Comprehensive identification and analysis of human accelerated regulatory DNA

Rachel M. Gittelman[1], Enna Hun[2], Ferhat Ay[1], Jennifer Madeoy[1], Len Pennacchio[2], William S. Noble[1], David R. Hawkins[1,2], Joshua M. Akey[1*]

[1]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA.

[2] Lawrence Berkeley National Laboratory, Genomics Division, Berkeley, California 94701, USA

[3]Department of Pediatrics, University of Washington, Seattle, WA 98195, USA.

* Correspondence to:

Joshua M. Akey, PhD

akeyj@uw.edu

Department of Genome Sciences

University of Washington School of Medicine

Box 355065

1705 NE Pacific Street

Seattle, WA 98195

Abstract

It has long been hypothesized that changes in gene regulation have played an important role in human evolution, but regulatory DNA has been much more difficult to study compared to protein-coding regions. Recent large-scale studies have created genome-scale catalogs of DNase I Hypersensitive Sites (DHS), which demark potentially functional regulatory DNA. To better define regulatory DNA that has been subject to human specific adaptive evolution, we performed comprehensive evolutionary and population genetics analyses on over 18 million DHS discovered in 130 cell types. We identified 524 DHS that are conserved in non-human primates, but accelerated in the human lineage (haDHS), and estimate that 70% of substitutions in haDHS are attributable to positive selection. Through extensive computational and experimental analyses, we demonstrate that haDHS are often active in brain or neuronal cell types, play an important role in regulating the expression of developmentally important genes, including many transcription factors such as *SOX6*, *POU3F2*, and *HOX* genes, and identify striking examples of adaptive regulatory evolution that may have contributed to human specific phenotypes. More generally, our results reveal new insights into conserved and adaptive regulatory DNA in humans, and refine the set of genomic substrates that distinguish humans from their closest living primate relatives.

Introduction

A number of traits distinguish humans from our closest primate relatives, including bipedalism, increased cognition, and complex language and social systems (reviewed in (O'Bleness et al. 2012). To date, the genetic basis of human specific phenotypes remains largely unknown, complicated by the difficulties in distinguishing between phenotypically significant and benign variation. Thus, evolutionary changes in protein-coding sequences have received considerable attention, as the phenotypic consequences of these mutations have historically been easier to interpret (Clark et al. 2003; Chimpanzee Sequencing and Analysis Consortium 2005; Nielsen et al. 2005; Arbiza et al. 2006; Dennis et al. 2012; Sudmant et al. 2013; Stedman et al. 2004). Although protein-coding evolution has clearly played a role in human evolution, proteins account for only ~1.5% of the human genome, most of which exhibit high sequence similarity between humans and chimpanzees (Chimpanzee Sequencing and Analysis Consortium 2005). However, between ~2.5-15% of the human genome is estimated to be functionally constrained (Chinwalla et al. 2002; Lunter et al. 2006; Asthana et al. 2007; Meader et al. 2010; Ponting and Hardison 2011). Thus, the mutational target size of non-coding DNA is considerably larger than protein-coding sequences, suggesting that regulatory DNA is also an important substrate of evolutionary change, as originally proposed four decades ago (Britten and Davidson 1969; King and Wilson 1975). In some cases, detailed studies of individual genes have revealed human-specific regulatory evolution, such as in *FOXP2*, which is thought to have influenced traits related to speech and language in humans (Enard et al. 2002).

Nonetheless, interpreting patterns of interspecific divergence and intraspecific polymorphism in non-coding DNA has been considerably more challenging compared to protein-coding sequences. An elegant and powerful way to identify evolutionary changes in non-coding DNA of potential significance, originally described by Pollard et al. (2006b) and extensively used thereafter (Pollard et al. 2006b; 2006a; Prabhakar et al. 2006; Kim and Pritchard 2007; Bush and Lahn 2008; McLean et al. 2010; Lindblad-Toh et al. 2011; Pertea et al. 2011), focuses on the discovery of sequences that are rapidly evolving or lost on the human lineage, but otherwise phylogenetically conserved and thus likely functional. This approach has led to the discovery of several regions with species-specific enhancer activity (Prabhakar et al. 2008; Kamm et al. 2013; Capra et al. 2013), as well as human-specific deletion of regulatory DNA (McLean et al. 2011).

However, phylogenetic conservation is an imperfect proxy for function, particularly for non-coding regulatory sequences that can exhibit significantly high rates of turnover (Dermitzakis and Clark 2002; Wray et al. 2003; Villar et al. 2014). To more directly identify regulatory DNA, recent studies such as the ENCODE (The Encode Project Consortium 2012) and Roadmap Epigenomics projects (Bernstein et al. 2010) have created genome-scale maps of DNase I hypersensitive sites (DHS) in a large number of cell types. DNase I preferentially cleaves regions of open and active DNA, making it a powerful assay to identify regulatory elements, regardless of their specific function (Galas and Schmitz 1978; Dorschner et al. 2004). Although high-resolution maps of DHS now exist, not all experimentally defined regulatory elements are expected to be functionally or phenotypically significant (Eddy 2012; Niu and Jiang 2013; Graur et al. 2013; Doolittle 2013).

Thus, we hypothesized that the synergistic combination of comparative and functional genomics would facilitate the high-resolution identification of conserved and human accelerated regulatory sequences. Here we describe the genome-wide architecture and characteristics of 113,577 DHS that are conserved in primates and 524 DHS that exhibit significantly accelerated rates of evolution in the human lineage (haDHS). We estimate that ~70% of substitutions within haDHS are attributable to positive selection, experimentally validated a large number of elements, and perform extensive bioinformatics analyses that integrates information across multiple functional genomics data sets to better understand the functional and biological characteristics of haDHS.

Results

Framework for identifying conserved and human accelerated regulatory DNA

To identify human accelerated regulatory DNA, we leveraged experimentally defined maps of DHS from 130 cell types identified in the ENCODE and Roadmap Epigenomics Projects (Supplementary Table 1). After merging DHS across cell types into 2,093,197 distinct loci (median size = 290 bp, sd = 159 bp), we used a whole genome alignment of six primates from the EPO pipeline (Paten et al. 2008) to obtain separate alignments for each DHS, using strict filtering criteria for alignment quality. We performed two likelihood ratio tests to distinguish between DHS that are evolving neutrally, conserved among primates, or conserved among primates but accelerated in the human lineage (Fig. 1). Specifically, we used a maximum likelihood test (Pollard et al. 2010), to first identify 113,577 DHS that exhibit significant evolutionary constraint across primates, which manifest as regions of low sequence divergence compared to

carefully defined putatively neutral flanking sequence (FDR=0.01; Fig. 1). Next, for DHS that are conserved in primates, we performed a second likelihood ratio test (Pollard et al. 2010) and identified 524 regulatory sequences that have experienced a significant acceleration of evolution in the human lineage and therefore exhibit an excess of human specific substitutions (FDR=0.05; Supplementary Table 2; Fig. 1). Importantly, to avoid biasing ourselves against identifying human acceleration, we excluded the human sequence in the first test for conservation.

Characteristics of primate conserved regulatory DNA

We first characterized the set of DHS conserved across primates. Approximately 93% of conserved DHS overlap a phastCons conserved element, but many also contain short segments of less conserved sequence, making them overall less conserved than those identified by phastCons (Fig. 2a). We hypothesize that these less conserved sequences interspersed within DHS may facilitate the rapid acquisition of novel transcription factor binding sites, as these regions are already actionable (i.e., accessible to proteins) and poised to evolve new functions compared to non-conserved sequences outside of DHS.

Patterns of conservation varied significantly across cell type category (Kruskal-Wallis test; P=5.08 x 10^{-8}; Fig. 2b; Methods), ranging from 5.0% of DHS in chronic lymphocyte leukemia cells to 20.4% in fetal brain cells. DHS active in fetal cell types showed the highest levels of conservation, consistent with the observation that gene regulation in developmental pathways is highly conserved (Lowe et al. 2011). Conversely, DHS in malignant cell types exhibited the fewest conserved DHS, which

may reflect ectopic activation of chromatin (Vernot et al. 2012). These patterns are also observed in cell-type specific DHS (Supplementary Fig. 1a).

Genomic landscape of human accelerated regulatory DNA

We next investigated the set of human accelerated DHS (haDHS). Overall, these elements have evolved at approximately four times the neutral rate in the human lineage, while other primate branches have evolved at less than half of the neutral rate (Fig. 3a). In total, 70 haDHS overlap previously identified human accelerated elements (HAEs) (Pollard et al. 2006b; Prabhakar et al. 2006; Bush and Lahn 2008; Lindblad-Toh et al. 2011), which is highly significant (permutation $P < 1 \times 10^{-5}$; Fig. 3b). Thus, by focusing on experimentally defined regulatory DNA, we identify 454 novel loci that show accelerated rates of evolution in the human lineage, increasing the set of 1,621 merged HAEs by 28%. The number of cell types each haDHS was active in varied substantially (Supplementary Fig. 2). Notably, 64% (337) of haDHS were identified in at least one brain or neural cell type, and 88.5% (464) were active in at least one developing fetal tissue.

In comparison to conserved non-accelerated DHS, haDHS are significantly enriched in non-coding regions ($P=1.16 \times 10^{-7}$, hypergeometric test, Fig. 3c). These data are consistent with the hypothesis that non-coding regions are more free to evolve and acquire new functions. Furthermore, we observed eight regions where four or more haDHS were clustered within a 1Mb window, suggesting coordinated changes in multiple regulatory elements (Fig. 3d). For instance, *TENM3*, which is required for establishing neuronal connections in vertebrate retinal ganglion cells (Antinucci et al. 2013; Merlin et

al. 2013), is the nearest gene to five haDHS, four of which are active in retinal pigment epithelial cells (Fig. 3d, inset).

Adaptive evolution is the primary determinant of rate acceleration in haDHS

Human acceleration can result from both adaptive and non-adaptive forces (Haygood et al. 2007; Taylor et al. 2008; Kostka et al. 2012). We therefore performed a number of analyses to better understand mechanisms governing rate acceleration of haDHS. First, to distinguish between relaxation of constraint and true rate acceleration on the human lineage we applied a novel permutation test (Supplementary Text) and found that 91.8% of haDHS were evolving faster than their surrounding neutral sequence, suggesting that most haDHS are not the consequence of relaxed functional constraint. In contrast, it has been estimated that only 55% of HAEs exceed the neutral rate (Kostka et al. 2012). Second, we investigated the contribution of GC-biased gene conversion (GC-BGC) to our data, which influences rate acceleration of HAEs (Duret and Galtier 2009; Pollard et al. 2006a; Galtier and Duret 2007; Kostka et al. 2012), and found that 9.7% (51 haDHS) show significant evidence of GC-BGC (Supplementary Text; Supplementary Figure 3a). Finally, we investigated patterns of human-macaque divergence around haDHS and found that local increases in mutation rate cannot explain rate acceleration in haDHS, although mutation rate heterogeneity has influenced previous inferences of HAEs (Supplementary Text; Supplementary Figure 3b).

To more directly quantify the proportion of substitutions in haDHS that can be attributed to positive selection, we used the McDonald-Kreitman framework and compared levels of polymorphism and divergence at haDHS. Specifically, we used

polymorphism data from the 1000 Genomes Project (The 1000 Genomes Project Consortium 2012) and calculated the statistic α, an estimate of the proportion of substitutions fixed by adaptive evolution. As a control, we first estimated α in conserved, non-accelerated DHS, which as expected was 0 (95% CI -0.02-0.007; Fig. 4a; Supplementary Fig. 4a). We estimate that 70.1% (95% CI 65.8%-73.7%) of substitutions can be attributed to positive selection in haDHS (Fig. 4a), and this number is robust to mutation rate heterogeneity in the presence of complex demographic history (Supplementary Text; Supplementary Fig. 4b). To evaluate the sensitivity of α to GC-BGC, we removed all weak to strong substitutions in haDHS and repeated the analysis. Although estimates of α decreased for haDHS subject to GC-BGC, α increased slightly for other haDHS and thus the overall estimate remained almost identical (69.9%, 95% CI 64.2%-75.2%; Fig. 4a). Of the remaining 29.9% of substitutions in haDHS not accounted for by positive selection, we estimate 9.0% are expected without human specific rate acceleration and 20.9% are attributable to additional factors such as relaxation of constraint (Fig. 4b). In support of this hypothesis, we find increased levels of nucleotide diversity in haDHS and HAEs (Supplementary Text, Supplementary Fig. 5).

haDHS are developmental enhancers that exhibit lineage specific activity

We performed extensive experimental studies to better understand the functional significance and potential regulatory roles of haDHS. We found that nine of our haDHS had previously been tested for *in vivo* enhancer activity using a transgenic mouse assay (Visel et al. 2007), and we tested nine additional loci. Overall, 13 out of 18 haDHS were positive for enhancer activity in one or more tissues at the single time point assayed

(e11.5; Supplementary Table 3). These 13 haDHS were active in a wide range of tissues (Fig. 5a), with the midbrain (n=7), forebrain (n=4), branchial arch (n=4), and limb (n=4) the most frequent tissues showing enhancer activity. Patterns of enhancer activity varied from very broad to very tissue specific (Fig. 5a). One interesting example is located on 11p15, and is only active in the branchial arch (Fig. 5a). This haDHS is located in an intron of *SOX6*, and as we describe below, we find evidence that it contacts the *SOX6* promoter. *SOX6* is a developmental transcription factor involved in brain, bone, and cartilage development (Lefebvre et al. 1998). Notably, the branchial arch develops into several structures, including the jaw and larynx (Graham 2003), making this haDHS an intriguing candidate that potentially influences traits such as facial morphology and speech.

We also performed luciferase assays to functionally test haDHS in a more high-throughput manner. Specifically, we experimentally tested 37 haDHS in SK-N-MC cells (derived from a neuroepithelioma) and 20 haDHS in IMR90 cells (fetal lung fibroblasts) by assaying for differences in regulatory activity of the human and chimpanzee orthologs using luciferase reporters. We chose SK-N-MC cells as a proxy for other neural cell type, and IMR90 cells because many haDHS were active in this cell type. Of the 37 pairs of haDHS tested in SK-N-MC, 14 showed significant enhancer activity ($P < 0.05$; Fig. 5b; Supplementary Fig. 6a), of which five (35%) exhibited significant differences between the human and chimpanzee haplotypes ($P < 0.05$; Fig. 5b; Supplementary Fig. 6a; Supplementary Table 4). In IMR90, 5 out of 20 haDHS showed significant evidence of enhancer function ($P < 0.05$; Fig. 5c; Supplementary Fig. 6b; Supplementary Table 4), one (20%) of which exhibited significant differences in expression between the human

and chimpanzee haplotypes. Human substitutions resulted in lower expression in four of the six haDHS with significant differences in reporter activity between human and chimpanzee sequences (Fig. 5b,c). The haDHS with the largest difference in regulatory activity between humans and chimpanzees (2.32-fold increase in chimpanzees; P=0.004) had five human-specific substitutions that overlapped several transcription factor binding motifs, and was located 186 base pairs upstream of *RNF145*, a zinc finger gene that is associated with variation in hematological traits (Soranzo et al. 2009) (Fig. 5d). Although this haDHS is likely part of the promoter for *RNF145*, as described below, it may target several other genes including *IL12B* and *CLINT1*.

Leveraging chromatin contact data to infer putative regulatory targets of haDHS

Delineating the set of target genes that haDHS regulate is key to determining their biological consequences and role in human evolution. However, identifying the targets of regulatory sequences poses a significant challenge. Enhancers often regulate distal genes, and in some cases these may not be the closest genes to the enhancer (van Arensbergen et al. 2014). Chromatin conformation technologies such as Hi-C (Lieberman-Aiden et al. 2009) identify physical contacts between distinct segments of DNA and have been shown to identify long-range interactions between promoters and enhancers (Sanyal et al. 2012). We leveraged high-coverage Hi-C data from human IMR90 fibroblast cells to identify putative regulatory targets of haDHS using a rigorous statistical method (Ay et al. 2014). We identified 9,000 significant contacts for the 524 haDHS at 40kb resolution (FDR=0.01, Fig 6a). On average, haDHS overlap transcription start sites for 3.5 genes, highlighting the potential benefit of using more sophisticated strategies than simply

identifying the nearest gene when inferring regulatory targets. We also found that haDHS contact fewer genes on average than conserved DHS (permutation P=0.004), suggesting adaptive regulation is more likely to occur when pleiotropic effects are minimized. Furthermore, 119 haDHS contact one or more transcription factors, and in total 132 distinct transcription factors are contacted by haDHS. These include *SOX6* (see Fig. 5a), *RUNX2*, and multiple *HOX* genes, all of which play important roles in development.

We performed a GO enrichment analysis on the set of genes whose transcriptional start sites are contacted by haDHS. Because haDHS are a subset of conserved DHS, we first performed the analysis on conserved DHS contact regions compared to the genomic background. We found that conserved DHS contacts are highly enriched for developmental genes, including those involved in neuron development (Supplementary Table 5), consistent with previous observations about conserved noncoding sequence (Lowe et al. 2011). Next, we tested for GO enrichments in haDHS contact genes using conserved DHS contact genes as the background and found a significant enrichment for developmental terms, including brain and neuron development (corrected $P < 0.05$; Supplementary Table 5). These results show that haDHS target genes are enriched for developmentally and neuronally important genes relative to conserved DHS, which themselves are already highly enriched for these categories.

Three examples of haDHS and their putative target regions are shown in Figure 6b-d. All contain transcription factor motifs that are dramatically strengthened or weakened by human-specific substitutions. These haDHS are likely targets of adaptive evolution as they show no evidence of GC-BGC and are evolving faster than surrounding neutral sequence. Moreover, all three are also active in only a small number of neuronal

cell types, such as fetal brain and fetal spinal cord, indicating a potential role in human-specific cognitive phenotypes. Of particular interest is an haDHS on Chromosome 6 that lies in a gene desert 300kb from *POU3F2*, a transcription factor that regulates *FOXP2* in a human-specific manner (Maricic et al. 2013) (Fig. 6c). Two of the substitutions in this haDHS strengthen a putative YY1 transcription factor binding site (Fig. 6c), which is known to mediate long distance DNA interactions (Atchison 2014).

Discussion

Advances in DNA sequencing technology have led to a vast catalogue of the variation in the genomes and epigenomes across many primates. However, interpreting the evolutionary, functional, and phenotypic significance of these differences and identifying the precise genetic changes that are causally related to human specific traits remains a formidable challenge. Here, we have leveraged extensive maps of experimentally defined regulatory DNA and comprehensive comparative and population genomics analyses to identify and delimit the characteristics of conserved and human accelerated regulatory DNA. In total, we discovered 113,577 DHS conserved in primates, 524 of which exhibit significant rates of acceleration in the human lineage.

We found marked heterogeneity in the distribution of conserved DHS across cell types (Fig. 2b), with fetal cell types showing the largest amount of constraint. Conversely, DHS in malignant cell types exhibited the lowest levels of conservation, an observation that may provide insight into cancer biology. For example, chromatin remodeling is disrupted in many cancers (Morin et al. 2010; Jiao et al. 2011). Previous work has shown that DHS in malignant cell types are more likely to be cell type specific and have levels of nucleotide diversity consistent with neutral evolution (Vernot et al.

2012). Thus, these observations combined with our results that DHS in malignant cell types have low levels of evolutionary conservation suggest that many malignant DHS may reflect ectopic chromatin activation.

Our results also provide new insights into human specific adaptive regulatory evolution. Of the 524 haDHS that we identified, 454 (87%) are novel and were not detected in previous studies of HAEs (Pollard et al. 2006b; Prabhakar et al. 2006; Bush and Lahn 2008; Lindblad-Toh et al. 2011). The haDHS that we discovered are significantly less affected by GC biased gene conversion and relaxation of functional constraint, and have a higher proportion of substitutions that are estimated to be due to positive selection compared to previous catalogs of HAEs (Supplementary Figure 3). We hypothesize these differences are largely the consequence of our study design that synergistically integrated experimentally defined regulatory sequences with phylogenetic conservation, which both focused our analyses to a subset of the genome enriched for functionally important sequence and limited the influence of confounding evolutionary forces. To support this hypothesis, we find that a higher proportion of haDHS overlap human-specific enhancer marks in the cortex (Reilly et al. 2015) than HAEs (P=7.62 x 10^{-5}; Fisher's exact test). Large catalogs of experimentally defined regulatory DNA did not exist when HAEs were initially discovered, and we anticipate that the continued development of functional genomics technology will enable even more refined evolutionary analyses than described here.

To help interpret the functional and potential phenotypic significance of haDHS, we performed extensive bioinformatics analyses and experimental validations. We found that haDHS were significantly enriched in non-coding regions, a large proportion of

experimentally tested elements showed enhancer activity, and many were active in brain or neural cell types and during fetal development. We also used Hi-C data to inform inferences of putative target genes that are regulated by haDHS. These analyses revealed that haDHS contact the transcriptional start sites of 132 transcription factors, suggesting that fine-tuning regulatory networks by tinkering with the sequences that govern the expression of regulatory proteins has been an important target of positive selection during human evolution. A number of transcription factors contacted by haDHS are strong candidates for influencing hominin or human specific traits. For example, *RUNX2* has been hypothesized to influence differential bone morphology in humans and Neanderthals (Green et al. 2010), and *HOX* genes play myriad roles in development. Another intriguing transcription factor contacted by a haDHS is *POU3F2*, which has recently been shown to regulate *FOXP2* in a human-specific manner (Maricic et al. 2013). *FOXP2* itself is a transcription factor that has previously been hypothesized to play a role in speech and language in humans (Enard et al. 2002). Our findings suggest that there may be additional levels of human-specific *FOXP2* regulation via differential expression of *POU3F2* expression. Furthermore, in addition to transcription factors, we identified other genes that are of significant biological interest. For instance, *PEX2* is contacted by a haDHS with two substitutions that create a SMAD4 motif (Fig. 6b). Mutations in *PEX2* can lead to Zellweger Syndrome, characterized by a constellation of features including impaired brain development and craniofacial abnormalities (Steinberg et al. 2006).

Our study has a number of important limitations. For example, the DHS we used were ascertained only in human tissues. Although experimentally defined regulatory

DNA has been generated in a limited number of non-human primates for a limited number of tissues (Shibata et al. 2012; Cotney et al. 2013), a more systematic and comprehensive effort would be of considerable value in understanding the evolution of regulatory sequences. Furthermore, we did not consider additional types of genetic variation, such as structural variation, that may influence human-specific phenotypes (Dennis et al. 2012; Sudmant et al. 2013). Furthermore, although there is evidence that chromatin conformation is relatively stable across cell types (Dixon et al. 2012), it would be of considerable interest to generate Hi-C or related data for a more comprehensive panel of cell types. These data, combined with gene expression profiles from the same tissue types, would provide further insights into the target genes regulated by haDHS. Finally, the transgenic mouse and luciferase assays that we performed are only a first step in the experimental characterization of these and other elements that potentially contribute to human specific phenotypes. Because the activity of a regulatory element may be highly cell type and developmental time point specific, and depend on the coordination of additional regulatory elements, more extensive *in vivo* experiments would be fruitful. Nonetheless, associating particular haDHS with specific phenotypes is complicated by the fact that the putative causal alleles are fixed in humans and thus refractory to traditional genetic mapping methods. However, if mutations at these sites are not lethal, given the current global population size of humans, such mutations are expected to exist and their discovery could provide valuable phenotypic insights.

In short, our data provide substantial new insights into sequences that have experienced human specific adaptive regulatory evolution, narrow the set of genetic

changes that may influence uniquely human phenotype, and facilitate more detailed experimental and animal models of the most promising human specific substitutions. Ultimately, delineating the suite of genetic changes that have causally influenced human specific phenotypes will provide insight into the evolutionary and molecular mechanisms that shaped our species evolutionary trajectory.

Methods

DNase I Hypersensitivity Sites

We used DnaseI Hypersensitivity peaks previously published as part of the ENCODE (The Encode Project Consortium 2012; Maurano et al. 2012) and Roadmap Epigenomics (Bernstein et al. 2010) projects. A list of cell types is available in Supplementary Table 1. All peaks were called using the hotspot algorithm (John et al. 2011), and represent the 150bp region of maximal DnaseI signal. We merged DHS across cell types using the BEDOPS package (Neph et al. 2012). Many DHS were very long after merging (>2000 bp), probably because they consist of distinct regulatory elements located in close succession along the genome. To avoid analyzing distinct, potentially independently evolving regulatory elements as a single unit, we segmented merged DHS according to the number of cell types each region was active in (Supplementary Text).

Primate Alignments

We downloaded the six primate EPO alignment from Ensembl version 70 (Flicek et al. 2014). Using this we obtained an alignment for each DHS and the surrounding 50kb of sequence. We masked all sites that were polymorphic in the 1000 Genomes Project

(The 1000 Genomes Project Consortium 2012) integrated phase 1 data (March 2012) at less than 95% allele frequency, all repeat masked bases (lower case mark up in the EPO alignment), and all sites that were part of a CpG in any species in the alignment. In the surrounding 50kb we additionally masked all segmental duplications (UCSC Table Browser), coding exons (UCSC RefSeq genes) padded by 10 base pairs in order to remove splice sites, promoters (500bp upstream of transcription start sites), other DNase I Hypersensitive sites, and phastCons Eutherian mammal and primate conserved elements (UCSC phyloP46way). This helped ensure that the 50kb surrounding region was a more appropriate approximation of the neutral evolutionary model for each DHS. We filtered any DHS in which a) fewer than 90% of the bases remained unmasked in the DHS, or b) fewer than 15kb remained unmasked in any of the 6 primates in the neutral region. Note, the EPO alignment is based on GRCh37 (hg19), and all subsequent analyses were done using GRCh37 coordinates. Given that we focus on conserved elements, which are by definition located in regions of the genome that are well resolved and alignable, we do not anticipate realigning to GRCh38 would significantly affect our results.

Identifying conserved and accelerated DHS

DHS that passed filtering were tested for overall conservation along the primate lineage with software from the PHAST package (Pollard et al. 2010; Hubisz et al. 2011). For each DHS we first ran phyloFit on the neutral alignment of the surrounding 50kb with the parameters –nrates 4 –subst-mod SSREV –EM. We used the newick tree provided with the 6 primate alignment in Ensembl. The resulting file was used as the neutral model while running phyloP. PhyloP was run with the parameters –-method LRT

–mode CON after removing human sequence from the alignment. DHS that were conserved at an FDR of 1% as determined with the Q-value package (Dabney and Storey) for R (R Core Team) were then tested for human acceleration. For this test we used the same neutral model of evolution, this time using the parameters –method LRT –mode ACC –subtree homo_sapiens. DHS significant for human acceleration at an FDR of 5% were considered in further analyses. We evaluated the accuracy of the FDR using a sampling approach (Supplementary Text).

To determine the overall rate of evolution in the neutral regions compared to haDHS, we first concatenated sequence from both sets of regions, and then conducted the same set of tests on the regions as a whole. To determine how much faster the human branch in the haDHS was evolving compared to the expected rate, we multiplied the estimated neutral human branch length by the estimated conservation scale factor, and divided the actual haDHS human branch length by this expected number.

Distribution of DHS across cell types and genomic location

To determine how conserved and accelerated DHS were distributed across cell types we used the bedmap program from the BEDOPS suite (Neph et al. 2012) to map DHS from individual cell types onto the set of merged DHS. We then calculated the proportion of DHS in each cell type that were called as conserved and the proportion of conserved DHS that were also called as accelerated (Fig. 2b; Supplementary Fig. 1a-c).

Distribution of DHS and haDHS across the genome was assessed using UCSC known gene annotations from the UCSC Genome Browser, downloaded on May 14, 2013. Annotations were filtered to contain only “canonical” transcripts from the

knownCanonical table. Promoters were defined as the 500bp upstream of a transcription start site. To identify physical clusters of haDHS we expanded each haDHS by 500kb on either side and then used the bedmap –count command from the BEDOPS suite (Neph et al. 2012) to count the number of haDHS and conserved DHS within each 1Mb region.

Other Human Accelerated Elements

We obtained previously identified human accelerated elements (HAEs) (Pollard et al. 2006b; 2006a; Prabhakar et al. 2006; Bush and Lahn 2008; Lindblad-Toh et al. 2011) and assessed overlap using the bedmap program from the BEDOPS package (Neph et al. 2012). When comparing our haDHS to these other HAEs, we merged all HAEs, again using the BEDOPS program. It was useful for us to compare haDHS to DHS that were conserved but not accelerated. In order to do similar analyses using the HAEs, we merged phastCons eutherian mammal and primate elements (UCSC Genome Browser) and considered any element that was longer than 100bp.

To determine if the amount of overlap between haDHS and other HAEs was significant, we created an empirical null distribution by randomly sampling 524 conserved DHS 10^4 times and determining overlap with HAEs for each sample.

Population genetics analyses

We downloaded the phase1 integrated release data from the 1000 Genomes Project (The 1000 Genomes Project Consortium 2012) and filtered sites according to several criteria (Supplementary Text). We calculated α as described previously (Charlesworth 1994), using the equation $1-(P_sF_n/P_nF_s)$ where P= number of polymorphic

sites, F= number of human specific substitutions, S= number of selected sites, N= number of neutral sites. We considered bases within haDHS to be putatively selected, and bases in the surrounding 4kb region to be putatively neutral.

Hi-C Analyses

We obtained raw paired-end Hi-C libraries for IMR90 fibroblasts two cell lines (Dixon et al. 2012). Although Hi-C data was also available from human embryonic stem cells, we chose not to include this cell type as it may have a more permissive chromatin landscape that is not representative of promoter/enhancer interactions (Dixon et al. 2012). We processed the Hi-C data for each cell line at 40 kb resolution as described in (Ay et al. 2014). Briefly, we mapped reads to the hg19 (GRCh 37) reference sequence, pairing mapped read ends, filtering duplicates, binning at 40 kb resolution, normalizing raw contact maps (Imakaev et al. 2012), and assigning statistical confidences for each contact bin pair using Fit-Hi-C with a refined null (Ay et al. 2014). We used a significance threshold of q-value <0.01 to determine regions that are contacted by haDHS containing 40 kb windows. We omitted contacts within the same window and between adjacent windows and only focused on intra-chromosomal contacts within 5 Mb of haDHS. Note that the binning at a coarse resolution and omission of inter-chromosomal contacts were done to identify only high confidence contacts with enough sequencing coverage. We used RefSeq gene annotations to obtain a list of transcription start sites that overlap contact regions and used these to perform GO analyses using the WebGestalt server (Wang et al. 2013) with the multiple testing method set to BH and the minimum number of genes per category to 10.

Transgenic Mouse Assays

Transgenic mouse assays were performed as previously described (Visel et al. 2007). Note, one of the previously tested assays was performed with the mouse ortholog (see Supplementary Table 3). Images of all the mouse assay replicates are available on the VISTA Enhancer Browser (Visel et al. 2007).

Luciferase Assays

We considered several factors when selecting which haDHS to experimentally study. First, because the luciferase assays detect enhancers, we prioritized haDHS showing evidence of enhancer activity. To this end, we identified a second set of haDHS that were within 500bp of an enhancer histone modification (H3K4me1, H3K27ac) signal identified in the same cell type. Histone modifications for this set of haDHS were downloaded from the UCSC Genome Browser or the Roadmap Epigenomics website. We included only DHS from the 20 cell types for which histone modification data was available (see Supplementary Table 6 for additional set of haDHS and the cell types used). There is a column identifying which haDHS were used in the luciferase assays in Supplementary Tables 2 and 6. Second, we prioritized haDHS that were active in IMR90, SK-N-MC or other similar cell types. Both cell types represent time points that are potentially interesting for studying human evolution: SK-N-MC is a brain cell type, and IMR90 is a fetal tissue. Finally, we prioritized haDHS that showed the greatest evidence for human-acceleration.

We used standard techniques for cloning, transfection, and performing luciferase assays. Details are provided in the supplement. For the luciferase assays, each allele and control had three to eight replicates. The positive control for each plate was cells transfected with the pGL3 control plasmid containing a minimal promoter with strong SV40 enhancer, while the negative control for each plate was cells transfected with the empty pGL3 plasmid with minimal promoter but no additional sequence cloned in.

To increase power to detect enhancer activity, negative control replicates were normalized by plate so that they could be directly comparable and combined. To accomplish this we used the lm() function in R (R Core Team) to create a linear model where the ratio of firefly to *Renilla* for all negative control replicates was a function of plate number. Then the coefficient for each plate was subtracted from all data points for that plate. Enhancer activity was determined using a one sided *t*-test, and haDHS were considered enhancers if either the chimp and/or human allele showed greater luciferase activity than the negative controls. We then tested enhancers for allelic differences with a two-sided *t*-test between the human and chimp alleles.

Acknowledgements

This work was supported the NIGMS grant GM110068 to JMA. RMG was supported by an NSF graduate research fellowship. L.A.P. was supported by NHGRI grants R01HG003988, and U54HG006997, and research was conducted at the E.O. Lawrence Berkeley National Laboratory and performed under Department of Energy Contract DE-AC02-05CH11231, University of California. WSB was supported by NIH grant U41HG007000.

Figure Legends

Fig. 1. Identifying evolutionarily conserved and accelerated human regulatory sequences. Schematic shows the framework for identifying DHS that are conserved in primates but accelerated in the human lineage. DHS appear as peaks of high coverage along the genome and are merged across cell types. An alignment (purple boxes) of six primates is obtained for each DHS and the neutral sequence surrounding them. Black bars represent any sequence that differs from the human sequence, except in the case where all species differ from human, which are represented as blue bars in the human sequence. Dotted red lines indicate the location of the DHS.

Fig. 2. Patterns of conservation vary across cell types. (a) Cumulative distribution of single base phyloP scores are shown for four-fold degenerate sites, conserved DHS, and phastCons elements. The dotted grey line indicates a cumulative distribution of zero or one. (b) The proportion of conserved DHS in each of the 130 cell types, ordered in increasing amounts of conservation. Colors denote four cell type categories: normal (purple), fetal (blue), pluripotent (yellow), or malignant (red). The inset violin plot shows the distribution of the proportion of conserved DHS for each cell type category. Cell type names at each end of the spectrum are shown for comparison.

Fig. 3. Characteristics of human-accelerated DHS. (a) Overlaid phylogenetic trees inferred in haDHS (maroon) versus their flanking neutral regions (grey). The human branch is highlighted by the dashed rectangle. (b) Venn diagram showing overlap of haDHS with human accelerated elements identified in previous studies (c) The proportion

of bases in haDHS and conserved DHS that are located in different functional classes of genomic sequence. (d) Distribution of haDHS across the genome. Each vertical bar on the chromosome ideogram represents a haDHS. The inset plot shows a region on Chromosome 4 near the *TENM3* gene that contains five haDHS. The 4th haDHS is enlarged to show that it is accessible in retinal pigment epithelial cells (blue), and is flanked by an H3K27ac signal (pink). Human substitutions are shown in red (weak to strong) and black (all others).

Fig. 4. Factors contributing to rate acceleration of haDHS. (a) Estimates of the proportion of adaptive substitutions, α, and 95% bootstrap confidence intervals for different classes of haDHS. Red and blue denote estimates that include or exclude weak to strong mutations, respectively. (b) Pie chart summarizing the proportion of substitutions in haDHS inferred to be influenced by different factors. Note, expected indicates the proportion of substitutions assuming rates of evolution in the human lineage were the same as that in non-human primates. Other denotes substitutions due to other factors such as relaxation of constraint or mutation rate heterogeneity.

Fig. 5. Experimental assays of enhancer activity in haDHS. (a) A schematic of the transgenic mouse model is depicted. Rows in the table correspond to each embryonic region, and numbers in parentheses indicate how many of the haDHS were positive in the region indicated. Columns represent the 13 haDHS that showed enhancer activity, and grey boxes indicate what tissues the haDHS was active in. Three examples of positive assays are shown above, along with a schematic depicting their location relative to nearby

genes. The haDHS tested is shown in red, and other haDHS in the region are shown in black. Panels (b) and (c) show results from Luciferase assays for haDHS that showed significant enhancer activity in SK-N-MC and IMR90 cells, respectively. Dotted lines indicate the mean relative expression from the negative controls, and the grey box indicates haDHS human and chimpanzee sequences that showed significantly different activity ($P < 0.05$). Bars indicate standard error. Stars below each plot indicate haDHS that were active in SK-N-MC or IMR90 (other haDHS were active in similar cell types, such as fetal brain or NHLF). (d) A schematic of the region surrounding haDHS12, which had the largest difference in enhancer activity. The haDHS is located just upstream of the alternatively spliced gene *RNF145*. Red substitutions are weak to strong, and all other substitutions are colored in blue. PhyloP scores are also shown across the region. This DHS was partitioned prior to statistical testing in to two distinct DHS. The red portion is human accelerated, and the black portion is not.

Fig. 6. Hi-C chomatin conformation data identify putative regulatory targets of haDHS. (a) Contacts are shown for all haDHS, and each row indicates the contacts for one haDHS, which is in he center. Black boxes indicate one 40kb contact region. The schematic above illustrates how chromatin conformation information gets translated in to the Hi-C contact data. Blue dots represent contact regions, and the red dot indicates an haDHS. (b-d) Three example haDHS are shown with their surrounding genes and a predicted transcription factor binding site that is affected by a human specific mutation(s). Genes that contact the haDHS in Hi-C data are highlighted in blue, with arrows pointing to their transcription start sites. Examples (b-c) depict substitutions that

create transcription factor binding sites, while (d) is a binding site that is predicted to be lost in humans. Human specific substitutions that go from a weak to a strong base are shown in red, while all other substitutions are shown in blue. Bar plots represent FIMO (Grant et al. 2011) log likelihood ratios of motif calls in each species.

References

Antinucci P, Nikolaou N, Meyer MP, Hindges R. 2013. Teneurin-3 specifies morphological and functional connectivity of retinal ganglion cells in the vertebrate visual system. *Cell Rep* **5**: 582–592.

Arbiza L, Dopazo J, Dopazo H. 2006. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS Comput Biol* **2:** e38.

Asthana S, Noble WS, Kryukov G, Grant CE, Sunyaev S, Stamatoyannopoulos JA. 2007. Widely distributed noncoding purifying selection in the human genome. *Proc Natl Acad Sci USA* **104**: 12410–12415.

Atchison ML. 2014. Function of YY1 in Long-Distance DNA Interactions. *Front Immunol* **5**: 45.

Ay F, Bailey TL, Noble WS. 2014. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Research* **24:** 999–1011.

Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, et al. 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28:** 1045–1048.

Britten RJ, Davidson EH. 1969. Gene regulation for higher cells: a theory. *Science* **165**: 349–357.

Bush EC, Lahn BT. 2008. A genome-wide screen for noncoding elements important in primate evolution. *BMC Evol Biol* **8**: 17.

Capra JA, Erwin GD, McKinsey G, Rubenstein JLR, Pollard KS. 2013. Many human accelerated regions are developmental enhancers. *Philos Trans R Soc Lond, B, Biol Sci* **368:** 20130025–20130025.

Charlesworth B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet Res* **63**: 213–227.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69–87.

Chinwalla AT, Cook LL, Delehaunty KD, Fewell GA, Fulton LA, Fulton RS, Graves TA, Hillier LW, Mardis ER, McPherson JD, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520–562.

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**: 1960–1963.

Cotney J, Leng J, Yin J, Reilly SK, DeMare LE, Emera D, Ayoub AE, Rakic P, Noonan JP. 2013. The evolution of lineage-specific regulatory activities in the human embryonic limb. *Cell* **154**: 185–196.

Dabney A, Storey JD. qvalue: Q-value estimation for false discovery rate control.

Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, Rosenfeld JA, Sajjadian S, Malig M, Kotkiewicz H, et al. 2012. Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **149**: 912–922.

Dermitzakis ET, Clark AG. 2002. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Molecular Biology and Evolution* **19**: 1114–1121.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**: 376–380.

Doolittle WF. 2013. Is junk DNA bunk? A critique of ENCODE. *Proc Natl Acad Sci USA* **110**: 5294–5300.

Dorschner MO, Hawrylycz M, Humbert R, Wallace JC, Shafer A, Kawamoto J, Mack J, Hall R, Goldy J, Sabo PJ, et al. 2004. High-throughput localization of functional elements by quantitative chromatin profiling. *Nat Meth* **1**: 219–225.

Duret L, Galtier N. 2009. Comment on "Human-specific gain of function in a developmental enhancer". *Science* **323**: 714–author reply 714.

Eddy SR. 2012. The C-value paradox, junk DNA and ENCODE. *Curr Biol* **22**: R898–9.

Enard W, Przeworski M, Fisher SE, Lai CSL, Wiebe V, Kitano T, Monaco AP, Pääbo S. 2002. Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* **418**: 869–872.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2014. Ensembl 2014. *Nucleic Acids Research* **42**: D749–55.

Galas D, Schmitz A. 1978. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Research* **5**: 3157–3170. http://pubget.com/site/paper/212715?institution=law.washington.edu.

Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet* **23**: 273–277.

Graham A. 2003. Development of the pharyngeal arches. *Am J Med Genet A* **119A**: 251–256.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics*. http://bioinformatics.oxfordjournals.org/content/27/7/1017.short.

Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E. 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biology and Evolution* **5**: 578–590.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, et al. 2010. A draft sequence of the Neandertal genome. *Science* **328**: 710–722.

Haygood R, Fedrigo O, Hanson B, Yokoyama K-D, Wray GA. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat Genet* **39**: 1140–1144.

Hubisz MJ, Pollard KS, Siepel A. 2011. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief Bioinformatics* **12**: 41–51.

Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Meth* **9**: 999–1003.

Jiao Y, Shi C, Edil BH, de Wilde RF, Klimstra DS, Maitra A, Schulick RD, Tang LH, Wolfgang CL, Choti MA, et al. 2011. DAXX/ATRX, MEN1, and mTOR pathway genes are frequently altered in pancreatic neuroendocrine tumors. *Science* **331**: 1199–1203.

John S, Sabo PJ, Thurman RE, Sung M-H, Biddie SC, Johnson TA, Hager GL, Stamatoyannopoulos JA. 2011. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet* **43**: 264–268.

Kamm GB, Pisciottano F, Kliger R, Franchini LF. 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 SY, Pritchard JK. 2007. Adaptive evolution of conserved noncoding elements in mammals. *PLoS Genet* **3**: 1572–1586.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**: 107–116. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=1090005&retmode=ref&cmd=prlinks.

Kostka D, Hubisz MJ, Siepel A, Pollard KS. 2012. The Role of GC-Biased Gene Conversion in Shaping the Fastest Evolving Regions of the Human Genome. *Molecular Biology and Evolution* **29**: 1047–1057.

Lefebvre V, Li P, de Crombrugghe B. 1998. A new long form of Sox5 (L-Sox5), Sox6 and Sox9 are coexpressed in chondrogenesis and cooperatively activate the type II collagen gene. *EMBO J* **17**: 5718–5733.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289–293.

Lindblad-Toh KK, Garber MM, Zuk OO, Lin MFM, Parker BJB, Washietl SS, Kheradpour PP, Ernst JJ, Jordan GG, Mauceli EE, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**: 476–482. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=21993624&retmode=ref&cmd=prlinks.

Lowe CB, Kellis M, Siepel A, Raney BJ, Clamp M, Salama SR, Kingsley DM, Lindblad-Toh K, Haussler D. 2011. Three Periods of Regulatory Innovation During Vertebrate Evolution. *Science* **333**: 1019–1024.

Lunter G, Ponting CP, Hein J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput Biol* **2**: e5.

Maricic T, Günther V, Georgiev O, Gehre S, Curlin M, Schreiweis C, Naumann R, Burbano HA, Meyer M, Lalueza-Fox C, et al. 2013. A recent evolutionary change affects a regulatory element in the human FOXP2 gene. *Molecular Biology and Evolution* **30**: 844–852.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. 2012. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**: 1190–1195.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**: 495–501.

McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, Indjeian VB, Lim X, Menke DB, Schaar BT, et al. 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**: 216–219.

Meader S, Ponting CP, Lunter G. 2010. Massive turnover of functional sequence in human and other mammalian genomes. *Genome Research* **20**: 1335–1343.

Merlin S, Horng S, Marotte LR, Sur M, Sawatari A, Leamey CA. 2013. Deletion of Ten-m3 induces the formation of eye dominance domains in mouse visual cortex. *Cereb Cortex* **23**: 763–774.

Morin RD, Johnson NA, Severson TM, Mungall AJ, An J, Goya R, Paul JE, Boyle M,

Woolcock BW, Kuchenbauer F, et al. 2010. Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. *Nat Genet* **42**: 181–185.

Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, et al. 2012. BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**: 1919–1920.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *Plos Biol* **3**: e170.

Niu D-K, Jiang L. 2013. Can ENCODE tell us how much junk DNA we carry in our genome? *Biochem Biophys Res Commun* **430**: 1340–1343.

O'Bleness M, Searles VB, Varki A, Gagneux P, Sikela JM. 2012. Evolution of genetic and genomic features unique to the human lineage. *Nat Rev Genet* **13**: 853–866.

Paten B, Herrero J, Beal K, Fitzgerald S. 2008. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome ....* http://genome.cshlp.org/content/18/11/1814.short.

Pertea M, Pertea GM, Salzberg SL. 2011. Detection of lineage-specific evolutionary changes among primate species. *BMC Bioinformatics* **12**: 274.

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

Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G, Baertsch R, et al. 2006a. Forces Shaping the Fastest Evolving Regions in the Human Genome. *PLoS Genet* **2**: e168.

Pollard KS, Salama SR, Lambert N, Lambot M-A, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A, et al. 2006b. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443**: 167–172.

Ponting CP, Hardison RC. 2011. What fraction of the human genome is functional? *Genome Research* **21**: 1769–1776.

Prabhakar S, Noonan JP, Paabo S, Rubin EM. 2006. Accelerated Evolution of Conserved Noncoding Sequences in Humans. *Science* **314**: 786–786.

Prabhakar S, Visel A, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Morrison H, Fitzpatrick DR, Afzal V, et al. 2008. Human-specific gain of function in a developmental enhancer. *Science* **321**: 1346–1350.

R Core Team. R: A Language and Environment for Statistical Computing.

http://www.R-project.org/.

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

Sanyal A, Lajoie BR, Jain G, Dekker J. 2012. The long-range interaction landscape of gene promoters. *Nature* **489**: 109–113.

Shibata Y, Sheffield NC, Fedrigo O, Babbitt CC, Wortham M, Tewari AK, London D, Song L, Lee B-K, Iyer VR, et al. 2012. Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. ed. J.M. Akey. *PLoS Genet* **8**: e1002789.

Soranzo N, Spector TD, Mangino M, Kühnel B, Rendon A, Teumer A, Willenborg C, Wright B, Chen L, Li M, et al. 2009. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat Genet* **41**: 1182–1190.

Stedman HH, Kozyak BW, Nelson A, Thesier DM, Su LT, Low DW, Bridges CR, Shrager JB, Minugh-Purvis N, Mitchell MA. 2004. Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature* **428**: 415–418.

Steinberg SJ, Dodt G, Raymond GV, Braverman NE, Moser AB, Moser HW. 2006. Peroxisome biogenesis disorders. *Biochim Biophys Acta* **1763**: 1733–1748.

Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, Mohajeri K, Kondova I, Bontrop RE, Persengiev S, et al. 2013. Evolution and diversity of copy number variation in the great ape lineage. *Genome Research* **23**: 1373–1382.

Taylor MS, Massingham T, Hayashizaki Y, Carninci P, Goldman N, Semple CAM. 2008. Rapidly evolving human promoter regions. *Nat Genet* **40**: 1262–3– author reply 1263–4.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.

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

van Arensbergen J, van Steensel B, Bussemaker HJ. 2014. In search of the determinants of enhancer-promoter interaction specificity. *Trends Cell Biol* **24**: 695–702.

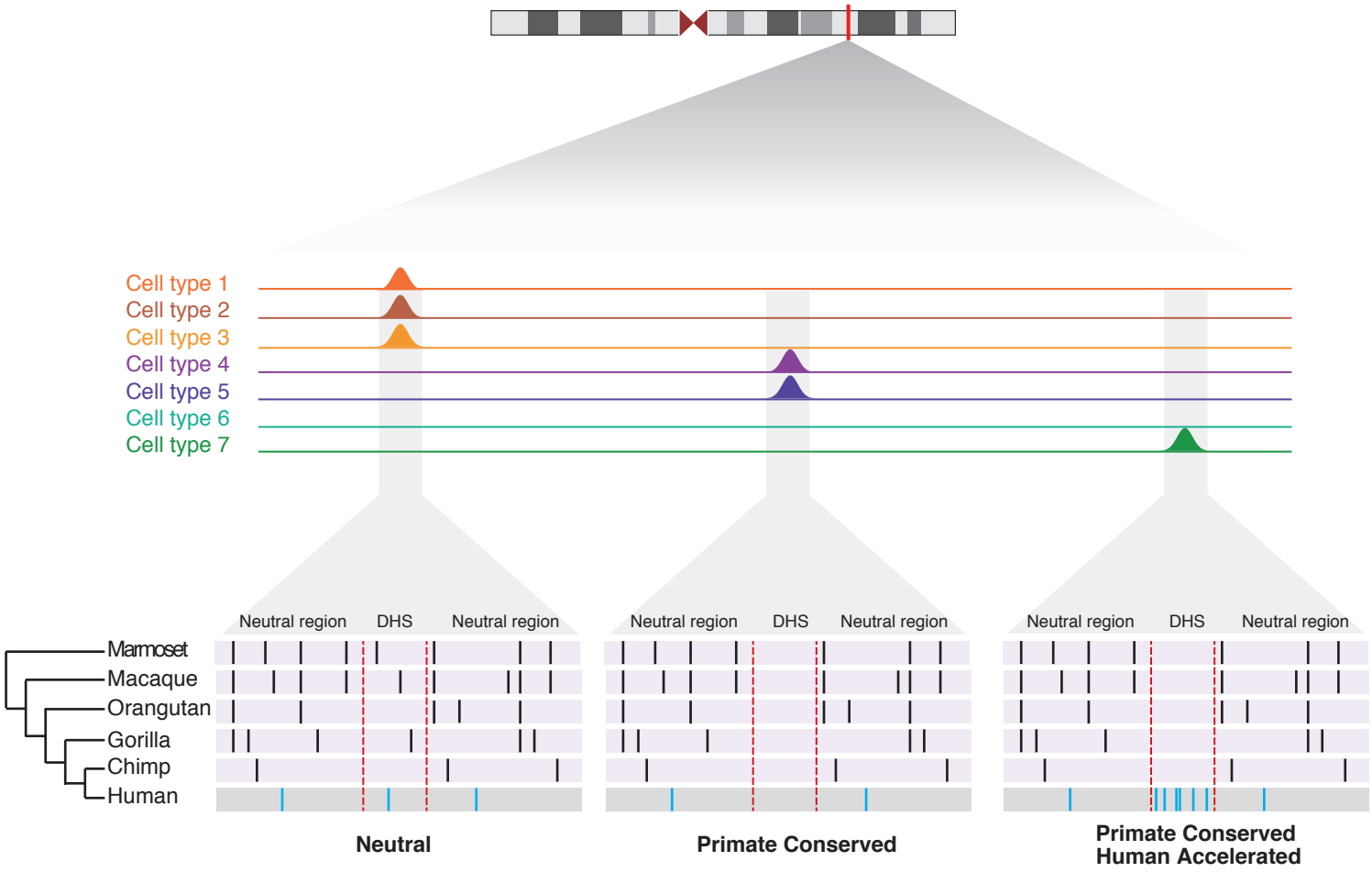
Vernot B, Stergachis AB, Maurano MT, Vierstra J, Neph S, Thurman RE, Stamatoyannopoulos JA, Akey JM. 2012. Personal and population genomics of human regulatory variation. *Genome Research* **22**: 1689–1697.

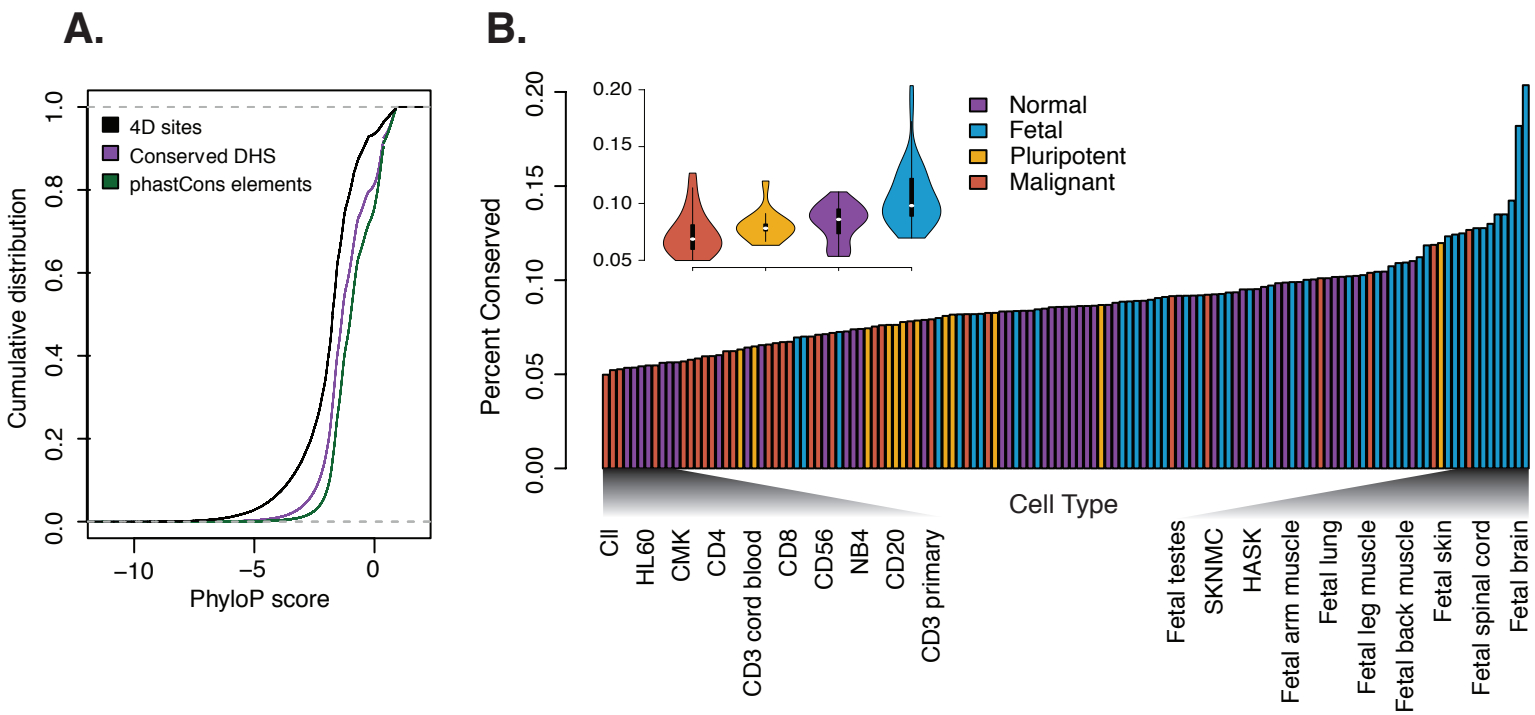
Villar D, Flicek P, Odom DT. 2014. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nat Rev Genet* **15**: 221–233.

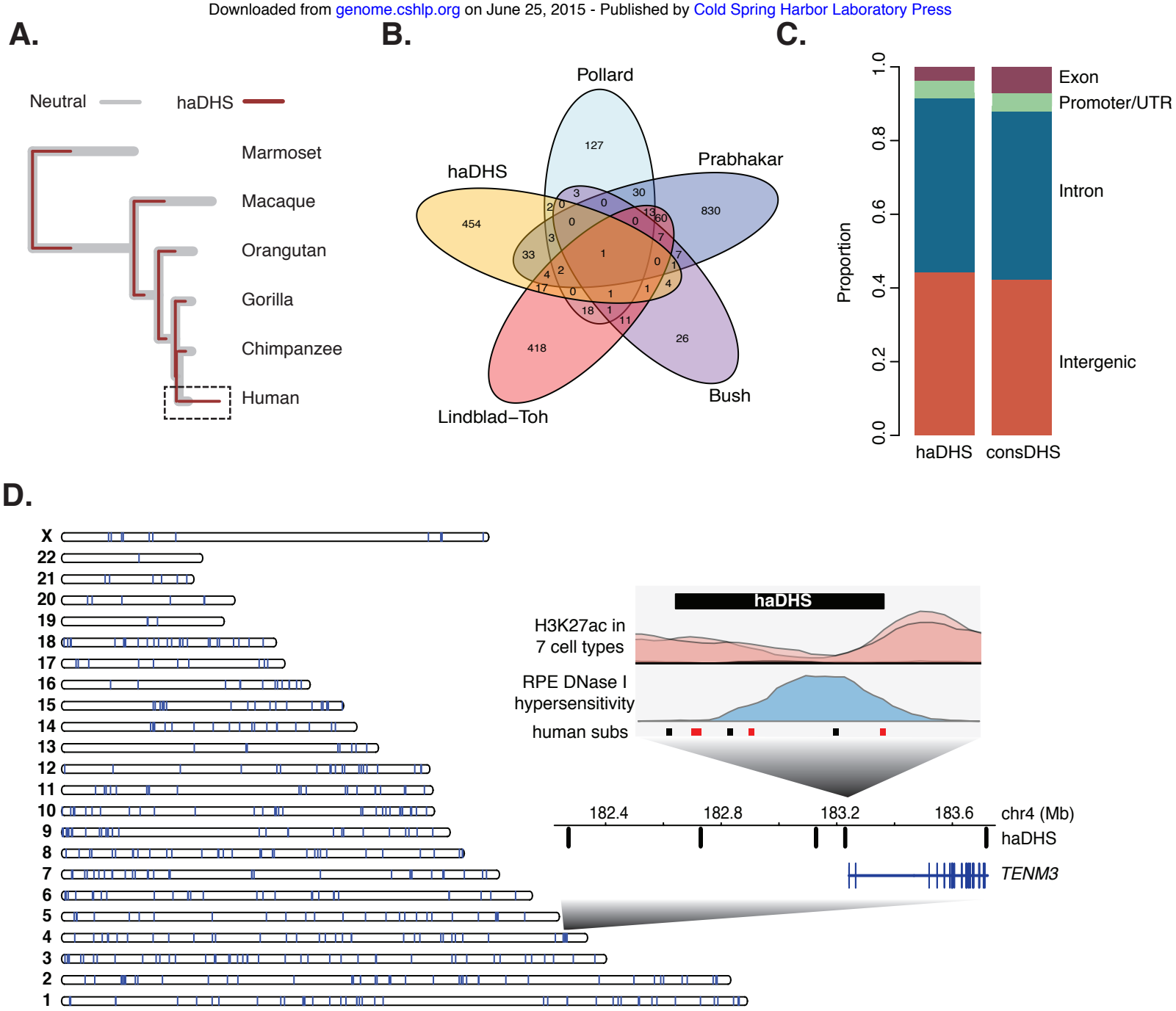
Visel A, Minovitsky S, Dubchak I, Pennacchio LA. 2007. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Research* **35**: D88–92.

Wang J, Duncan D, Shi Z, Zhang B. 2013. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Research* **41**: W77–83.

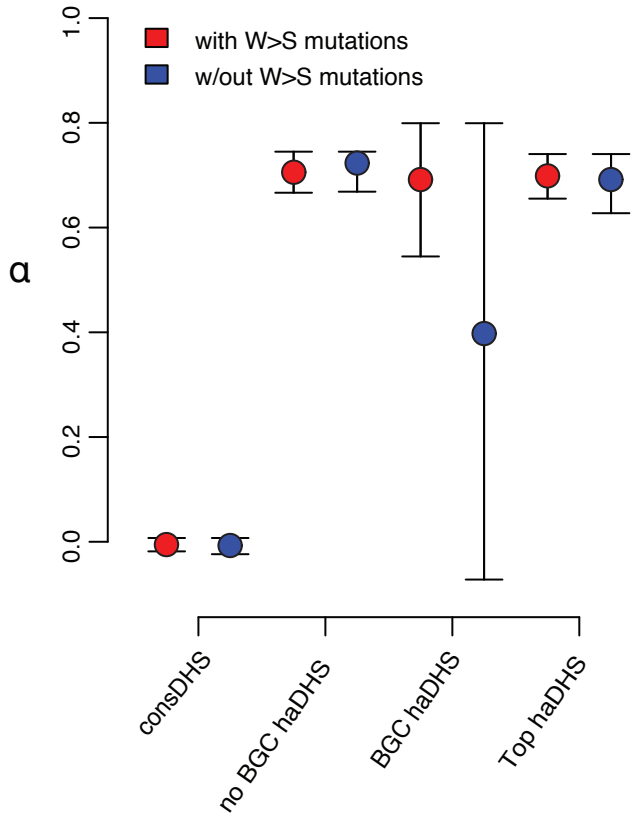
Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA. 2003. The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution* **20**: 1377–1419.



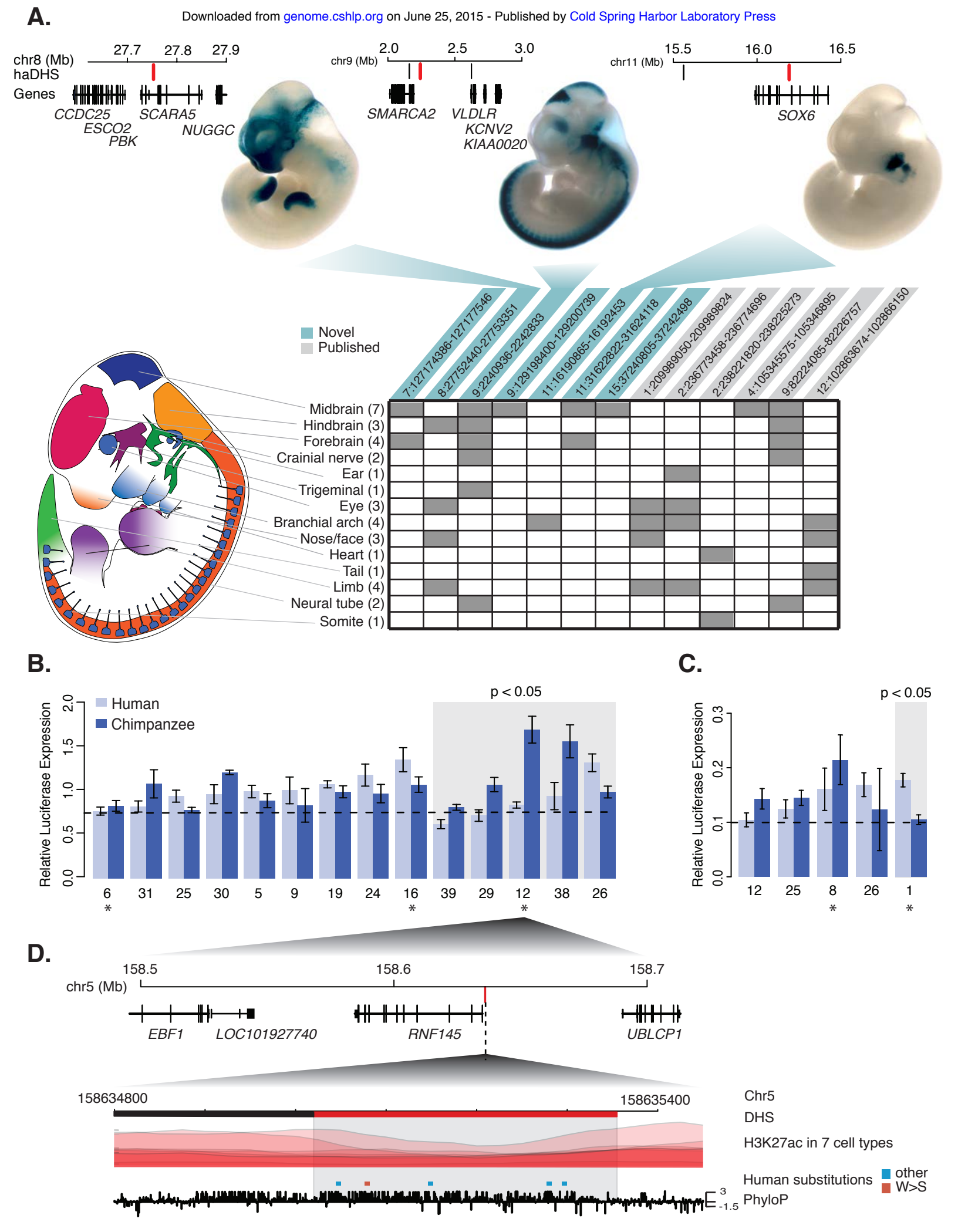


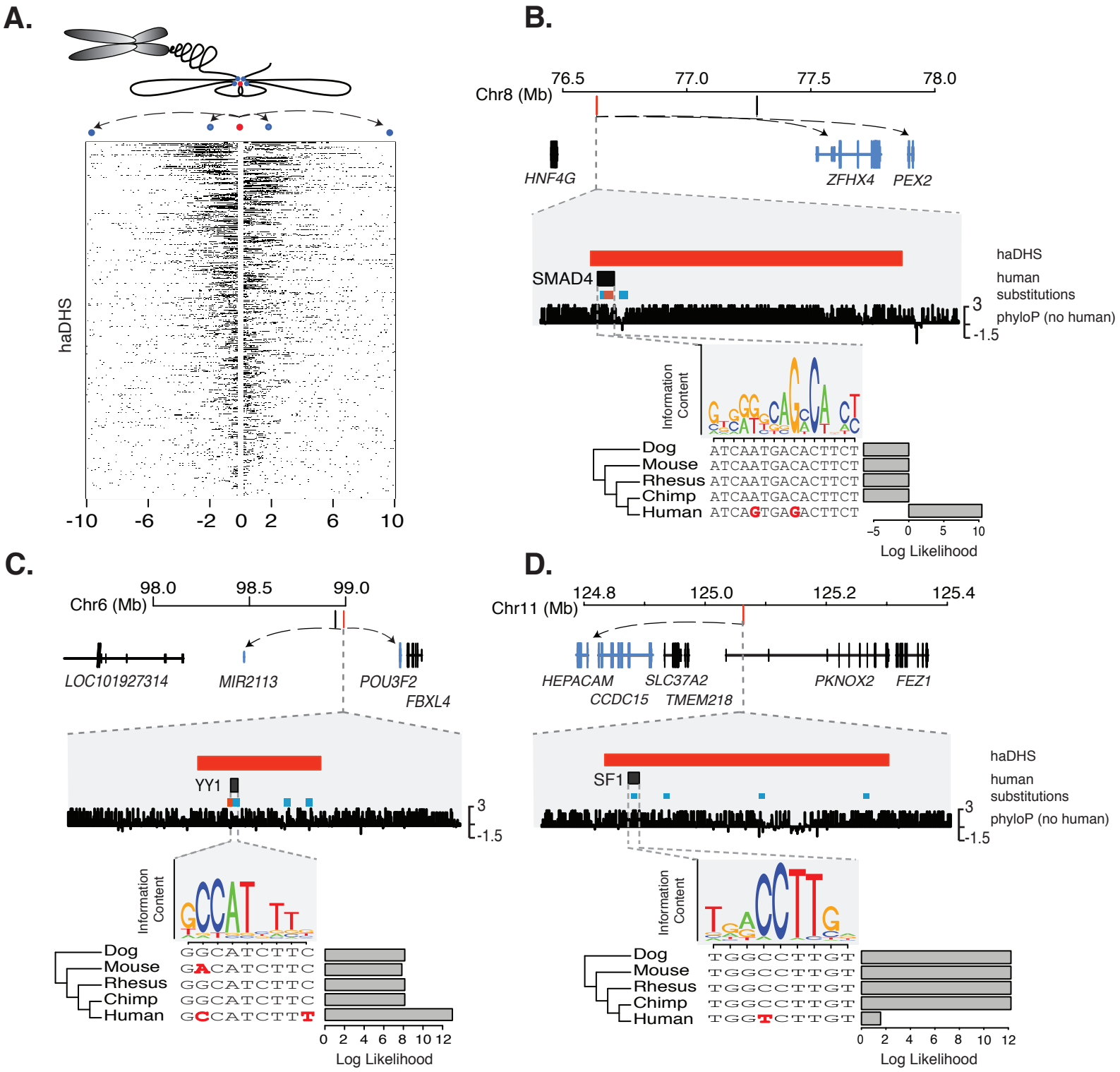


A.



B.





Comprehensive identification and analysis of human accelerated regulatory DNA

Rachel M. Gittelman, Enna Hun, Ferhat Ay, et al.

Genome Res. published online June 23, 2015
Access the most recent version at doi:10.1101/gr.192591.115