evaluate and control for any effects that differences in the cellular environment between the two species might have on regulatory activity. Together, the lentiMPRA and iPSC technologies open the door to high-throughput functional characterization of HARs in primate cells.

We took advantage of these emerging technologies to assay 714 HARs (from (Lindblad-Toh et al., 2011; Pollard et al., 2006)) for their neuronal enhancer activity in multiple biological and technical replicates of both human and chimp iPSC-derived neurons. We find that 43% (306/714) of HARs function as active neuronal enhancers. Of these, two-thirds (204/306) show differential enhancer activity between the human and chimpanzee sequence, very consistently across human and chimpanzee cells and two developmental stages. By synthesizing and testing all permutations of human mutations for seven HAR enhancers (i.e., all possible evolutionary intermediates), we find that multiple human-specific nucleotide changes interact both positively and negatively to generate the net difference in enhancer activity between each pair of human and chimpanzee sequences. This single set of experiments substantially increases our understanding of HAR function and the interplay between multiple human-specific mutations in HARs.

RESULTS

Massively parallel characterization of all HARs in primate neural progenitor cells

We assessed the enhancer function of 714 HARs using lentiMPRA in human and chimpanzee iPSC-derived neural progenitor cells fated for the telencephalon (**Figure 1**). These HARs are the union of all HARs from our prior studies (Lindblad-Toh et al., 2011; Pollard et al., 2006) that were fully covered in the most current human (hg19) and chimpanzee (panTro2) genome assemblies at the time we initiated this study. Oligonucleotides (oligos) were designed to cover the human and chimpanzee sequences for each HAR (**Table S1**). Positive and negative control oligos were designed from ENCODE controls used for luciferase assays across human cell lines (provided by the lab of Dr. Richard M. Myers), as well as additional negative controls from H3K27me3 ChIP-seq peaks in human iPSC-derived neural progenitors (data generated in the Ahituv lab) (**Table S1**). We also individually validated two negative and two positive control sequences with luciferase assays in chimpanzee and human neural progenitors (28 passages) (**Figure S1**). All oligos were array-synthesized at a length of 171 bp. The median HAR length is 227 bp, and a third of them could be synthesized using a single oligo (including flanking sequences if the HAR is less than 171 bp). For the remaining HARs, we used multiple overlapping oligos tiled across the HAR which were separately quantified for enhancer activity, assessed for agreement, and then merged for downstream analysis producing one activity measurement per HAR (see Methods). The library was PCR amplified and cloned into the pLS-mP lentiviral enhancer assay vector (Inoue et al., 2017) and then used to generate lentivirus.

As we wanted to test HARs for enhancer activity in neurodevelopment, we generated neural progenitor cell lines from iPSCs derived from two different human and chimpanzee individuals. Neural induction was initiated with noggin, a BMP inhibitor, and cells were cultured in retinoic acid-free media supplemented with growth factors FGF and EGF in order to generate early (N2; 12-18 passages) and late (N3: 20-28 passages) telencephalon-fated neural progenitors (**Figure 1**). All human and chimpanzee lines exhibited normal cell morphology (**Figure 2A,G**) and normal karyotypes (**Figure 2D,J**), as well as neural rosette morphology at an early induction stage and neural progenitor cell morphology at later stages of differentiation (**Figure 2B-C, H-I**). Characterization through immunohistochemistry assays showed that both human and chimpanzee N2 and N3 cells express neural and glial progenitor proteins such as PAX6 and GFAP (**Figure 2E-F,K-L**). We assessed the heterogeneity of human and chimpanzee N2 and N3 cells through single cell RNA-seq (scRNA-seq) and observed comparable patterns of telencephalon and radial glia marker expression (**Figure 2M**). Each cell line was infected with the HAR lentiMPRA library in triplicate and assayed via DNA-normalized barcode RNA-seq, yielding a total of 24 measurements of each HAR or control sequence’s enhancer activity (3 technical replicates x 2 biological replicates x 2 species x 2 stages).

Many HARs function as neural enhancers

To robustly identify HARs that are active neural enhancers, we first compared reproducibility of lentiMPRA enhancer activity measurements across technical and biological replicates. HARs and control sequences have highly correlated activity levels across technical replicates (**Figure S2**). We therefore defined significant activity as the ability to drive expression above the 75th percentile of negative controls in all three technical replicates for at least two N2 or N3 lines (see Methods). We found that 306 out of 714 HARs meet this stringent criterion for the human and/or chimp sequence. The heatmap for these 306 HARs (**Figure 3A**) shows that lentiMPRA activity levels are strikingly similar across biological replicates, including different cell species and cell stages, but often differ between human and chimpanzee sequences, an effect that we quantify more rigorously below. Compared to conserved non-coding elements that are not accelerated in humans (phastCons elements), the 306 HAR neural enhancers we identified are enriched in loci annotated with Gene Ontology terms related to transcription, brain development and serine metabolism and reside near genes expressed in the brain (**Table S2**). Since these processes are already known to be enriched for HARs compared to conserved non-coding elements, we also compared the 306 HAR neural enhancers to all 714 HARs and found them to be slightly enriched near genes annotated with the biological processes “behavior” and “negative regulation of transcription” (**Table S2**). These associations are consistent with the 306 active HARs functioning as neurodevelopmental enhancers.

Fifty-one human HAR sequences from our lentiMPRA library have previously been tested for enhancer activity using mouse transgenic enhancer assays [mostly at embryonic day (E) 11.5, a developmental stage similar to N2] (Capra et al., 2013; Prabhakar et al., 2008; Visel et al., 2007). These assays can reveal spatial and temporal differences in enhancer activity but are not quantitative. Our lentiMPRA quantitatively measures enhancer activity but lacks the spatial information of *in vivo* reporter assays. Despite these differences, we found a fairly high overlap between active enhancers from the two types of assays (odds ratio = 2.09, hypergeometric p=0.35) (**Figure 3B**). Twenty-nine of the HARs tested with both assays were prioritized for *in vivo* assays because they had epigenomic signatures of developmental enhancers (Capra et al., 2013), and all of these that showed activity in mouse embryos (7/29) are also active in our lentiMPRA (p=0.024). These findings provide independent validation of lentiMPRA enhancer activity measurements while also highlighting differences between MPRAs and mouse experiments.

To further explore the similarities and differences between enhancer activity as measured in transgenic mice versus lentiMPRA, we performed *in vivo* reporter experiments for four HARs that drive expression in N2 and/or N3 cells and are located nearby neurodevelopmental genes: HAR152 (3' UTR of *NEUROG2*), 2xHAR.133 (same topological domain as *MEIS2* in several cell types), 2xHAR.518 (intron of *NRXN3)*, and 2xHAR.548 (same topological domain as *FOXP1*). Neural expression patterns were validated via lacZ staining in mouse embryos for the human and chimpanzee sequences of HAR152 (E10.5) and 2xHAR.548 (E13.5) (**Figure 3C, D**). 2xHAR.518 showed enhancer activity in the eye (E13.5), while HAR.133 did not show any consistent expression patterns (E11.5) (**Figure S3**). None of these validated HAR enhancers showed clear spatial differences in lacZ staining between human and chimpanzee sequences, although the human sequence for 2xHAR.548 showed more consistent expression patterns in the midbrain and hindbrain compared to more consistent expression in the eye and neural tube for the chimpanzee sequence (**Figure 3C**). HAR152 showed consistent enhancer activity for both the chimpanzee and human sequence in the developing neural tube and hindbrain (**Figure 3D**). Consistent with previous *in vivo* HAR enhancer experiments, these results generally validate our results but also illustrate the quantitative differences that can be obtained using lentiMPRA.

cis regulatory features are stronger drivers of HAR enhancer activity than the cell-line environment

We next compared activity of the human and chimpanzee alleles of the 306 HAR enhancers in human and chimpanzee N2 and N3 cells. A key feature of our experimental design is the inclusion of both alleles of each HAR in the same lentiMPRA library, so that activity of human and chimpanzee sequences can be directly compared without confounding from batch effects. Leveraging this comparability and our observation of consistent activity levels across cell species and cell stage for most HARs (**Figure 3A**), we modeled average differential activity across all conditions (Methods) and discovered 204 HARs with significant differences in enhancer activity between the human and chimpanzee sequences at a false discovery rate (FDR) less than 1% (**Figure 4A**; **Table S3**). These differentially active HARs are divided almost evenly into 100 HARs where the human allele is more active than the chimpanzee allele and 104 where the human allele is less active. It is important to note that the fold changes between human and chimpanzee alleles are relatively modest for most of these HARs (range of human:chimpanzee activity 0.43 to 2.23 with 90% between 0.85 and 1.17). However, experimental variation is low between technical replicates for both human and chimpanzee sequences (**Figure S2**) so that fold changes are highly consistent across biological conditions (**Figure 4A**; on average variance is less than 3% of mean), and we performed enough replicates to have good power to detect small quantitative differences in enhancer activity. Therefore, *in vivo* assays are unlikely to detect most of these quantitative human-chimpanzee differences. Indeed, all four of the HARs included in our study that have shown differences in neural expression domains between human and chimpanzee sequences in transgenic mice (2xHAR.114, 2xHAR.164, 2xHAR.170, 2xHAR.238) are also differentially active in lentiMPRAs (**Table S3**)(Capra et al., 2013), but five additional HARs that are differentially active in lentiMPRAs do not show differences *in vivo*, including the two we tested in this study (HAR152, 2xHAR.548).

Compared to the large number of HARs that have significant differences in enhancer activity between human and chimpanzee sequences ("*cis* effects"), we found very little evidence that these sequence effects depend on the cell line ("*trans* effects"). Most HARs with *cis* effects are significant across all four combinations of cell species and stage (**Figure 4B**). For example, most *cis* effect HARs are still significant using only N2 samples (169/204) or only N3 (131/204) samples, despite lower power with smaller sample sizes. Furthermore, only six HARs had *cis* effects that were significantly different between the N2 and N3 stages: 2xHAR.319, HAR5, 2xHAR.28, 2xHAR.238, 2xHAR.1, and 2xHAR.49. Similarly, just three HARs (2xHAR.518, HAR51, 2xHAR.264) had *cis* effects that were significantly different in human cell lines versus chimpanzee cell lines. These HARs with *trans* effects do not share any obvious characteristics that distinguish them from other HARs (**Table S4**). The consistency of *cis* effects across conditions also provides strong support for the reproducibility of our lentiMPRA, despite the fact that constructs randomly integrate into the genome. Combined, these results show that the effects of human-specific nucleotide changes on neurodevelopmental enhancer activity can be assayed via lentiMPRA in human or chimpanzee cells and that cells at related developmental stages produce similar enhancer activity measurements.

HARs contain fixed differences that disrupt TFBS motifs

To further dissect the regulatory architecture of HARs, we looked at predicted transcription factor binding sites (TFBS) that were lost or gained due to human substitutions in each of the 204 differentially active HAR enhancers. Using all vertebrate TRANSFAC (Matys et al., 2006) motifs with a p-value threshold of 10^{-5}, we characterized TFBS that were only present in the human allele (98 TFBS) or only in the chimp allele (137 TFBS) (**Table S5**). Some notable transcription factors that appear to have gained binding sites as a result of human substitutions in HAR enhancers are zinc finger proteins from the early growth response (*EGR*)*1*/*2*/*3*/*4* families which play a role in neuronal plasticity (Liu et al., 2000; Knapska et al., 2004; Lu et al., 2011), *POU1F1* which activates growth hormone genes (Sobrier et al., 2016)*,* and *FOXP1* which plays a role in radial migration and morphogenesis of cortical neurons and associated with autism and speech disorders (Li et al., 2015; Lozano et al., 2015; Teramitsu et al., 2004). It is worth noting that 2xHAR.548, one of our top scoring differentially active HARs (**Table S3**), is located within the topologically associating domain (TAD;(Dixon et al., 2012)) that encompasses *FOXP1*. Transcription factors that appear to have lost TFBS due to human mutations in HARs include several *POU2/3/4/5* family transcription factors which are important neurodevelopmental regulators (Schonemann et al., 1998), as well as zinc finger protein 263 (*ZNF263*). Misregulation of and mutations in *ZNF263* have been linked with autism and hypothalamic hamartoma (Ning et al., 2015; Saitsu et al., 2016). These transcriptional regulators with TFBS losses and gains in HARs point to specific pathways and molecular processes that may have played a role in human brain evolution.

HAR enhancers are associated with variants linked to neuropsychiatric disorders

To further explore the biological role of functionally characterized neural HAR enhancers, we investigated the overlap between HAR enhancers and neuropsychiatric disorder single nucleotide polymorphisms (SNPs) from the National Human Genome Research Institute (MacArthur et al., 2017) and the Psychiatric Genome-Wide Association Study (GWAS) Consortium (Sullivan, 2010). We associated HARs to SNPs if they fall in the same chromatin contact domain in Hi-C data from cortical plate (CP) or germinal zone (GZ) brain regions (Won et al., 2016). We confirmed these HAR-SNP associations using TADs derived from the GM12878 cell line, which is less biologically relevant but has very high coverage Hi-C data and hence high resolution TADs (Rao et al., 2014). Many HAR enhancers are in TADs with SNPs associated with ASD, schizophrenia, bipolar disorder, attention deficit hyperactivity disorder, or major depressive disorder (**Table S6**). We also checked if HARs are in significant chromatin interactions (FDR<10%) with GWAS SNPs and found CP and GZ chromatin loops that link 2xHAR.37 to rs10149407 and rs2068012, both associated with schizophrenia (Consortium, 2014). These results build on the findings of (Doan et al., 2016), showing that many neurodevelopmental HAR enhancers are in genomic loci that are associated with neuropsychiatric disease.

In fact, a few HAR neural enhancers contain neuropsychiatric disorder associated SNPs. For example, 2xHAR.170 shows differential enhancer activity between human and chimpanzee sequences and contains a SNP associated with schizophrenia (rs2434531)(Consortium, 2014)(**Figure 5A**). Another HAR with significant *cis* effects, 2xHAR.502, overlaps a SNP (rs10249234) associated with both educational attainment (Okbay et al., 2016) and age of first birth (Barban et al., 2016). Additionally, variants in 2xHAR.502 are in linkage disequilibrium with GWAS SNPs associated with schizophrenia (Consortium, 2014) and age of first birth. Other HARs, such as 2xHAR.141 and 2xHAR.214, contain GWAS SNPs but are not active enhancers in N2 or N3 cells, suggesting that they might be active in other developmental stages or cell types or have a different function. Further dissection of the effects of disease-associated sequence variants in and nearby HARs may shed light on mechanisms through which nucleotide changes during human evolution altered neurodevelopmental traits.

Combinations of nucleotide substitutions drive functional changes in HAR enhancers

We next wanted to investigate how nucleotide changes within HARs lead to differential enhancer activity. Seven HARs (2xHAR.1, HAR34, 2xHAR.65, 2xHAR.142, 2xHAR.164, 2xHAR.170, 2xHAR.238) were selected based on differential enhancer activity in our lentiMPRAs and prior evidence of enhancer activity and/or activity differences between human and chimpanzee sequences in transgenic mice (Capra et al., 2013). To dissect the effects of each nucleotide difference in these HARs and how they interact, we designed a second lentiMPRA library that contained a series of oligos for each HAR carrying every individual nucleotide difference from chimpanzee and their combination/s up to the sequence carrying all nucleotide differences in the human reference genome. The oligos for each HAR represent all potential evolutionary intermediates between its human and chimpanzee reference genome sequences. These “permutations” were tested in human and chimpanzee N2 and N3 cells (3 technical replicates of 1 biological replicate from each species) using lentiMPRA, as done for the previous library. Measurements of enhancer activity from technical replicates were again highly reproducible (**Figure S4**). For each permutation, we computed the log-ratio comparing its activity to that of the chimpanzee sequence. These log-ratios were modeled as a function of the nucleotide differences they contain, as well as the *trans* environment (cell species and stage) using penalized regression to avoid over-fitting (Methods).

We first used the penalized model for each HAR to assess if interactions between nucleotides are needed to account for activity differences across permutations carrying different combinations of nucleotide changes. These analyses strongly suggest that nucleotide differences within HARs do not have strictly additive effects on enhancer activity. First, the top features in our fitted models rarely correspond to individual nucleotides. Second, by fitting a series of models with increasingly complex interactions (i.e., including higher order interactions between increasing numbers of nucleotides), we found that models allowing up to 4-way interactions provide the best fit to HAR activity log-ratios (**Figure 5B**) and again rarely include effects of individual sites. Fitted models were compared to each other and to the model with no interactions, in which each nucleotide has an independent association with HAR enhancer activity (main effects only). Models with higher order interactions (5-way to 9-way) do not significantly improve fit and coefficients for specific interactions are stable over models of varying complexity (**Figure S5**), suggesting that penalized regression succeeded in preventing over-fitting, with 4-way interactions indicating the likely complexity of the underlying biology. The significant interactions we discovered mostly involve two or more nucleotides (**Table S7**); cell species and stage are rarely important by themselves, but rather interact with nucleotide effects. Finally, we observed a few cases where modeling indicates that the effect of a nucleotide on HAR enhancer activity is different in the presence of a second nucleotide change. For example, 2xHAR.170 contains a variant in the human genome that is associated with higher activity than the chimp allele but its activity is reduced when another variant occurs with it. This activity can then increase when a third variant appears along with the other two (**Figure 5C**). Together these results show that human-specific mutations in HARs both buffer and amplify each other's effects on enhancer activity during neurodevelopment.

DISCUSSION

How nucleotide changes in gene regluatory elements lead to differences in phenotypes remains largely unknown. Here, we used HARs as a test case to address this. The function of HARs has been an intriguing question since their discovery more than a decade ago (Pollard et al., 2006). Until recently, we have lacked the tools to comprehensively determine if HARs are regulatory elements, whether human sequence variants in HARs alter their function and why HARs harbor so many human-specific variants. Despite evolutionary, genome location, and epigenomic data suggesting that many HARs are developmental enhancers, less than one hundred hand-picked examples have been tested for enhancer activity, primarily in transgenic mice using low-throughput reporter assays. Of these, just two dozen have been assayed using both the human and chimpanzee sequence to test for differences in expression domains in mouse embryos, an experiment that cannot detect quantitative differences in expression levels. Effects of the species *trans* environment on HAR enhancer activity have not been explored beyond a few HARs being tested in mice and zebrafish, due to obvious limitations on human and non-human primate experimentation. To address all of these challenges in a single set of experiments, we performed lentiMPRA to assay the enhancer activity of the human and chimpanzee sequences of 714 HARs in human and chimpanzee derived neural progenitor cells. This investigation revealed that nearly half of all HARs function as enhancers in this cell type.

By including the human and chimpanzee sequence of each HAR in our lentiMPRA library, we could quantify the expression driven by both alleles side-by-side, providing highly accurate measurements of differential activity. Specifically, any technical noise or differences in the cellular environment cancel out when computing the log-ratio of human versus chimpanzee RNA/DNA for a given sample. Consequently, we detected small magnitude but statistically significant differences in enhancer activity between human and chimpanzee sequences for more than two hundred HARs. We were also able to quantitatively dissect the interacting effects of human-specific substitutions in seven HARs on their enhancer activity by assaying all combinations of substitutions, covering every possible evolutionary intermediate. This is important, because it has not been clear if HAR variants compensate or otherwise interact with each other versus simply being neutral variants that hitchhiked on a haplotype with one causal variant. Modeling these data with penalized regression revealed strong evidence of both positive and negative interactions between multiple substitutions within each HAR. This suggests that changes in HAR sequences during human evolution may have amplified and dampened the functional consequences of other nearby substitutions. For example, cases like 2xHAR.170 where one substitution changes activity compared to the chimpanzee sequence while another brings activity back closer to the chimpanzee level (**Figure 5C**), could represent compensatory evolution and may provide some explanation for the unexpectedly large number of substitutions in HARs.

Since we performed lentiMPRA in triplicate from two human biological replicates and two chimpanzee biological replicates, we could quantify the contribution of the *trans* environment to HAR enhancer function. Strikingly, we found that the activity of HARs and the differential activity of human versus chimpanzee HAR sequences are very consistent across human and chimpanzee cell lines, as well as two developmental time points (N2 versus N3) separated by approximately eight passages. On one hand, this result might be expected given the similarity of the human and chimpanzee proteome and prior evidence that human enhancers assayed in zebrafish and mice are largely concordant (Ritter et al., 2010). On the other hand, one might predict that differences in expression of human and chimpanzee proteins or even batch effects (if not appropriately modeled) would alter HAR enhancer activity across our samples. The consistency we observe therefore indicates that our lentiMPRA is highly reproducible and that *cis* effects of human-specific nucleotide substitutions can be accurately assayed in cells from either species, which will be useful given the large number of available cell types currently available from humans.

It is important to note that our approach has several limitations. As our cloning technique is based on oligo synthesis, our assayed sequences were 171bp in length. For a third of the HARs, this length allowed us to clone the entire HAR, but for some HARs we had to assay two or more overlapping oligos. For these HARs, we summed DNA and RNA read counts across both oligos before taking the RNA/DNA ratio, which normalizes for the additional coverage in the library. Further advances in DNA synthesis or cloning using DNA capture methods could allow us in the future to increase our assayed sequence size and carry out MPRAs that capture the entire HAR length. As we could not carry out these experiments in chimpanzees or humans, we used neural cells derived from iPSCs. As expected and also shown by our scRNA-seq (**Figure 2M**), our population of cells is heterogeneous. The use of cell specific markers could allow for sorting of specific cellular populations in future lentiMPRA experiments. In addition, we only tested telencephalon-fated neural progenitors in this study. Additional nervous system cells, such as astrocytes, oligodendrocytes, glia and other cell types, as well as brain organoids, could be tested using this approach. Despite the heterogeneity, we did observe a strong effect of *cis* factors on enhancer activity, suggesting that the sequence itself has a stronger effect than the *trans* environment.

Looking ahead, it will be exciting to expand upon these initial lentiMPRA investigations of HARs. One straightforward extension will be assaying additional sequences that were associated with human evolution, including regions with human-specific epigenomic marks (Prescott et al., 2015; Reilly et al., 2015; Vermunt et al., 2016). It will also be very important to measure HAR enhancer activity in additional cell types, both iPSC derived (e.g., glia, astrocytes, cardiomyocytes, hepatocytes) and others that cannot currently be derived from iPSCs, as well as organoids. The consistent activity we observed between human and chimpanzee samples indicates that working with cells only derived from human individuals may be sufficient, which broadens the types of cells where lentiMPRA could be performed. Another promising future direction is to apply the permutation approach to more HARs in order to confirm that our results regarding interactions generalize. This approach should also be applied to non-HAR regulatory elements in order to decipher the role of interactions between sites in regulatory grammar more broadly. Finally, since we focused here on sequence differences between the human and chimpanzee reference genomes, the effects of polymorphic human variants on HARs and other conserved non-coding elements, including disease associated variants would be of extreme interest for future investigation. This study lays the groundwork for these future studies.

ACKNOWLEDGMENTS

This work was supported in part by the National Institute of Mental Health grant number 1R01MH109907 (N.A. and K.S.P.), National Heart Lung and Blood Institute grant number UM1HL098179 (K.S.P.), National Human Genome Research Institute (NHGRI) and the National Cancer Institute grant number 1R01CA197139 (N.A. and J.S.), NHGRI grant number 1UM1HG009408 (N.A. and J.S.), and ORIP/OD grant number P51OD011132 (A.P.). N.A. is also supported in part by the National Institute of Child Health and Human Development grant number 1P01HD084387 and the National Hearth Lung and Blood Institute grant number 1R01HL138424.

AUTHOR CONTRIBUTIONS

H.R., A.W., N.A. and K.S.P. designed experiments, H.R., F.I., B.M., B.A., A.P. carried out wet lab experiments, H.R. and F.I. did lentiMPRA, H.R., B.A. and A.P. carried out scRNA-seq and its analyses, B.M. did lentiMPRA sequencing, H.R., S.W., M.K., S.T., K.S.P. analyzed lentiMPRA data, H.R. and M.A.H.S. did TFBS analyses, H.R. and K.K. carried out GWAS analyses, H.R., F.I., N.A. and K.S.P analyzed transgenic embryos, A.K., J.S., A.P., N.A. and K.S.P. provided resources for the study, H.R., S.W., A.P., N.A. and K.S.P. wrote the manuscript and all authors contributed to its editing.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

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., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68-74.

Babbitt, C.C., Warner, L.R., Fedrigo, O., Wall, C.E., and Wray, G.A. (2011). Genomic signatures of diet-related shifts during human origins. Proceedings Biological sciences / The Royal Society *278*, 961-969.

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

Barban, N., Jansen, R., de Vlaming, R., Vaez, A., Mandemakers, J.J., Tropf, F.C., Shen, X., Wilson, J.F., Chasman, D.I., Nolte, I.M.*, et al.* (2016). Genome-wide analysis identifies 12 loci influencing human reproductive behavior. Nat Genet *48*, 1462-1472.

Bershteyn, M., Nowakowski, T.J., Pollen, A.A., Di Lullo, E., Nene, A., Wynshaw-Boris, A., and Kriegstein, A.R. (2017). Human iPSC-Derived Cerebral Organoids Model Cellular Features of Lissencephaly and Reveal Prolonged Mitosis of Outer Radial Glia. Cell Stem Cell *20*, 435-449.e434. doi: 410.1016/j.stem.2016.1012.1007. Epub 2017 Jan 1019.

Boyd, J.L., Skove, S.L., Rouanet, J.P., Pilaz, L.J., Bepler, T., Gordan, 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.

Burns, J.K. (2004). An evolutionary theory of schizophrenia: cortical connectivity, metarepresentation, and the social brain. The Behavioral and brain sciences *27*, 831-855; discussion 855-885.

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. doi: 20130010.20131098/rstb.20132013.20130025. Print 20132013 Dec 20130019.

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421-427. doi: 410.1038/nature13595.

Conway, J.R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics *33*, 2938-2940.

Crow, T.J. (1997). Is schizophrenia the price that Homo sapiens pays for language? Schizophrenia research *28*, 127-141.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376-380.

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*, 341-354.e312. doi: 310.1016/j.cell.2016.1008.1071.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell systems *3*, 95-98.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC bioinformatics *10*, 48.

Franchini, L.F., and Pollard, K.S. (2017). Human evolution: the non-coding revolution. BMC biology *15*, 89.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017-1018.

Hubisz, M.J., and Pollard, K.S. (2014). Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution. Curr Opin Genet Dev *29*, 15-21.

Inoue, F., and Ahituv, N. (2015). Decoding enhancers using massively parallel reporter assays. Genomics *10*, 30008-30002.

Inoue, F., Kircher, M., Martin, B., Cooper, G.M., Witten, D.M., McManus, M.T., Ahituv, N., and Shendure, J. (2017). A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. Genome Res *27*, 38-52. doi: 10.1101/gr.212092.212116.

Kamm, G.B., Pisciottano, F., Kliger, R., and Franchini, L.F. (2013). The developmental brain gene NPAS3 contains the largest number of accelerated regulatory sequences in the human genome. Molecular biology and evolution *30*, 1088-1102.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat Methods *12*, 357-360.

Kircher, M. (2012). Analysis of high-throughput ancient DNA sequencing data. Methods Mol Biol *840*, 197-228.

Li, X., Xiao, J., Frohlich, H., Tu, X., Li, L., Xu, Y., Cao, H., Qu, J., Rappold, G.A., and Chen, J.G. (2015). Foxp1 regulates cortical radial migration and neuronal morphogenesis in developing cerebral cortex. PLoS One *10*, e0127671.

Liao, Y., Smyth, G.K., and Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res *41*, e108.

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.

Lozano, R., Vino, A., Lozano, C., Fisher, S.E., and Deriziotis, P. (2015). A de novo FOXP1 variant in a patient with autism, intellectual disability and severe speech and language impairment. Eur J Hum Genet *23*, 1702-1707.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J.*, et al.* (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res *45*, D896-d901.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K.*, et al.* (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res *34*, D108-110.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference 51-56.

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 *28*, 495-501.

Miyaoka, Y., Chan, A.H., Judge, L.M., Yoo, J., Huang, M., Nguyen, T.D., Lizarraga, P.P., So, P.L., and Conklin, B.R. (2014). Isolation of single-base genome-edited human iPS cells without antibiotic selection. Nat Methods *11*, 291-293.

Ning, Z., McLellan, A.S., Ball, M., Wynne, F., O'Neill, C., Mills, W., Quinn, J.P., Kleinjan, D.A., Anney, R.J., Carmody, R.J.*, et al.* (2015). Regulation of SPRY3 by X chromosome and PAR2-linked promoters in an autism susceptibility region. Hum Mol Genet *24*, 5126-5141.

Okbay, A., Beauchamp, J.P., Fontana, M.A., Lee, J.J., Pers, T.H., Rietveld, C.A., Turley, P., Chen, G.B., Emilsson, V., Meddens, S.F.*, et al.* (2016). Genome-wide association study identifies 74 loci associated with educational attainment. Nature *533*, 539-542.

Okita, K., Yamakawa, T., Matsumura, Y., Sato, Y., Amano, N., Watanabe, A., Goshima, N., and Yamanaka, S. (2013). An efficient nonviral method to generate integration-free human-induced pluripotent stem cells from cord blood and peripheral blood cells. Stem cells (Dayton, Ohio) *31*, 458-466.

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.

Prabhakar, S., Noonan, J.P., Paabo, S., and Rubin, E.M. (2006). Accelerated evolution of conserved noncoding sequences in humans. Science *314*, 786.

Prabhakar, S., Visel, A., Akiyama, J.A., Shoukry, M., Lewis, K.D., Holt, A., Plajzer-Frick, I., Morrison, H., Fitzpatrick, D.R., Afzal, V.*, et al.* (2008). Human-specific gain of function in a developmental enhancer. Science *321*, 1346-1350.

Prescott, S.L., Srinivasan, R., Marchetto, M.C., Grishina, I., Narvaiza, I., Selleri, L., Gage, F.H., Swigut, T., and Wysocka, J. (2015). Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. Cell *163*, 68-83.

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

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

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

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.

Ritter, D.I., Li, Q., Kostka, D., Pollard, K.S., Guo, S., and Chuang, J.H. (2010). The importance of being cis: evolution of orthologous fish and mammalian enhancer activity. Molecular biology and evolution *27*, 2322-2332.

Saitsu, H., Sonoda, M., Higashijima, T., Shirozu, H., Masuda, H., Tohyama, J., Kato, M., Nakashima, M., Tsurusaki, Y., Mizuguchi, T.*, et al.* (2016). Somatic mutations in GLI3 and OFD1 involved in sonic hedgehog signaling cause hypothalamic hamartoma. Annals of clinical and translational neurology *3*, 356-365.

Schonemann, M.D., Ryan, A.K., Erkman, L., McEvilly, R.J., Bermingham, J., and Rosenfeld, M.G. (1998). POU domain factors in neural development. Adv Exp Med Biol *449*, 39-53.

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N.A., Huber, W., C, H.H., Mirny, L.*, et al.* (2017). Two independent modes of chromatin organization revealed by cohesin removal. Nature *551*, 51-56.

Sobrier, M.L., Tsai, Y.C., Perez, C., Leheup, B., Bouceba, T., Duquesnoy, P., Copin, B., Sizova, D., Penzo, A., Stanger, B.Z.*, et al.* (2016). Functional characterization of a human POU1F1 mutation associated with isolated growth hormone deficiency: a novel etiology for IGHD. Hum Mol Genet *25*, 472-483.

Sullivan, P.F. (2010). The psychiatric GWAS consortium: big science comes to psychiatry. Neuron *68*, 182-186.

Teramitsu, I., Kudo, L.C., London, S.E., Geschwind, D.H., and White, S.A. (2004). Parallel FoxP1 and FoxP2 expression in songbird and human brain predicts functional interaction. J Neurosci *24*, 3152-3163.

Tretyakov, K. (2014). pyliftover.

van der Walt, S., Colbert, C.S., and Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science and Engineering *13*, 22-30.

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

Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L.A. (2007). VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Res *35*, D88-92.

Wang, X., and McManus, M. (2009). Lentivirus production. J Vis Exp *(32).* 1499. doi: 1410.3791/1499.

Won, H., de la Torre-Ubieta, L., Stein, J.L., Parikshak, N.N., Huang, J., Opland, C.K., Gandal, M.J., Sutton, G.J., Hormozdiari, F., Lu, D.*, et al.* (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. Nature *538*, 523-527. doi: 510.1038/nature19847.

FIGURE LEGENDS

Figure 1. Massively parallel characterization of HARs

Human and chimpanzee sequences for all 714 HARs were assayed using lentiMPRA for enhancer activity in human and chimpanzee iPSC-derived N2 and N3 cells. Two biological replicates were tested per species and all four cell lines were assayed in triplicate at the N2 and N3 stages of differentiation (8 conditions; 24 samples). Differential activity was measured between sequences (human versus chimpanzee) in each sample.

Figure 2. Characterization of human and chimpanzee cells

(A-C) Brightfield images of human iPSCs (A), iPSC differentiated into neural rosettes (B) and N2 cells (C) demonstrating typical morphology.

(D) Human iPSCs demonstrate normal karyotypes.

(E) Human N2 cells express Paired Box 6 (*PAX6*), a neural marker.

(F) Human N3 cells express Glial Fibrillary Acidic Protein (GFAP), a glial marker.

(G-I) Brightfield images of chimpanzee iPSCs (G), iPSC differentiated into neural rosettes (H) and N2 cells (I) demonstrating typical morphology.

(J) Chimpanzee iPSCs demonstrate normal karyotypes.

(K) Chimpanzee N2 cells express *PAX6*.

(L) Chimpanzee N3 cells express GFAP.

(M) Single cell gene expression analysis from human and chimpanzee N2 and N3 cells show comparable marker expression for radial glia and telencephalon. In both human and chimpanzee cell lines at the N2 and N3 stage, 50-90% of cells expressed FOXG1, a marker of the telencephalon.

Figure 3. Many HARs are active enhancers in human and chimpanzee N2 and N3 cells

(A) A heatmap depicting z-scores as relative levels of enhancer activity for all 306 human and chimpanzee HAR sequences with enhancer activity greater than the 75th percentile of negative control sequences in N2 or N3 cells. Each row represents a HAR that is active in all three replicates for at least two N2 lines or at least two N3 lines. Column annotations for sequence origin, cell species and cell stage are shown below the heatmap. Four active HARs (2xHAR.548, HAR152, 2xHAR.133, and 2xHAR.518) that were tested in mice are labeled on the right next to their corresponding rows.

(B) Overlap of active HAR enhancers tested with both lentiMPRA and reporter assays in transgenic mouse embryos.

(C-D) *In vivo* validated neurodevelopmental enhancers 2xHAR.548 (E13.5)(C) and HAR152 (E10.5)(D).

Figure 4. HARs show species-specific enhancer activity across cell types

(A) The relative activity of human and chimpanzee HAR enhancers are depicted as z-scores for 204 differentially active HARs (adjusted p-value<0.01 across all samples). Most of these *cis* effects are consistent across *trans* environments, except for three HARs that demonstrated significant cell species effects (2xHAR.518, HAR51, 2xHAR.264) and six HARs with significant cell stage effects (2xHAR.319, HAR5, 2xHAR.28, 2xHAR.238, 2xHAR.1, 2xHAR.49).

(B) HAR enhancer activity is separated by four conditions - human N2, human N3, chimp N2, chimp N3. The number of HARs that exhibit activity in any given set of conditions is depicted by the histogram and the intersection of conditions are represented by the connected dots below the histogram. Most HARs that have enhancer activity are active in all four conditions.

Figure 5. Impact of individual or combinatorial mutations in HARs with species-specific function

(A) UCSC genome browser snapshot of 2xHAR.170 with three fixed differences highlighted (green, yellow and blue rectangles). The first fixed variant (green) is predicted to introduce a POU3F2 TFBS. Representative 2xHAR.170 transgenic mouse embryos on the bottom right of the panel show expanded forebrain and midbrain enhancer expression of the human sequence compared to the chimpanzee sequences at E11.5. Adapted from (Capra et al., 2013).

(B) Degree versus deviance ratio plotted for permutation data on seven HARs that show species-specific function in mouse embryos: 2xHAR.170, 2xHAR.238, 2xHAR.164, 2xHAR.142, 2xHAR.1, HAR34, 2xHAR.65. The null deviance is the residual error of the null model, using only the mean of the response variable (the log2 rna/dna ratio). The deviance ratio, or the percent null deviance explained, is relative to the null deviance. The number of degrees determines how many interactions between features were allowed in each model. For most HARs, a few degrees of interactions beyond the main effects are required to explain a substantial portion of the null deviance. The largest positive and negative coefficients per HAR are shown.

(C) Colormap showing coefficients corresponding with increases (orange) or decreases (purple) in the response variable per 2xHAR.170 permutation across human and chimpanzee N3 cells.

SUPPLEMENTAL METHODS

Cell lines

We performed lentiMPRA in N2 and N3 cells derived from four separate iPSC lines from two human and two chimpanzee males. All lines were reprogrammed from fibroblasts using episomal plasmids according to a recently published protocol (Okita et al., 2013). One iPSC line was previously described (WTC; (Miyaoka et al., 2014)), and three were generated from low passage fibroblasts (P3 – P7) from Coriell Cell Repository (Hs1: 2 year old human male, catalog AG07095; Pt2: 6 year old chimpanzee male, Maverick, catalog: S003611; Pt5: 8 year old chimpanzee male, catalog PR00738). We electroporated three micrograms of episomal expression plasmid mixture encoding OCT3/4, SOX2, KLF4, L-MYC, LIN28, and shRNA for TP53 into 300,000 fibroblasts from each individual with a Neon Electroporation Device (Invitrogen), using a 100 µL kit, with setting of 1,650V, 10ms, and three pulses (Bershteyn et al., 2017). After 5 – 8 days, cells were detached and seeded onto irradiated SNL feeder cells. The culture medium was replaced the next day with primate ESC medium (Reprocell) containing 5 – 20 ng/mL of βFGF. Colonies were picked after 20 – 30 days, and selected for further cultivation. After three to five passages, colonies were transferred to Matrigel-coated dishes and maintained in mTeSR1 medium (Stem Cell Technologies, 05850) supplemented with Penicillin/Streptomycin/Gentomycin. Further passaging was performed using calcium and magnesium free PBS to gently disrupt colonies. Each line showed a normal karyotype, and will be described further in a forthcoming paper (Pollen et al., in preparation). The UCSF Committee on Human Research and the UCSF GESCR (Gamete, Embryo, and Stem Cell Research) Committee approved all human iPSC experiments.

Neural differentiation of human and chimpanzee iPSCs

Human and chimpanzee iPSCs were cultured in Matrigel-coated plates with mTeSR media in an undifferentiated state. Cells were propagated at a 1:3 ratio by treatment with 200 U/mL collagenase IV and mechanical dissection. To trigger neural induction, iPSCs were split with EDTA at 1:5 ratios in culture dishes coated with matrigel and culture in N2B27 medium (comprised of DMEM/F12 medium (Invitrogen) supplemented with 1% MEM-nonessential amino acids (Invitrogen), 1 mM L-glutamine, 1% penicillin-streptomycin, 50 ng/mL bFGF (FGF-2) (Millipore), 1x N2 supplement, and 1 x B27 supplement (Invitrogen)) supplemented with 100 ng/ml mouse recombinant Noggin (R&D systems). Cells at passages 1-3 were split by collagenase into small clumps, similar to hESC culture, and continuously cultured in N2B27 medium with Noggin. After passage 3, cells were plated at the density of 5E4 cells/cm^2 after disassociation by TrypLE express (Invitrogen) into single-cell suspension, and cultured in N2B27 medium supplemented with 20 ng/mL bFGF and EGF. Cells were maintained under this culture condition for a minimum of three months with a stable proliferative capacity. N2 cells were collected at P12-18 and N3 cells at P20-28.

Validation of N2 and N3 markers through immunostaining

Human and chimpanzee N2 and N3 cells were examined using immunostaining against neural and glial progenitor markers. Cells were cultured in chambered Millipore EZ slides, rinsed with PBS, fixed with 4% paraformaldehyde in PBS for 15 minutes at room temperature, washed three times with ice cold PBS, and permeabilized through incubation for 10 min with PBS containing 0.1% Triton X-100. Cells were washed in PBS three times and incubated with 10% donkey serum for 30 minutes to block unspecific binding of antibodies. Cells were next incubated with diluted primary antibodies against Nestin (monoclonal mouse, Abcam, AB6142), Pax6 (polyclonal rabbit, Abcam, AB5790), and GFAP (polyclonal rabbit, Chemicon, AB5804) in 10% donkey serum for 1 hour at room temperature. The cells were then washed three times in PBS, 5 minutes each wash, then incubated with secondary antibody (Alexa 488 donkey anti rabbit, Life technologies; Alexa 546 donkey anti mouse, Life technologies) in donkey serum for 1 hour at room temperature in the dark. Cells were then washed three times with PBS in the dark, then covered with a coverslip in Cytoseal mounting media (Thermo Scientific).

Single Cell RNA-Sequencing

To determine the composition of cell types in human and chimpanzee cell lines used for lentiMPRA, we generated single cell gene expression (scRNA-seq) data and clustered cells from each line based on expression. Cells were captured using the C1TM Single-Cell Auto Prep Integrated Fluidic Circuit (IFC), which uses a microfluidic chip to capture the cells, perform lysis, reverse transcription and cDNA amplification in nanoliter reaction volumes. The details of the protocol are described in PN100-7168 (http://www.fluidigm.com/). Sequencing libraries were prepared after the cDNA was harvested from the C1 microfluidic chip using the Nextera XT Sample Preparation Kit (Illumina), following its protocol with minor modifications. The single cell libraries from each C1 capture were then pooled, cleaned twice with 0.9X Agencourt AMPure XP SPRI beads (Beckman Coulter), eluted in DNA suspension buffer (Teknova) or EB buffer (Qiagen) buffer and quantified using High Sensitivity DNA Chip (Agilent).

scRNA-seq paired-end reads were generated per library. Sequencing data is available through dbGaP (phs000989). We trimmed reads for quality using cutadapt under the Trim Galore! wrapper (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with the default settings, and Nextera transposase sequences were removed. Reads shorter than 20 bp were discarded. Read level quality control was then assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were aligned to the NCBI human reference assmebly GRCh38 by HiSat2 (Kim et al., 2015) using the prefilter-multihits option and a guided alignment via the human Gencode Basic v20 transcriptome. Expression for RefSeq genes was quantified by the featureCounts routine, in the subRead library (Liao et al., 2013), using only uniquely mapping reads and discarding chimeric fragments and unpaired reads. Gene expression values were normalized based on library size as counts per million reads (CPM). We used visual image calls to remove any libraries that originated from C1 chambers with multiple cells. To further identify outlier cells, we removed libraries with fewer than 1,000 genes detected, or with greater than 20% of reads aligning to mitochondrial or ribosomal genes. Gene expression was analyzed using a threshold of detection for each gene at 2 CPM. We then calculated the percentage of cells expressing regional identity genes (e.g., FOXG1 for telencephalon, DLX6-AS1 for GABAergic neurons, MKI67 for dividing cells, SLC1A3 for radial glia). In both human and chimpanzee cell lines at the NPC and GPC stage, 50-90% of cells expressed telencephalon (FOXG1) and radial glia/astrocyte markers.

Luciferase assays

To generate pLS-mP-Luc vector (Addgene 106253), minimal promoter and Luciferase gene fragment was amplified using pGL4.23 (promega) as a template and inserted into pLS-mP (Addgene 81225) replacing with mP-EGFP. To generate pLS-SV40-mP-Rluc (Addgene106292), renilla luciferase gene was amplified using pGL4.74 (promega) as a template and inserted into pLS-SV40-mP vector (Inoue et al., 2017) replacing with EGFP gene. Negative 1 (chr2:238,336,485-238,336,655; hg19), Negative 2 (chr7:96,637,215-96,637,385; hg19), Positive 1 (chrX:55,041,354-55,041,524; hg19) and Positive 2 (chr6:10,147,166-10,147,336; hg19) were cloned into the pLS-mP-luc using In-Fusion (Clontech). Lentivirus was generated using standard methods (Wang and McManus, 2009), as described below for the library, individually for each clone with pLS-SV40-mP-Rluc spiked in at 10% of the total amount of plasmid used. 2x10^4 Chimpanzee P2 N3 and human WTC N3 cells per well were seeded in a 96-well plate and were infected with virus 24 hours later. Three independent replicate cultures were transfected per plasmid and two biological replicates were done in different days. Firefly and Renilla luciferase activities were measured on a Synergy 2 microplate reader (BioTek) using the Dual-Luciferase Reporter Assay System (Promega). Enhancer activity was calculated as the fold change of each construct's firefly luciferase activity normalized to renilla luciferase activity.

MPRA library design

All human and chimpanzee sequences for 714 HARs from our prior studies (Lindblad-Toh et al., 2011; Pollard et al., 2006) that were present in both the human (hg19) and chimpanzee (panTro2) reference genome sequences were included in the library design. For each HAR, we designed 171-bp MPRA oligos representing the orthologous human and chimpanzee sequences. Genomic sequence was added to HARs shorter than 171 bp, and HARs longer than 171 bp were tiled with multiple oligos having variable overlap depending on the length of the HAR. For controls, oligos were also tiled across ENCODE positive and negative controls commonly used in various cell lines for luciferase assays (provided by the Myers lab), as well as H3K27ac and H3K27me3 ChIP-seq peaks in human iPSC-derived N2 and N3 neural progenitor cell lines (data generated in the Ahituv lab). All HAR and control sequences were scanned for restriction sites (for SbfI and EcoRI) and modified to avoid problems in synthesis and cloning. The final array design included 2,440 unique 171-bp sequences, each with 100 uniquely assigned 15-bp barcodes for a total of 244,000 oligos. Using 100 barcoded replicates per candidate enhancer ensures robustness to integration site and other sources of technical variability.

MPRA library synthesis and cloning

All MPRA sequences were array-synthesized as 230-bp oligos (Agilent Technologies) containing universal priming sites (AGGACCGGATCAACT…CATTGCGTGAACCGA), a 171-bp candidate enhancer sequence, spacer (CCTGCAGGGAATTC), and 15-bp barcode. The amplification and cloning of the enhancers and barcodes into the pLS-mP lentiviral vector was performed as previously described (Inoue et al., 2017). Briefly, pLS-mP was cut with SbfI and

EcoRI taking out the minimal promoter and EGFP reporter gene. The oligos containing the HAR, spacer, and barcode (**Table S1**) were amplified with adaptor primers (pLSmP-AG-f and pLSmPAG-r) that have overhangs complementary to the cut vector backbone, and the products were cloned using NEBuilder HiFi DNA Assembly mix (NEB, E2621). The cloning reaction was transformed into electro-competent cells (NEB C3020) and multiple transformations were pooled and midiprepped (Chargeswitch Pro Filter Plasmid Midi Kit, Invitrogen CS31104). The library was then cut using *Sbf*I and *Eco*RI sites contained within the spacer, so that the minimal promoter and EGFP could be reintroduced via a sticky end ligation (T4 DNA Ligase, NEB M0202). This library was transformed and purified, as previously described, and sequenced to determine complexity.

Lentivirus library preparation and infection

Lentivirus packaging of the HAR MPRA library was performed by the UCSF Viracore using standard techniques (Wang and McManus, 2009). Twelve million HEK293T cells were plated in a 15-cm dish and cultured for 24 hours. The cells were co-transfected with 8 µg of the HAR library and 4 µg of packaging vectors using jetPRIME (Polyplus-transfections). The transfected cells were cultured for 3 days and lentiviruses were harvested and concentrated as previously described (Wang and McManus, 2009). For all human and chimpanzee cell lines and cell stages, about twelve million cells were plated in 15-cm dishes and cultured for 24-48 hours. Cells were infected with a multiplicity of infection (MOI) of 50. Infected cells were washed daily with PBS. They were then harvested and washed again with PBS three times before cell lysis in order to remove any non-integrated lentivirus.

RNA & DNA isolations and sequencing

Genomic DNA and total RNA were extracted using the AllPrep DNA/RNA mini kit (Qiagen). Messenger RNA was purified from the total RNA using Oligotex mRNA mini kit (Qiagen) and treated with Turbo DNAseq to remove contaminating DNA. The RT-PCR, amplification and sequencing of RNA and DNA were performed as previously described (Inoue et al., 2017), with some alterations for adding Unique Molecular Identifiers (UMIs) in the process. In brief, mRNA was reverse transcribed with SuperScript II (Invitrogen) using a primer downstream from the barcode. The resulting cDNA was split into multiple reactions to reduce PCR jack-potting effects and cDNA amplification performed with Kapa Robust polymerase for three cycles, incorporating unique molecular identifiers (UMIs) of 10 bp length. PCR products were cleaned with AMPure XP beads (Beckman Coulter) to remove primers and concentrate samples. These products underwent a second round of amplification in 8 reactions per replicate for 15 cycles, switching from the UMI-incorporating reverse primer to one containing only the P7 flow cell sequence. All reactions were pooled and run on agarose gels for size selection and submitted for sequencing. For DNA, each replicate was amplified for 3 cycles with UMI-incorporating primers, just as the RNA. First round products were cleaned up with AMPure XP beads, and amplified in split reactions, each for 20 cycles. Again, reactions were pooled and gel-purified.

RNA and DNA for all three replicates for all samples were sequenced on an Illumina NextSeq instrument (2x15 bp barcodes + 10bp UMI + 10bp sample index) and are available through the Short Read Archive (SRA) with BioProject accession numbers PRJNA428580 (chimpanzee cells) and PRJNA428579 (human cells). Illumina Paired End reads each sequenced the barcodes from the forward and reverse direction and allowed for adapter trimming and consensus calling of tags (Kircher, 2012). Barcode or UMI sequences containing unresolved bases (N) or not matching the designed length of 15bp were excluded. In data analysis, each barcode x UMI pair is counted only once and only barcodes matching perfectly to those included in the above oligo design were considered.

Normalization of RNA/DNA ratios and quantification of enhancer activity

RNA/DNA ratios per HAR per sample were calculated by taking the sum of RNA counts for all barcodes assigned to all oligo(s) tiling across each HAR, divided by the sum of all DNA counts for all barcodes across all oligo(s) per HAR, and using only barcodes with >0 counts in DNA. We summed counts across oligos for HARs tiled using two or more oligos, because we observed generally good agreement between oligos for the same HAR. HAR sequences were defined as active enhancers if they had a human and/or chimpanzee allele with an RNA/DNA ratio above the 75th percentile of all negative control oligo RNA/DNA ratios within the same sample, across all three technical replicates for at least two N2 lines or at least two N3 lines. To quantify enhancer activity differences between human and chimpanzee sequences, we computed log2(human [RNA/DNA]/chimpanzee [RNA/DNA]) for all expressed HARs. Importantly, it was not necessary to normalize RNA or DNA counts or these log-ratios by sequencing depth or other batch effects, because the human and chimpanzee sequences of each oligo were always assayed together in the same sample, leading to these biases cancelling out when computing log ratios. We did convert RNA and DNA counts into counts per million for plotting heatmaps of separate human and chimpanzee enhancer activity.

Modeling sequence origin, cell species, and cell stage effects on enhancer activity

To identify HARs with different enhancer activity between human and chimpanzee sequences (“cis effects”), we used the R limma package to fit a linear model for the log2(human [RNA/DNA]/chimpanzee [RNA/DNA]) of each HAR across all 24 samples (human and chimpanzee cells, N2 and N3 stages) and tested for mean log-ratios significantly different from zero. P-values were adjusted for multiple testing using the false discovery rate (FDR<1%). We also modeled HAR log2 (human [RNA/DNA]/chimpanzee [RNA/DNA]) ratios as a function of cell species of origin (human versus chimpanzee cells) and cell stage (N2 versus N3) with the R limma package (“trans effects”). We used these models to test for cis effects that depend on cell species or stage. The UpSetR package (Conway et al., 2017) was used to compare the different subsets of significant HARs from these analyses.

Modeling impact of HAR mutations on species-specific enhancer activity

For each of seven selected HARs with significant *cis* effects in our lentiMPRAs and prior evidence of enhancer activity (2xHAR.1, HAR34, 2xHAR.65, 2xHAR.142, 2xHAR.164, 2xHAR.170, 2xHAR.238), we designed a second MPRA library containing oligos carrying each single mutation, pair of mutations, et cetera ("permutations"). These represent all possible evolutionary intermediates between the human reference genome sequence and the chimpanzee reference genome sequence. Permutation oligos were assayed with lentiMPRA in three technical replicates of N2 and N3 cell lines from one human (WTC) and one chimpanzee (Pt2) following the same protocols described above. RNA and DNA count data was quantified as above, comparing each permutation oligo to the chimpanzee sequence: log2(permutation [RNA/DNA]/chimpanzee [RNA/DNA]). The one-hot encoded oligo sequence, along with cell species and stage, were used to model the log-ratios for a given HAR with penalized linear regression (R/glmnet 2.0-13, alpha = 1 [LASSO]) for feature selection. This allows the effect of each nucleotide with a human-chimp sequence difference, interactions between nucleotides, and interactions between nucleotides and cell species or stage (batch effects) to be estimated. Feature interactions from orders 1 to 9 were pre-computed using the PolynomialFeatures function in scikit-learn (0.19.1). The importance of interactions (orders 2 to 9) was assessed by examining features assigned the largest positive and negative coefficients by the penalized model. The value of glmnet's penalization strength (lambda) was chosen using the "1-SE" rule where the selected feature subset is the smallest of those within one standard error of the best performing subset.

Gene Ontology analysis

Gene Ontology (GO) terms associated with the 306 active HAR enhancers were compared to two background sets: (i) a random set of 20,000 phastCons elements, and (ii) all 714 tested HARs. Each element (phastCons or HAR) was mapped to the nearest gene and the resulting lists of genes were used as input to the GOrilla website (Eden et al., 2009) in the “two unranked lists” mode. We also tested for associations with GO terms and tissue-specific gene expression using GREAT (McLean et al., 2010).

Transcription factor binding site analysis

We scanned all differentially active HARs (p-value of differential expression between human and chimpanzee orthologs < 0.01) for presence of TRANSFAC vertebrate motifs (Matys et al., 2006). We first converted the TRANSFAC motifs to .meme format using the transfac2meme tool from the MEME suite (Bailey et al., 2009) and then scanned the sequences using FIMO (Grant et al., 2011) with a p-value cutoff of 10^{-5} for significance of motif hits. We then aligned the HAR orthologs and identified the TFBS that are specific to either species.

Analysis of HAR proximity to neurological phenotype-associated variants

Variants associated with neurological phenotypes were aggregated from the National Human Genome Research Institute (MacArthur et al., 2017) and the Psychiatric GWAS Consortium (Sullivan, 2010) datasets. These include GWAS SNPs from the following studies: PMC3880556, PMC21368711, PMC3925336, PMC3810676, PMC4033708, PMC4033708, PMC3714010, PMC20889312, PMC3714010, PMC3637176, PMC3714010, PMC3303194, PMC3827979, PMC4112379, PMC4940340, PMC4522619, PMC4667957, PMC3896259, PMC5695684, PMC4883595, PMC4879186.

TADs were determined using Juicer (Durand et al., 2016) on cortical plate and germinal zone Hi-C data from the Geschwind lab (Won et al., 2016). TADs were called at 5 Kb and 10 Kb resolution. We also called significant chromatin interactions between specific genomic regions at FDR<10%. For GM12878 cell line DpnII Hi-C data, we downloaded Juicer called TADs from (Rao et al., 2014) and also used lavaburst (Schwarzer et al., 2017) to independently call TADs on the same data. Python v3.5.2 with packages numpy (v 1.11.13; (van der Walt et al., 2011)), pandas (v 0.20.1; (McKinney, 2010)) and pyliftover (v 0.3; (Tretyakov, 2014)) were used for coordinate conversion to the hg19 assembly.

Plink v1.9 (Purcell et al., 2007) was used to calculate linkage disequilibrium (LD) per 1000 Genomes Project (Auton et al., 2015) phase 3 super-population for all bi-allelic SNPs with minimum minor allele frequencies of 5 percent. Python v3.5.2 with packages numpy (v 1.11.13; (van der Walt et al., 2011)) and pandas (v 0.20.1; (McKinney, 2010)) was used to determine GWAS SNPs that overlapped or were in LD with SNPs in HARs at various levels of genome-wide significance. SNPs with R^2 greater than or equal to 0.6 were considered to be in LD. Overlap and interaction analyses for TADs, genes, GWAS SNPs and HARs were completed using BEDTools v2.26.0 (Quinlan and Hall, 2010) and Python v3.5.2 with package pandas (v0.20.1; (McKinney, 2010)).

SUPPLEMENTAL FIGURES

Figure S1. Luciferase assays in chimpanzee and human N3 cells for positive and negative controls

(A) Relative luciferase activity in chimpanzee P2 N3 cells for Negative 1 (chr2:238,336,485-238,336,655), Negative 2 (chr7:96,637,215-96,637,385), Positive 1 (chrX:55,041,354-55,041,524) and Positive 2 (chr6:10,147,166-10,147,336).

(B) Relative luciferase activity in human WTC N3 cells for Negative 1 (chr2:238,336,485-238,336,655), Negative 2 (chr7:96,637,215-96,637,385), Positive 1 (chrX:55,041,354-55,041,524) and Positive 2 (chr6:10,147,166-10,147,336).

All coordinates are based on the hg19/GRCh37 build of the human reference genome.

Figure S2. Reproducibility of RNA/DNA ratios across technical replicates

RNA/DNA ratios for pairs of technical replicates of the 714-HAR lentiMPRA library show high reproducibility, with correlations above 0.95 for most pairs of replicates. For each cell line and stage, the three replicates are plotted pairwise in three panels. On a given panel (e.g., technical replicate 1 versus technical replicate 2), one dot is plotted for each human (blue) or chimpanzee (orange) HAR, with the 75th percentile of negative controls shown using purple lines. The Pearson correlation (“cor”) between the two replicates is shown. Human cell lines are “wtc” and “hs1”, chimpanzee cell lines are “pt2a” and “pt5c”; differentiation stages are N2 and N3; r1= first technical replicate, r2 = second technical replicate, r3 = third technical replicate.

Figure S3. Transgenic mouse embryos for HAR152, 2xHAR.133, 2xHAR.518 and 2xHAR.548

Figure S4. Reproducibility across technical replicates of permutation library

RNA/DNA ratios for pairs of technical replicates of the HAR permutation lentiMPRA library show high reproducibility. For each cell line and stage, the three replicates are plotted pairwise in three panels. On a given panel (e.g., technical replicate 1 versus technical replicate 2), one dot is plotted for HAR variant. The Pearson correlation (“cor”) between the two replicates is shown. Human cell line is “wtc”, chimpanzee cell line is “pt2a”; differentiation stages are N2 and N3; r1= first technical replicate, r2 = second technical replicate, r3 = third technical replicate.

Figure S5. Fitted models compared to the model with no interactions

Results of penalized regression modeling of HAR enhancer activity from the permutation lentiMPRA as a function of variants, cell stage, and cell species.

(A) Coefficient values (color scale) quantifying the importance of each feature (rows) to explaining variation in HAR enhancer activity across permutations carrying different variants and assayed in two cell stages and two cell species. Positive coefficients are associated with increased activity, negative coefficients are associated with decreased activity and zero indicates no association. The baseline (intercept) is all chimp nucleotides, cell stage=N3, and cell species=chimpanzee; NPC is cell stage=N2; WTC is cell species=human; other variables are variants: number before underscore indicates variant number in the HAR with the variant nucleotide listed after the underscore. Coefficient values are plotted across models (columns) allowing increasingly higher order interactions (degree 1 = each variable on its own: “main effects” only, mean = average across model degrees). Only the main effects with the largest absolute coefficients in the degree 1 models (left) are shown. As higher order feature interactions are added (moving to right), most main effects retain similar coefficients.

B) Similar to (A) but starting with the most complex model (degree 9; left) and showing only the interactions from that model with the largest absolute coefficients. Combinations of variants, cell species and cell stage that have large coefficients in the most complex model retain their importance when included in simpler models (moving to right), down to degree 3 to 5 when the maximum degree allowed becomes smaller than the interaction degree. ^ denotes a variable raised to a power (i.e., times itself) which should be interpreted similarly to the variable itself since variables are zeros and ones denoting if an oligo contains a variant or belongs to a cell type or stage.

SUPPLEMENTAL TABLES

Table S1. Oligonucleotides used for lentiMPRA

Table S2. Gene ontology enrichment results using either GOrilla or GREAT

Table S3. Results of HAR differential enhancer activity analysis for all expressed HARs

Table S4. HARs showing differential activity due to *trans* effects

Table S5. TFBS of differentially expressed HARs

Table S6. GWAS variants in the same TADs as HAR enhancers

Table S7. lentiMPRA permutation library results from penalized regression models allowing up to 4-way interactions

NOTE: Tables S1, S3, S5, S6, and S7 are in formats not supported by biorxiv and available by request from the authors.

Figure 1

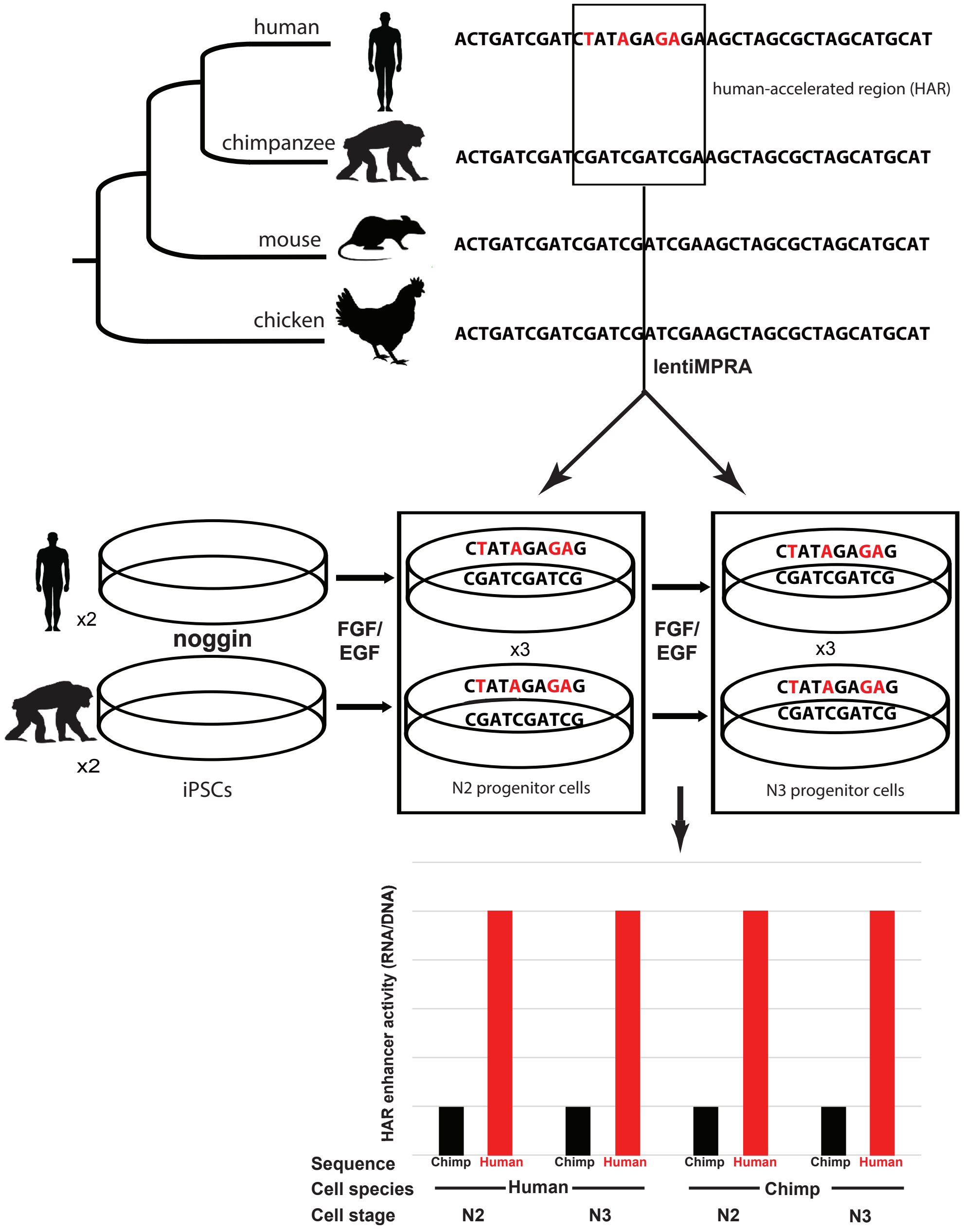

human
chimpanzee
mouse
chicken
ACTGATCGATCTATAGAGAGAAGCTAGCGCTAGCATGCAT
human-accelerated region (HAR)
ACTGATCGATCGATCGATCGAAGCTAGCGCTAGCATGCAT
ACTGATCGATCGATCGATCGAAGCTAGCGCTAGCATGCAT
ACTGATCGATCGATCGATCGAAGCTAGCGCTAGCATGCAT
lentiMPRA
x2
noggin
x2
iPSCs
FGF/ EGF
CTATAGAGAG
CGATCGATCG
x3
CTATAGAGAG
CGATCGATCG
N2 progenitor cells
FGF/ EGF
CTATAGAGAG
CGATCGATCG
x3
CTATAGAGAG
CGATCGATCG
N3 progenitor cells
HAR enhancer activity (RNA/DNA)
Sequence
Chimp Human
Chimp Human
Chimp Human
Chimp Human
Cell species
Human
Chimp
Cell stage
N2
N3
N2
N3

Figure 2

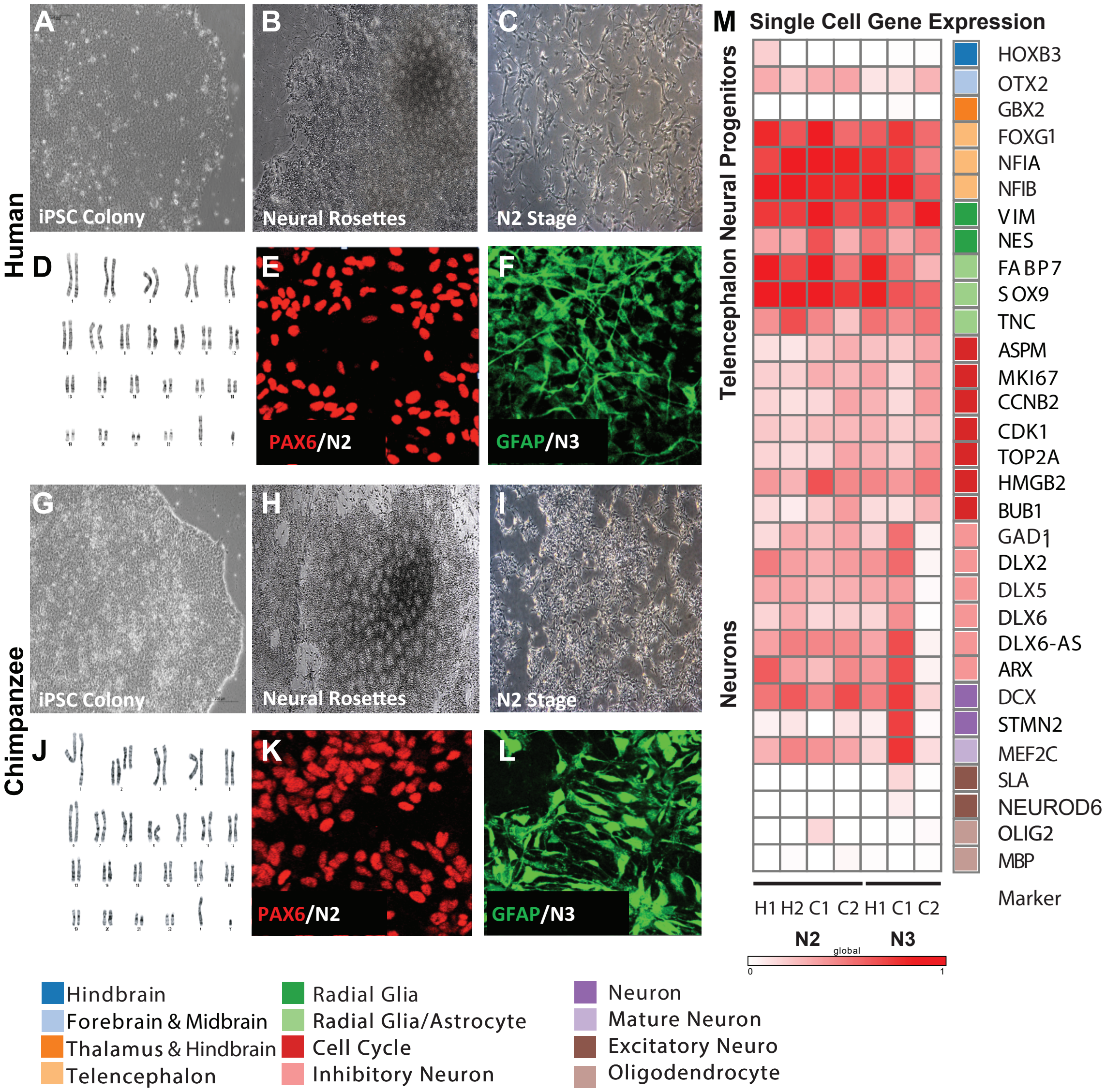

A
iPSC Colony
B
Neural Rosettes
C
N2 Stage
Human
D
E
PAX6/N2
F
GFAP/N3
G
iPSC Colony
H
Neural Rosettes
I
N2 Stage
Chimpanzee
J
K
PAX6/N2
L
GFAP/N3
M Single Cell Gene Expression
Telencephalon Neural Progenitors
Neurons
HOXB3
OTX2
GBX2
FOXG1
NFIA
NFIB
VIM
NES
FABP7
SOX9
TNC
ASPM
MKI67
CCNB2
CDK1
TOP2A
HMGB2
BUB1
GAD1
DLX2
DLX5
DLX6
DLX6-AS
ARX
DCX
STMN2
MEF2C
SLA
NEUROD6
OLIG2
MBP
H1 H2 C1 C2 H1 C1 C2
N2
N3
Marker
global
0
1
Hindbrain
Forebrain & Midbrain
Thalamus & Hindbrain
Telencephalon
Radial Glia
Radial Glia/Astrocyte
Cell Cycle
Inhibitory Neuron
Neuron
Mature Neuron
Excitatory Neuro
Oligodendrocyte

Figure 3

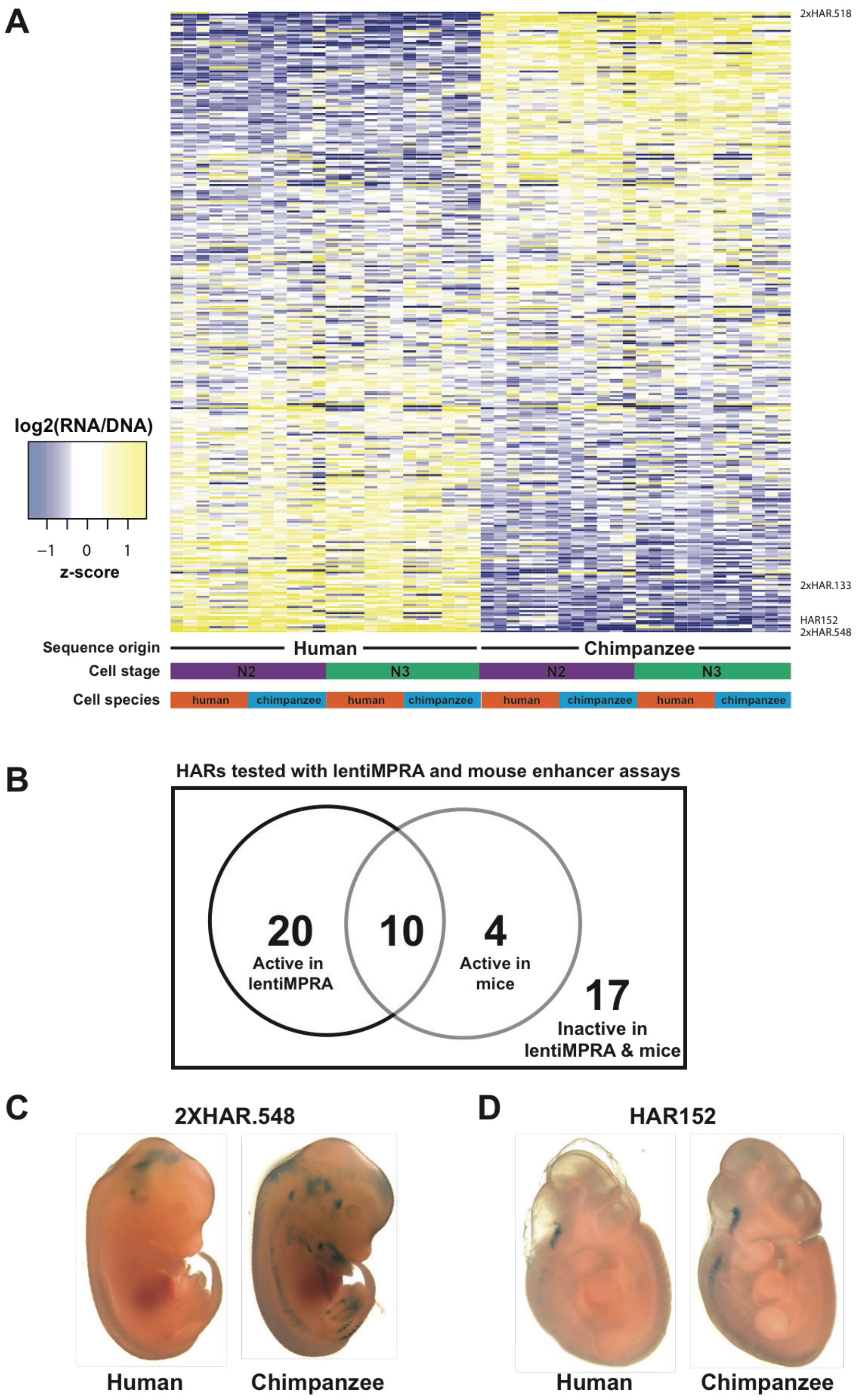

A
2xHAR.518
log2(RNA/DNA)
−1
0
1
z-score
2xHAR.133
HAR152
2xHAR.548
Sequence origin
Human
Chimpanzee
Cell stage
N2
N3
N2
N3
Cell species
human
chimpanzee
human
chimpanzee
human
chimpanzee
human
chimpanzee
B
HARs tested with lentiMPRA and mouse enhancer assays
20
Active in lentiMPRA
10
4
Active in mice
17
Inactive in lentiMPRA & mice
C
2XHAR.548
Human
Chimpanzee
D
HAR152
Human
Chimpanzee

Figure 4

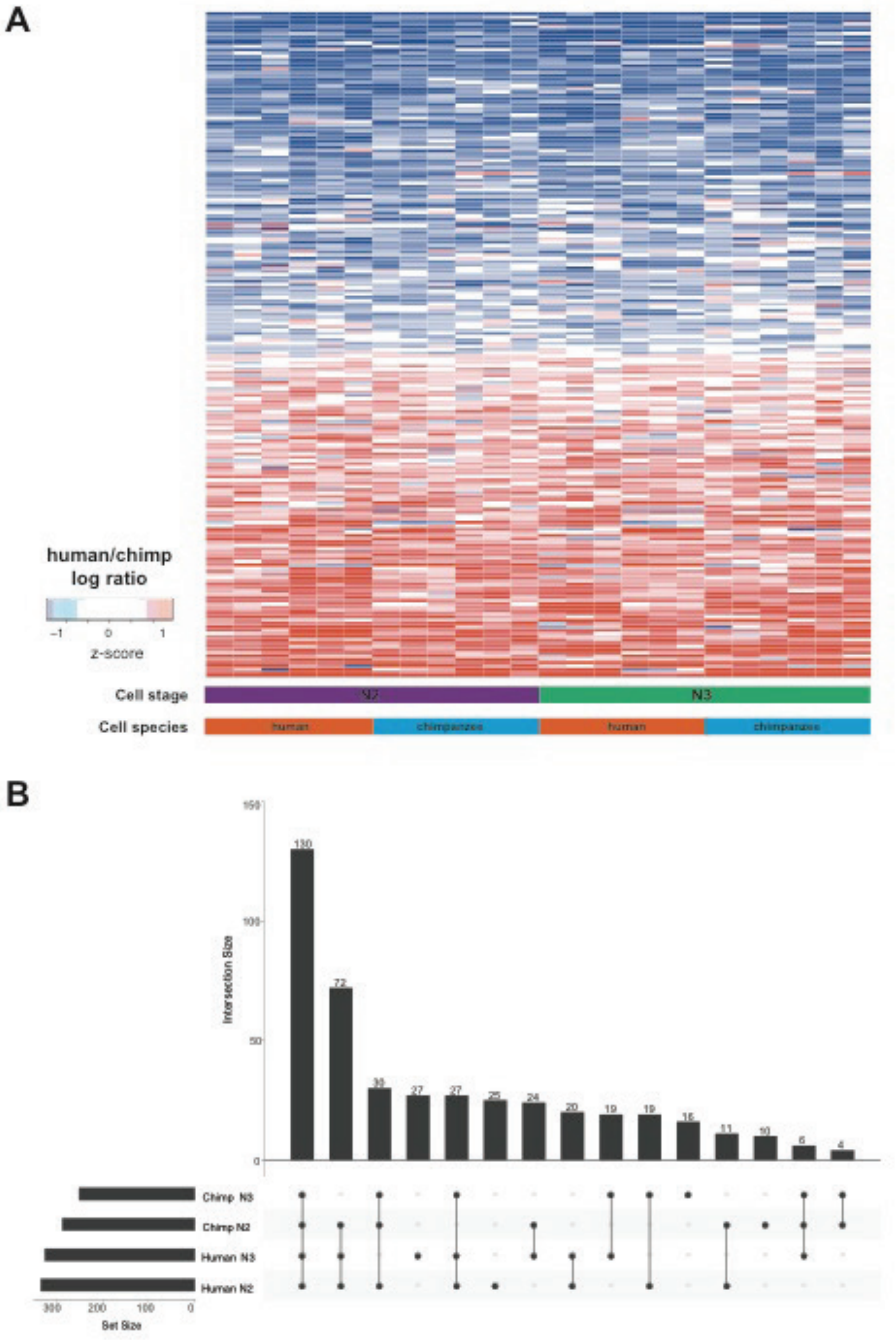
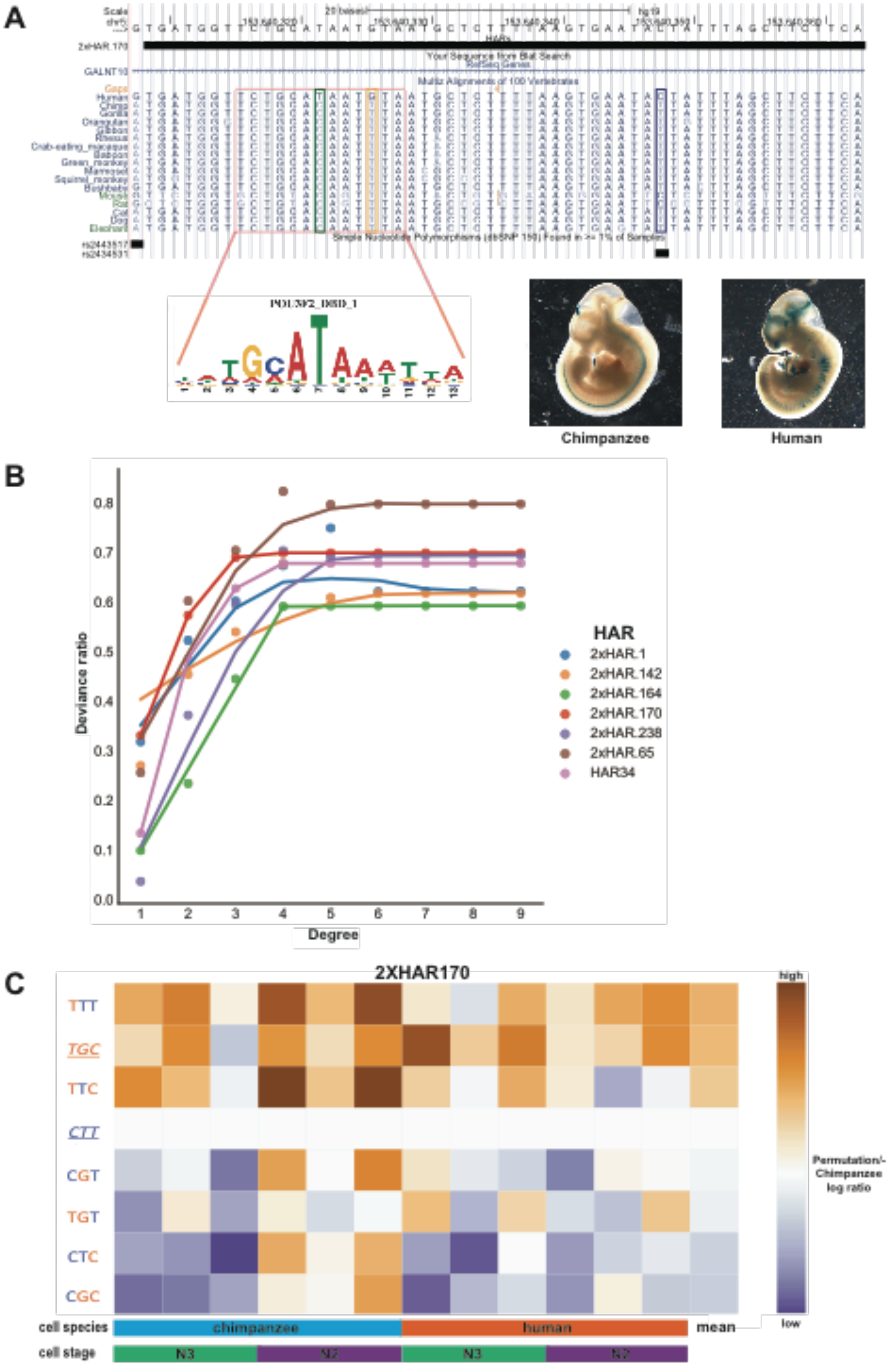

A
human/chimp
log ratio
−1
0
1
z-score
Cell stage
N2
N3
Cell species
human
chimpanzee
human
chimpanzee
B
150
100
50
0
Intersection Size
130
72
30
27
27
25
24
20
19
19
16
11
10
6
4
Chimp N3
Chimp N2
Human N3
Human N2
300
200
100
0
Set Size

Figure 5


A
Scale
chr5
20 bases
hg19
2xHAR.170
HARs
Your Sequence from Blat Search
GALNT10
RefSeq Genes
Multiz Alignments of 100 Vertebrates
Gaps
Human
Chimp
Gorilla
Orangutan
Gibbon
Rhesus
Crab-eating_macaque
Baboon
Green_monkey
Marmoset
Squirrel_monkey
Bushbaby
Mouse
Rat
Cat
Dog
Elephant
Simple Nucleotide Polymorphisms (dbSNP 150) Found in >= 1% of Samples
POU3F2_DBD_1
Chimpanzee
Human
B
Deviance ratio
Degree
HAR
2xHAR.1
2xHAR.142
2xHAR.164
2xHAR.170
2xHAR.238
2xHAR.65
HAR34
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
1
2
3
4
5
6
7
8
9
C
2XHAR170
TTT
TGC
TTC
CTT
CGT
TGT
CTC
CGC
high
Permutation/-Chimpanzee log ratio
low
cell species
chimpanzee
human
mean
cell stage
N3
N7
N3
N7