Positive selection in the human lineage across time scales

Rachel M. Gittelman

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:
Joshua Akey, Chair
Willie Swanson
Phil Green

Program Authorized to Offer Degree:
Department of Genome Sciences

University of Washington

Abstract

Positive selection in the human lineage across time scales

Rachel M. Gittelman

Chair of the Supervisory Committee:
Professor Joshua M. Akey
Department of Genome Sciences

Understanding the evolutionary history of humankind has long been a central goal of biology. This history extends over millions of years- beginning with the events that led to our divergence from chimpanzees and other primate relatives, and ending with the ongoing migration patterns that give rise to the diverse human populations we see today. DNA sequencing and other molecular techniques for ascertaining genetic variation have been instrumental in reconstructing the timeline of this history. However, these technologies can also be leveraged to answer more fundamental questions about the ways that natural selection has shaped human evolution. For instance, studying patterns of DNA variation can elucidate the genetic basis of traits that were selected for in the human lineage across time scales. Although many methods have been developed to identify these signatures that natural selection has left on our genomes, there are still many cases of positive selection that current methods are not designed to detect. Here I present novel genome-wide approaches for detecting natural selection at two different points in human history. In the first, I extended methods to identify regulatory elements that show elevated rates of human-specific substitutions, and may thus underlie unique human phenotypes that evolved after our divergence from other apes. In the second, I

characterized haplotypes inherited from interbreeding events with archaic hominin species that facilitated modern human adaptation to out of Africa environments. Both of these approaches build significantly on previous work to identify positive selection in the human genome, and provide an extensive catalogue of loci to study using more targeted hypotheses in the future.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank Phil Green, David Hawkins, Willie Swanson, Katie Peichel, Bill Noble, Stan Fields, Celeste Berg, Maitreya Dunham, and Adam Leache for their guidance throughout my graduate work, as well as members of the Akey lab for their patience and willingness to answer my questions. I would also like to thank Brian Giebel and the GSIT staff for helping me manage the day-to-day challenges of graduate school. I must also thank my friends, family, and the other members of my 2011 student cohort for their continued love and support.

Finally, I wish to especially thank my advisor Josh Akey for sharing his abundant scientific knowledge, for his positive mentorship style and steadfast advocacy, and for ensuring my time in grad school was just as fun as it was challenging.

Chapter 1

INTRODUCTION

1.1 Identifying signatures of natural selection facilitates the study of evolution

Evolution occurs when the frequency of a heritable trait changes in a population over time. This process underlies the great phenotypic diversity we see both between and within species, and understanding the general mechanisms by which evolution acts has long been a central goal of biology. Additionally, studying the specific trajectory of human evolution can provide insight in to the characteristics that make us unique as a species, as well as provide context for human disease and diversity.

DNA mutations, the basis of heritable traits, may change in frequency across generations for a number of reasons. For instance, genetic drift occurs when frequencies change due to random chance. However, if a mutation affects the fitness of an individual it may also change in frequency, an idea originally proposed by Darwin and Wallace as part of the theory of natural selection[1]. Negative selection refers to mutations that decrease fitness, and thus decrease in frequency over time, while positive selection acts to increase the frequency of beneficial mutations in a population. Thus it is this last phenomenon of positive selection that may be the most likely to facilitate the evolution of novel traits by promoting changes in populations, making it key in understanding the evolutionary processes that have shaped humankind. Because of this, considerable effort has been put in to studying the history of positive selection on the human genome by developing methods to distinguish the signatures of genetic variation that positive selection creates.

1.2 Methods for genomic scans for selection

There are many different approaches for identifying signatures of positive natural selection, and the specific method chosen depends on the time scale and type of selection being studied (Fig. 1.1). Across the most ancient time scale, comparisons of variants that are fixed in the human lineage but absent in other primates can identify the genetic basis of unique human traits. Although it is difficult to speculate about the significance of single substitutions, if positive selection acts repeatedly on a single gene or other genomic sequence, a scenario that is possible over a long time scale, this can increase the fixation rate of new substitutions to levels detectably greater than neutrality. These methods require some way of classifying neutral vs. functional substitutions. The most widely used statistic, the dN/dS test, compares rates of synonymous and nonsynonymous DNA substitutions, as these are most readily classified based on their effect on protein sequence[2,3]. Early applications of the dN/dS test identified numerous candidate genes, including many involved in immune defense, sensory perception, and spermatogenesis[4-6].

On more recent time scales, when a beneficial new mutation appears in a population, it rapidly rises to fixation (or near fixation), bringing closely linked variants along for the ride. This phenomenon, known as a hard selective sweep, leaves temporary signatures on the genome that numerous statistics have been developed to characterize. First, the site frequency spectrum is shifted compared to neutral expectations. Because chromosomes not carrying the selected allele are lost, sweeps lead to regions of low diversity and an excess of rare alleles. Additionally, because linked derived alleles are swept to fixation, there is an excess of high-frequency derived alleles. Several statistics

were developed to identify shifts in the site frequency spectrum, including Tajima's D[7], Fu and Li's D*[8], and Fay and Wu's H[9]. I present one published article in which I calculated these statistics genome-wide in order to identify pigmentation candidate genes in East Asians in Appendix A. Second, because selected alleles rise in frequency more rapidly than alleles that drift to high frequency, recombination has less time to break down the association with nearby variants, leading to a region of extended linkage disequilibrium (LD). Numerous methods have been developed to identify regions of extended LD[10-13].

Another set of methods aims to identify selection that has acted on distinct populations in response to the unique environmental pressures associated with different geographic regions. Population-specific selection leads to large differences in allele frequencies between populations, and several statistics can be used to quantify this[14-16].

The advent of high-throughput sequencing technology has allowed many studies to apply these various statistics genome-wide, across large cohorts of diverse populations. Although there are considerable differences in the relative study populations, genotyping technology, and methods used, commonalities have emerged that begin to elucidate the pathways and phenotypes that were subject to selection and played an important role as humans dispersed across the globe and adapted to diverse environments. To begin, several loci are repeatedly identified as outliers, and follow up studies have identified functional variants that cause changes to organismal phenotypes. For instance, LCT, the lactase gene, shows extreme population differentiation and patterns of LD[11,17,18]. Variation in upstream regulatory regions is associated with the lactase persistence phenotype, and this phenotype correlates strongly with the domestication of cattle by

different populations[19]. Overall, genes involved in immunity, skin and hair pigmentation, survival in high altitude climates, and adaptation to local diet often appear overrepresented in genomics scans for selection, providing a long list of plausible candidate genes to follow up on in the future[11,20-23].

1.3 The challenges of characterizing signatures of selection

There are significant challenges to identifying cases of positive selection in the human genome. Most importantly, methods must aim to distinguish true positive selection and patterns of variation caused by genetic drift, the phenomenon where changes in allele frequencies occur by chance. Genetic drift can be strongly influenced by demographic history, including population structure, size, and growth rates. In order to accomplish this and assign P values to candidate loci, many early studies made use of extensive coalescent simulations in order to determine how likely a pattern of variation is under different demographic scenarios[19,24]. However, there is still much uncertainty in what demographic parameters are appropriate, especially for understudied populations, making it difficult to draw conclusions from simulations. Additionally, some studies aim to compare candidate loci to neutrally evolving regions of the genome, such as ancestral repeats or non-coding sequence[25,26]. However, it is difficult to confidently define neutral regions, and even variation in non-functional sequence is often influenced by background selection[27]. In a complimentary approach, genome-wide studies can build an empirical null distribution and identify loci that are clear outliers when compared to the rest of the genome. However, there are known limitations to the sensitivity and specificity of this approach, yielding numerous false positives while missing more subtle signatures of

selection[28,29]. Recently methods have aimed at combining scores from different types of statistics, reasoning that true cases of positive selection should be robust to different types of tests[30].

All methods for identifying positive selection also suffer from power issues. In the case of detecting human-specific selection, methods like dN/dS won't work if only one or a small number of variants were selected, as may often be the case[31]. Methods to detect selective sweeps also suffer from multiple issues. First, the signatures described are only temporary, so older sweeps (> 30kya) may not be detectable via methods like long-range haplotype tests. Second, these methods have low power to identify anything besides a hard sweep, in which a novel allele rapidly fixes or nearly fixes. Other scenarios where selection acts on standing variation that is already segregating on multiple haplotypes, or where the selected allele doesn't rise to fixation, will often go undetected. Although methods have been developed to better identify these cases[30,32], this remains an active area of research.

1.4 Integrating genomic annotations with scans for selection

Another important way to increase the utility of both scans for recent selective sweeps and long term species-specific selection is to improve genome annotations through the integration of functional and comparative genomic data. The methods that identify recent selective sweeps have very poor resolution- typically identifying regions on the order of hundreds of kilobases long, and only in rare cases does a segment of that length contain a single, well-annotated gene that lends itself well to hypotheses about its function in human evolution. Even after identifying candidate genes, it is often difficult to identify

specific variants that are likely to be functional. Protein-coding variation can be difficult to interpret, and even nonsynonymous changes may have little effect on protein function. Indeed, predicting the functional effects of protein-coding mutations is an active area of research[33-36]. On top of this, coding sequences make up only ~1% of the genome[5], while ~2-15% of the genome is phylogenetically conserved across species[37-41]. 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[42,43]. There is mounting evidence that this is the case, and several examples of selection on regulatory DNA have been documented[19,44].

Methods for annotating non-coding DNA and predicting the effects of non-coding variation have improved dramatically in recent years. An increased catalog of draft genomes across a wide variety of species has enabled high resolution estimates of phylogentic conservation, a clear indicator of function[45,46]. However, phylogenetic conservation is an imperfect proxy for function, particularly for non-coding regulatory sequences that can exhibit significantly high rates of turnover[47-49]. To more directly identify regulatory DNA, recent studies such as the ENCODE[50] and Roadmap Epigenomics projects[51] have leveraged new functional, genome-wide assays that characterize the specific biochemical signatures that mark regulatory elements, such as histone modification, nucleosome positioning, and transcription factor binding. Grossman et al[52] recently leveraged these improved annotations to obtain a higher resolution catalog of putative causal variants from a genomic scan for selection. They were able to present strong experimental data to show that nonsynonymous variants in

TLR5 and *EDAR* have clear functional consequences and were likely the causal variants underlying the longer signatures detected in their scans[52,53]. Still, both cases involved selection on protein-coding changes, and much more work will be needed to characterize regions with non-coding variation.

There is also a clear need for integrating recent genomic annotations with scans for species-specific selection. In contrast to most methods that detect recent selective sweeps, which are agnostic to annotations about the underlying sequence function, methods to detect species-specific positive selection require that one can make *a priori* hypotheses about what substitutions are likely to be functional. This is why the dN/dS test, which simply assumes nonsynonymous mutations are functional, has been widely used. However, non-coding DNA has likely played an important role in long-term human evolution as well. For instance, 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[54].

In order to extend the dN/dS test to non-coding sequence, Pollard et al[55] described an elegant and powerful approach that discovers sequences that are rapidly evolving or lost on the human lineage, but otherwise phylogenetically conserved and thus likely functional. The approach has been used extensively since[26,45,56-60]. This approach has led to the discovery of several regions with species-specific enhancer activity[61-63], as well as human-specific deletion of regulatory DNA[59]. Despite these improvements, studies have yet to incorporate information from the functional genomics datasets that more directly identify regulatory elements. I hypothesized that the synergistic combination of

comparative and functional genomics would facilitate the high-resolution identification of conserved and human accelerated regulatory sequences.

1.5 Leveraging sequence from archaic hominin genomes to study positive selection

Finally, studies of positive selection have largely ignored scenarios in which adaptive variants are introgressed from closely related species. Thus another important way in which methods for identifying positive selection can be extended is the integration of genome sequence from other archaic hominins. Neandertals diverged from modern humans ~500kya and inhabited wide ranges across Eurasia, before going extinct ~40kya[64]. Although we have long studied their fossil remains, recent advances in sequencing technology and DNA extraction have enabled the high quality reconstruction of the Neandertal draft genome[65,66]. One main finding of this work is that Neandertals share more genetic variants with non-Africans than they do with Africans, an observation that is now well accepted as evidence that Neandertals interbred with modern humans[67,68]. As a result, all non-Africans inherit ~2% of their DNA from Neandertals[65]. Furthermore, in 2008 small numbers of fossil remains were recovered from a single cave in Siberia that were later sequenced to high coverage and determined to belong to a distinct archaic sister group to Neandertals, now known as Denisovans[69]. Denisovans contributed an additional 2-4% of the genomes of Melanesian populations[70].

Since these discoveries, considerable progress has been made in cataloging Neandertal and Denisovan sequences that persist in modern individuals[71-75], but the consequences of hybridization remain poorly understood. Recent studies suggest that surviving Neandertal sequences influence susceptibility to a broad spectrum of

diseases[72,76]. In addition to introducing deleterious alleles into the modern human gene pool, archaic admixture may have also resulted in the acquisition of advantageous alleles that allowed modern humans to adapt to emergent selective pressures as they dispersed into new environments. This hypothesis is especially appealing as Neandertals and Denisovans had resided in Eurasia for thousands of years before encountering modern humans, and likely already evolved adaptations to the local environments. This phenomenon of adaptive introgression has been increasingly recognized in other species, including beak morphology in Darwin's finches[77] and butterfly mimicry[78]. Several examples of adaptive introgression have been hypothesized in humans as well[79,80], including a Denisovan like haplotype of the *EPAS1* gene that confers adaptation to high-altitude in Tibetans[81]. Nonetheless, studies have yet to conduct a comprehensive scan for adaptive introgression, and many important questions remain including the number of loci subjected to adaptive introgression, the population genetics characteristics of such loci, and what the functional and phenotypic consequences of adaptive Neandertal and Denisovan sequences are in modern humans. Most importantly, it is still unclear how large a part adaptive introgression has played in human evolution relative to other modes of positive selection.

1.6 Objectives

In the following chapters I develop frameworks to extend methods for identifying positive selection throughout the human genome at different time scales. These frameworks aim primarily at characterizing signatures of selection that remain poorly understood compared to those of hard selective sweeps and human-specific protein

evolution. Key to these frameworks is the incorporation of novel datasets that allow more targeted hypotheses about the mechanisms harnessed by natural selection throughout human evolution. Specific objectives for these methods are:

1. To synergistically combine comparative and functional genomics data in order to conduct a scan for regulatory elements that have undergone human-specific adaptive evolution.

2. To develop a simulation framework for characterizing neutral expectations for the population genetic dynamics of introgressed loci. This will allow me to comprehensively identify adaptively introgressed loci throughout the genome.

3. To investigate the functional and phenotypic effects of adaptively introgressed loci, and place introgressed loci within the context of all other positively selected loci.

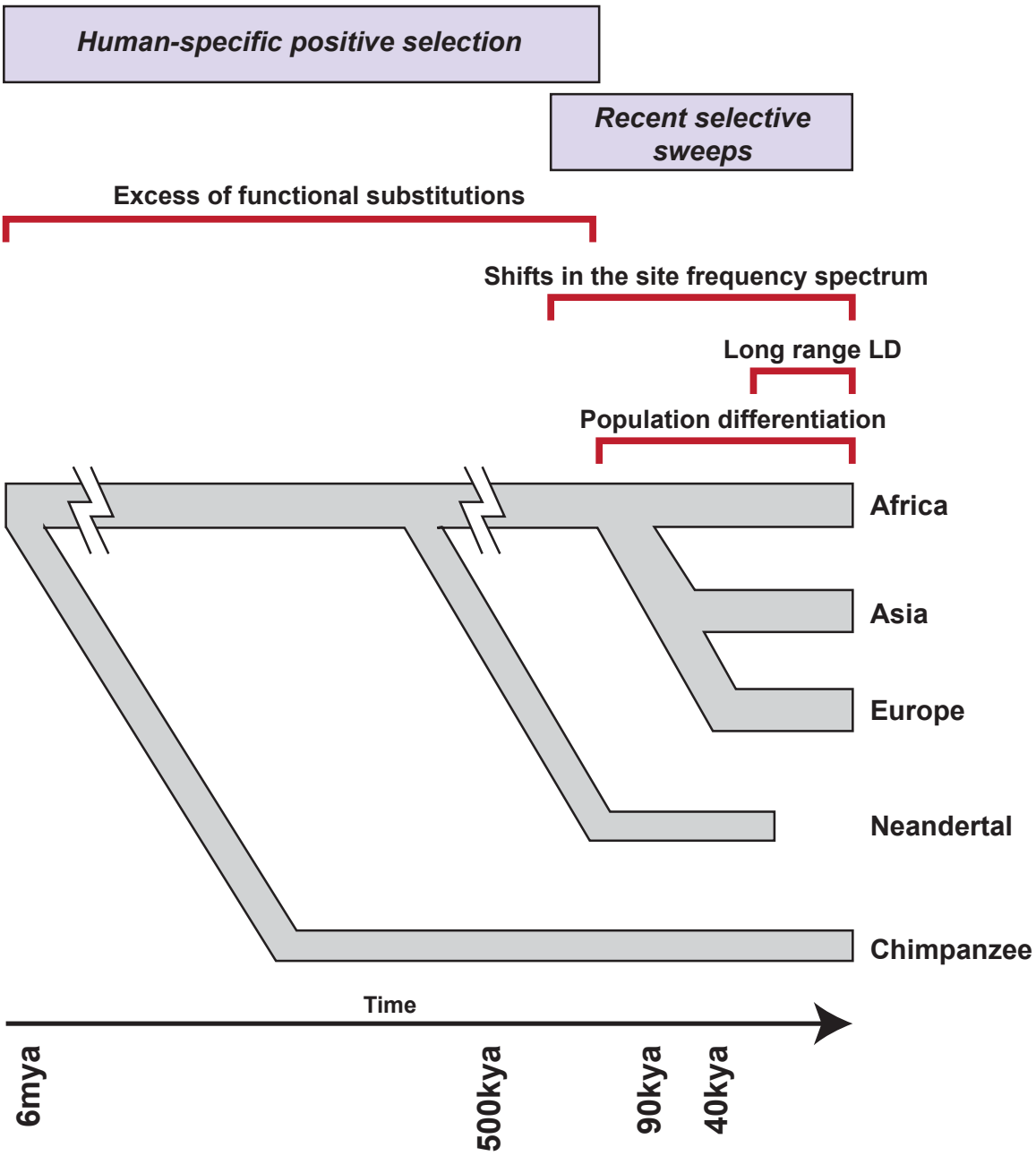

Fig. 1.1. Time scales for the signatures of selection. Signatures of selection are detectable across different time scales, indicated by the red bars. This figure was adapted[21].

Chapter 2

COMPREHENSIVE IDENTIFICATION AND ANALYSIS OF HUMAN ACCELERATED REGULATORY DNA

This chapter contains material published in Gittelman et al[82]

2.1 Rationale

A number of traits distinguish humans from our closest primate relatives, including bipedalism, increased cognition, and complex language and social systems (for review[83]. To date, the genetic basis of human specific phenotypes remains largely unknown, although a number of studies have identified protein-coding changes that likely play a role in some of these unique traits[4-6,31,84,85]. Several recent studies have aimed at identifying human-specific positive selection in regulatory elements, which may have also had a significant impact on human evolution[42,43], by discovering sequences that are rapidly evolving or lost on the human lineage, but otherwise phylogenetically conserved and thus likely functional[26,45,55-60]. However, studies have yet to incorporate information from recent functional genomics datasets such as that more directly identify regulatory elements. Because not all experimentally defined regulatory elements are expected to be functionally or phenotypically significant[86-89], I hypothesized that the synergistic combination of comparative and functional genomics would facilitate the high-resolution identification of conserved and human accelerated regulatory sequences.

In order to perform a comprehensive scan for regulatory elements that have undergone human-specific positive selection, I analyzed genome-scale maps of DNaseI hypersensitive sites (DHS) in 130 cell types from ENCODE[50] and Roadmap

Epigenomics[51]. DNaseI preferentially cleaves regions of open and active DNA, making it a powerful assay to identify regulatory elements, regardless of their specific function[90,91]. We identified 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.

2.2 Results

2.2.1 Identifying conserved and human accelerated regulatory DNA

To identify human accelerated regulatory DNA, I leveraged experimentally defined maps of DHS from 130 cell types identified in the ENCODE and Roadmap Epigenomics Projects (Table B.1). After merging DHS across cell types into 2,093,197 distinct loci (median size = 290 bp, sd = 159 bp), I used a whole genome alignment of six primates from the EPO pipeline[92] to obtain separate alignments for each DHS, using strict filtering criteria for alignment quality. I 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. 2.1). Specifically, I used a maximum likelihood test[25], 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. 2.1). Next, for DHS that are conserved in primates, I performed a

second likelihood ratio test[25] 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=0.01; Table B.2; Fig. 2.1). Importantly, to avoid biasing ourselves against identifying human acceleration, I excluded the human sequence in the first test for conservation.

2.2.2 Characteristics of primate conserved regulatory DNA

I 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. 2.2a). I 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. 2.2b; 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[93]. Conversely, DHS in malignant cell types exhibited the fewest conserved DHS, which may reflect ectopic activation of chromatin[94]. These patterns are also observed in cell-type specific DHS (Fig. B.1a).

2.2.3 Genomic landscape of human accelerated regulatory DNA

I 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 branches have evolved at less than half of the neutral rate (Fig. 2.3a). In total, 70 haDHS overlap previously identified human accelerated elements (HAEs)[26,45,56,58], which is highly significant (permutation $P < 1 \times 10^{-5}$; Fig. 2.3b). Thus, by focusing on experimentally defined regulatory DNA, I 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 (Fig. B.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 x 10^{-7}, hypergeometric test, Fig. 2.3c). These data are consistent with the hypothesis that non-coding regions are more free to evolve and acquire new functions. Furthermore, I observed eight regions where four or more haDHS were clustered within a 1Mb window, suggesting coordinated changes in multiple regulatory elements (Fig. 2.3d). For instance, *TENM3*, which is required for establishing neuronal connections in vertebrate retinal ganglion cells[95,96], is the nearest gene to five haDHS, four of which are active in retinal pigment epithelial cells (Fig. 2.3d, inset).

2.2.4 Adaptive evolution is the primary determinant of rate acceleration in haDHS

Human acceleration can result from both adaptive and non-adaptive forces[97-99]. I 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 I applied a novel permutation test (Appendix B) 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[99]. Second, I investigated the contribution of GC-biased gene conversion (GC-BGC) to my data, which influences rate acceleration of HAEs[56,99,100], and found that 9.7% (51 haDHS) show significant evidence of GC-BGC (Appendix B; Figure B.3a). Finally, I 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 (Appendix B; Figure B.3b).

To more directly quantify the proportion of substitutions in haDHS that can be attributed to positive selection, I used the McDonald-Kreitman framework and compared levels of polymorphism and divergence at haDHS. Specifically, I used polymorphism data from the 1000 Genomes Project[101] and calculated the statistic α, an estimate of the proportion of substitutions fixed by adaptive evolution. As a control, I first estimated α in conserved, non-accelerated DHS, which as expected was 0 (95% CI -0.02-0.007; Fig. 2.4a; Fig. B.4a). I estimate that 70.1% (95% CI 65.8%-73.7%) of substitutions can be attributed to positive selection in haDHS (Fig. 2.4a), and this number is robust to mutation rate heterogeneity in the presence of complex demographic history (Appendix B; Fig. B.4b). To evaluate the sensitivity of α to GC-BGC, I 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. 2.4a). Of the remaining 29.9% of substitutions in haDHS not accounted for by positive selection, I estimate 9.0% are expected without human specific rate acceleration and 20.9% are attributable to additional factors such as relaxation of constraint (Fig. 2.4b). In support of this hypothesis, I find increased levels of nucleotide diversity in haDHS and HAEs (Appendix B, Fig. B.5).

2.2.5 haDHS are developmental enhancers that exhibit lineage specific activity

I performed extensive experimental studies to better understand the functional significance and potential regulatory roles of haDHS. I found that nine haDHS had previously been tested for *in vivo* enhancer activity using a transgenic mouse assay[102], and I 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; Table B.3). These 13 haDHS were active in a wide range of tissues (Fig. 2.5a), 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. 2.5a). One interesting example is located on 11p15, and is only active in the branchial arch (Fig. 2.5a). This haDHS is located in an intron of *SOX6*, and as I describe below, I find evidence that it contacts the *SOX6* promoter. *SOX6* is a developmental transcription factor involved in brain, bone, and cartilage development[103]. Notably, the branchial arch develops into several structures, including the jaw and larynx[104], making

this haDHS an intriguing candidate that potentially influences traits such as facial morphology and speech.

I also performed luciferase assays to functionally test haDHS in a more high-throughput manner. Specifically, I 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. I 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. 2.5b; Fig. B.6a), of which five (35%) exhibited significant differences between the human and chimpanzee haplotypes ($P < 0.05$; Fig. 2.5b; Fig. B.6a; Supplementary Table 4). In IMR90, 5 out of 20 haDHS showed significant evidence of enhancer function ($P < 0.05$; Fig. 2.5c; Fig. B.6b; Table B.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. 2.5b,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[105] (Fig. 2.5d). Although this haDHS is likely part of the promoter for *RNF145*, as described below, it may target several other genes including *IL12B* and *CLINT1*.

2.2.6 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[106]. Chromatin conformation technologies such as HiC[107] identify physical contacts between distinct segments of DNA and have been shown to identify long-range interactions between promoters and enhancers[108]. I leveraged high-coverage HiC data from human IMR90 fibroblast cells to identify putative regulatory targets of haDHS using a rigorous statistical method[109]. I identified 9,000 significant contacts for the 524 haDHS at 40kb resolution (FDR=0.01, Fig 2.6a). 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. I 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. 2.5a), *RUNX2*, and multiple *HOX* genes, all of which play important roles in development.

I 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, I first performed the analysis on conserved DHS contact regions compared to the genomic background. I found that conserved DHS contacts are highly enriched for developmental

genes, including those involved in neuron development (Table B.5), consistent with previous observations about conserved noncoding sequence[93]. Next, I 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 $P < 0.05$; Table B.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 2.6b-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[110] (Fig. 2.6c). Two of the substitutions in this haDHS strengthen a putative YY1 transcription factor binding site (Fig. 2.6c), which is known to mediate long distance DNA interactions[111].

2.3 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, I 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, I discovered 113,577 DHS conserved in primates, 524 of which exhibit significant rates of acceleration in the human lineage.

I found marked heterogeneity in the distribution of conserved DHS across cell types (Fig. 2.2b), 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[112,113]. 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[94]. Thus, these observations combined with my results that DHS in malignant cell types have low levels of evolutionary conservation suggest that many malignant DHS may reflect ectopic chromatin activation.

My results also provide new insights into human specific adaptive regulatory evolution. Of the 524 haDHS that I identified, 454 (87%) are novel and were not detected in previous studies of HAEs[26,45,56,58]. The haDHS that I 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 (Figure B.3). I hypothesize these differences are largely the consequence of my study design that synergistically integrated experimentally defined regulatory sequences with phylogenetic conservation, which both focused my

analyses to a subset of the genome enriched for functionally important sequence and limited the influence of confounding evolutionary forces. To support this hypothesis, I find that a higher proportion of haDHS overlap human-specific enhancer marks in the cortex[114] 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 I 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, I performed extensive bioinformatics analyses and experimental validations. I 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. I also used HiC 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 Neandertals[66], 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[110]. *FOXP2* itself is a transcription factor that has previously been hypothesized to play a role in speech and language in humans[54]. My

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, I 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. 2.6b). Mutations in *PEX2* can lead to Zellweger Syndrome, characterized by a constellation of features including impaired brain development and craniofacial abnormalities[115].

My study has a number of important limitations. For example, the DHS I 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[116,117], a more systematic and comprehensive effort would be of considerable value in understanding the evolution of regulatory sequences. Furthermore, I did not consider additional types of genetic variation, such as structural variation, that may influence human-specific phenotypes[84,85]. Furthermore, although there is evidence that chromatin conformation is relatively stable across cell types[118], it would be of considerable interest to generate HiC 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 I 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, my 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.

2.4 Methods

2.4.1 DNaseI Hypersensitivity Sites

I used DnaseI Hypersensitivity peaks previously published as part of the ENCODE[50] and Roadmap Epigenomics[51] projects. A list of cell types is available in Table B.1. All peaks were called using the hotspot algorithm[119], and represent the 150bp region of maximal DnaseI signal. I merged DHS across cell types using the Bedops package[120]. 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, I

segmented merged DHS according to the number of cell types each region was active in (Appendix B).

2.4.2 Primate Alignments

I downloaded the six primate EPO alignment from Ensembl version 70[121]. Using this I obtained an alignment for each DHS and the surrounding 50kb of sequence. I masked all sites that were polymorphic in the 1000 Genomes Project integrated phase 1 data (March 2012)[101] 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 I 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 DNaseI 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. I 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.

2.4.3 Identifying conserved and accelerated DHS

DHS that passed filtering were tested for overall conservation along the primate lineage with software from the PHAST package[25,122]. For each DHS I first ran phyloFit on the neutral alignment of the surrounding 50kb with the parameters –nrates 4 –subst-mod

SSREV –EM. I 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[123] were then tested for human acceleration. For this test I 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. I evaluated the accuracy of the FDR using a sampling approach (Appendix B).

To determine the overall rate of evolution in the neutral regions compared to haDHS, I 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, I 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.

2.4.4 Distribution of DHS across cell types and genomic location

To determine how conserved and accelerated DHS were distributed across cell types I used the bedmap program from the bedops suite[120] to map DHS from individual cell types onto the set of merged DHS. I 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. 2.2b; Fig. B.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 I expanded each haDHS by 500kb on either side and then used the bedmap –count command from the bedops suite[120] to count the number of haDHS and conserved DHS within each 1mb region.

2.4.5 Other Human Accelerated Elements

I obtained previously identified human accelerated elements (HAEs)[26,45,56,58] and assessed overlap using the bedmap program from the bedops package[120]. When comparing my haDHS to these other HAEs, I 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, I 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, I created an empirical null distribution by randomly sampling 524 conserved DHS 10^4 times and determining overlap with HAEs for each sample.

2.4.6 Population genetics analyses

I downloaded the phase1 integrated release data from the 1000 genomes project[101] and filtered sites according to several criteria (Appendix A). I calculated α as described

previously[124], 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. I considered bases within haDHS to be putatively selected, and bases in the surrounding 4kb region to be putatively neutral.

2.4.7 HiC Analyses

My coauthor, Ferhat Ay, obtained raw paired-end Hi-C libraries for IMR90 fibroblasts two cell lines[118]. Although Hi-C data was also available from human embryonic stem cells, I chose not to include this cell type as it may have a more permissive chromatin landscape that is not representative of promoter/enhancer interactions[118]. He processed the Hi-C data for each cell line at 40 kb resolution as described in[109]. Briefly, he mapped reads to the hg19 (GRCh 37) reference sequence, pairing mapped read ends, filtering duplicates, binning at 40 kb resolution, normalizing raw contact maps[125], and assigning statistical confidences for each contact bin pair using Fit-Hi-C with a refined null[109]. 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. I 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[126] with the multiple testing method set to BH and the minimum number of genes per category to 10.

2.4.8 Transgenic Mouse Assays

Transgenic mouse assays were performed as previously described[102]. Note, one of the previously tested assays was performed with the mouse ortholog (Table B.3). Images of all the mouse assay replicates are available on the VISTA enhancer browser[102].

2.4.9 Luciferase Assays

I considered several factors when selecting which haDHS to experimentally study. First, because the luciferase assays detect enhancers, I prioritized haDHS showing evidence of enhancer activity. To this end, I 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. I included only DHS from the 20 cell types for which histone modification data was available (see Table B.6 for additional set of haDHS and the cell types used). Second, I 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, I prioritized haDHS that showed the greatest evidence for human-acceleration.

I 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 I used the lm() function in R 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. I then tested enhancers for allelic differences with a two-sided t-test between the human and chimp alleles.

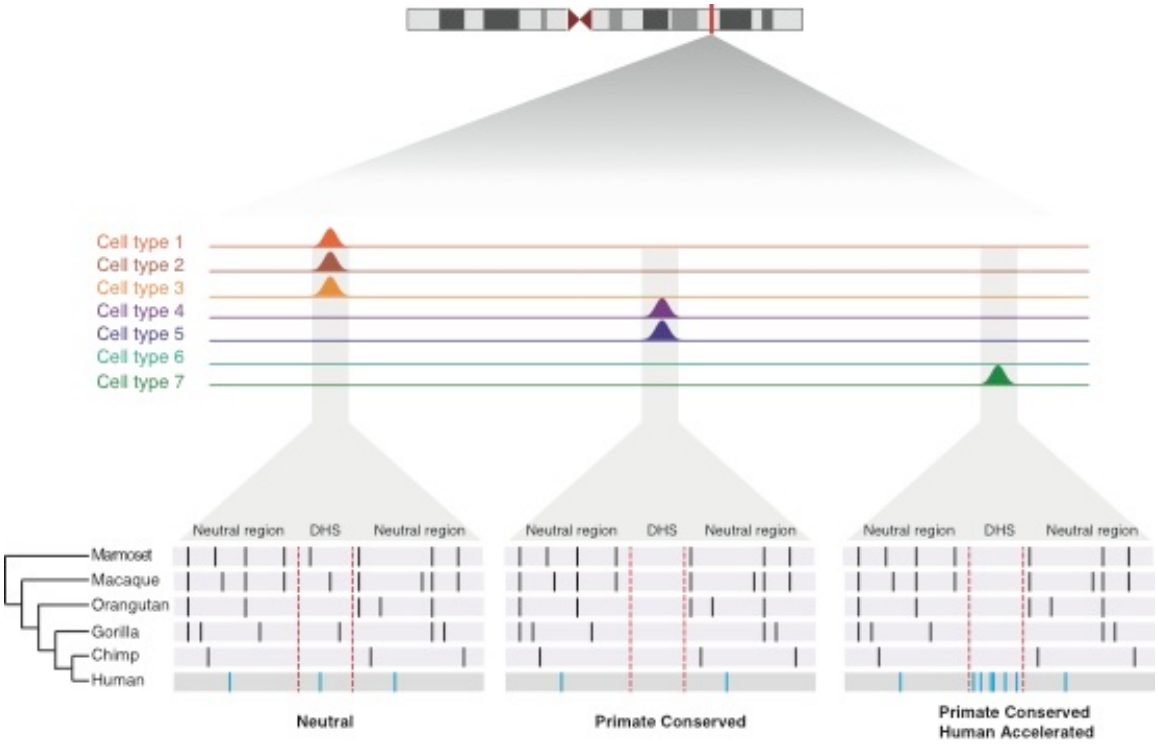

Fig. 2.1. Identifying evolutionarily conserved and accelerated human regulatory sequences. Schematic shows the framework for identifying DHS that are conserved in primates but accelerated in the human lineage. DHS appear as peaks of high coverage along the genome and are merged across cell types. An 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.

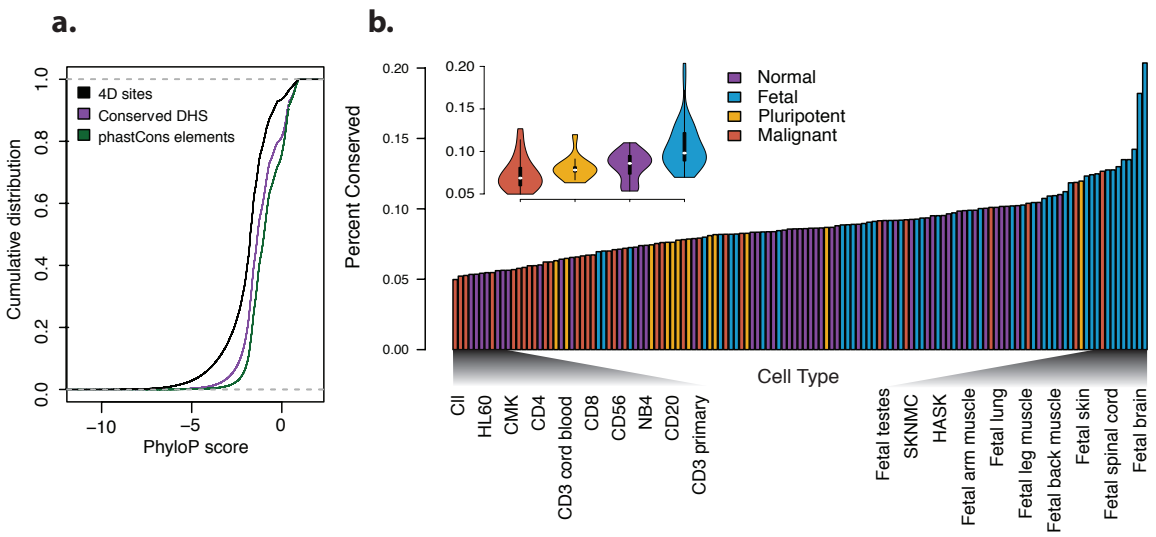

Fig. 2.2. Patterns of conservation vary across cell types. (a) Cumulative distribution of single base phyloP scores are shown for four-fold degenerate sites, conserved DHS, and phastCons elements. The dotted grey line indicates a cumulative distribution of zero or one. (b) The proportion of conserved 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.

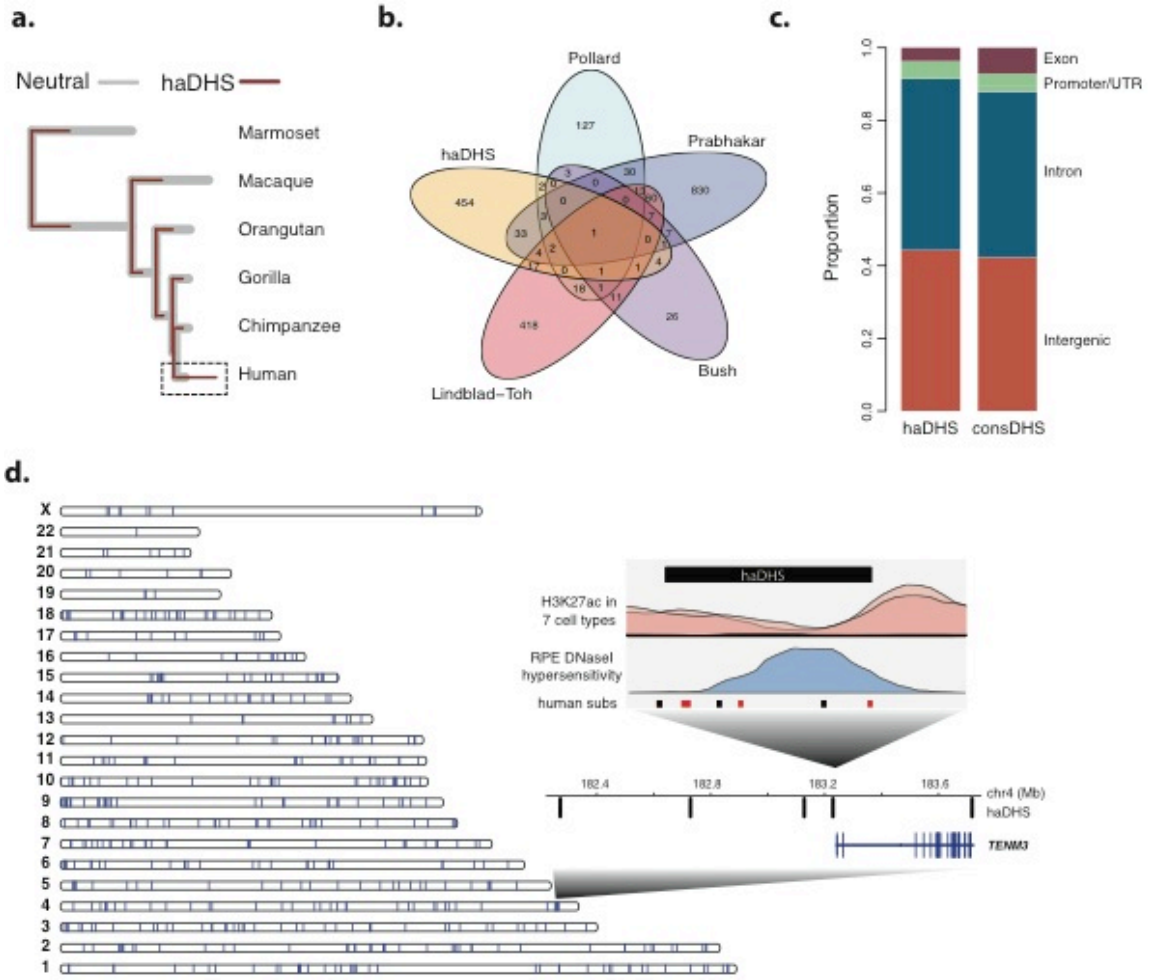

Fig. 2.3. Characteristics of human-accelerated DHS. (a) Overlaid phylogenetic trees inferred in haDHS (maroon) versus their flanking neutral regions (grey). The human branch is highlighted by the dashed rectangle. (b) Venn diagram showing overlap of haDHS with human accelerated elements identified in 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).

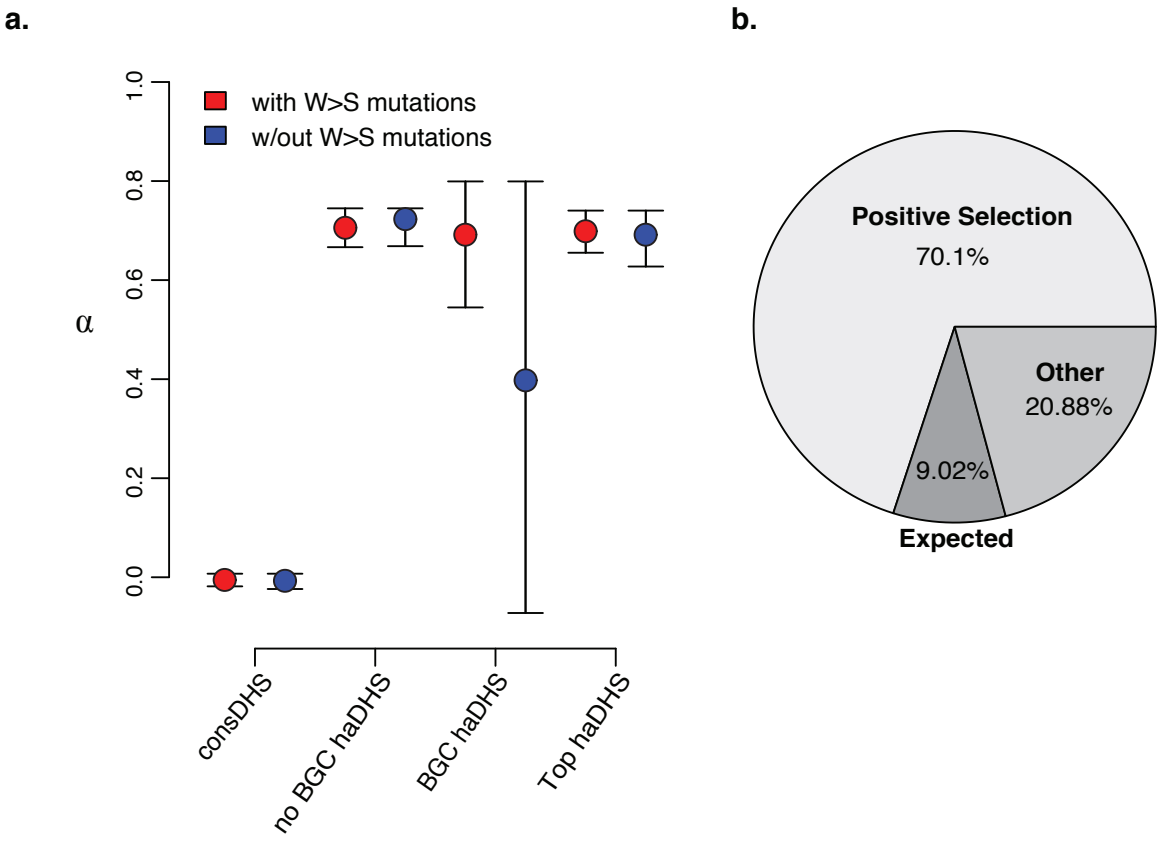

Fig. 2.4. Factors contributing to rate acceleration of haDHS. (a) Estimates of the proportion of adaptive substitutions, α, and 95% bootstrap confidence intervals for different classes of haDHS. Red and blue denote estimates that include or exclude weak to strong mutations, respectively. (b) Pie chart 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.

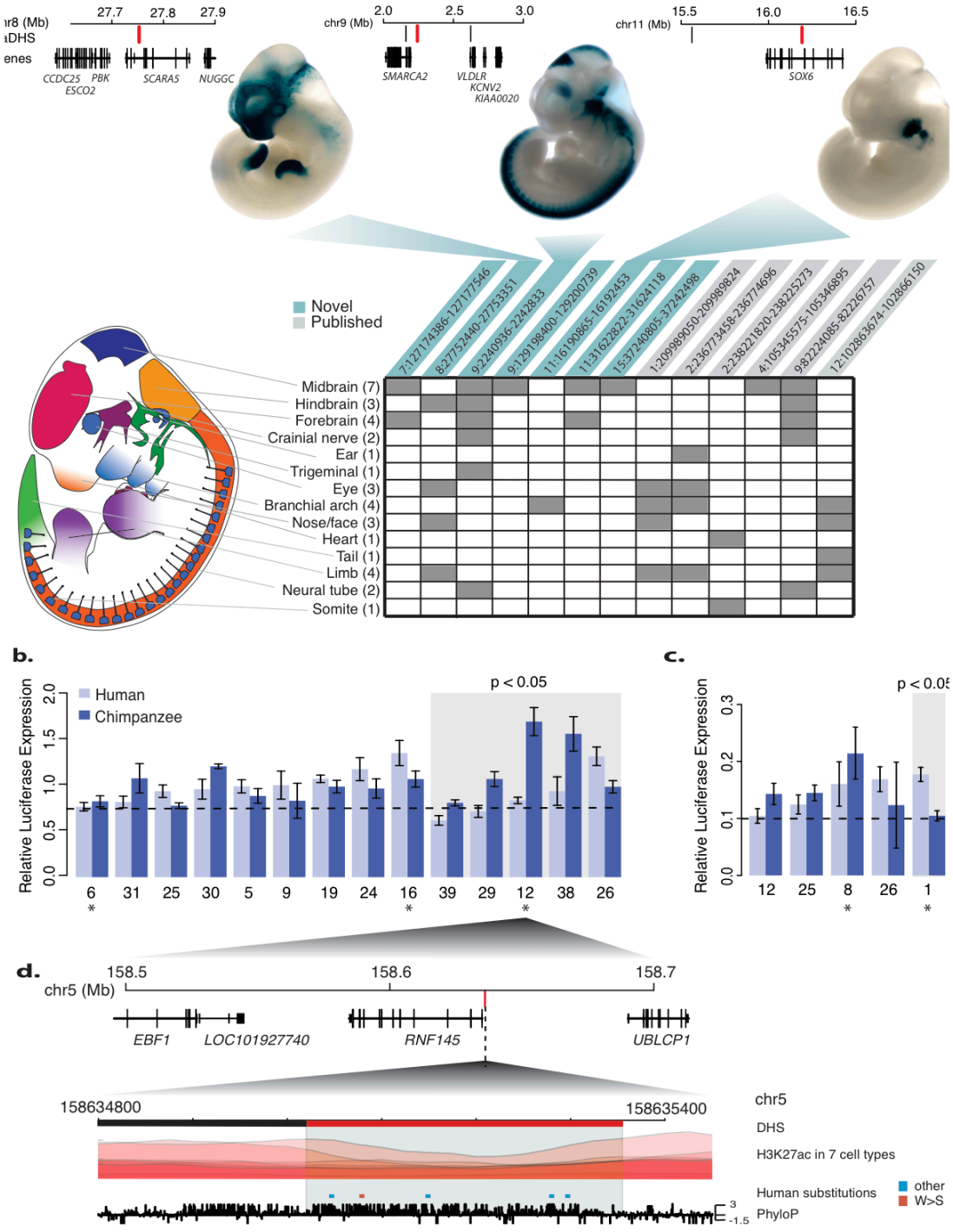
27.7
27.8
27.9
CCDC25
PBK
ESCO2
SCARA5
NUGGC
chr9 (Mb)
2.0
2.5
3.0
SMARCA2
VLDLR
KCNV2
KIAA0020
chr11 (Mb)
15.5
16.0
16.5
SOX6
Novel
Published
7:127174386-127177546
8:27752440-27753351
9:2240936-2242833
9:129198400-129200739
11:16190865-16192453
11:31622822-31624118
15:37240805-37242498
1:209890050-209989824
2:236773458-236774696
2:238221820-238225273
4:105345575-105346895
9:82224085-82226757
12:102863674-102866150
Midbrain (7)
Hindbrain (3)
Forebrain (4)
Crainial nerve (2)
Ear (1)
Trigeminal (1)
Eye (3)
Branchial arch (4)
Nose/face (3)
Heart (1)
Tail (1)
Limb (4)
Neural tube (2)
Somite (1)
b.
Human
Chimpanzee
p < 0.05
Relative Luciferase Expression
0.0
0.5
1.0
1.5
2.0
6
31
25
30
5
9
19
24
16
39
29
12
38
26
c.
p < 0.05
Relative Luciferase Expression
0.0
0.1
0.2
0.3
12
25
8
26
1
d.
chr5 (Mb)
158.5
158.6
158.7
EBF1
LOC101927740
RNF145
UBLCP1
158634800
158635400
chr5
DHS
H3K27ac in 7 cell types
Human substitutions
other
W>S
PhyloP
3
-1.5

Fig. 2.5. 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 SKNMC 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 SKNMC 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.

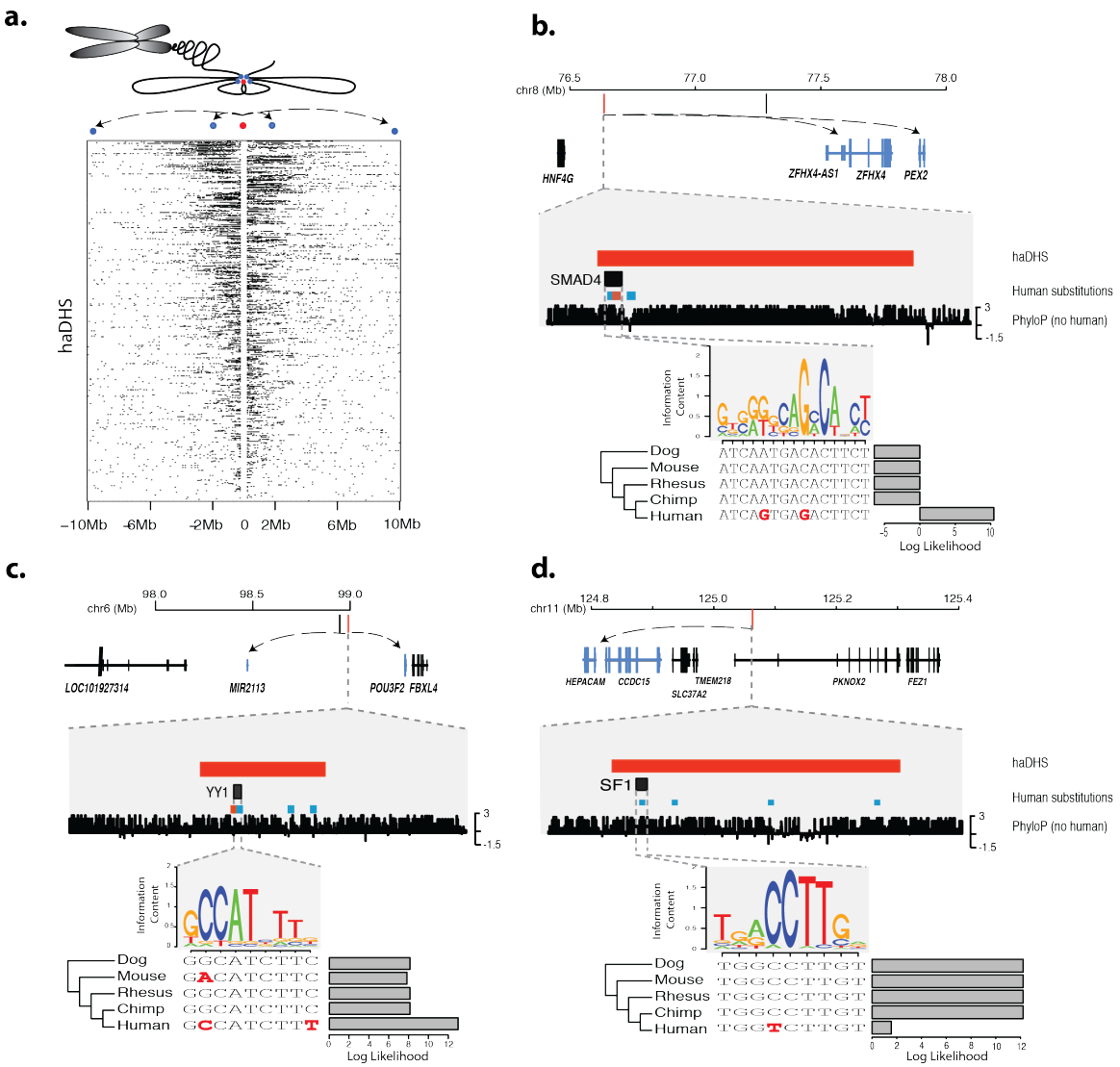

Fig. 2.6. HiC chomatin conformation data identify putative regulatory targets of haDHS. (a) Contacts are shown for all haDHS, and each row indicates the contacts for one haDHS, which is in he center. Black boxes indicate one 40kb contact region. The schematic above illustrates how chromatin conformation information gets translated in to the HiC 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 HiC 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.

Chapter 3

ARCHAIC HOMININ ADMIXTURE FACILITATED ADAPTATION TO OUT-OF-AFRICA ENVIRONMENTS

This chapter contains material from Gittelman et al[127]

3.1 Rationale

Modern humans interbred with archaic humans as they left Africa, and as a result all non-Africans have inherited ~2% of their genomes from Neandertals[65], and Melanesian populations inherited an additional ~2-4% from Denisovans[70]. Recent studies have made considerable progress constructing genome-wide catalogues of introgressed sequence[71-74], but little is known about the functional consequences of these regions. I sought to identify introgressed sequences that may have facilitated modern human adaptation to out-of-Africa environments. A key characteristic of such sequences would have been an increase in allele frequency compared to neutrally evolving loci. Thus, in the following chapter I conducted a scan for high-frequency introgressed haplotypes, and integrate numerous large-scale data sets to characterize and annotate putatively adaptive haplotypes.

3.2 Results

3.2.1 A framework for identifying adaptively introgressed loci

To more comprehensively understand how adaptive introgression has shaped patterns of human genomic variation, I leveraged recently constructed genome-scale maps of Neandertal and Denisovan sequences identified in 1,523 geographically diverse individuals[73] (Fig. 1a). Collectively, I analyzed 1.34Gb and 303Mb of Neandertal and Denisovan sequences, respectively, that segregates in 504 East Asian (EAS), 503

European (EUR), 489 South Asian (SAS), and 27 individuals from Island Melanesia (MEL). I first carefully identified SNPs that "tag" archaic haplotypes (see Methods), and performed extensive coalescent simulations to compare the number of observed high frequency haplotypes to neutral expectations across a range of demographic models (Methods). Consistent with previous studies[28,29], my simulations suggest that simple outlier approaches are effective in enriching for positively selected loci, although the false discovery rate (FDR) can be high (Fig. 3.1b). For example, at a FDR ≤50% (corresponding to extreme frequency outliers in the >99th percentile) I identify 126 candidate adaptively introgressed loci (44, 45, 22, and 38 in EAS, EUR, MEL, and SAS, respectively; Table C.1). Thus, I estimate that there are on the order of 10-20 true cases of adaptive introgression per population, and this estimate is robust to different FDR thresholds (see Methods). Unless otherwise noted, I focus subsequent analyses on this set of 126 loci.

Of the 126 distinct archaic haplotypes that are found at unusually high frequencies, seven have previously been highlighted as putative targets of adaptive introgression[73,79,80]. High frequency archaic haplotypes either span or are proximal to 7 genes involved in skin traits and 31 genes involved in immunity, with significant GO enrichments for defense response (Benjamini and Hochberg corrected P =8x10^{-4}) and cytokine receptor activity (Benjamini and Hochberg corrected P =3.64x10^{-6}) among others (Table C.2). I estimate the strength of selection acting on these loci to be ~10^{-3}, which is an order of magnitude lower than selection coefficients associated with strong recent selective sweeps, such as loci that confer lactase persistence and malaria resistance[22,128] (Fig. C.1).

3.2.2 Patterns of adaptive introgression across populations

107 of the 126 distinct regions are at high frequency in only one population (Fig. 3.1c). For instance, 66% and 58% of archaic haplotypes are population specific in Europeans and South Asians, respectively, whereas 84% and 86% of haplotypes are population specific in East Asians and Melanesians. These data are consistent with additional distinct pulses of introgression into East Asians and Melanesians[73] (Fig. 3.1c).

I next analyzed the number of high frequency archaic haplotypes that were inherited from Neandertals or Denisovans. As described in Vernot et al[5], some archaic haplotypes exhibit high sequence similarity to both the Neandertal and Denisovan reference genomes and thus cannot be confidently labeled; I refer to these haplotypes as ambiguous. As expected, all high-frequency archaic haplotypes in Europeans, East Asians, and South Asians are of Neandertal origin. Strikingly, however, 59% of high frequency haplotypes in Melanesians are inherited solely from Neandertal, despite the fact that these individuals have considerably more Denisovan compared to Neandertal ancestry[73] (Fig. 3.1c). I also identified five regions segregating both Neandertal and Denisovan sequence (Fig. C.2), including archaic haplotypes that are mostly Denisovan and span the *TNFAIP3* (Fig. 3.2), a ubiquitin-editing enzyme involved in the attenuation of cytokine-induced innate immune responses[129].

3.2.3 Adaptively introgressed loci influence pigmentation traits and disease.

To better understand the phenotypic consequences of high frequency archaic haplotypes, I analyzed previously published GWAS results[130]. I found that 10 haplotypes are

associated with 17 traits, including breast carcinoma, nasopharyngeal carcinoma, bone abnormalities (Paget's disease), Celiac Disease, rheumatoid arthritis, optic disk size, and Atopic dermatitis ($P \leq 1x10^{-8}$). Notably, the highest frequency introgressed haplotype in East Asians (62%) encompasses a 29.7kb region of the *OCA2* gene (Fig. 3.3a), and is also found at appreciable frequencies in South Asians (29%), Europeans (20%), and Melanesians (35%). *OCA2* encodes a transmembrane protein involved in iris, skin, and hair pigmentation[131], and both coding and non-coding variants are under strong selection in Europeans and East Asians[44]. Interestingly, the highest frequency introgressed haplotype found in Europeans (68%) encompasses *BNC2* (Table C.1), which is also a known pigmentation gene[132].

The introgressed *OCA2* haplotype contains an average of 10.6 differences with Neandertal (Fig. 3.3b). In contrast, African haplotypes contain an average of 74.3 differences, with the exception of four haplotypes that appear introgressed, likely a result of recent gene flow[133] (Fig. 3.3b). I used a coalescent framework to calculate the probability that a haplotype of this length and divergence from Neandertal is caused by incomplete lineage sorting as opposed to introgression (Methods), and across a wide range of parameters can robustly reject the hypothesis of ILS ($P < 0.003$). The introgressed haplotype does not overlap, nor is it in LD (maximum $r^2 < 0.03$), with any of the variants that are under selection and are most strongly associated with pigmentation traits (Fig. 3.3a). However, it does contain a variant that is associated with blue vs. brown eyes in Europeans ($P = 4x10^{-10}$)[134] (Fig. 3.3a), suggesting that it represents another distinct locus influencing pigmentation traits. Finally, although there are no coding variants on the Neandertal haplotype, 25 introgressed variants overlap regulatory

elements active in melanocytes (Fig. 3.3a). Thus, these data are consistent with the hypothesis of recurrent positive selection acting on multiple variants of *OCA2*, some of which arose in modern humans and some that were inherited through hybridization with Neandertals.

3.2.4 Adaptively introgressed loci modify gene expression

The median length of putative adaptively introgressed haplotypes is 81kb. Thus, compared to analyses of recent selective sweeps that typically identify large genomic regions, adaptively introgressed sequences often result in single gene resolution[75]. 59% of loci overlap protein-coding genes, and there are 49 protein-coding variants in 36 distinct genes. However, 80% of high frequency haplotypes contain no protein-coding variants, indicating that regulatory evolution is the prominent mechanism that adaptively introgressed sequences act through. Given the likelihood that many high frequency archaic haplotypes influence patterns of gene expression, I leveraged extensive RNA-seq data from the GTEx[135] and Geuvadis[136] projects to identify expression quantitative trait loci (eQTL; Methods). Of the 48 high-frequency introgressed haplotypes that could be tested, 13 act as eQTLs to 34 different genes across multiple tissues (Permutation FDR = 0.05; Fig. 3.1a).

My eQTL analyses provide significant new insights into both novel and previously hypothesized targets of adaptive introgression. For instance, Mendez et al.[137] proposed a Neandertal haplotype encompassing the genes *OAS1, OAS2,* and *OAS3*, which encode for antiviral proteins[138], was driven to high frequency by positive selection. Although this haplotype is only an outlier in Europeans, it is also found at appreciable

frequencies in other populations (17%, 13%, and 13% in East Asians, Melanesians, and South Asians, respectively). This introgressed haplotype is strongly associated with expression of all three *OAS* genes across various tissues (Fig. 3.4). When looking at expression aggregated across isoforms, the eQTL is specific to *OAS2* and *OAS3* in transformed fibroblasts and *OAS1* and *OAS3* in transformed LCLs, with individuals containing the introgressed allele show lower expression in all cases (Fig. 3.4). Further analyses of exon-level expression suggests the Neandertal haplotype results in differential splicing of *OAS1* and *OAS2* (Figs. C.3-4). In particular, the introgressed haplotype contains a 3' splice variant between exons 6 and 7 of *OAS1*, leading to the production of a higher activity isoform of *OAS1*[139]. It is important to note that the introgressed haplotype also harbors protein-coding variants that could be targets of selection (Table C.3).

3.2.5 New insights in the adaptively introgressed TLR1/6/10 haplotype

My eQTL analyses reveal novel insights into the recently discovered *TLR1/6/10* haplotype[79,140] inherited from Neandertals, which is at high frequency (39%) in East Asians and intermediate frequency in other populations (22%, 6%, and 17% in Europeans, Melanesians, and South Asians, respectively; Fig. 3.5a). Toll-like receptors play a key role in the innate immune system, and *TLR1* and *TLR6* are well-characterized non-viral receptors[141], whereas *TLR10* has only recently been identified as a possible receptor for influenza[142] and other pathogens[143]. The introgressed haplotype contains 25 variants that have been associated with several immune traits, including helicobacter pylori status[144], allergy burden[145], and cellular response to Pam3CSK4[30] (a *TLR1*

agonist). Consistent with previous results[79], the introgressed haplotype results in significantly higher expression ($P < 2x10^{-5}$) of all three genes in transformed lymphoblast cell lines (LCLs; Fig. 3.5b, Fig. C.5). However, the Neandertal haplotype is associated with significantly lower expression of *TLR6* ($P < 0.019$) in transformed fibroblasts and primary B cells from healthy volunteers[31] (Fig. 3.5b; Fig. C.5). The differential effect in B cells is particularly interesting because they are closely related to the LCLs, differing only in their lack of viral transformation. Finally, *TLR6* expression in other GTEx cell types does not show significant association with the Neandertal haplotype (Fig. C.6).

I hypothesized that the tissue specific patterns of eQTL observed for the Neandertal haplotype reflects differential states of innate immune activation. To test this hypothesis, I measured expression levels in whole blood samples from healthy volunteers before and after stimulation with LPS, a *TLR4* agonist (Methods). Before stimulation, the Neandertal haplotype showed no association with levels of *TLR1*, *TLR6*, or *TLR10* expression ($P > 0.05$; Fig. 3.5c, Fig. C.5). However, after stimulation both *TLR10* and *TLR6* show a significant positive association between expression and number of Neandertal alleles ($P < 0.003$; Fig. 3.5c, Fig. C.5). The effect is modest, but is likely attenuated due to the low proportion of immune cells in whole blood. These data are consistent with the hypothesis that the introgressed Neandertal haplotype influences TLR expression in a context-dependent manner, increasing *TLR6* expression specifically in stimulated immune cells (Fig. 3.5b). Fine-scale mapping suggests that the causal regulatory variant, or variants, may fall within the promoter of *TLR1* or *TLR10* (Fig. C.7-9).

3.3 Discussion

My results provide a comprehensive study of loci that were adaptively introgressed in diverse populations. By combining a scan of genome-wide introgressed loci with a wide range of coalescent simulations, I was able to directly compare the empirical distribution of haplotype frequencies with those expected under neutral evolution. This allowed me to estimate the numbers of true cases of adaptive introgression present among the outliers of the empirical frequency distribution, which is an important improvement over traditional genome-wide scans[28]. My simulations also showed that introgressed loci were subject to weaker selection than regions identified in scans for recent selective sweeps, demonstrating the need for studies such as this, which have increased power to detect signatures of selection that are typically missed by traditional genome-wide scans.

My results show that immune and pigmentation traits were frequent substrates of adaptive introgression, and that in many cases adaptive archaic haplotypes also contribute to the disease susceptibility in contemporary individuals (Fig. 3.1a). By leveraging large-scale catalogs of genomic annotations and functional data[130,135,136] I was also able to make detailed hypotheses about the function of numerous haplotypes, including *OCA2, TLR1/6/10, OAS1/2/3, and TNFAIP3*. However, even with this wealth of data many haplotypes have yet to be associated with specific traits or molecular phenotypes. For instance, almost all of the association data was collected using European samples, rending the data less useful for characterizing haplotypes found in other populations, such as the Melanesian samples. Future studies that characterize functional variation in more diverse populations will be of great use when interpreting these adaptively introgressed loci. Finally, even some loci found at a high frequency in Europeans will need to be characterized with more targeted data sets. For instance, the haplotype encompassing

BNC2, a known pigmentation gene[132], is the highest frequency locus in Europeans, making it an intriguing candidate for follow-up studies. However, variation within the haplotype was not associated with pigmentation traits, and it didn't act as an eQTL in any of the gene expression data. If the haplotype does indeed cause a different phenotype than the modern human allele, it may be restricted to an un-assayed cell type or developmental time point.

In summary, hybridization with Neandertals and Denisovans provided an important reservoir of advantageous mutations for modern humans that enabled adaptation to emergent selective pressures as they dispersed out-of-Africa.

3.4 Methods

3.4.1 Description of samples

I analyzed recently constructed genome-scale maps of Neandertal and Denisovan sequences identified in 1,523 geographically diverse individuals[73]. These individuals included 504 East Asian (EAS), 503 European (EUR), 489 South Asian (SAS) samples sequenced as part of phase 3 release of the 1000 Genomes Project[146], as well as 27 additional individuals from 11 sampling locations in the Bismark Archipelago of Northern Island Melanesia, Papua New Guinea. In total I analyzed 1.34Gb and 303Mb of Neandertal and Denisovan sequences, respectively.

3.4.2 Estimating introgressed haplotype frequencies

I began with the complete set of SNPs that tag archaic haplotypes identified in any individual, analyzing each population separately. I filtered these SNPs to include only

those that match Neandertal (or the correct archaic sequence in the case of Melanesians), and required that each of these tag SNPs belong to a 50kb window with at least two other tag SNPs. Then, in order to aggregate SNPs in to cohesive haplotypes, I calculated LD among tag SNPs using 1000 Genomes data[146] for each population and clustered SNPs with $r^2 \geq 0.3$. I used VCFtools[147] to calculate r^2 statistics. After obtaining these initial haplotypes, I extended each haplotype by identifying all additional SNPs at $r^2 \geq 0.8$ with any tag SNPs on the haplotype. Finally, I filtered any haplotypes that had less than 5 tag SNPs, less than 10 total SNPs, or were shorter than 10kb. Unless otherwise specified, when referring to "variants on a haplotype", I am referring to this final set of tag SNPs and variants in LD with tag SNPs. To estimate the frequency of each haplotype, I calculated the median frequency of all tag SNPs on a haplotype in each population separately.

3.4.3 Coalescent Simulations

I developed a two-part approximate likelihood coalescent simulation framework to a) estimate the false discovery rate (FDR) for adaptive introgression across haplotype frequencies, and b) estimate selection coefficients for adaptively introgressed haplotypes. Briefly, in the first step I simulate introgression across a wide variety of demographic models in order to identify the demographic model that best matches the observed data in each population, and thus allowed us to calculate a FDR for the observed data. In the second step, I use this demographic model to simulate introgression under a range of selection coefficients to estimate the maximum likelihood estimate of the selection coefficient across all frequencies.

I began by simulating introgression at a single locus with MSMS[148] given a base demographic model[149] as follows: a) Ancestral N_e of 10000, b) splitting of archaic and modern human lineages 700,000 years ago, with an archaic N_e of 1500, and modern human N_e of 10000, c) Splitting of Africans and non-Africans at 95,000 years ago, d) a single 500 year pulse of archaic migration in to the out of Africa population, e) population growth in the out of African population starting at 23,000 years ago to an N_e of 10000 at 5,115 years ago, f) exponential population growth starting at 5,115 years ago to a final N_e of 700,000 in the out of Africa population, and 424,000 in the African population. Within this base model I varied several parameters as follows: a) The out of Africa N_e ranged across a grid of N_e = [2000, 4000, 6000, 8000, 10000, 12000], b) The time of introgression ranged across a grid T_I = [40kya, 50kya, 60kya, 70kya, 80kya, 90kya], c) The archaic migration rate in to modern humans varied for each population, with 0.00095 for East Asians, 0.000867 for Europeans, 0.00214 for Melanesians, and 0.000867 for South Asians, d) The number of chromosomes sampled for each population matched the number of chromosomes sampled in empirical data: 1008 for East Asians, 1006 for Europeans, 54 for Melanesians, and 978 for South Asians e) The archaic population harbored a mildly deleterious (s=-0.000021) variant at the time of introgression. The negative selection coefficient was determined based on estimates of the average deleteriousness of introgressed alleles[150]. Briefly, I conservatively use $-3\text{x}10^{-8}$ as the strength of selection per introgressed exonic base, 70kb as the length of the average introgressed haplotype, and 1% as the fraction of exonic bases in the genome. Then the average selection against an introgressed haplotype is:

$$3\times10^{-8} \times 70000 \times 0.01 = 2.1\times10^{-5}$$

The frequency of this variant in the archaic population is described at the end of this section. This base demographic model is depicted in figure C.10. I ran 5 million simulations for each of the 36 distinct demographic models and recorded the frequency of introgressed chromosomes in the final Out-of-Africa population, removing all simulations where the introgressed frequency is zero. I then identified the model with the best fit to the observed data by maximizing the following equation:

$$P(D|X) = \prod_{i}^{N} P(D_i|X)$$

Here $P(D_i|X)$ is simply the proportion of simulations from a particular model X at the frequency of the ith haplotype in the observed data (D). I noticed that the observed data contains an excess of low frequency haplotypes, likely due to the difficulties of aggregating low frequency tag SNPs in to coherent haplotypes. I thus conservatively exclude all haplotypes below 2% frequency from my calculations. Importantly, I note that the chosen parameters may not represent the true demographic history of each population, but are instead the best model given the biases inherent to my dataset, thus allowing us to simulate data that most closely resemble the dataset. Likelihoods for each of the 36 models are depicted for each population in figure C.11.

In the second step, I run additional simulations using the maximum likelihood demographic model, while varying the selection coefficient in the archaic population across a grid of s = [-0.000021, 0, 0.0005, 0.00075, 0.001, 0.0015, 0.002, 0.003, 0.004, 0.005]. Information on the frequency of the selected allele at the time of introgression can be found in the note at the bottom of this section. I ran 25 million simulations for each selection coefficient, again removing replicates where the introgression frequency is zero

in the final population. I next constructed a likelihood landscape for each possible allele frequency by determining the proportion of simulations at that frequency for each selection coefficient. Finally, I conducted a likelihood ratio test between a model with the selection coefficient at the highest likelihood and the "null" model of a mildly deleterious selection coefficient (s=-0.000021). To determine which tests were significant, I compared test statistics to a chi-square null distribution with one degree of freedom.

Note, in order to accurately simulate the frequency of the selected allele in the archaic population at the time of introgression, I ran an initial set of simulations in the archaic population in which a selected allele arose at a frequency of $1/N_e$, randomly between the time of introgression and 500kya. I then recorded the frequency of the selected allele at the time of introgression, discarding simulations in which the frequency went to zero. For subsequent simulations, I randomly sampled the starting frequencies from these distributions (Fig. C.12).

3.4.4 Estimates of true positives are robust to FDR threshold

At a FDR threshold of 50%, my simulations suggest the number of true positives is on the order of 10-20 per population. This estimate is generally robust to varying the FDR threshold. For example, at a FDR threshold of 30%, there are 6, 29, 7, and 19 loci significant and thus 4, 19, 5, and 13 true positives in EAS, EUR, MEL, and SAS, respectively. Similarly, at a FDR threshold of 70%, there are 86, 181, 112 AND 107 loci significant and thus 26, 56, 35, and 32 true positives in EAS, EUR, MEL, and SAS, respectively.

3.4.5 Gene Ontology Enrichments

To obtain a list of genes proximal to, or encompassed by, all high frequency introgressed haplotypes, I downloaded the complete set of UCSC genes from the UCSC genome browser[151] (https://genome.ucsc.edu/) on 10/14/2014 and took the unique set of the nearest gene to each SNP on a haplotype. I input this list as the "foreground" set of genes on the WebGestalt gene ontology browser[126] (http://bioinfo.vanderbilt.edu/webgestalt/), with the following parameters: Enrichment Analysis: GO Analysis; Reference set: hsapiens_genome; Statistical Method: Hypergeometric; Multiple Test Adjustment: BH; Significance Level: 0.05; Minimum Number of Genes: 10.

3.4.6 Coalescent approach to estimate probability of ILS

To provide additional confidence that an archaic haplotype was the result of introgression and not incomplete lineage sorting (ILS) my coauthor Josh Schraiber developed a coalescent likelihood model. Specifically, assume a haplotype spans L bases and contains K mutations relative to the archaic reference sequence. In the following description, he focuses on ILS with respect to Neandertals given that most putative adaptively introgressed sequences are Neandertal in origin. To compute the joint probability of L and K, condition on the fragment coalescing with Neandertal time $T = t$ (in coalescent time units). Then, the distribution of L is exponential with rate ρt, where $\rho = 4Nr$, N is the effective population size and r is the per-base pair recombination rate. Given the length $L = l$, the number of differences K is Poisson distributed with mean θlt, where $\theta =$

$4N\mu$, and μ is the per-base pair mutation rate. Thus, given $T = t$, the joint distribution of K and L is

$$f_{K,L|T}(k,l|t) = \frac{(\theta l t)^k}{k!} e^{-\theta l t} \rho t e^{-\rho t l}.$$

Next, integrate over the distribution of coalescence times. Assuming that coalescence can begin at some time t^* (e.g. the time of introgression, or the population split time), the distribution of coalescence times is simply a shifted exponential distribution with rate 1. Therefore,

$$f_{K,L}(k,l;t^*) = \int_{t^*}^{\infty} f_{K,L|T}(k,l|t) e^{-(t-t^*)} dt$$

$$= \frac{(\theta l)^k}{k!} \frac{\rho \Gamma(2+k, (1+l(\rho+\theta))t^*)}{(1+l(\rho+\theta))^{k+2}} e^{t^*}.$$

Now use Bayes' theorem to compute the probability that a fragment is introgressed by supposing that the introgression proportion is p, and that introgression happened at time t_{GF} in the past while population divergence occurred at time t_D. Implicitly, $t_{GF} < t_D$. So,

$$P(\square\square\square\square\square\square\square\square\square\square\square\square \mid K,L) = \frac{f_{K,L(k,l;t_{GF})p}}{f_{K,L(k,l;t_{GF})p} + f_{K,L(k,l;t_D)(1-p)}}.$$

I used this equation to calculate $P(\square\square\square\square\square\square\square\square\square\square\square\square \mid 11{,}29747)$ for the *OCA2* locus. To ensure that my estimate is robust to different demographic models and mutation rates, I randomly sampled parameters to obtain 100 distinct parameters sets and calculated $P(\square\square\square\square\square\square\square\square\square\square\square\square \mid 11{,}29747)$ with each set. Sampling was from the following uniform distributions: N=[1000-10000]; μ =[5x10^{-9}-5x10^{-8}]; t_{GF} =[40kya-80kya]; t_D=[400kya-600kya], with fixed r=3.19x10^{-8} (the average recombination rate in the region[151]).

3.4.7 RNA-seq normalization

For GTEx data, I began with un-normalized gene read counts provided with the version 4 of GTEx pilot data and applied a processing pipeline similar to that of the GTEx Project. I removed all samples with <10,000,000 mapped reads and summed expression values for technical replicates. I removed 32 individuals identified as non-Europeans by examining a PCA plot of sample genotypes. I removed any genes in which fewer than 10 individuals had at least 5 reads. I also removed samples that were two or more standard deviations below the mean D statistic[152], a measure of overall expression similarity to other samples. I then used the R DESeq package[153] to normalize samples based on library size, and $\log_2$ transformed these data. Finally, I performed outlier detection by mapping expression values for each gene to a standard normal distribution. I ran PEER[154] with these normalized data, using the first two genotype principal components and sex as covariates in order to identify hidden sources of expression heterogeneity. I used the first 15 PEER factors and known covariates in a linear model with the expression data using the lm() function in R[155], and used the residual expression data as input for association testing.

For the Geuvadis Project data, I downloaded pre-normalized gene RPKM expression data from http://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/analysis_results/GD462.GeneQuantRPKM.50FN.samplename.resk10.norm.txt.gz. These data were pre-normalized based on library size and learned PEER factors, and contain only expressed genes. I kept only samples of European origin.

3.4.8 eQTL analyses

I performed eQTL tests using GTEx and Geuvadis data for haplotypes in the top 99th percentile for any population studied. Because these data sets contain only European samples, I only tested haplotypes that reach a frequency of ≥15% in Europeans. Importantly, because SNP content can vary somewhat between the same haplotype in different populations, I also only tested SNPs that were shared between the introgressed haplotypes in Europeans and the population being tested. For GTEx data, I also limited analyses to tissues with 60 or more samples, which included subcutaneous adipose, aorta, tibial artery, transformed fibroblasts, esophagus mucosa, esophagus muscle, left heart ventricle, lung, skeletal muscle, tibial nerve, sun exposed skin, thyroid, and whole blood. Finally, for power considerations, I only ran tests in which there were at least three samples homozygous for each allele.

For each haplotype, I tested for eQTLs across all combinations of SNPs, tissues, and genes within 500kb of the haplotype by building linear models using the lm() function in R[155]. I retained the minimum *P* value from all SNPs tested as the single test statistic for each tissue-gene combination. Then, to determine which tests were significant, I ran permutations by shuffling genotypes 1000 times and repeating tests. Using these data, I chose the maximum P-value cutoff that gave an FDR ≤ 0.05; which is the cutoff at which the ratio of the number of significant permutation tests to the number of significant tests on the true data is ≤ 0.05. Importantly, I calculated a single cutoff across all haplotypes, rather than controlling for FDR on a per haplotype basis.

3.4.9 Analysis of B Cell eQTLs from Fairfax et al 2012

I normalized raw array expression data using the R packages Lumi, Limma, and a variance stabilizing transformation. I mitigated potential batch effects with SVA and removed a single sample that was an outlier in a principal component analysis performed on the genotype data. I then tested expression probes for each *TLR* gene (one probe for *TLR1*, one for *TLR6*, and two from *TLR10*) for association with each of the three SNPs on the Neandertal haplotype that were genotyped in the study, and retained the most significant association for each gene. Tests were performed using standard linear regression models from the lm() function in R[155].

3.4.10 Analysis of whole blood LPS stimulation cohort

My coauthors Carmen Mikacenic and Mark Wurfel recruited healthy Caucasian volunteers from the Seattle metropolitan area. Exclusions to enrollment were active smoking, recent antibiotic use, symptoms consistent with an infection, a history of autoimmune disease, immunodeficiency, cancer, pregnancy, or use of immunosuppressive medications. This cohort has been previously described[156]. The study was approved by the Human Subjects committee at the University of Washington. They isolated genomic DNA and genotyped subjects using the Illumina Human 1M Beadchip array (San Diego, CA). They imputed genotypes using EUR genotypes from 1000 Genomes as a reference population. From the same subjects, whole blood was stimulated with ultrapure LPS from *Salmonella minnesota* R595 (List Biological Laboratories, Inc., Campbell, CA). RNA extraction was done with the AB 6100 Nucleic Acid Prep Station (ABI/Life Technologies, Foster City, CA), and RNA quality was analyzed with Experion Automated Electrophoresis System (Bio-

Rad, Hercules, CA). Gene expression was quantified using the Illumina HumanRef-8 v3.0 Gene Expression BeadChip array (Illumina, San Diego, CA). Expression data quality control was performed using GenomeStudio (Illumina). Duplicate arrays were removed and data were log2 transformed in BASE[157]. For this study, 252 subjects were used for the analyses, which were limited to expression for probes specific to TLR 1,TLR10 and TLR6.

I then built standard linear regression models using the lm() function in R[155] to test for association between all SNPs on the TLR haplotype that were genotyped (81 SNPs) and all normalized TLR expression probes (two for *TLR6* and *TLR10*, and one for *TLR1*), both before and after stimulation with LPS. I used age, sex, and the first three genotyping principal components as covariates.

3.4.11 Roadmap Epigenomics Data

I downloaded H3K27ac, H3K4ME1, and DNaseI narrowPeak calls for consolidated melanocyte epigenomes E059 and E061 from the Roadmap Epigenomics project[51] (http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak/) and used bedops[120] to determine which variants on the OCA2 haplotype overlapped melanocyte regulatory elements.

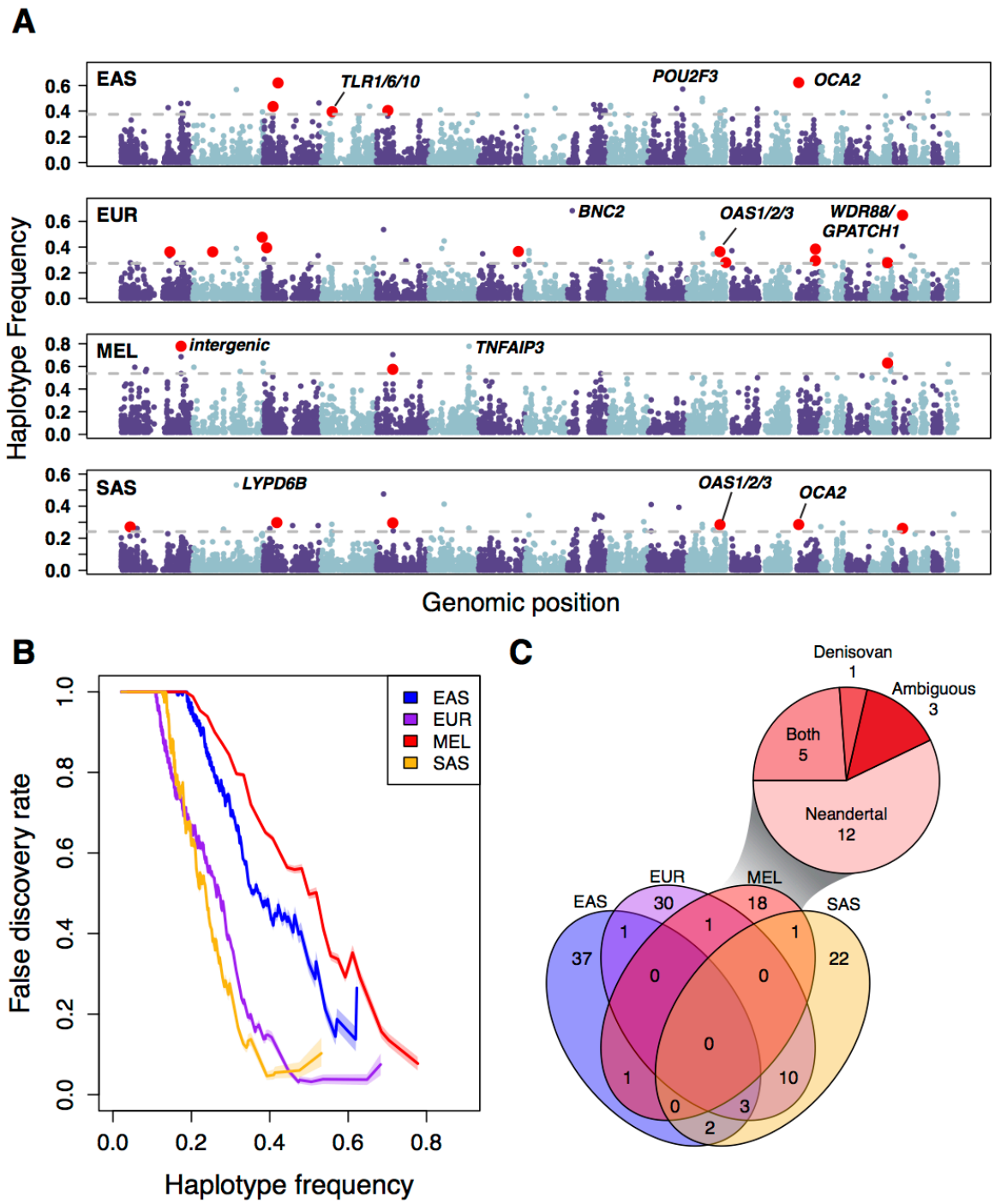

Fig. 3.1. Genomic distribution and characteristics of high-frequency archaic haplotypes in geographically diverse populations. A) Each dot represents the frequency and genomic position of an introgressed archaic haplotype. Loci above the grey lines correspond to putative targets of adaptive introgression (outliers in the ≥99th percentile; FDR ≤ 50%). Outlier loci that had a significant phenotypic association (GWAS or eQTL) are highlighted in red. **B)** Relationship between the archaic haplotype

frequency threshold for identifying adaptively introgressed loci and FDR for each population. Shaded regions delimit 95% confidence intervals. **C)** Venn diagram showing overlap of high frequency archaic haplotypes between populations. The inset pie chart shows how many of the Melanesian high frequency haplotypes are Neandertal, Denisovan, or Ambiguous.

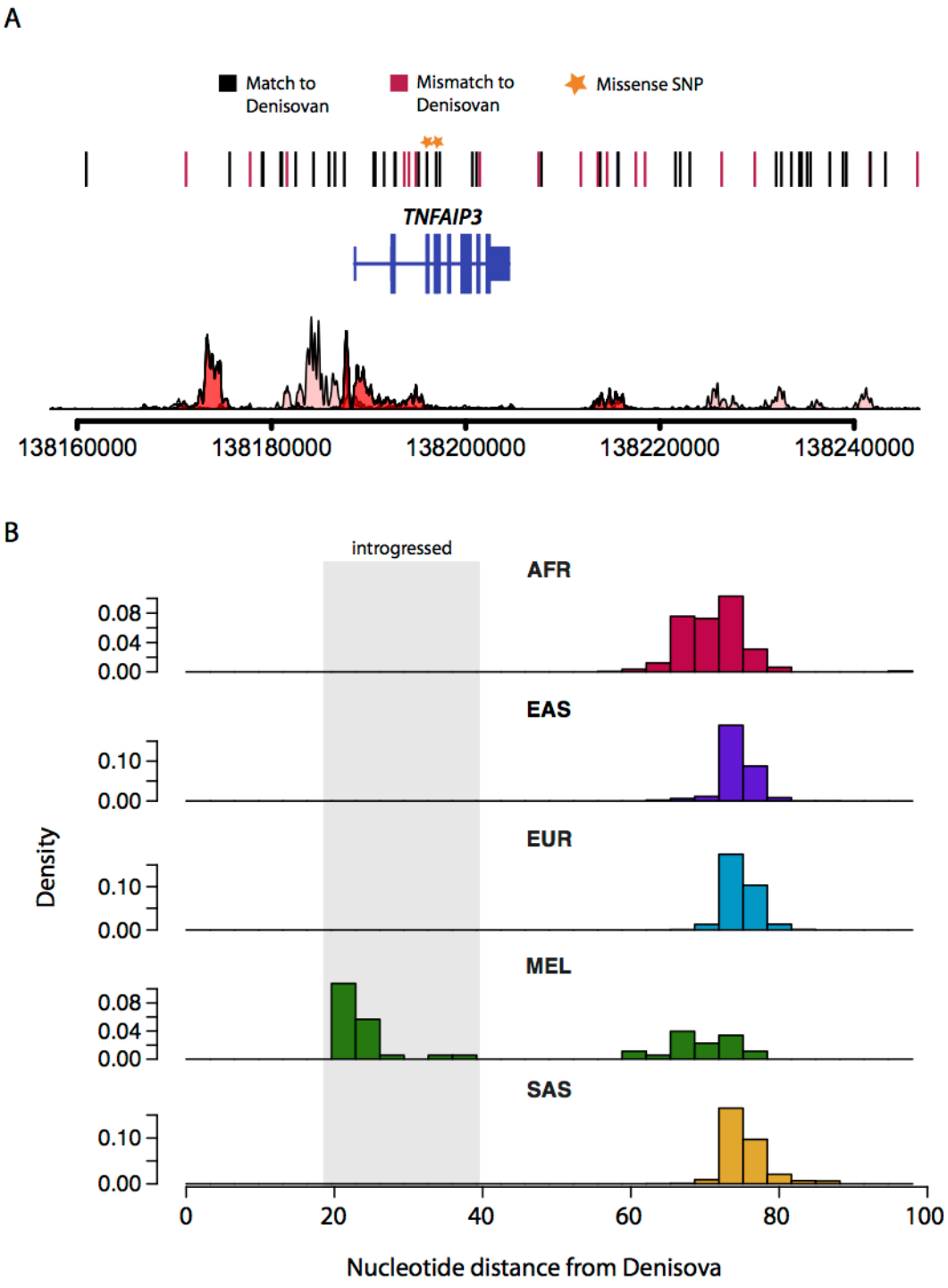

Fig. 3.2. Adaptive introgression of Denisovan sequence at the *TNFAIP3* locus. A. Schematic of the *TNFAIP3* region is shown with vertical bars indicating introgressed SNPs. Black and red denote matches and mismatches to the Denisovan reference genome, respectively. Two missense SNPs are highlighted with stars. The track along the bottom depicts H3K27ac signal from seven ENCODE cell types[50]. **B.** Distributions of

absolute genetic distance to the Denisovan reference genome for all four populations studied, as well as Africans. The grey box indicates the portion of the distribution comprised of introgressed sequence.

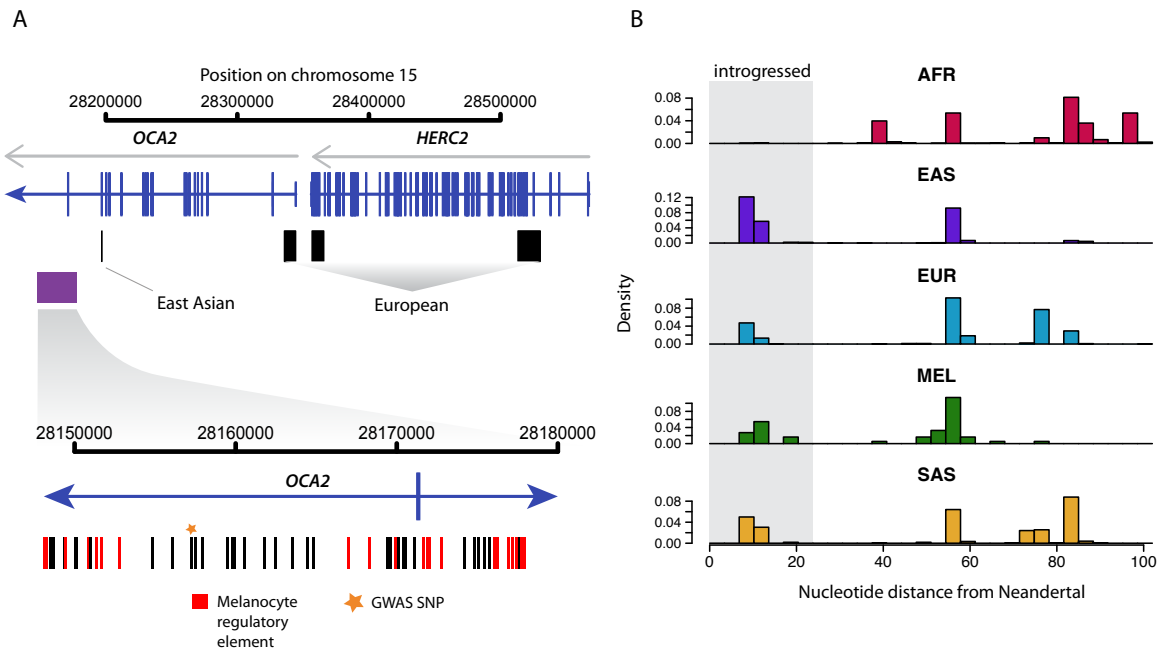

Fig. 3.3. Adaptive introgression at the *OCA2* locus. A. Schematic of the *OCA2/HERC2* region. The purple box indicates the introgressed region, and the black boxes indicate previously identified positively selected andpigment associated regions in East Asians and Europeans. Below, a zoomed in view of the introgressed region is shown and vertical bars indicate introgressed variants. Variants that overlap melanocyte regulatory elements are shown in red and GWAS study variants are indicated with a yellow star. **B.** Distributions of absolute genetic distance to Neandertal for all four populations studied as well as Africans. The grey box indicates the portion of the distribution comprised of introgressed sequence.

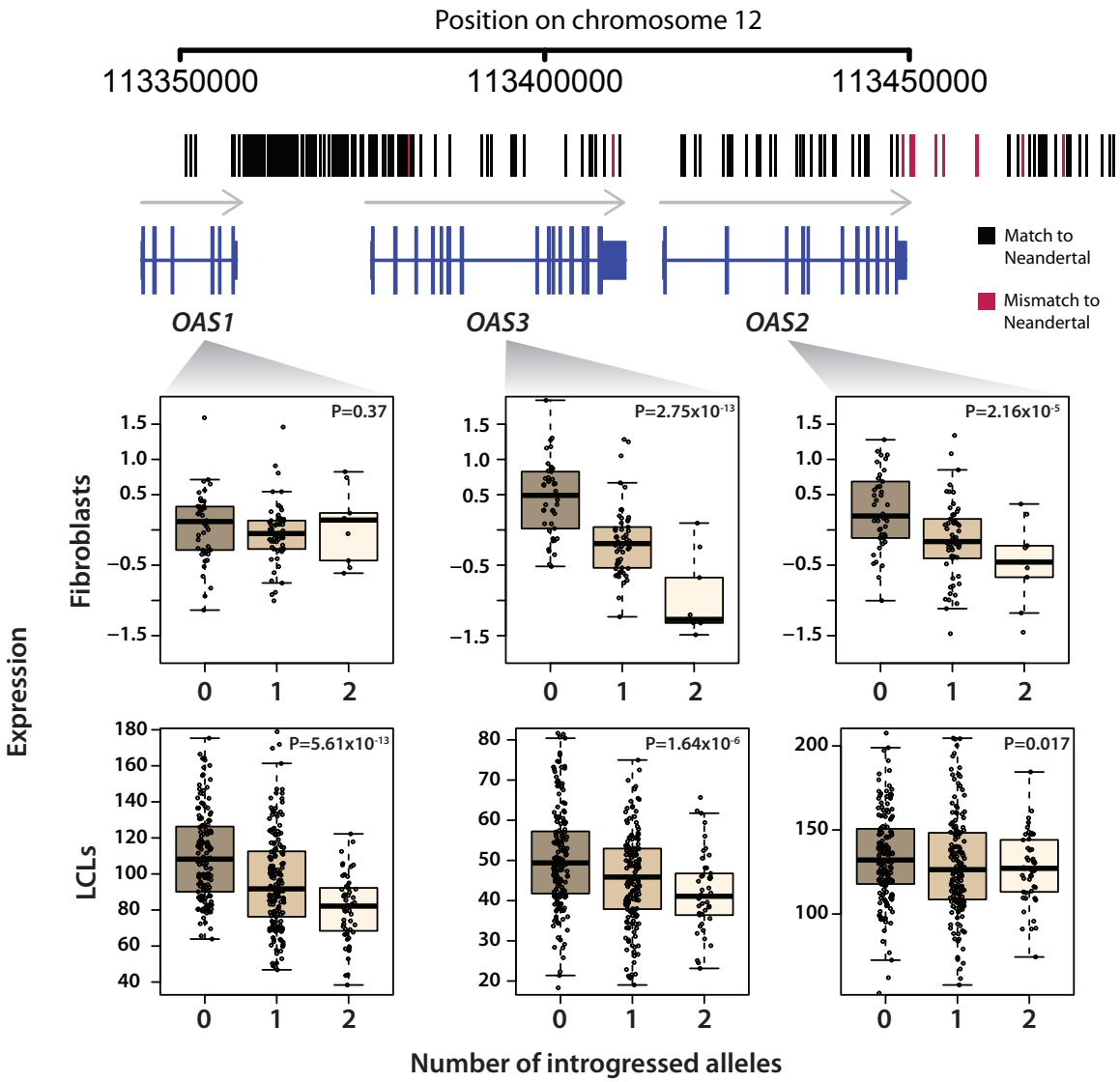

Fig. 3.4. Introgression at the *OAS* locus and its impact on gene expression. A schematic of the *OAS1/2/3* region is shown with vertical bars indicate introgressed variants. Black and red denote matches and mismatches to the Neandertal reference genome, respectively. Below, gene expression for *OAS1*, *OAS2*, and *OAS3* is shown stratified by the number of Neandertal haplotypes an individual has in fibroblasts and LCLs.

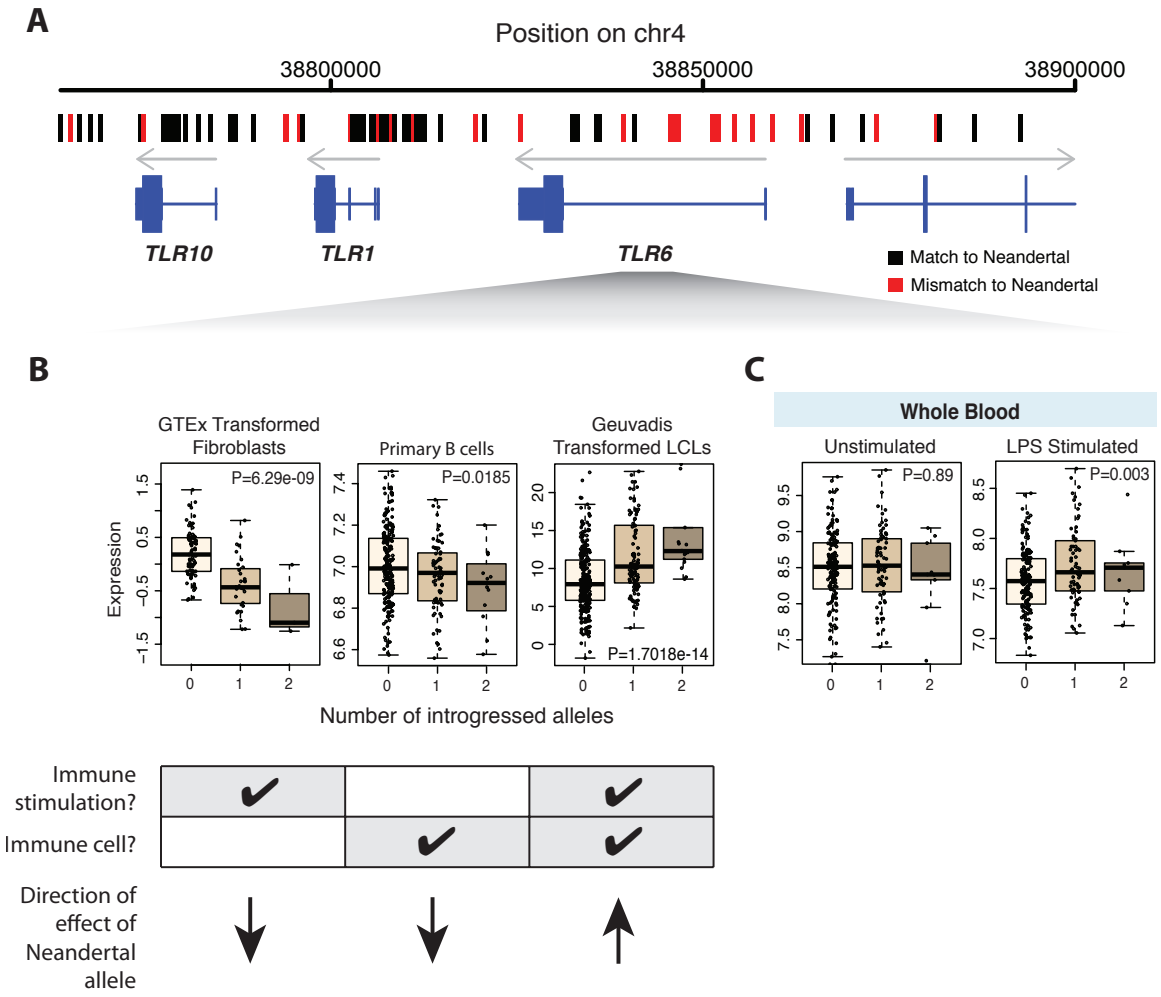

Fig. 3.5. Patterns of eQTL effects for the introgressed *TLR1/6/10* haplotype across multiple cell types. Schematic of the *TLR1/6/10* region is shown with vertical bars indicating introgressed variants. Black and red denote matches and mismatches to the Neandertal reference genome, respectively.. Below, gene expression for *TLR1*, *TLR6*, and *TLR10* is shown stratified by the number of Neandertal alleles each sample has. P values are indicated in each plot.

Chapter 4

CONCLUDING REMARKS

Next-generation sequencing technology has enabled large-scale catalogs of genetic variation, both within and between species, ushering in great advances in human evolutionary and population genomics. Many of these advances come from computational and statistical tools that have been developed to identify signatures that natural selection has left on this genetic variation, and my thesis work has extended existing methods to identify signatures of human-specific selection in regulatory DNA, as well as positive selection on introgressed loci. Despite these important advances, or perhaps because of them, much work remains on elucidating the impact of positive natural selection on the human genome.

Most importantly, the growing catalogs of candidate positively selected loci have vastly outpaced their validation through functional follow-up studies. As mentioned in the introduction, scans for selection often yield large regions that are difficult to interpret, and testing putatively functional variants can be costly and time-consuming. In series of recent papers, Grossman et al[52] and Kamberov et al[53] provide strong evidence that two nonsynonymous substitutions underlie the signatures of selection at two different candidate loci identified through genome-wide scans for selection. However, even a majority of regions that repeatedly show up in studies remain difficult to interpret. For instance, the *OCA2* locus has repeatedly been shown to be under selection, and several variants in the region are associated with different pigmentation traits[44,131]. However, to this day, the genetic architecture and precise identity of the functional variant(s)

underlying these signatures are unknown. As I show in chapter 3, this may be complicated by the presence of multiple functional haplotypes, each present at different frequencies in different populations. Another reason to focus on functional studies is that they can provide evidence for positive selection in the absence of detectable signatures of selection. For instance, Stedman et al.[31] report a single frameshift mutation in the myosin heavy chain protein between humans and chimps that correlates with our decreased jaw musculature. This change would not have been detected with dN/dS methods, but functional testing and correlation of the date of the mutation with fossil records was able to provide a compelling case for the mutation's role in evolution.

In the future, leveraging new technology that catalogs function in a higher throughput manner can increase the rate of validation. This is especially important for selection on non-coding regions, which is difficult to interpret. There are several new approaches for validating enhancer activity in a genome-wide manner[158,159], as well as methods used for dissecting the function of multiple variants within regulatory elements[160]. Genome-editing technology, such as the CRISPR/Cas9 system, holds promise for functional validation in a variant's native context[161]. These technologies have yet to be used for large-scale validations of positively selected regions. One of the main impediments to their use is the necessity of carefully chosen cell types and regions to be tested. Transgenic mouse enhancer assays have shown that enhancers often act in exquisitely specific manners, promoting gene expression in just one small part of the body, at one developmental time point[102]. The luciferase assays I performed in chapter 2, a more medium-throughput approach, would likely yield more positive results if the

haDHS had been tested in more targeted cell types. Thus it may be difficult to test large amounts of sequences in one assay.

In addition to testing for function of non-coding regions using reporter assays, there is an important need to associate regulatory elements with the genes that they control. For instance, connecting haDHS with their target genes is essential to characterizing their function in human evolution. Although we were able to leverage Hi-C data to gain a low-resolution picture of these connections, more specific maps of promoter-DHS connections will improves the inferences we can make. A new medium-throughput technology that couples DNaseI to Hi-C shows promise for more confidently associating enhancers and promoters[162].

Important advances in technology will also improve the applications for ancient DNA in inferences of positive selection. Direct inferences of selection, in which we observe allele frequencies rise rapidly over time were previously impossible for humans. However, as more ancient DNA samples are sequenced, we can begin to sample allele frequencies at different parts of a selective sweep, enabling more confident assessments of an allele's trajectory. More extensive catalogs of ancient genomes are currently being constructed[163-166], and data have already been used to confirm selection at many previously reported loci, as well as identify novel candidates for selection[164]. These ancient genomes will also be key in refining demographic models, which will in turn facilitate inferences of selection. For instance, the degree of population structure in ancient Europe is currently under debate[163,165]. More ancient genomes will also help us characterize our history of introgression. The overall decline in Neandertal ancestry across time is clearly observed in ancient Eurasian genomes[163], and can help estimate the

strength of selection against Neandertal haplotypes, as well as pinpoint the timing of introgression events. Finally, additional sequencing of Neandertal genomes will have numerous impacts on the field, including improving power to detect introgression, facilitating comparisons of fixed differences between humans and Neandertals, and reconstructing the population history of Neandertals themselves.

There is also still much work to be done to characterize types of selection that leave more subtle signatures on the genome. As I mention in the introduction, recent hard sweeps and long term protein-coding selection have been widely studied because they leave dramatic patterns on genetic variation. However, there are several other modes by which selection can act, and my work identifying human-specific regulatory adaptation and selection on introgressed variants represents only two of these additional avenues. Selection on standing variation remains challenging to identify, but may be a primary mechanism for evolution, especially as new alleles become favored following environmental changes or colonization of new locations[167,168]. Furthermore, in polygenic selection, evolution may often act on complex traits by fine-tuning allele frequencies at multiple loci across the genome[169]. Polygenic adaptation has been detected for traits such as height[167,170,171], but more work will be needed to examine traits that remain poorly characterized by GWAS studies. Cases of polygenic adaptation will also be especially difficult to model with functional studies, because they require detecting small effects jointly.

Improving methods and validating predictions will enable us to answer fundamental questions regarding the role of natural selection in human evolution. We still have a poor understanding of the relative frequency with which different types of

selection act on genetic variation. Many have hypothesized that evolution more often works through subtle soft sweeps and polygenic adaptation than immediate selection on novel variation[169,172]. Indeed, my findings on adaptive introgression demonstrate that interbreeding with closely related species can introduce pre-existing beneficial alleles to a gene pool, a process that is surely faster than the evolution of novel alleles. Future work will help place adaptive introgression in the larger context of all positive selection, and help determine how often adaptive introgression has played a role in humans and other species. Another important question to be explored further is the relative contributions of protein-coding and regulatory mutations in evolution. As I mentioned in the introduction, regulatory evolution may play a dominant role in humans and other species, and my work certainly highlights the potential for adaptive changes in gene expression. My work in chapter 2 was specifically designed to detect regulatory changes on the human lineage. However, my work on adaptive introgression in chapter 3 was agnostic to underlying sequence function, but still uncovered numerous instances where an introgressed haplotype acted to control nearby gene expression.

These unanswered questions pertaining to positive selection throughout evolution are only a small part of our overall goal of understanding the human genome. Positive selection is constantly working along side the other forces of purifying selection, mutation, and genetic drift to shape genetic variation. Understanding all of these forces will be necessary if we are to reconstruct the trajectory that human evolution has taken, and learn the genetic basis of the numerous traits that make us who we are, both as a species, and as individuals.

BIBLIOGRAPHY

1. Darwin, C. & Wallace, A. On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection. *Journal of the Proceedings of the Linnean Society of London* **3,** 45–62 (1858).
2. Hughes, A. L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335,** 167–170 (1988).
3. Li, W. H., Wu, C. I. & Luo, C. C. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2,** 150–174 (1985).
4. Nielsen, R. *et al.* A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3,** e170 (2005).
5. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
6. Arbiza, L., Dopazo, J. & Dopazo, H. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS Comput. Biol.* **2,** e38 (2006).
7. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–595 (1989).
8. Fu, Y. X. & Li, W. H. Statistical tests of neutrality of mutations. *Genetics* **133,** 693–709 (1993).
9. Fay, J. C. & Wu, C. I. Hitchhiking under positive Darwinian selection. *Genetics* **155,** 1405–1413 (2000).
10. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837 (2002).
11. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449,** 913–918 (2007).
12. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4,** e72 (2006).
13. Kim, Y. & Nielsen, R. Linkage disequilibrium as a signature of selective sweeps. *Genetics* **167,** 1513–1524 (2004).
14. Lewontin, R. C. & Krakauer, J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74,** 175–195 (1973).
15. Shriver, M. D. *et al.* The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Hum. Genomics* **1,** 274–286 (2004).
16. Akey, J. M., Zhang, G., Zhang, K., Jin, L. & Shriver, M. D. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* **12,** 1805–1814 (2002).
17. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74,** 1111–1120 (2004).
18. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
19. Tishkoff, S. A. *et al.* Convergent adaptation of human lactase persistence in

Africa and Europe. *Nat. Genet.* **39,** 31–40 (2007).
20. Lachance, J. & Tishkoff, S. A. Population Genomics of Human Adaptation. *Annu Rev Ecol Evol Syst* **44,** 123–143 (2013).
21. Sabeti, P. C. *et al.* Positive natural selection in the human lineage. *Science* **312,** 1614–1620 (2006).
22. Fu, W. & Akey, J. M. Selection and adaptation in the human genome. *Annu Rev Genomics Hum Genet* **14,** 467–489 (2013).
23. Ronald, J. & Akey, J. M. Genome-wide scans for loci under selection in humans. *Hum. Genomics* **2,** 113–125 (2005).
24. Tishkoff, S. A. *et al.* Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* **293,** 455–462 (2001).
25. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20,** 110–121 (2010).
26. Prabhakar, S., Noonan, J. P., Pääbo, S. & Rubin, E. M. Accelerated evolution of conserved noncoding sequences in humans. *Science* **314,** 786–786 (2006).
27. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* **5,** e1000471 (2009).
28. Kelley, J. L., Madeoy, J., Calhoun, J. C., Swanson, W. & Akey, J. M. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res.* **16,** 980–989 (2006).
29. Teshima, K. M., Coop, G. & Przeworski, M. How reliable are empirical genomic scans for selective sweeps? *Genome Res.* **16,** 702–712 (2006).
30. Grossman, S. R. *et al.* A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327,** 883–886 (2010).
31. Stedman, H. H. *et al.* Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature* **428,** 415–418 (2004).
32. Garud, N. R., Messer, P. W., Buzbas, E. O. & Petrov, D. A. Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. *PLoS Genet.* **11,** e1005004 (2015).
33. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31,** 3812–3814 (2003).
34. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7,** Unit7.20–7.20.41 (2013).
35. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11,** 801–807 (2014).
36. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7,** 741–746 (2010).
37. Mouse Genome Sequencing Consortium *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
38. Lunter, G., Ponting, C. P. & Hein, J. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput. Biol.* **2,** e5 (2006).
39. Asthana, S. *et al.* Widely distributed noncoding purifying selection in the human

genome. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 12410–12415 (2007).
40. Meader, S., Ponting, C. P. & Lunter, G. Massive turnover of functional sequence in human and other mammalian genomes. *Genome Res.* **20,** 1335–1343 (2010).
41. Ponting, C. P. & Hardison, R. C. What fraction of the human genome is functional? *Genome Res.* **21,** 1769–1776 (2011).
42. Britten, R. J. & Davidson, E. H. Gene regulation for higher cells: a theory. *Science* **165,** 349–357 (1969).
43. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188,** 107–116 (1975).
44. Donnelly, M. P. *et al.* A global view of the OCA2-HERC2 region and pigmentation. *Hum. Genet.* **131,** 683–696 (2012).
45. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–482 (2011).
46. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).
47. Dermitzakis, E. T. & Clark, A. G. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* **19,** 1114–1121 (2002).
48. Wray, G. A. *et al.* The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* **20,** 1377–1419 (2003).
49. Villar, D., Flicek, P. & Odom, D. T. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nat. Rev. Genet.* **15,** 221–233 (2014).
50. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
51. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
52. Grossman, S. R. *et al.* Identifying recent adaptations in large-scale genomic data. *Cell* **152,** 703–713 (2013).
53. Kamberov, Y. G. *et al.* Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell* **152,** 691–702 (2013).
54. Enard, W. *et al.* Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* **418,** 869–872 (2002).
55. Pollard, K. S. *et al.* An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443,** 167–172 (2006).
56. Pollard, K. S. *et al.* Forces shaping the fastest evolving regions in the human genome. *PLoS Genet.* **2,** e168 (2006).
57. Kim, S. Y. & Pritchard, J. K. Adaptive evolution of conserved noncoding elements in mammals. *PLoS Genet.* **3,** 1572–1586 (2007).
58. Bush, E. C. & Lahn, B. T. A genome-wide screen for noncoding elements important in primate evolution. *BMC Evol. Biol.* **8,** 17 (2008).
59. McLean, C. Y. *et al.* Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471,** 216–219 (2011).
60. Pertea, M., Pertea, G. M. & Salzberg, S. L. Detection of lineage-specific evolutionary changes among primate species. *BMC Bioinformatics* **12,** 274 (2011).

61. Prabhakar, S. *et al.* Human-specific gain of function in a developmental enhancer. *Science* **321,** 1346–1350 (2008).
62. Capra, J. A., Erwin, G. D., McKinsey, G., Rubenstein, J. L. R. & Pollard, K. S. Many human accelerated regions are developmental enhancers. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **368,** 20130025–20130025 (2013).
63. Kamm, G. B., Pisciottano, F., Kliger, R. & Franchini, L. F. The developmental brain gene NPAS3 contains the largest number of accelerated regulatory sequences in the human genome. *Mol. Biol. Evol.* **30,** 1088–1102 (2013).
64. Higham, T. *et al.* The timing and spatiotemporal patterning of Neanderthal disappearance. *Nature* **512,** 306–309 (2014).
65. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505,** 43–49 (2014).
66. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328,** 710–722 (2010).
67. Yang, M. A., Malaspinas, A.-S., Durand, E. Y. & Slatkin, M. Ancient structure in Africa unlikely to explain Neanderthal and non-African genetic similarity. *Mol. Biol. Evol.* **29,** 2987–2995 (2012).
68. Sankararaman, S., Patterson, N., Li, H., Pääbo, S. & Reich, D. The date of interbreeding between Neandertals and modern humans. *PLoS Genet.* **8,** e1002947 (2012).
69. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338,** 222–226 (2012).
70. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468,** 1053–1060 (2010).
71. Vernot, B. & Akey, J. M. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343,** 1017–1021 (2014).
72. Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507,** 354–357 (2014).
73. Vernot, B. *et al.* Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* aad9416 (2016). doi:10.1126/science.aad9416
74. Sankararaman, S., Mallick, S., Patterson, N. & Reich, D. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Curr. Biol.* (2016). doi:10.1016/j.cub.2016.03.037
75. Vattathil, S. & Akey, J. M. Small Amounts of Archaic Admixture Provide Big Insights into Human History. *Cell* **163,** 281–284 (2015).
76. Simonti, C. N. *et al.* The phenotypic legacy of admixture between modern humans and Neandertals. *Science* **351,** 737–741 (2016).
77. Grant, B. R. & Grant, P. R. Fission and fusion of Darwin's finches populations. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **363,** 2821–2829 (2008).
78. Pardo-Diaz, C. *et al.* Adaptive introgression across species boundaries in Heliconius butterflies. *PLoS Genet.* **8,** e1002752 (2012).
79. Dannemann, M., Andrés, A. M. & Kelso, J. Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. *Am. J. Hum. Genet.* **98,** 22–33 (2016).
80. Racimo, F., Sankararaman, S., Nielsen, R. & Huerta-Sánchez, E. Evidence for archaic adaptive introgression in humans. *Nat. Rev. Genet.* **16,** 359–371 (2015).

81. Huerta-Sánchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512,** 194–197 (2014).
82. Gittelman, R. M. *et al.* Comprehensive identification and analysis of human accelerated regulatory DNA. *Genome Res.* **25,** 1245–1255 (2015).
83. O'Bleness, M., Searles, V. B., Varki, A., Gagneux, P. & Sikela, J. M. Evolution of genetic and genomic features unique to the human lineage. *Nat. Rev. Genet.* **13,** 853–866 (2012).
84. Dennis, M. Y. *et al.* Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **149,** 912–922 (2012).
85. Sudmant, P. H. *et al.* Evolution and diversity of copy number variation in the great ape lineage. *Genome Res.* **23,** 1373–1382 (2013).
86. Eddy, S. R. The C-value paradox, junk DNA and ENCODE. *Curr. Biol.* **22,** R898–9 (2012).
87. Niu, D.-K. & Jiang, L. Can ENCODE tell us how much junk DNA we carry in our genome? *Biochem. Biophys. Res. Commun.* **430,** 1340–1343 (2013).
88. Graur, D. *et al.* On the immortality of television sets: 'function' in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol* **5,** 578–590 (2013).
89. Doolittle, W. F. Is junk DNA bunk? A critique of ENCODE. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 5294–5300 (2013).
90. Galas, D. J. & Schmitz, A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.* **5,** 3157–3170 (1978).
91. Dorschner, M. O. *et al.* High-throughput localization of functional elements by quantitative chromatin profiling. *Nat. Methods* **1,** 219–225 (2004).
92. Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.* **18,** 1814–1828 (2008).
93. Lowe, C. B. *et al.* Three periods of regulatory innovation during vertebrate evolution. *Science* **333,** 1019–1024 (2011).
94. Vernot, B. *et al.* Personal and population genomics of human regulatory variation. *Genome Res.* **22,** 1689–1697 (2012).
95. Antinucci, P., Nikolaou, N., Meyer, M. P. & Hindges, R. Teneurin-3 specifies morphological and functional connectivity of retinal ganglion cells in the vertebrate visual system. *Cell Rep* **5,** 582–592 (2013).
96. Merlin, S. *et al.* Deletion of Ten-m3 induces the formation of eye dominance domains in mouse visual cortex. *Cereb. Cortex* **23,** 763–774 (2013).
97. Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K.-D. & Wray, G. A. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat. Genet.* **39,** 1140–1144 (2007).
98. Taylor, M. S. *et al.* Rapidly evolving human promoter regions. *Nat. Genet.* **40,** 1262–3– author reply 1263–4 (2008).
99. Kostka, D., Hubisz, M. J., Siepel, A. & Pollard, K. S. The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. *Mol. Biol. Evol.* **29,** 1047–1057 (2012).
100. Galtier, N. & Duret, L. Adaptation or biased gene conversion? Extending the null

hypothesis of molecular evolution. *Trends Genet.* **23,** 273–277 (2007).
101. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
102. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35,** D88–92 (2007).
103. Lefebvre, V., Li, P. & de Crombrugghe, B. 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 (1998).
104. Graham, A. Development of the pharyngeal arches. *Am. J. Med. Genet. A* **119A,** 251–256 (2003).
105. Soranzo, N. *et al.* A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat. Genet.* **41,** 1182–1190 (2009).
106. van Arensbergen, J., van Steensel, B. & Bussemaker, H. J. In search of the determinants of enhancer-promoter interaction specificity. *Trends Cell Biol.* **24,** 695–702 (2014).
107. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).
108. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489,** 109–113 (2012).
109. Ay, F., Bailey, T. L. & Noble, W. S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.* **24,** 999–1011 (2014).
110. Maricic, T. *et al.* A recent evolutionary change affects a regulatory element in the human FOXP2 gene. *Mol. Biol. Evol.* **30,** 844–852 (2013).
111. Atchison, M. L. Function of YY1 in Long-Distance DNA Interactions. *Front Immunol* **5,** 45 (2014).
112. Morin, R. D. *et al.* Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. *Nat. Genet.* **42,** 181–185 (2010).
113. Jiao, Y. *et al.* DAXX/ATRX, MEN1, and mTOR pathway genes are frequently altered in pancreatic neuroendocrine tumors. *Science* **331,** 1199–1203 (2011).
114. Reilly, S. K. *et al.* Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347,** 1155–1159 (2015).
115. Steinberg, S. J. *et al.* Peroxisome biogenesis disorders. *Biochim. Biophys. Acta* **1763,** 1733–1748 (2006).
116. Shibata, Y. *et al.* Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. *PLoS Genet.* **8,** e1002789 (2012).
117. Cotney, J. *et al.* The evolution of lineage-specific regulatory activities in the human embryonic limb. *Cell* **154,** 185–196 (2013).
118. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).
119. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43,** 264–268 (2011).
120. Neph, S. *et al.* BEDOPS: high-performance genomic feature operations.

Bioinformatics **28,** 1919–1920 (2012).
121. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42,** D749–55 (2014).
122. Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinformatics* **12,** 41–51 (2011).
123. Storey, J. D. & Dabney, A. qvalue: Q-value estimation for false discovery rate control.
124. Charlesworth, B. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet. Res.* **63,** 213–227 (1994).
125. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9,** 999–1003 (2012).
126. Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res.* **41,** W77–83 (2013).
127. Gittelman, R. M. *et al.* Archaic hominin admixture fascilitated adaptation to out-of-Africa environments.
128. Vitti, J. J., Grossman, S. R. & Sabeti, P. C. Detecting natural selection in genomic data. *Annu. Rev. Genet.* **47,** 97–120 (2013).
129. Heyninck, K. *et al.* The zinc finger protein A20 inhibits TNF-induced NF-kappaB-dependent gene expression by interfering with an RIP- or TRAF2-mediated transactivation signal and directly binds to a novel NF-kappaB-inhibiting protein ABIN. *J. Cell Biol.* **145,** 1471–1482 (1999).
130. Li, M. J. *et al.* GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.* **44,** D869–76 (2016).
131. Sulem, P. *et al.* Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat. Genet.* **39,** 1443–1452 (2007).
132. Hider, J. L. *et al.* Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry. *BMC Evol. Biol.* **13,** 150 (2013).
133. Sikora, M. *et al.* Population genomic analysis of ancient and modern genomes yields new insights into the genetic ancestry of the Tyrolean Iceman and the genetic structure of Europe. *PLoS Genet.* **10,** e1004353 (2014).
134. Sulem, P. *et al.* Two newly identified genetic determinants of pigmentation in Europeans. *Nat. Genet.* **40,** 835–837 (2008).
135. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).
136. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).
137. Mendez, F. L., Watkins, J. C. & Hammer, M. F. Neandertal origin of genetic variation at the cluster of OAS immunity genes. *Mol. Biol. Evol.* **30,** 798–801 (2013).
138. Hornung, V., Hartmann, R., Ablasser, A. & Hopfner, K.-P. OAS proteins and cGAS: unifying concepts in sensing and responding to cytosolic nucleic acids. *Nat. Rev. Immunol.* **14,** 521–528 (2014).
139. Bonnevie-Nielsen, V. *et al.* Variation in antiviral 2',5'-oligoadenylate synthetase (2‘5’AS) enzyme activity is controlled by a single-nucleotide polymorphism at a

splice-acceptor site in the OAS1 gene. *Am. J. Hum. Genet.* **76,** 623–633 (2005).
140. Deschamps, M. *et al.* Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. *Am. J. Hum. Genet.* **98,** 5–21 (2016).
141. Kawai, T. & Akira, S. The role of pattern-recognition receptors in innate immunity: update on Toll-like receptors. *Nat. Immunol.* **11,** 373–384 (2010).
142. Lee, S. M. Y. *et al.* Toll-like receptor 10 is involved in induction of innate immune responses to influenza virus infection. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 3793–3798 (2014).
143. Nagashima, H. *et al.* Toll-like Receptor 10 in Helicobacter pylori Infection. *J. Infect. Dis.* **212,** 1666–1676 (2015).
144. Mayerle, J. *et al.* Identification of genetic loci associated with Helicobacter pylori serologic status. *JAMA* **309,** 1912–1920 (2013).
145. Hinds, D. A. *et al.* A genome-wide association meta-analysis of self-reported allergy identifies shared and allergy-specific susceptibility loci. *Nat. Genet.* **45,** 907–911 (2013).
146. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
147. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).
148. Ewing, G. & Hermisson, J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26,** 2064–2065 (2010).
149. Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 11983–11988 (2011).
150. Juric, I., Aeschbacher, S. & Coop, G. The Strength of Selection Against Neanderthal Introgression. *bioRxiv* 030148 (2015). doi:10.1101/030148
151. Rosenbloom, K. R. *et al.* The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.* **43,** D670–81 (2015).
152. 't Hoen, P. A. C. *et al.* Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat. Biotechnol.* **31,** 1015–1022 (2013).
153. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
154. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7,** 500–507 (2012).
155. R Core Team. *R: A language and environment for statistical computing*. *R Foundation for Statistical Computing* (2013). at <http://www.R-project.org>
156. Wurfel, M. M. *et al.* Toll-like receptor 1 polymorphisms affect innate immune responses and outcomes in sepsis. *Am. J. Respir. Crit. Care Med.* **178,** 710–720 (2008).
157. Saal, L. H. *et al.* BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biol.* **3,** SOFTWARE0003 (2002).
158. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339,** 1074–1077 (2013).

159. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30,** 265–270 (2012).
160. Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* **23,** 800–811 (2013).
161. Findlay, G. M., Boyle, E. A., Hause, R. J., Klein, J. C. & Shendure, J. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* **513,** 120–123 (2014).
162. Ma, W. *et al.* Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nat. Methods* **12,** 71–78 (2015).
163. Fu, Q. *et al.* The genetic history of Ice Age Europe. *Nature* **534,** 200–205 (2016).
164. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528,** 499–503 (2015).
165. Seguin-Orlando, A. *et al.* Paleogenomics. Genomic structure in Europeans dating back at least 36,200 years. *Science* **346,** 1113–1118 (2014).
166. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513,** 409–413 (2014).
167. Field, Y. *et al. Detection of human adaptation during the past 2,000 years*. *bioRxiv* 052084 (Cold Spring Harbor Labs Journals, 2016). doi:10.1101/052084
168. Przeworski, M., Coop, G. & Wall, J. D. The signature of positive selection on standing genetic variation. *Evolution* **59,** 2312–2323 (2005).
169. Pritchard, J. K., Pickrell, J. K. & Coop, G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.* **20,** R208–15 (2010).
170. Berg, J. J. & Coop, G. A population genetic signal of polygenic adaptation. *PLoS Genet.* **10,** e1004412 (2014).
171. Turchin, M. C. *et al.* Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat. Genet.* **44,** 1015–1019 (2012).
172. Hernandez, R. D. *et al.* Classic selective sweeps were rare in recent human evolution. *Science* **331,** 920–924 (2011).
173. Capra, J. A., Hubisz, M. J., Kostka, D., Pollard, K. S. & Siepel, A. A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. *PLoS Genet.* **9,** e1003684 (2013).
174. Green, P. & Ewing, B. Comment on "Evidence of abundant purifying selection in humans for recently acquired regulatory functions". *Science* **340,** 682–682 (2013).
175. Eyre-Walker, A. The genomic rate of adaptive evolution. *Trends Ecol. Evol. (Amst.)* **21,** 569–575 (2006).
176. Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18,** 337–338 (2002).
177. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337,** 64–69 (2012).
178. Fu, W., Gittelman, R. M., Bamshad, M. J. & Akey, J. M. Characteristics of neutral and deleterious protein-coding variation among individuals and populations. *Am. J. Hum. Genet.* **95,** 421–436 (2014).
179. You, F. M. *et al.* BatchPrimer3: a high throughput web application for PCR and

sequencing primer design. *BMC Bioinformatics* **9,** 253 (2008).

Appendix A

This appendix contains published material[132]

Hider *et al. BMC Evolutionary Biology* 2013, **13**:150
http://www.biomedcentral.com/1471-2148/13/150

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry

Jessica L Hider[1], Rachel M Gittelman[2], Tapan Shah[3], Melissa Edwards[1], Arnold Rosenbloom[3], Joshua M Akey[2] and Esteban J Parra[1*]

Abstract

Background: Currently, there is very limited knowledge about the genes involved in normal pigmentation variation in East Asian populations. We carried out a genome-wide scan of signatures of positive selection using the 1000 Genomes Phase I dataset, in order to identify pigmentation genes showing putative signatures of selective sweeps in East Asia. We applied a broad range of methods to detect signatures of selection including: 1) Tests designed to identify deviations of the Site Frequency Spectrum (SFS) from neutral expectations (Tajima's D, Fay and Wu's H and Fu and Li's D* and F*), 2) Tests focused on the identification of high-frequency haplotypes with extended linkage disequilibrium (iHS and Rsb) and 3) Tests based on genetic differentiation between populations (LSBL). Based on the results obtained from a genome wide analysis of 25 kb windows, we constructed an empirical distribution for each statistic across all windows, and identified pigmentation genes that are outliers in the distribution.

Results: Our tests identified twenty genes that are relevant for pigmentation biology. Of these, eight genes (*ATRN, EDAR, KLHL7, MITF, OCA2, TH, TMEM33* and *TRPM1,*) were extreme outliers (top 0.1% of the empirical distribution) for at least one statistic, and twelve genes (*ADAM17, BNC2, CTSD, DCT, EGFR, LYST, MC1R, MLPH, OPRM1, PDIA6, PMEL (SILV)* and *TYRP1*) were in the top 1% of the empirical distribution for at least one statistic. Additionally, eight of these genes (*BNC2, EGFR, LYST, MC1R, OCA2, OPRM1, PMEL (SILV)* and *TYRP1*) have been associated with pigmentary traits in association studies.

Conclusions: We identified a number of putative pigmentation genes showing extremely unusual patterns of genetic variation in East Asia. Most of these genes are outliers for different tests and/or different populations, and have already been described in previous scans for positive selection, providing strong support to the hypothesis that recent selective sweeps left a signature in these regions. However, it will be necessary to carry out association and functional studies to demonstrate the implication of these genes in normal pigmentation variation.

Background

The major out of Africa migrations of anatomically modern humans took place within the last 60,000-50,000 years [1]. As a result of these migrations, humans encountered novel environments, with varying climates, pathogens and foods and adapted to these new conditions via natural selection. One of the climatic factors showing clear geographic patterns is ultraviolet (UV) radiation, which is more intense and shows less seasonality in equatorial and tropical areas than in high latitude regions [2]. Skin pigmentation, which is primarily determined by the amount, type and distribution of cutaneous melanin, also shows a clear latitudinal gradient, and this has been explained as the result of natural selection [3-8]. In agreement with this hypothesis, genome-wide scans have shown that many genes involved in the pigmentation pathway show signatures of positive selection [9-21]. Interestingly, most of the putative selection signatures have been identified in European and East Asian populations, indicating that the majority of the selective sweeps took place after the out-of-Africa migration of modern humans. Although some of the pigmentation

* Correspondence: esteban.parra@utoronto.ca
[1]Department of Anthropology, University of Toronto at Mississauga, Mississauga, Ontario, Canada
Full list of author information is available at the end of the article

genes show signatures of selection that are shared between European and East Asian populations (e.g. *KITLG*) [12,18] many genes show positive selection signals in only one population (e.g. *SLC24A5* and *SLC45A2* in Europe, *DCT* in East Asia) or independent signals in European and East Asian groups (*OCA2*). These findings support an evolutionary model in which the most important changes in pigmentary traits occurred after the migration out-of-Africa and the separation of the lineages that gave rise to contemporary European and East Asian populations [5,12,22-25].

Most of the surveys of signatures of selection published to date have been based on data from the HapMap project or the Human Genome Diversity Project (HGDP), which primarily captured common genetic variants. The recent availability of the 1000 Genomes project Phase I data, which includes full genome sequences (based on a combination of low-coverage whole genome sequencing and targeted deep exome sequencing) for more than 1,000 individuals from 14 populations, has opened new opportunities to study genetic variation in our species [26]. In particular, the improved representation of rare variants and the decreased bias in variant detection would be expected to increase the power of some of the tests used to identify selective sweeps. In this study, we applied a range of genome-wide methods to detect signatures of selection in the 1000 Genomes Phase I dataset. The methods employed include: 1) Tests designed to identify deviations of the Site Frequency Spectrum (SFS) from neutral expectations (Tajima's D, Fay and Wu's H and Fu and Li's D* and Fu's F*), 2) Tests focused on the identification of high-frequency haplotypes with extended Linkage Disequilibrium (LD) (iHS and Rsb) and 3) Tests based on genetic differentiation between populations (LSBL). The main goal of the study was to identify pigmentation genes that have been the target of positive selection in East Asia. To date, the overwhelming majority of the genetic association studies focused on pigmentary traits have been carried out in European populations and as a result the last decade has brought a much better understanding of the genetic basis of normal pigmentation variation in this group [6,12,27,28]. In contrast to the long list of genes that have been associated with pigmentary traits in European populations, there is very limited knowledge concerning the genes involved in pigmentary traits in East Asia. Notable exceptions are the genes *OCA2* and *MC1R*, which harbor non-synonymous mutations, rs1800414 (His615Arg) in OCA2 and rs885479 (Arg163Gln) in MC1R, that have been associated with skin pigmentation in East Asian populations [23,29,30]. These polymorphisms are present in high frequency in East Asian populations, and are absent or present at low frequencies in European and African populations, suggesting again that there has been convergent evolution towards depigmentation in Europe and East Asia. Additional research efforts in East Asian populations, or admixed populations showing a substantial East Asian contribution [31], will be necessary in order to elucidate the genetic architecture of pigmentation in East Asian populations, and more generally, the evolutionary events responsible for the pigmentary changes that took place after the out-of-Africa migration of modern humans. By identifying pigmentation genes showing putative signatures of selective sweeps in East Asia, we will be able to prioritize a list of genes for subsequent association studies in East Asian samples characterized with quantitative methods (e.g. skin reflectometry).

Methods

Samples

All the statistical analyses were completed using the 1,000 Genomes Phase 1 data, which includes approximately 38 million Single Nucleotide Polymorphisms (SNPs) [26]. Indels were excluded from all the analyses. The 1000 genomes data set includes samples Japanese from Tokio (JPT), Han Chinese from Beijing (CHB) and Southern Han Chinese (CHS).

Statistics used to identify putative signatures of positive selection

1-Statistics based on the Site Frequency Spectrum (SFS)

These statistics compare different estimators of the population mutation rate $\Theta = 4N\mu$, which have the same expectation under neutrality.

Tajima's D [32]. This test compares estimates of Θ derived from the average number of pairwise differences (π) and the number of segregating sites (S).
Fu and Li's D* [33]. This test compares estimates of Θ derived from the number of segregating sites (S) and the number of singleton mutations (η_s, alleles appearing only once in the sample).
Fu's F* [34]. This test compares estimates of Θ derived from the average number of pairwise differences (π) and the number of singleton mutations (η_s).

For these three tests, negative values indicate an excess of rare polymorphisms, and positive values an excess of intermediate-frequency alleles with respect to neutral expectations.

Fay and Wu's H [35]. In contrast to the previous three tests, H requires information on allele state (ancestral vs. derived). This test compares estimates of Θ derived from the average number of pairwise differences (π) with another estimate derived from the frequency of derived alleles at segregating sites (Θ_H). H is negative when

derived alleles are found at high frequency, with respect to neutral expectations.

These four statistics were estimated for non-overlapping 25 kilobase windows, using a Python script. The statistics were calculated independently in three East Asian samples from the 1000 Genomes Phase 1 panel: Han Chinese from Beijing (CHB) (97), Southern Han Chinese (CHS) (100) and Japanese from Tokyo (JPT) (89). Variants that did not pass the 1000 genomes project filtering metrics were masked, as well as any variants in which the ancestral allele could not be determined. Windows with less than 10 markers were excluded from further analyses.

2-Tests based on genetic differentiation

We estimated genetic differentiation using the Locus Specific Branch Length (LSBL), as described in Shriver et al., 2004 [36]. In this case, we focused on the identification of regions with high East Asian LSBL values, indicating strong differentiation between East Asia and Europe/Africa. For these analyses, we used the combined 1000 Genomes Phase I East Asian (Han Chinese from Beijing, Southern Han Chinese and Japanese from Tokyo, N = 286), European (Tuscans from Italy, British, Finnish, Iberians, and Utah residents with Western European ancestry, N = 379) and African (Yoruba from Nigeria and Luhya from Kenya, N = 185) samples. East Asian LSBL values were estimated from the East Asian-African, East Asian-European and African-European pairwise F_{ST} distances for each locus using the formula LSBL(Eas) = (Eas-Eur F_{ST} + Eas-Afr F_{ST} −Afr-Eur F_{ST})/2. F_{ST} values were calculated with the program VCFTOOLS using Weir and Cockerham (1984) unbiased estimator [37]. Negative F_{ST} values were converted to zero. Using a combination of shell and python scripts, we created non-overlapping windows of 25 kilobases, and reported for each window the maximum LSBL. Windows with less than 10 markers were excluded from further analyses.

3-Long-range haplotype tests

We employed two approaches based on haplotype diversity. For these tests, we restricted the analyses to markers with minor allele frequencies equal or higher than 5%. The statistics are based on the combined 1000 Genomes Phase I East Asian samples (iHS test), and the combined 1000 Genomes Phase I East Asian, European and African samples (Rsb tests). More details about the statistics are described below.

iHS (Integrated Haplotype Score) iHS compares integrated EHH (Extended Haplotype Homozygosity) values between alleles at a given SNP. EHH quantifies the breakdown of LD at increasing distances from each allele (ancestral or derived). Large negative iHS values are indicative of unusually long haplotypes carrying the derived allele and large positive values are associated with long haplotypes carrying the ancestral allele [38]. iHS values were estimated using the program rehh [39].

Rsb Rsb is a standardized ratio of iES (Integrated EHHS) from two populations. iES integrates the area under the curve of site-specific EHH (EHHS) [11]. Extreme values of Rsb indicate slower haplotype homozygosity decay in one population versus another. This test was designed to identify potential sweeps that have occurred only in one population. Given that we are primarily interested in identifying sweeps that are specific to East Asian populations, we focused on the comparison between East Asian and European populations, and East Asian and African populations. In this particular situation, extreme positive values of Rsb will indicate longer haplotypes in East Asian populations than in European or African populations. Rsb(Eas-Eur) and Rsb(Eas-Afr) were estimated using the program rehh [39].

After obtaining the iHS, Rsb(Eas-Eur), and Rsb(Eas-Afr) statistics for each locus, we used a combination of shell and python scripts to report the results for non-overlapping windows of 25 kilobases, indicating for each window the maximum absolute value of iHS (or the maximum value of Rsb). Windows with less than 10 markers were excluded from further analyses.

Construction of empirical distribution of p-values based on results for 25 kb windows and identification of putative pigmentation genes under positive selection.

Based on the results obtained in the analyses of 25 kb windows (e.g. values obtained for each of the SFS statistics, and maximum values for LSBL, iHS and Rsb, see above for additional information), we sorted the windows in descending order based on the values of the relevant statistics, and identified pigmentation genes that are outliers in the empirical distribution (top 0.1% or 1% of the distribution), following the approach detailed below:

1-Identification of extreme outliers with empirical p-values < 0.001 and annotation of the relevant windows using the DAVID database

We used the ENSEMBL genome browser (http://useast.ensembl.org/index.html) to identify genes overlapping with the top 0.1% of the 25 kb windows for each statistic. These genes were then annotated using the DAVID database (Database for Annotation, Visualization and Integrated Discovery) [40] in order to identify genes involved in the pigmentation pathway. Briefly, we used the ENSEMBL gene IDs retrieved from the ENSEMBL genome browser as input to perform a functional annotation of each gene using DAVID Functional Annotation

Tool. This tool provides different types of annotations for each gene, including annotations based on functional categories, gene ontology (e.g. GOTERM, PANTHER), pathways (e.g. BIOCARTA, KEGG_PATHWAY), protein domains (e.g. INTERPRO) and protein interactions.

2-Identification of known pigmentation genes with empirical p-values < 0.01

We prepared a list of known pigmentation genes that 1) Have been associated with pigmentary traits in association studies or 2) Have been reported as outliers in previous scans of positive selection in human populations. The list included the following genes:

ADAM17, ADAMTS20, AP3D1, ASIP, ATRN, BLOC1S6 (PLDN), BNC2, CTSD, DCT, DRD2, DTNBP1, EDAR, EDN2, EGFR, HPS1, IRF4, KIT, KITLG, LYST, MATP (SLC45A2), MC1R, MITF, MLPH, MYO5A, MYO7A, OCA2/HERC2, OPRM1, PAX3, PDIA6, PMEL (SILV), POMC, PPARD, RAB27A, RAD50, RGS19, SLC24A4, SLC24A5, TYR, TYRP1, TP53BP1, TRPM1 and *TPCN2*.

We retrieved the results of the 8 statistics analyzed in this study for all the 25 Kb windows overlapping the aforementioned genes, and identified the genes with windows in the top 1% of the empirical distributions.

Therefore, all the genes reported in the Results and Discussion section are outliers that show statistics in the top 1% of the empirical distribution, and in some cases, the top 0.1% of the empirical distribution.

Results and discussion

We carried out genome-wide scans for signatures of selection in East Asians, with a major focus on the identification of pigmentation genes that have been under positive selection in this population group. We used three types of statistics: Statistics based on the Site Frequency Spectrum (SFS) (D, D*, F* and H), statistics based on genetic differentiation (LSBL) and long-range haplotype tests (iHS and Rsb) (See the Materials and Methods section for more details about each statistic). Importantly, these statistics are based on different characteristics of the data and are powered to identify different types of selective sweeps. For example, tests based on the SFS are primarily powered to identify older selective events and recently completed sweeps, whereas long-range haplotype tests are more useful to identify more recent events (<30,000 years ago) and incomplete or partial sweeps [41]. All statistical analyses were based on the 1,000 Genomes Phase 1 reference samples, which include approximately 38 million SNPs. For the statistical analyses, we created non-overlapping windows of 25 kb, which were used to construct an empirical distribution for each statistic. We identified genes located within the top 0.1% of the windows for each statistic and these genes were then annotated using the DAVID database in order to select genes that may potentially be involved in the pigmentation pathway. In addition to these extreme outliers, we also explored if a list of genes that have been previously associated with pigmentary traits in association studies or reported as outliers in previous scans of positive selection in human populations were located in the top 1% of the empirical distributions.

The top 0.1% 25Kb windows identified for the different statistics (957 windows) are depicted in Additional file 1: Table S1, and the basic annotations for the genes retrieved using the DAVID database (422 genes) are provided as Additional file 2: Table S2. Many of the genes reported to harbor signatures of positive selection in previous studies including East Asian populations, based on a wide range of methods, such as the Composite of Multiple Signals (CMS) test [42], the XP-EHH test [15,17] the iHS test [17,43], the LRH test [43], the Rsb test [11] and the XP-CLR test [20] are also outliers in our study. Overall, 79 genes described in these studies were also identified in our analyses, and these genes are highlighted in red in Additional file 2: Table S2. It is important to note that among the genes reported in Additional file 2: Table S2, many are located on the same genomic regions. In fact, several genomic regions are characterized by the presence of large numbers of outlier windows, such as the *EDAR* region on chromosome 2, or a genomic region on chromosome 17 characterized by extreme values for several SFS statistics (Additional file 1: Table S1 and Additional file 2: Table S2). Presumably, these are genomic regions that have been under strong and relatively recent positive selection, and the selective sweeps left a strong signature in these regions, encompassing several genes. Further analyses would be needed to determine which genes were targets of positive selection in these regions. Our primary goal in this study has been to identify pigmentation genes showing putative signals of positive selection.

Table 1 shows the list of outlier genes that are relevant for pigmentation biology. There are 20 genes in the list. Of these, 8 genes (*ATRN, EDAR, KLHL7, MITF, OCA2, TH, TMEM33* and *TRPM1,*) were extreme outliers (top 0.1% of the empirical distribution) for at least one statistic, and 12 genes (*ADAM17, BNC2, CTSD, DCT, EGFR, LYST, MC1R, MLPH, OPRM1, PDIA6, PMEL (SILV)* and *TYRP1*) were in the top 1% of the empirical distribution for at least one statistic. Most of the genes are outliers for more than one statistic, and show multiple significant windows. It is important to note that, with the exception of *TH*, *KLHL7* and *CTSD*, these genes have already been described in genome-wide scans of signatures of selection in previous studies [12-17,20,21,38, 44-49], lending strong support to the hypothesis that positive selection has substantially shaped the patterns of variation of these genes. Additionally, eight of these

Table 1 Pigmentation genes that are outliers based on the empirical distribution of different tests of positive selection

Gene	iHS	LSBL	Rsb-Eas-Afr	Rsb-Eas-Eur	Tajima D	H	D*	F*
Top 0.1%								
EDAR	+(0.0019, 2)	++ (9.3E-06, 5)	++ (9.6E-04, 2)	+(0.0049, 1)	++(6.2E-04, 3) CHB			
MITF					++ (9.9E-04, 4) CHB/CHS/JPT	+(0.0037, 3) CHB/CHS/JPT	++ (2.1E-04, 5) CHS/JPT	++ (1.8E-04, 6) CHS/JPT
ATRN					+(0.0077, 2) CHB		++ (9.6E-04, 3) CHB	++ (9.4E-04, 6) CHB/CHS
OCA2		++(5.6E-05, 5)						
TRPM1			+(0.0012, 6)	+(0.0019, 5)	++(8.9E-04, 1) CHS			
TH	++(7.9E-04, 1)		+(0.0073, 1)	+(0.0028, 1)				
KLHL7							++(8.81E-04, 5) CHB/CHS	+(0.0010, 5) CHB/CHS
TMEM33					++(1.72E-04, 9) CHB/CHS/JPT		+(0.0029, 3) CHB	+(0.0017, 3) CHB
Top 1%								
PMEL (SILV)							+(0.0051, 2) CHS	+(0.0072, 2) CHS
BNC2					+ (0.0026, 2) CHB/CHS	+ (0.0056, 2) CHS/JPT	+ (0.0051, 1) CHS	+ (0.0039, 1) CHS
EGFR		+ (0.0029, 3)	+ (0.0056, 1)	+ (0.0028, 1)				
LYST			+ (0.0019, 1)				+ (0.0072, 1) CHS	+ (0.0057, 1) CHS
DCT		+ (0.0099, 1)					+(0.0064, 1) CHB	+(0.0067, 1) CHB
OPRM1					+ (0.0079, 1) JPT			
TYRP1						+(0.0081, 1) CHS		
MC1R	+(0.0073, 1)		+ (0.0027, 1)					
MLPH			+(0.0015, 4)			+(0.0014, 3) CHB/CHS/JPT		
ADAM17			+(0.0015, 2)	+(0.0044, 2)	+(0.0021, 3) JPT			
CTSD							+(0.0045, 1) JPT	+(0.0044, 1) JPT
PDIA6			+(0.0069, 1)			+(0.0072, 1) JPT		

The analysis was based on windows of 25 kilobases. In each cell, we indicate if the gene is in the top 0.1% (++) or 1% (+) of the empirical distribution for the relevant statistics. In parenthesis, we report the smallest p-value observed for any of the windows overlapping the gene, and the number of windows with p-values < 0.01. For the Site Frequency Spectrum tests (D, H, D*, and F*) we also indicate the East Asian populations in which we identified outlier 25 kb windows.

genes (*BNC2, EGFR, LYST, MC1R, OCA2, OPRM1, PMEL (SILV)* and *TYRP1*) have been associated with pigmentary traits in association studies. The *OCA2* gene is of particular interest, because different haplotypes are associated with pigmentary traits in Europeans and East Asians. Variants located in the nearby *HERC2* gene, which affect the transcription of the *OCA2* gene, are strongly associated with blue eye color in European populations [50-55]. Another non-synonymous variant, which is common in East Asian populations but absent or very rare in Europe, has been associated with skin pigmentation in East Asia [23,29]. Several polymorphisms in the *MC1R* gene show a strong association with red hair/fair skin in European populations (Asp84Glu, Arg151Cys, Arg160Trp and Asp294His), and other variants also show a weak association with these traits (Val60Leu, Val92Met and Arg163Gln) [56]. Interestingly, the derived 163Gln allele, which is present in very high frequencies in East Asian populations (>60%), but very low frequencies in European and African populations, has been recently associated with lighter skin in an East Asian sample [30]. Polymorphisms in the *TYRP1* gene have been associated with hair and iris

color in European populations [57], and a non-synonymous mutation that is found only in Oceania was recently associated with blond hair in Melanesians [58]. Frudakis et al. [59] reported association of haplotypes in the *PMEL (SILV)* gene with iris color. The *BCN2* gene has been associated with skin pigmentation and freckling in European populations [60,61]. Variants in the *LYST* gene have been associated with eye color in a Dutch sample [53]. Finally, polymorphisms in the genes *EGFR* and *OPRM1* have been recently associated with skin pigmentation in admixed samples from the New World [49]. We provide more detailed information about each gene, its relevance in the pigmentation pathway, and a description of previous natural selection scans or association studies with pigmentary phenotypes, if relevant, as Additional file 3.

One of the methods employed in our study (LSBL) was designed to highlight genomic regions showing extreme differentiation in one population, with respect to other groups. In this study, we were primarily interested in identifying genomic regions that differentiate East Asian populations with respect to Europeans and Africans. In particular, we would like to find regions in which positive selection may have driven the reduction of melanin levels specifically in East Asia. In order to explore this in more detail, we used the program Haploview [62] to compare the haplotype structure of the candidate pigmentation genes showing large LSBL values in East Asians (*EDAR, OCA2, EGFR* and *DCT*) with the haplotype structure observed in European populations, which are also characterized by reduced melanin content. For the top LSBL windows found for each of these genomic regions, we identified overlapping common markers (frequency higher than 1%) between the East Asian and European 1000 Genomes reference samples, and used the program Haploview to generate the haplotype block structure in each population, using the default algorithm [63]. As expected, we observed very large differences in haplotype frequencies in these regions between the East Asian and European populations. These haplotype differences range between 59% (*DCT*) and 90% (*EDAR*) and these contrasting patterns indicate that positive selection may have favored specific haplotypes in East Asian populations. Figure 1 shows the haplotype structure observed for the *OCA2* gene in East Asian and European populations. The largest differences in frequency are observed for East Asian haplotype blocks 4 and 5, which span slightly more than 10 kilobases, from position 28,187,772 to 28,199,863 on chromosome 15. Interestingly, the non-synonymous variant that has been associated with skin pigmentation in East Asians, rs1800414 [23,29] is located within this region (genomic position 28,197,037 in Genome Built 37.3). Other studies have also reported distinct signatures of positive selection and different haplotype distributions for OCA2 in Europe and East Asia [12,14,24,64,65]. The haplotype structure of the genes *EDAR*, *DCT*, and *EGFR* in European and East Asian populations is provided as Additional file 4. These analyses confirm that some genes relevant for pigmentation biology show extreme haplotype differentiation between European and East Asian populations, and suggest that a careful analysis of haplotype variation, in combination with a detailed annotation of the variants present in the relevant windows, may help to identify the genetic variants responsible for the selective sweeps in East Asians.

In summary, we have carried out a genome-wide analysis of selection signatures in East Asian populations, focusing on genes that are relevant for pigmentation biology. This analysis allowed us to identify a number of genes that show extremely unusual patterns of genetic variation in East Asia. It is in principle possible that some of these findings are false positives and are not due to the action of recent positive selection in East Asian populations. However, most of these genes are outliers for different tests and/or different populations (CHB, CHS, JPT), and have been described in previous scans for positive selection, providing strong support to the hypothesis that recent selective sweeps left a signature in these regions. It is important to note that, even if selective sweeps are responsible for these unusual patterns of variation, it is possible that the selective factors involved did not have any effect on melanin levels in East Asian populations. Many of these genes are expressed widely and have a broad range of functions, and these selective sweeps may be related to phenotypes other than pigmentation. For example, certain mutations of the Ectodysplasia A receptor gene *(EDAR)*, which is an extreme outlier based on three different types of test, are associated with pigmentary phenotypes in mice (http://www.informatics.jax.org/), and for this reason *EDAR* is a pigmentation candidate gene. However, this gene is also important in the development of hair, teeth, and other ectodermal derivatives, and mutations in this gene have been associated with several traits in humans, including hypohidrotic ectodermal dysplasia [66] shovel-shaped incisors [67] and hair thickness [68]. In a recent study [69], researchers generated a knock-in mouse to test the phenotypic consequences of the EDARV370A (370A) polymorphism, which has been associated with hair thickness and incisor shoveling in East Asian populations. The researchers found that 370A homozygous mice had thicker hair than 370V homozygous mice, similarly to the patterns observed in human populations. Importantly, they also observed that the 370A mice had smaller mammary fat pads and increased eccrine gland numbers. An association study in individuals of Han descent showed that the 370A allele was associated with shoveling of the upper incisors and also eccrine gland density. These findings suggest that the dramatic

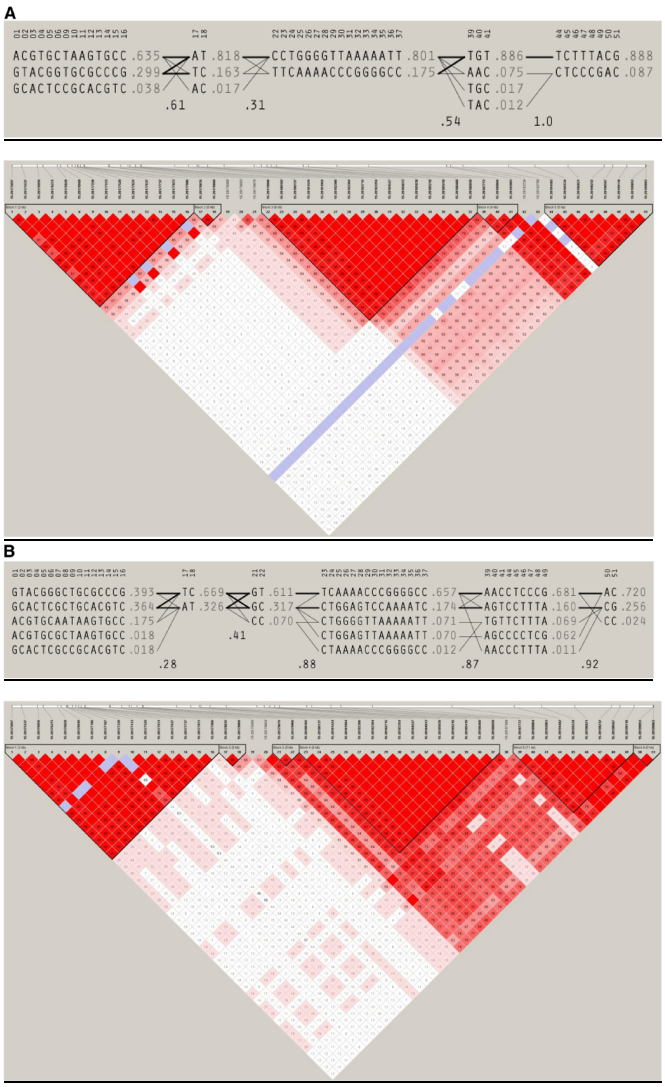

Figure 1 Haplotype block structure of OCA2 gene (window chr15:28,175,001-28,200,000) A. In East Asian 1000 Genomes samples. B. In European 1000 Genomes samples.

increase in the frequency of the 370A allele in East Asia could have been driven by selection favoring more efficient evapo-transpiration, although the authors also mentioned the possibility that reduced mammary fat pad size could also have been adaptive, or, given the clear pleiotropic effects of the 370A mutation, that selection

acted on multiple traits. This fascinating example highlights the challenges encountered when trying to explain the ultimate selective factors responsible for some of the selective sweeps observed in human populations, and emphasizes the importance of association studies, functional studies, and studies in animal models to complement genome-wide scans of selection signatures.

Conclusions

Our study has identified a list of genes that could potentially explain the reduction of melanin levels that took place in East Asia after the out-of-Africa migration of anatomically modern humans. The application of recently developed methods, such as the Composite of Multiple Signals (CMS) test [42] or Approximate Bayesian Computation (ABC) tests [70], and a more extensive annotation of the polymorphisms present within and around these genes, may be useful to narrow down the genic regions that were the target of positive selection, and to distinguish between selection that has acted on newly arisen mutations or standing variation. However, it will be necessary to carry out association studies in samples for which quantitative data on pigmentary traits are available, and functional studies in melanocytes to confirm the implication of these genes in normal pigmentation variation.

Additional files

Additional file 1: Table S1. List of the top 0.1% windows for all the statistics used in this study.

Additional file 2: Table S2. Basic annotation of the genes overlapping the top 0.1% windows identified in this study. The annotations were obtained with the DAVID database. The table reports the gene, the gene name, chromosome location, statistical tests for which the gene is an outlier, other references that have reported signatures of selection in the relevant genes, disease class for which genetic associations have been reported, OMIM_disease, KEGG pathway and GOTERM Biological Pathway.

Additional file 3. Description of the pigmentation genes identified in the scan for signatures of selection in East Asia.

Additional file 4. Haplotype block structure of the genes DCT, EDAR and EFGR in the East Asian and European 1000 Genomes samples.

Abbreviations

HGDP: Human Genome Diversity Project; LD: Linkage Disequilibrium; SNP: Single Nucleotide Polymorphism; JPT: Japanese; CHB: Han Chinese from Beijing; CHS: Southern Han Chinese; LSBL: Locus Specific Branch Length; iHS: Integrated Haplotype Score; SFS: Site Frequency Spectrum; DAVID: Database for Annotation, Visualization and Integrated Discovery; CMS: Composite of Multiple Signals; ABC: Approximate Bayesian Computation.

Competing interests

The authors have declared that no competing interests exist.

Authors' contributions

JLH participated in acquisition of data, contributed to the analysis and interpretation of data, and wrote the first draft of the manuscript, TS, RG, AR and JMA contributed to the analysis and interpretation of data, ME contributed to the acquisition of data, EJP was responsible for study conception and design, contributed to the analysis and interpretation of data and participated in the preparation of the final version of the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgements

EJP was supported by an Early Researcher Award from the Ministry of Research and Innovation, Government of Ontario and by NSERC (NSERC Discovery grant).

Author details

[1]Department of Anthropology, University of Toronto at Mississauga, Mississauga, Ontario, Canada. [2]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [3]Department of Mathematical and Computational Sciences, University of Toronto at Mississauga, Mississauga, Ontario, Canada.

Received: 25 April 2013 Accepted: 5 July 2013
Published: 12 July 2013

References

1. Henn BM, Cavalli-Sforza LL, Feldman MW: **The great human expansion.** *Proc Natl Acad Sci USA* 2012, **109:**17758–17764.
2. Jablonski NG, Chaplin G: **Human skin pigmentation as an adaptation to UV radiation.** *Proc Natl Acad Sci* 2010, **107**(Suppl 2):8962–8968.
3. Scherer D, Kumar R: **Genetics of pigmentation in skin cancer.** *Mutat Res* 2010, **705:**141–153.
4. Jablonski NG, Chaplin G: **The evolution of human skin coloration.** *J Hum Evol* 2000, **39:**57–106.
5. Parra EJ: **Human pigmentation variation: evolution, genetic basis, and implications for public health.** *AM J Phys Anthropol Supp* 2007, **45:**85–105.
6. Juzeniene A, Setlow R, Porojnicu, Steindal AH, Moan J: **Development of different human skin colors: a review of highlighting photobiological and photobiophysical aspects.** *Photochem Photobiol B Biol* 2009, **96:**93–100.
7. Elias PM, Menon G, Wetzel BK, Williams JW: **Barrier requirements as the evolutionary driver of epidermal pigmentation in humans.** *Am J Hum Biol* 2010, **22:**526–537.
8. Jablonski NG, Chaplin G: **Human skin pigmentation, migration and disease susceptibility.** *Phil Trans R Soc B* 2012, **367:**785–792.
9. Izagirre N, Garcia I, Junquera, de la Rua C, Alonso S: **A scan for signatures of positive selection in candidate loci for skin pigmentation in humans.** *Mol Biol Evol* 2006, **23:**1697–1706.
10. Nakayama K, Soemantri A, Jin F, Dashnyam B, Ohtsuka R, Duanchang P, Isa MN, Settheetham-Ishida W, Harihara S, Ishida T: **Identification of novel functional variants of the melanocortin 1 receptor gene originated from Asians.** *Hum Genet* 2006, **119:**322–330.
11. Tang K, Thornton KR, Stoneking M: **A New approach for using genome scans to detect recent positive selection in the human genome.** *PLOS Bio* 2007, **5:**1587–1602.
12. McEvoy B, Beleza S, Shriver MD: **The genetic architecture of normal variation in human pigmentation: an evolutionary perspective and model.** *Hum Mol Genet* 2006, **15**(suppl 2):176–181.
13. Myles S, Somel M, Tang K, Kelso J, Stoneking M: **Identifying genes underlying skin pigmentation differences among human populations.** *Hum Genet* 2007, **120:**613–621.
14. Lao O, De Gruijter JM, Van Duijn K, Navarro A, Kayser M: **Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms.** *Ann Hum Genet* 2007, **71:**354–369.
15. Sabeti PC, Varilly P, Fry B, Lohmueller J, Elizabeth Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, The International HapMap Consortium: **Genome-wide detection and characterization of positive selection in human populations.** *Nature* 2007, **449:**913–918.
16. Alonso S, Izagirre N, Smith-Zubiaga I, Gardeazabal J, Díaz-Ramón JL, Díaz-Pérez JL, Zelenika D, Boyano MD, Smit N, de la Rúa C: **Complex signatures of selection for the melanogenic loci *TYR, TYRP1* and *DCT* in humans.** *BMC Evol Biol* 2008, **8:**1471–2148.
17. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Meyers RM, Feldman MW, Pritchard JK: **Signals of recent positive selection in a worldwide sample of human populations.** *Genome Res* 2009, **19:**826–837.

18. Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, *et al*: **The role of geography in human adaptation.** *PLoS Genet* 2009, **5:**e1000500.
19. Hancock AM, Witonsky DB, Alkorta-Aranburu G, Beall CM, Gebremedhin A, Sukernik R, Utermann G, Pritchard JK, Coop G, Di Rienzo A: **Adaptations to climate-mediated selective pressures in humans.** *PLoS Genet* 2011, **7:**1–16.
20. Chen H, Patterson N, Reich D: **Population differentiation as a test for selective sweeps.** *Genome Res* 2010, **20:**393–402.
21. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R: **Localizing recent adaptive evolution in the human genome.** *PLoS Genet* 2007, **3:**e90.
22. Norton HL, Kittles RA, Parra E, McKeigue P, Mao X, Cheng K, Canfield VA, Bradley DG, McEvoy B, Shriver MD: **Genetic evidence for convergent evolution of light skin in European and East Asians.** *Mol Bio Evol* 2007, **24:**710–722.
23. Edwards M, Bigham A, Tan J, Li S, Gozdzik A, Ross K, Jin L, Parra EJ: **Association of the *OCA2* Polymorphism His615Arg with Melanin Content in East Asian Populations: Further Evidence of Convergent Evolution of Skin Pigmentation.** *PLoS Genet* 2010, **6:**e1000897.
24. Donnelly MP, Paschou P, Grigorenko E, Gurwitz D, Barta C, Lu RB, Zhukova OV, Kim JJ, Siniscalco M, New M, Li H, Kajuna SL, Manolopoulos VG, Speed WC, Pakstis AJ, Kidd JR, Kidd KK: **A global view of the OCA2-HERC2 region and pigmentation.** *Hum Gen* 2012, **13:**683–696.
25. Beleza S, Alonso SM, McEvoy, Alves I, Martinho C, Cameron E, Shriver MD, Parra EJ, Rocha J: **The timing of pigmentation lightening in europeans.** *Mol Evol Bio* 2013, **1:**24–35.
26. 1000 genomes project consortium: **An integrated map of genetic variation from 1092 human genomes.** *Nature* 2012, **491:**56–65.
27. Rees JL, Harding RM: **Understanding the evolution of human pigmentation: recent contributions from population genetics.** *J Invest Dermatol* 2012, **132:**846–853.
28. Sturm RA, Teasdale RD, Box NF: **Human pigmentation genes: identification, structure and consequences of polymorphic variation.** *Gene* 2001, **277:**49–62.
29. Abe Y, Tamiya G, Nakamura T, Hozumi Y, Suzuki T: **Association of melanogenesis genes with skin color variation among Japanese females.** *J Dermatol Sci* 2013, **69:**167–172.
30. Yamaguchi K, Watanabe C, Kawaguchi A, Sato T, Naka I, Shindo M, Moromizato K, Aoki K, Ishida H, Kimura R: **Association of melanocortin 1 receptor gene (MC1R) polymorphisms with skin reflectance and freckles in Japanese.** *J Hum Genet* 2012, **57:**700–708.
31. Ang KC, Ngu MS, Reid KP, Teh MS, Aida ZS, Koh DX, Berg A, Oppenheimer S, Salleh H, Clyde MM, Md-Zain BM, Canfield VA, Cheng KC: **Skin color variation in Orang Asli tribes of Peninsular Malaysia.** *PLoS One* 2012, **7:**e42752.
32. Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.** *Genetics* 1989, **123:**585–595.
33. Fu YX, Li WH: **Statistical tests of neutrality of mutations.** *Genetics* 1993, **133:**693–709.
34. Fu X: **Statistical tests of neutrality of mutations against population growth, hitch-hiking, and background selection.** *Genetics* 1997, **147:**915–925.
35. Fay JC, Wu CI: **Hitchhiking under positive Darwinian selection.** *Genetics* 2000, **155:**1405–1413.
36. Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, Huang, Akey JM, Jones KW: **The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs.** *Hum Genomics* 2004, **4:**274–286.
37. Weir BS, Cockerham CC: **Estimating F-statistics for the analysis of population structure.** *Evolution Int J Or Evol* 1984, **38:**1358–1370.
38. Voight BF, Kudaravalli S, Wen X, Pritchard JK: **A map of recent positive selection in the human genome.** *PLoS Biol* 2006, **4:**e72.
39. Gautier M, Vitalis R: **rehh: An R package to detect footprints of selection in genome-wide SNP data from haplotype structure.** *Bioinformatics* 2012, **28:**1176–1177.
40. Dennis G, Sherman BT, Hosack DA, Yang J, Baseler MW, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** *Genome Biol* 2003, **4:**P3. Epub.
41. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen S, Altshuler D, Lander ES: **Positive natural selection in the human lineage.** *Science* 2006, **312:**614–620.
42. Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH, Cabili M, Adegbola RA, Bamezai RNK, Hill AVS, Vannberg FO, Rinn JL, 1000 Genomes Project, Lander ES, Schaffner SF, Sabeti PC: **Identifying recent adaptations in large-scale genomic data.** *Cell* 2013, **4:**703–713.
43. The International HapMap Consortium: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449:**851–861.
44. Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM: **Genomic signatures of positive selection in humans and the limits of outlier approaches.** *Genome Res* 2006, **16:**980–989.
45. Kimura R, Fujimoto A, Tokunaga K, Ohashi J: **A practical genome scan for population-specific strong selective sweeps that have reached fixation.** *PLoS One* 2007, **2:**e286.
46. Duffy DL, Montgomery GM, Chen W, Zhao ZZ, Le L, James MR, Hayward NK, Martin NG, Sturm RA: **A Three–Single-Nucleotide Polymorphism Haplotype in Intron 1 of *OCA2* Explains Most Human Eye-Color Variation.** *Am J Hum Genet* 2007, **80:**241–252.
47. Zhong M, Lange K, Papp JC, Fan R: **A powerful score test to detect positive selection in genome-wide scans.** *Europ J Hum Genet* 2010, **18:**1148–1159.
48. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD: **Interrogating a High-Density SNP Map for Signatures of Natural Selection.** *Genome Res* 2002, **12:**1805–1814.
49. Quillen EE, Bauchet M, Bigham AW, Delgado-Burbano ME, Faust FX, Klimentidis YC, Mao X, Stoneking M, Shriver MD: **OPRM1 and EGFR contribute to skin pigmentation differences between Indigenous Americans and Europeans.** *Hum Genet* 2011, **13:**1073–1080.
50. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, Manolescu A, Karason A, Palsson A, Thorleifsson G, Jakobsdottir M, Steinberg S, Palsson S, Jonasson F, Sigurgeirsson B, Thorisdottir K, Ragnarsson R, Benediktsdottir KR, Aben KK, Kiemeney LA, Olafsson JH, Gulcher J, Kong A, Thorsteinsdottir U, Stefansson K: **Genetic determinants of hair, eye and skin pigmentation in Europeans.** *Nat Genet* 2007, **39:**1443–1452.
51. Kayser M, Liu F, Janssens AC, Rivadeneira F, Lao O, van Duijn K, Vermeulen M, Arp P, Jhamai MM, van Ijcken WF, Den Dunnen JT, Heath S, Zelenika D, Despriet DD, Klaver CC, Vingerling JR, De Jong PT, Hofman A, Aulchenko YS, Uitterlinden AG, Oostra BA, Van Duijn CM: **Three genome-wide association studies and a linkage analysis identify *HERC2* as a human iris color gene.** *Am J Hum Genet* 2008, **82:**411–423.
52. Sturm RA, Duffy DL, Zhao ZZ, Leite FPN, Stark MS, Hayward NK, Martin NG, Montgomery GW: **A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color.** *Am J Hum Genet* 2008, **82:**424–431.
53. Liu F, Wollstein A, Hysi PG, Ankra-Badu GA, Spector TD, Park D, Zhu G, Larsson M, Duffy DL, Montgomery GW, Mackey DA, Walsh S, Lao O, Hofman A, Rivadeneira F, Vingerling JR, Uitterlinden AG, Martin NG, Hammond CJ, Kayser M: **Digital quantification of human eye color highlights genetic association of three new loci.** *PLoS Genet* 2010, **6:**e1000934.
54. Cook AL, Chen W, Thurber AE, Smit DJ, Smith AG, Bladen TG, Brown DL, Duffy DL, Pastorino L, Bianchi-Scarra G, Leonard JH, Stow JL, Sturm RA: **Analysis of cultured human melanocytes based on polymorphisms within the SLC45A2/MATP, SLC24A5/NCKX5, and OCA2/P loci.** *J Invest Dermatol* 2009, **129:**392–405.
55. Visser M, Kayser M, Palstra RJ: **HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter.** *Genome Res* 2012, **22:**446–455.
56. Duffy D, Box N, Chen W, Palmer JS, Montgomery GW, James MR, Hayward NK, Martin NG, Sturm RA: **Interactive effects of MC1R and OCA2 on melanoma risk phenotypes.** *Hum Mol Genet* 2004, **13:**447–461.
57. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Jakobsdottir M, Steinberg S, Gudjonsson SA, Palsson A, Thorleifsson G, Pálsson S, Sigurgeirsson B, Thorisdottir K, Ragnarsson R, Benediktsdotttir KR, Aben KK, Vermeulen SH, Goldstein AM, Tucker MA, Kiemeney LA, Olafsson JH, Gulcher J, Kong A, Thorsteinsdottir U, Stefansson K: **Two newly identified genetic determinants of pigmentation in Europeans.** *Nat Genet* 2008, **40:**835–837.
58. Kenny EE, Timpson NJ, Sikora M, Yee MC, Moreno-Estrada A, Eng C, Huntsman S, Burchard EG, Stoneking M, Bustamante CD, Myles S: **Melanesian blond hair is caused by an amino acid change in TYRP1.** *Science* 2012, **336:**554.
59. Frudakis T, Thomas M, Gaskin Z, Venkateswarlu K, Suresh Chandra K, Ginjupalli S, Gunturi S, Natrajan S, Ponnuswamy VK, Ponnuswamy KN: **Sequences associated with human iris pigmentation.** *Genetics* 2003, **165:**2071–2083.
60. Jacobs LC, Wollstein A, Lao O, Hofman A, Klaver CC, Uitterlinden AG, Nijsten T, Kayser M, Liu F: **Comprehensive candidate gene study highlights UGT1A and BNC2 as new genes determining continuous skin color variation in Europeans.** *Hum Genet* 2013, **132:**147–158.
61. Eriksson N, Macpherson JM, Tung JY, Hon LS, Naughton B, Saxonov S, Avey L, Wojcicki A, Pe'er I, Mountain J: **Web-Based Participant-Driven Studies Yield Novel Genetic Associations for Common Traits.** *PLoS Genet* 2010, **6:**e1000993.

62. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps.** *Bioinformatics* 2005, **21:**263–265.
63. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D: **The structure of haplotype blocks in the human genome.** *Science* 2002, **296:**2225–2229.
64. Yuasa HJ, Takubo M, Takahashi A, Hasegawa T, Noma H, Suzuki T: **Evolution of vertebrate indoleamine 2,3-dioxygenases.** *J Mol Evol* 2007, **65:**705–714.
65. Anno S, Abe T, Yamamoto T: **Interactions between SNP alleles at multiple loci contribute to skin color differences between Caucasoid and Mongoloid subjects.** *Int J Biol Sci* 2008, **4:**81–86.
66. Cluzeau C, Hadj-Rabia S, Jambou M, Mansour S, Guigue P, Masmoudi S, Bal E, Chassaing N, Vincent MC, Viot G, Clauss F, Maniere MC, Toupenay S, Le Merrer M, Lyonnet S, Cormier-Daire V, Amiel J, Faivre L, de Prost Y, Munnich A, Bonnefont JP, Bodemer C, Smahi A: **Only four genes (EDA1, EDAR, EDARADD, and WNT10A) account for 90% of hypohidrotic/anhidrotic ectodermal dysplasia cases.** *Hum Mutat* 2011, **32:**70–72.
67. Kimura R, Yamaguchi T, Takeda M, Kondo O, Toma T, Hanejk K, Hanihara T, Matsukusa H, Kawamura S, Maki K, Osawa M, Ishida H, Oota H: **A common variation in EDAR is a genetic determinant of shovel-shaped incisors.** *Am J Hum Genet* 2009, **85:**528–535.
68. Fujimoto A, Kimura R, Ohashi J, Omi K, Yuliwulandari R, Batubara L, Mustofa MS, Samakkarn U, Settheetham-Ishida W, Ishida T: **A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness.** *Hum Mol Genet* 2008, **17:**835–843.
69. Kamberov YG, Wang S, Tan J, Gerbault P, Wark A, Tan L, Yang Y, Li S, Tang K, Chen H, Powell A, Itan Y, Fuller D, Lohmueller J, Mao J, Schachar A, Paymer M, Hostetter E, Byrne E, Burnett M, McMahon AP, Thomas MG, Lieberman DE, Jin L, Tabin CJ, Morgan BA, Sabeti PC: **Modeling recent human evolution in mice by expression of a selected EDAR variant.** *Cell* 2013, **152:**691–702.
70. Peter BM, Huerta-Sanchez E, Nielsen R: **Distinguishing between selective sweeps from standing variation and from a de novo mutation.** *PLoS Genet* 2012, **8:**e1003011.

doi:10.1186/1471-2148-13-150
Cite this article as: Hider *et al.*: Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry. *BMC Evolutionary Biology* 2013 **13**:150.

Appendix B

SUPPLEMENT FOR CHAPTER 2: COMPREHENSIVE IDENTIFICATION AND ANALYSIS OF HUMAN ACCELERATED REGULATORY DNA

This appendix contains material published in Gittelman et al[82]

B.1 Supplementary Text

B.1.1 Segmenting merged DHS

To segment long, merged DHS that may represent multiple regulatory elements, for every 5th base pair in the genome I counted the number of cell types that contained a DHS at that base. Based on this information, I identified local minima in the number of cell types, and partitioned a DHS at the local minima if the minima was less than half of the previous maximum, and if the number of cell types again increased by at least 5 (Fig. B.7). This resulted in an increase from the initial 2,026,055 merged DHS to the final 2,093,197 DHS, so the vast majority of DHS in the analysis were not partitioned. 8,929 of the final 113,577 conserved DHS were part of a larger DHS that got partitioned, and 68 of the 524 haDHS were partitioned.

B.1.2 Calibrating the False discovery rate for conserved and human accelerated DHS

To evaluate whether the FDR was appropriately calibrated in the phyloP tests, I randomly sampled regions throughout the genome from the same size distribution as the final set of DHS. I applied the same filtering strategy to this set, arriving at a final set of 47,317 simulated DHS and ran the same phyloP tests. Of the randomly sampled regions, 1.1% were called as conserved at the same p value threshold (0.00242) for the real DHS,

demonstrating that my estimate of a 1% FDR for the conserved DHS is accurate. Similarly, I found that simulated DHS were called as conserved and human accelerated at 5.7% the rate that DHS from the real data set were. These data indicated that my estimated FDRs were well calibrated.

B.1.3 Testing for acceleration relative to the neutral rate

I applied a non-parametric permutation strategy to determine if the human branch in each haDHS was evolving more rapidly than the estimated neutral rate. For each haDHS I first used phyloFit to fit a phylogenetic model to the alignment of the haDHS. Next, I sampled contiguous regions of the alignment of the same size as the haDHS from the local neutral region (the 50kb block of sequence surrounding the haDHS) 10^3 times, each time fitting a phylogenetic model to the alignment. haDHS with a substitution rate greater than 95% of the null distribution were determined to be evolving more rapidly than the neutral rate.

B.1.4 GC-biased gene conversion

I determined whether each mutation in my consDHS and haDHS was from weak to strong or other (strong to weak, weak to weak, or strong to strong) based on the human ancestral sequence inferred as part of the 6 primate EPO alignment. I computed an empirical null distribution to non-parametrically assess whether there was an enrichment of weak to strong mutations in haDHS. To this end, I randomly sampled 524 conserved (but not human accelerated) DHS 10^4 times and determined how often a sample contained more than the proportion of weak to strong mutations seen in the haDHS. This analysis yielded significant evidence of GC-BGC, with a higher proportion of weak to

strong mutations in haDHS compared to the conserved non-accelerated DHS samples (0.45 and 0.38, respectively, permutation $P = 0.006$; Fig. B.3a). Capra et al[173] recently wrote a software package, phastBias, that identifies discrete regions of the genome undergoing GC-biased gene conversion. I found that 51 haDHS overlapped these regions. To reevaluate the effect of GC-BGC after removing haDHS that overlap phastBias regions, I repeated the sampling procedure by randomly selecting 473 conserved DHS each replicate, and determined how often samples had a higher proportion of weak to strong substitutions that in the non GC-biased haDHS. I found that removing the 51 haDHS (9.7%) that overlap regions subject to GC-BGC eliminates the observed weak to strong bias (permutation $P = 0.32$; Fig. B.3a), indicating a small but significant effect of GC-BGC in my data.

B.1.5 Human-macaque divergence

I wanted to test the hypothesis that increased mutation rates could explain the rate acceleration in haDHS. To accomplish this, I calculated human-macaque divergence in the surrounding 4kb of sequence for each haDHS in order to approximate the neutral mutation rate. To do this, I obtained an alignment and used the same filtering strategy I used for the neutral regions in the LRT tests. I calculated divergence using the phyloFit program in the PHAST package[122]. I found that only the 51 haDHS that overlap GC-biased elements show a significant increase in human-macaque divergence relative to conserved non-accelerated DHS, indicating that mutation rate heterogeneity is not a major factor in non GC-BGC haDHS (Fig. B.3b). In contrast, the entire set of HAEs showed increased divergence relative to other highly conserved elements (Fig. B.3b).

B.1.6 Population genetics analyses

To perform population genetic analyses on my DHS, I downloaded the phase1 integrated release data from the 1000 genomes project[101]. I used the ancestral and derived calls in these data to create an ancestralized human reference genome in which any polymorphic base that is derived in the reference was replaced with the ancestral base. This ancestralized reference therefore still had any fixed substitutions on the human lineage, but allowed to us to identify 1000 genomes variants that occurred in ancestral CpG sites. I filtered any variants that occurred within CpG sites in the ancestralized reference as well as any variants called only in the high coverage exome data, so as not to bias rare variant detection in regions with different coverage, which has been shown to be an issue in previous applications of the 1000 genomes data[174].

Because α can be negative due to mildly deleterious segregating sites[175], when calculating α I filtered any 1000 genome variants below a 10% derived allele frequency in all combined African populations. I filtered repeat masked sequence from all sites, and additionally filtered exons, DHS, and phastCons elements from the neutral class of sites.

To determine how my estimates of α might be affected by demographic history, I simulated 10^4 haDHS and surrounding neutral regions (324 bp; the average length of all haDHS; 4000bp neutral regions) using the program ms[176] and a demographic model used previously[177]. To this I added an outgroup to model divergence from chimpanzees with the switch -ej 8.207934. I repeated these simulations with different mutation rate heterogeneity parameters, which describe how quickly the simulated haDHS evolved compared to the neutral region. Note, I did not simulate selection and thus the mutation

rate parameter should affect polymorphism and divergence equally. The following is the ms command line used to simulate an haDHS evolving at the neutral mutation rate (1e-08):

ms 493 1000 -I 3 492 0 1 -t 0.0947376 -r 0.0944452 324 -n 1 58.002735978 -n 2 70.041039672 -eg 0 1 482.46 -eg 0 2 570.18 -em 0 1 2 0.7310 -em 0 2 1 0.7310 -eg 0.006997264 1 0 -eg 0.006997264 2 89.7668 -en 0.006997264 1 1.98002736 -en 0.031463748 2 0.141176471 -en 0.03146375 2 0.254582763 -em 0.03146375 1 2 4.386 -em 0.03146375 2 1 4.386 -em 0.069767442 1 2 0 -em 0.069767442 2 1 0 -ej 0.069767442 2 1 -en 0.069767442 1 1.98002736 -en 0.20246238 1 1 -ej 8.207934 3 1

I calculated nucleotide diversity as $\pi = \frac{n}{n-1}\left(\sum_{i=1}^{S} 2p_i\,(1-p_i)\right)$, where n is the number of chromosomes and p_i is the frequency of the major allele for th *i*th segregating site, *S,* and divided by the number of bases considered in each DHS. For comparison I calculated π in four fold degenerate sites that I defined using NCBI-called reading frames with single base phyloP scores less than zero. I also calculated π in the set of previously described HAEs and phastCons elements. Coverage at each variant was obtained from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/supporting/ ALL.wgs.project_consensus_vqsr2b.20101123.snps.low_coverage.sites.vcf.gz. I used PhyloP scores that I previously calculated without the human lineage[178]. I normalized all estimates of π by the human-macaque divergence estimated previously (see section "human-macaque divergence"), since I have shown that divergence varies substantially across loci, and will thus significantly effect estimates of π. I calculated nucleotide

diversity separately for each population, as well as in aggregate with all populations combined.

B.1.7 Luciferase assays

I obtained Human DNA from Novagen and chimpanzee DNA from Coriell (S003487), and amplified both human and chimpanzee alleles for each haDHS using the following PCR reaction: 200uM dNTPs, 10uM forward primer, 10uM reverse primer, 3% DMSO, 20ng DNA, 1u/50uL phusion polymerase (NEB) and 1x phusion buffer with an annealing temperature of 63 degrees Celsius. Primers were designed using BatchPrimer3 v1.0[179] and then run through In-Silico PCR downloaded from the UCSC Genome Browser to ensure they would produce one unique amplicon. PCR fragments were cloned into a pGL3-promoter (Promega) backbone with minimal promoter using the In-Fusion HD Cloning Kit from Clontech and transformed into Stellar competent cells provided with the kit. I obtained this pGL3-promoter backbone from Bing Wren's lab, which was modified to include an EcoRV restriction site downstream of the luciferase gene for cloning. Jennifer Madeoy assisted in the cloning.

Enna Hun conducted the rest of steps in the luciferase assay. Briefly, SKNMC and IMR90 cells were grown in EMEM with 10% FBS and Penicillin-Streptomycin. Cells were grown in T175 flasks with media changes every two days until 80-85% confluency. Cells were counted and replated onto 96-well plates at 30,000 cells per well (SKNMC) or 6500 cells per well (IMR90) and allowed to settle for 24 hours. Luciferase assays were performed following the Dual-Luciferase Reporter Assay System protocol (Promega). Cells were transfected with 500ng firefly luciferase plasmid with reporter

construct, and 10ng plasmid with *Renilla* luciferase. After 24 hours cells were lysed and luciferase activity was measured using a Perkin Elmer VICTOR 3 V 1420 Multi-label counter with injector. Raw firefly luciferase values were normalized by dividing by the *Renilla* luciferase values for each well.

A close inspection of the IMR90 data revealed a relationship between the raw *Renilla* luciferase value and the firefly/ *Renilla* ratio (Fig. B.8). As *Renilla* expression approaches zero, and the reading approaches that of an empty well without any transfected plasmid, the ratio of firefly to *Renilla* will approach one, which will lead to false positives. I therefore filtered any replicates in the IMR90 data in which the *Renilla* reading was less than 2500 relative light units. Any haDHS with less than two replicates, or any plates in which the negative control had less than two remaining replicates were filtered.

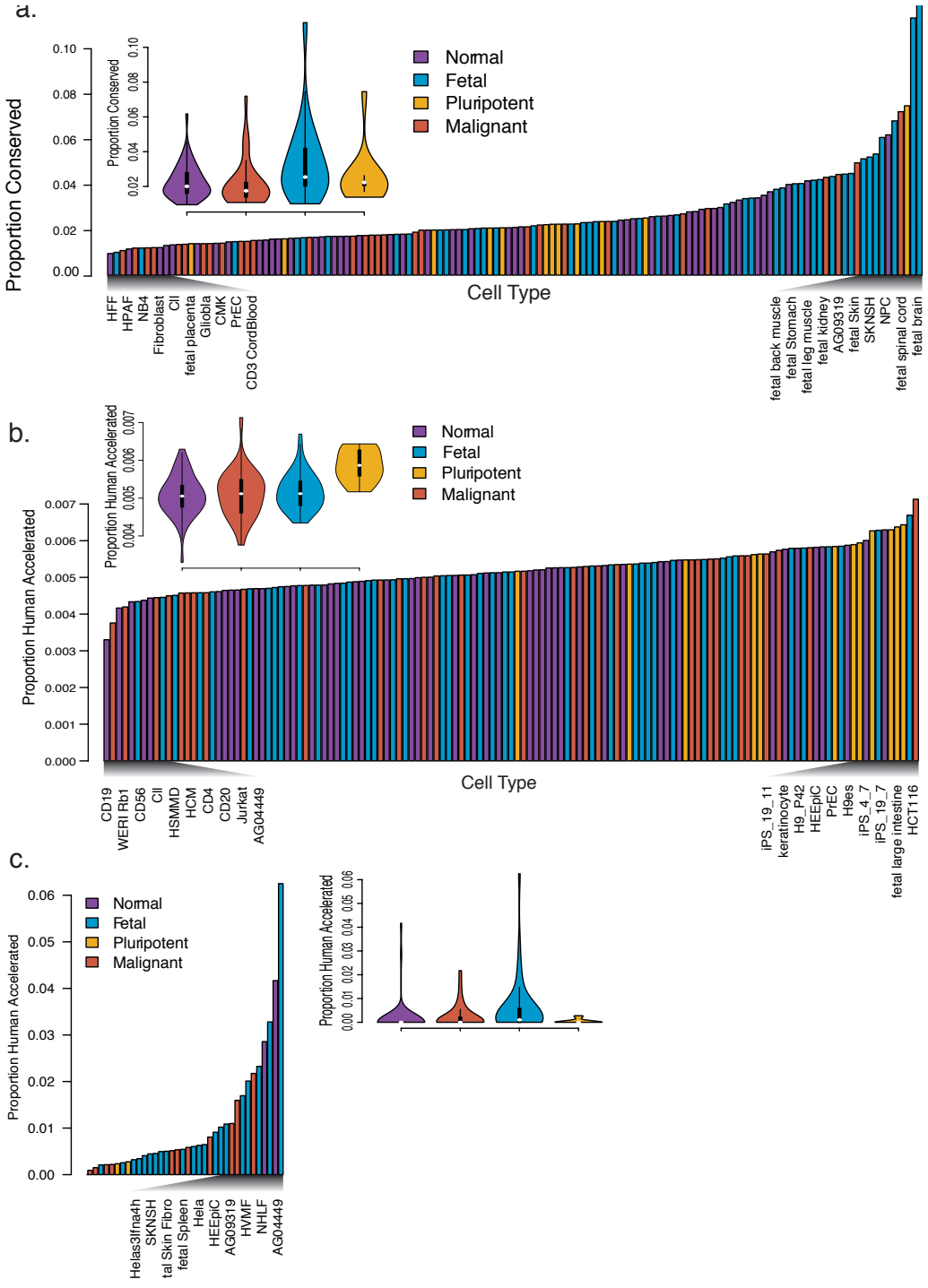

Fig. B.1. Distribution of conservation and acceleration across DHS. (a) The proportion of cell-type specific DHS (DHS that were active in only one cell type) that were called as conserved is shown for all 130 cell types. Cell types are colored according

to category. Inset violin plot summarizes bar plot **(b)** The proportion of conserve DHS that were called as human-accelerated is shown for all 130 cell types. Cell types are colored according to category. Inset shows data for each category combined in a violin plot. **(c)** The proportion of cell-type specific DHS that were called as human-accelerated is shown for all 130 cell types. Cell types are colored according to category. Only cell types with a proportion greater than zero are shown. Inset violin plot summarizes bar plot.

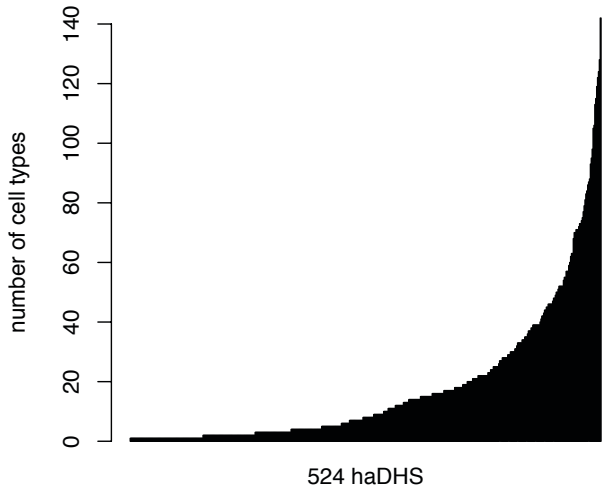

Fig. B.2. Number of cell types each haDHS is active in. All 524 haDHS are shown along the x axis, ranked from completely cell type specific to active in all or nearly all cell types. The range on the Y-axis exceeds the number of cell types analyzed (130), because in some cases multiple DHS from the same cell type were merged.

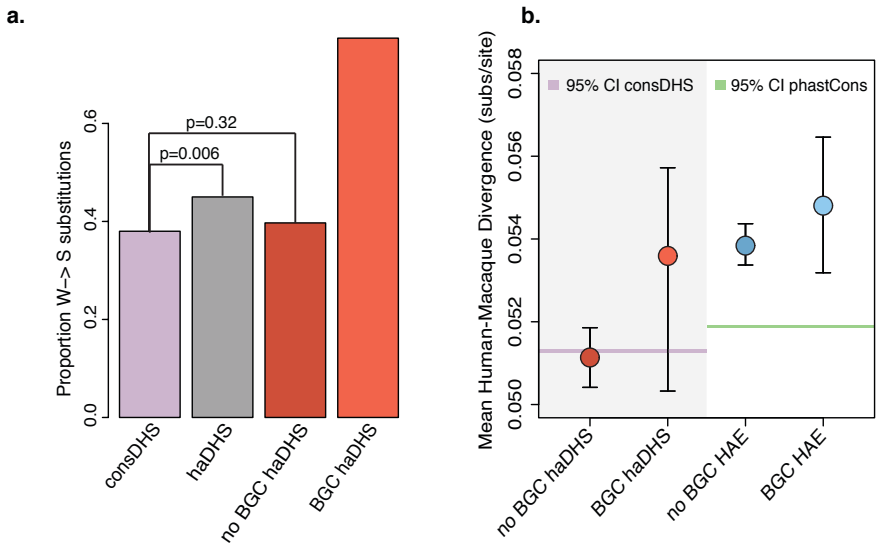

Fig. B.3. Additional forces contributing to rate acceleration in haDHS and HAEs. (a) The proportion of weak to strong substitutions in conserved DHS (consDHS), all haDHS, the set of haDHS that do not overlap regions predicted to experience GC-BGC (no BGC haDHS), and the set of haDHS that do overlap predicted regions of GC-BGC (GCB haDHS). **(b)** Mean and 95% bootstrap confidence intervals of human-macaque divergence in regions surrounding haDHS and HAEs that do or do not overlap predicted GC-BGC elements. Horizontal bars denote the 95% confidence interval of the mean human-macaque divergence for conserved DHS and phastCons elements.

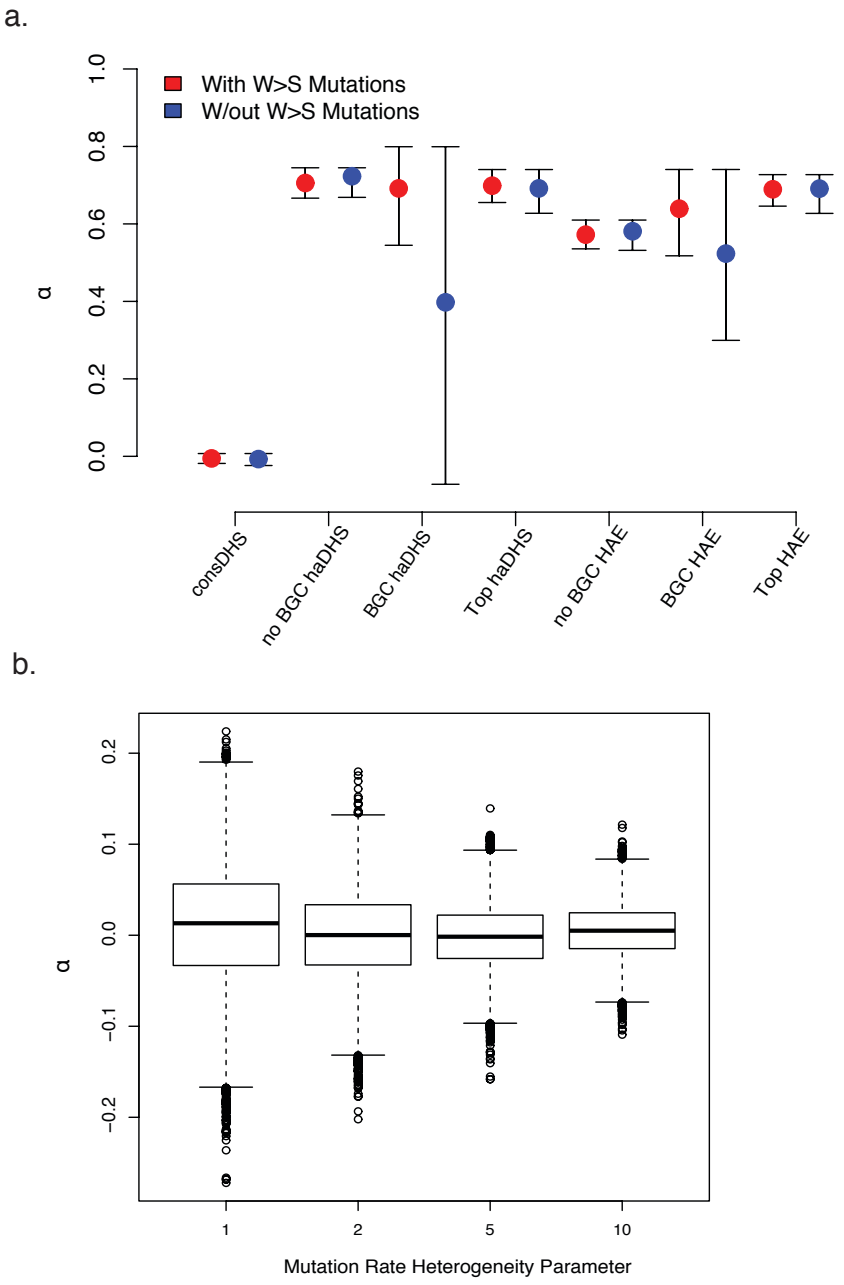

Fig. B.4. Estimating the proportion of substitutions subject to positive selection. (a) The bootstrapped 95% confidence interval for α is shown for different groups of elements. "Top" refers to elements that are evolving more rapidly than neutral, and do not overlap BGC regions. **(b)** Values of α for 10^4 simulations are shown for different mutation rate heterogeneity parameters. The mutation rate parameter denotes how much more rapidly the haDHS is evolving compared to the neutral region.

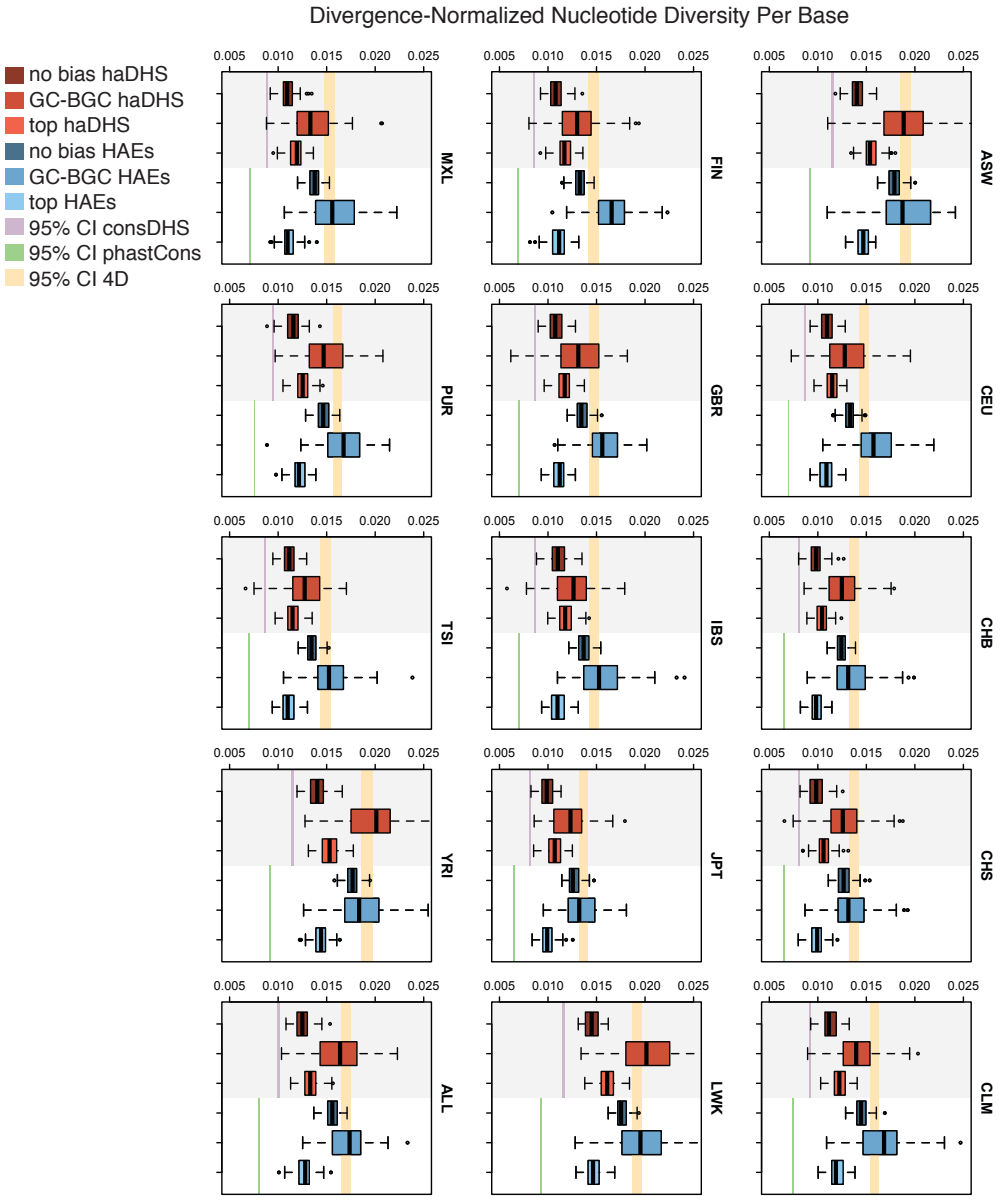

Fig. B.5. Nucleotide diversity across 1000 genomes populations. Boxplots of the bootstrap estimates of the mean π is shown for each of the 1000 genomes populations, for different groups of elements, indicated in the figure legend. "Top" refers to elements that are evolving more rapidly than neutral, and do not overlap BGC regions.

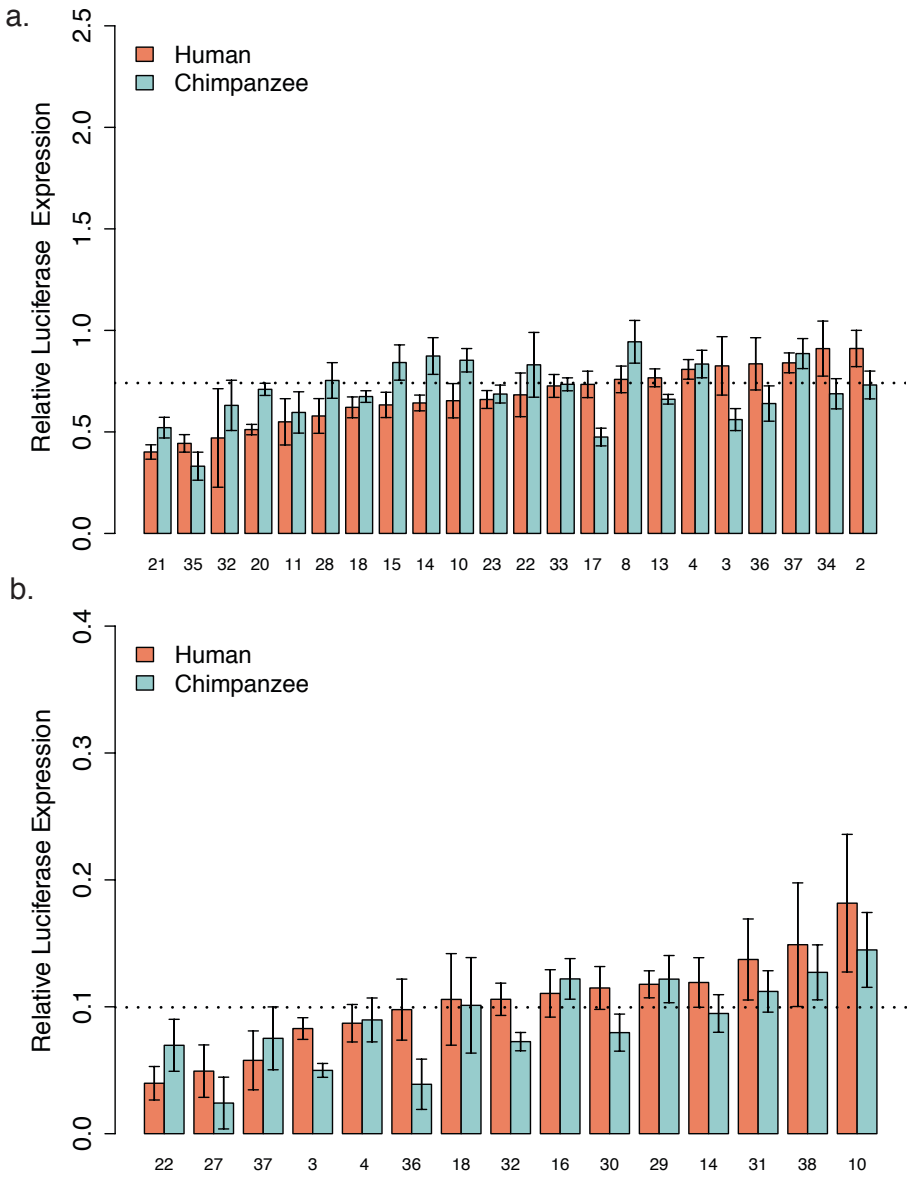

Fig. B.6. Luciferase reporters that did not show enhancer activity. (a) Luciferase data for all SKNMC tests in which neither allele acted as an enhancer is shown. The dashed line indicates the mean of the negative control replicates. (b) Luciferase data for all IMR90 tests that passed the 2500 Renilla threshold but did not act as enhancers is shown. The dashed line indicates the mean of the negative control replicates.

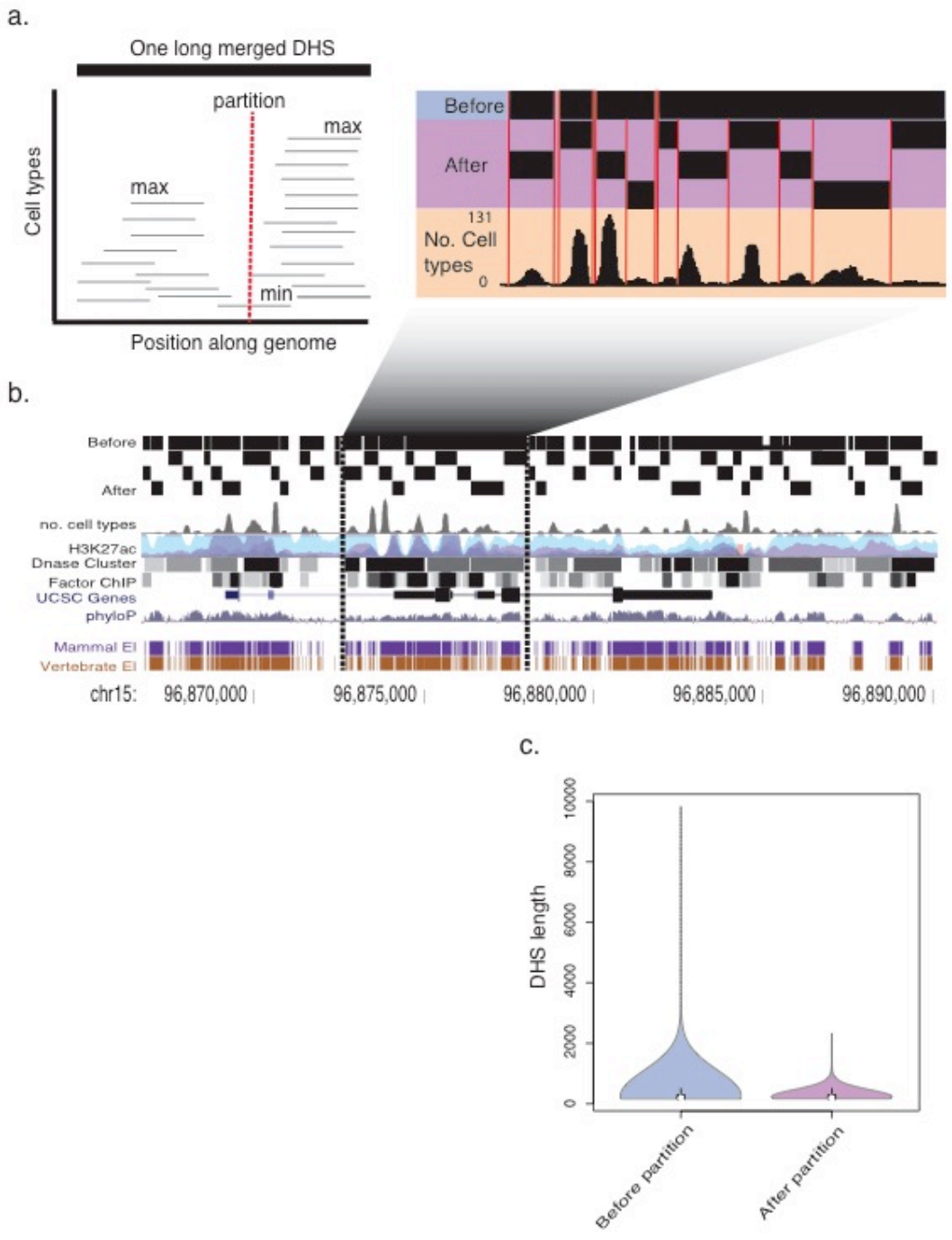

Fig. B.7. Segmenting long merged DHS. (a) Schematic of segmenting. This DHS would get partitioned into two distinct DHS at the minimum, because the minimum number of cell types is less than half the previous maximum, and the rest of the region climbs by at least 5 DHS. **(b)** Segmenting of the NR2F2 region is shown as an example. **(c)** The size distribution of DHS before applying the partitioning algorithm and after is shown.

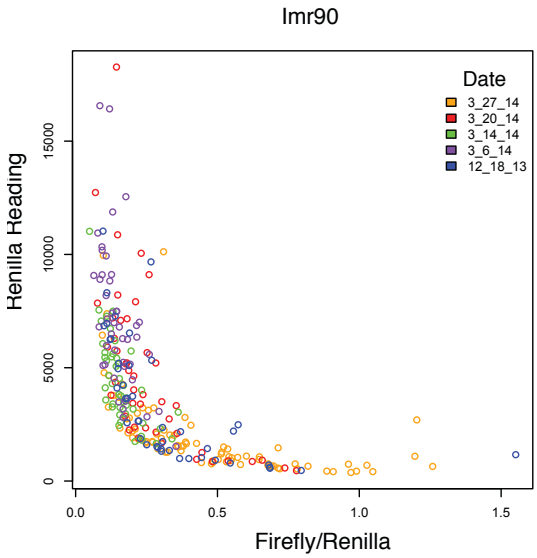

Fig. B.8. IMR90 raw data. The ratio of firefly to renilla luciferase expression is plotted against raw renilla values. Colors indicate the date that each data point was collected.

Table B.1. Cell types used. This table contains information about all of the cell types used in the DHS data.

source	cell type	category	index	used in 2nd set
roadmap epigenomics	CD19	normal	0	
roadmap epigenomics	CD20	normal	1	
roadmap epigenomics	CD3_CordBlood	normal	2	
roadmap epigenomics	CD4	normal	3	
roadmap epigenomics	CD56	normal	4	
roadmap epigenomics	CD8	normal	5	
roadmap epigenomics	fAdrenal	fetal	6	
roadmap epigenomics	fIntestine_Lg	fetal	7	
roadmap epigenomics	fIntestine_Sm	fetal	8	
roadmap epigenomics	fMuscle_arm	fetal	9	
roadmap epigenomics	fMuscle_back	fetal	10	
roadmap epigenomics	fMuscle_leg	fetal	11	
roadmap epigenomics	fPlacenta	fetal	12	
roadmap epigenomics	fSkin	fetal	13	
roadmap epigenomics	fSkin_fibro_abdomen	fetal	14	
roadmap epigenomics	fSkin_fibro_back	fetal	15	
roadmap epigenomics	fSpinal_cord	fetal	16	
roadmap epigenomics	fSpleen	fetal	17	
roadmap epigenomics	fStomach	fetal	18	
roadmap epigenomics	fTestes	fetal	19	
roadmap epigenomics	fThymus	fetal	20	
roadmap epigenomics	H1_P18	pluripotent	21	
roadmap epigenomics	H9_P42	pluripotent	22	
roadmap epigenomics	iPS_19_11	pluripotent	23	
roadmap epigenomics	iPS_19_7	pluripotent	24	
roadmap epigenomics	iPS_4_7	pluripotent	25	
roadmap epigenomics	iPS_6_9	pluripotent	26	
roadmap epigenomics	NPC	pluripotent	27	
roadmap epigenomics	FETAL_HEART	fetal	28	yes
roadmap epigenomics	BREAST_vHMEC	malignant	29	yes
roadmap epigenomics	FETAL_KIDNEY	fetal	30	yes
roadmap epigenomics	CD34_PRIMARY	normal	31	yes
roadmap epigenomics	H1_BMP4_MESENDODERM	fetal	32	yes
roadmap epigenomics	CD3_PRIMARY	normal	33	yes
roadmap epigenomics	FIBROBLAST	normal	34	yes
roadmap epigenomics	FETAL_LUNG	fetal	35	yes
roadmap epigenomics	IMR90	normal	36	yes
roadmap epigenomics	KERATINOCYTE	normal	37	yes
roadmap epigenomics	FETAL_BRAIN	fetal	38	yes
ENCODE	HRCE	normal	39	
ENCODE	HMVEC_dNeo	normal	40	
ENCODE	NHA	normal	41	yes

ENCODE	HPAEC	normal	42	
ENCODE	HRGEC	fetal	43	
ENCODE	HMEC	normal	44	yes
ENCODE	AoAF	normal	45	
ENCODE	HAc	fetal	46	
ENCODE	AG04449	fetal	47	
ENCODE	HNPCEpiC	fetal	48	
ENCODE	CD20.2	normal	49	
ENCODE	HUVEC	normal	50	yes
ENCODE	hTH2	normal	51	
ENCODE	AG09309	normal	52	
ENCODE	SkMC	fetal	53	
ENCODE	HIPEpiC	fetal	54	
ENCODE	HMVEC_dLyAd	normal	55	
ENCODE	PrEC	normal	56	
ENCODE	HAh	fetal	57	
ENCODE	BJ	normal	58	
ENCODE	HMVEC_dBlNeo	normal	59	
ENCODE	HMVEC_dBlAd	normal	60	
ENCODE	WI_38	fetal	61	
ENCODE	HFF	normal	62	
ENCODE	RPTEC	normal	63	
ENCODE	HPF	fetal	64	
ENCODE	HPAF	fetal	65	
ENCODE	HCF	fetal	66	
ENCODE	HCFaa	normal	67	
ENCODE	HCM	fetal	68	
ENCODE	HMVEC_LLy	normal	69	
ENCODE	HSMM	normal	70	yes
ENCODE	AG09319	normal	71	
ENCODE	HGF	normal	72	
ENCODE	NHDF_Neo	normal	73	
ENCODE	HMVEC_dLyNeo	normal	74	
ENCODE	HMF	normal	75	
ENCODE	HSMM_D	normal	76	
ENCODE	hTH1	normal	77	
ENCODE	HAEpiC	fetal	78	
ENCODE	HAsp	fetal	79	
ENCODE	AG04450	fetal	80	
ENCODE	HEEpiC	fetal	81	
ENCODE	SAEC	normal	82	
ENCODE	NHLF	normal	83	yes
ENCODE	HVMF	fetal	84	
ENCODE	HCPEpiC	fetal	85	
ENCODE	NHEK	normal	86	yes
ENCODE	HBMEC	fetal	87	
ENCODE	WI_38_TAM	fetal	88	
ENCODE	NHDF_Ad	normal	89	
ENCODE	HMVEC_dAd	normal	90	
ENCODE	HConF	fetal	91	

ENCODE	HPdLF	normal	92	
ENCODE	HMVEC_LBl	normal	93	
ENCODE	AG10803	normal	94	
ENCODE	HRE	normal	95	
ENCODE	HRPEpiC	fetal	96	
ENCODE	SKNSH	malignant	97	
ENCODE	Helas3Ifna4h	malignant	98	
ENCODE	PANC1	malignant	99	
ENCODE	CACO2	malignant	100	
ENCODE	Ishikawa_E	malignant	101	
ENCODE	HL60	malignant	102	
ENCODE	NT2_D1	malignant	103	
ENCODE	Medullo	malignant	104	
ENCODE	T47d	malignant	105	
ENCODE	LNCap	malignant	106	
ENCODE	HCT116	malignant	107	
ENCODE	LncapAndro	malignant	108	
ENCODE	Mcf7Hypoxlac	malignant	109	
ENCODE	HepG2	malignant	110	yes
ENCODE	CMK	malignant	111	
ENCODE	MCF7	malignant	112	
ENCODE	K562	malignant	113	yes
ENCODE	Cll	malignant	114	
ENCODE	NB4	malignant	115	
ENCODE	SK_N_MC	malignant	116	
ENCODE	8988t	malignant	117	
ENCODE	WERI_Rb1	malignant	118	
ENCODE	Jurkat	malignant	119	
ENCODE	A549	malignant	120	
ENCODE	HFF_MyC	malignant	121	
ENCODE	Ishikawa_T	malignant	122	
ENCODE	Hela	malignant	123	
ENCODE	Gliobla	malignant	124	
ENCODE	BE_2_C	malignant	125	
ENCODE	H9es	pluripotent	126	
ENCODE	hESCT0	pluripotent	127	
ENCODE	HESC	pluripotent	128	
ENCODE	H1hesc	pluripotent	129	yes
ENCODE	Ips	pluripotent	130	

Table B.2: haDHS. This table contains information about all of the haDHS identified in this study.

chr	start	stop	lnL human acc	lnL primate cons	p value human acc	p value primate cons	phastBias?
chr8	1649165	1649870	81.77141	32.62441	0	0	phastBias
chr7	3682200	3682370	27.27594	16.29147	0	0	
chr3	11641980	11642395	25.68122	18.37153	0	0	
chr9	3968920	3969210	19.68761	11.53517	0	0	phastBias
chr10	127190180	127190350	18.84089	5.16675	0	0.00065	
chr7	9304080	9304315	18.37978	8.5726	0	2.00E-05	
chr2	236773805	236774115	15.40915	32.83619	0	0	phastBias
chr3	2783280	2783850	15.08353	5.34825	0	0.00054	phastBias
chr14	37608120	37608470	14.85935	23.74542	0	0	
chr18	908420	909170	14.6353	4.50669	0	0.00134	phastBias
chr3	170819800	170820100	14.44437	11.99648	0	0	phastBias
chr1	103345220	103345450	13.62502	30.79517	0	0	
chr10	234780	235210	13.62397	12.56184	0	0	phastBias
chr10	80423600	80423770	12.96168	17.53088	0	0	
chr6	2396360	2396830	12.91487	21.50385	0	0	phastBias
chr9	98811345	98812075	12.62179	32.52967	0	0	
chr2	147187000	147187170	12.60065	18.06198	0	0	
chr5	91951840	91952110	12.56736	13.21338	0	0	
chrX	137792645	137793565	12.50837	35.30704	0	0	phastBias
chr17	33395320	33395870	12.10691	12.49613	0	0	
chr9	101866180	101866525	12.00747	19.56163	0	0	
chr1	116854360	116854940	11.98675	10.74158	0	0	
chr2	237776600	237776890	11.86629	14.58629	0	0	
chr3	141883300	141883515	11.42323	4.2644	0	0.00175	
chr6	1287320	1287870	11.24727	13.6731	0	0	phastBias
chr12	108876280	108876740	11.23679	4.45011	0	0.00143	
chr12	97931360	97931530	11.23103	23.98123	0	0	
chr2	119067240	119067880	11.21731	43.85904	0	0	phastBias
chr17	5264465	5264710	11.12699	13.57776	0	0	phastBias
chr2	113992940	113993455	11.0507	28.30117	0	0	phastBias
chr1	88393120	88393370	11.05032	21.82661	0	0	
chr18	3499005	3499550	10.99665	33.45076	0	0	phastBias
chr4	124317660	124318050	10.88304	9.27693	0	1.00E-05	
chr2	27716160	27716650	10.88208	11.48427	0	0	
chr10	33735040	33735190	10.86224	16.7929	0	0	
chr12	130765765	130766030	10.76447	8.24932	0	2.00E-05	phastBias
chr15	37242100	37242250	10.74565	8.16377	0	3.00E-05	
chr21	15825740	15825930	10.55654	8.14757	0	3.00E-05	phastBias
chr11	124616780	124617190	10.51549	11.64839	0	0	phastBias
chr10	11077660	11077890	10.42587	19.48001	0	0	
chr3	2141600	2142075	10.37101	4.42949	0	0.00146	phastBias
chr8	56805800	56806060	10.29046	8.0105	0	3.00E-05	
chr18	42879000	42879290	10.23646	13.43162	0	0	

chr1	87835465	87835710	10.23134	21.57876	0	0	
chr12	114845100	114845780	10.18108	6.16872	0	0.00022	
chr8	27752600	27753030	10.1046	9.40597	0	1.00E-05	
chr8	72094400	72094550	10.09718	10.83189	0	0	
chr9	82225960	82226320	10.09214	16.11731	0	0	
chr10	126812845	126812995	10.0819	6.54403	0	0.00015	
chr20	39513665	39513815	10.06611	6.08772	0	0.00024	
chr7	4057880	4058030	9.89678	10.85619	0	0	
chr14	33811460	33811810	9.84998	4.82887	0	0.00094	phastBias
chr18	33570460	33570610	9.80885	9.84507	0	0	
chr3	56037220	56037530	9.78014	25.5321	0	0	
chr4	141856440	141856690	9.77534	14.73433	0	0	
chr3	159359220	159359370	9.62411	8.40927	1.00E-05	2.00E-05	
chr9	71819680	71820175	9.62104	11.44756	1.00E-05	0	
chr3	81267645	81267795	9.60267	7.92939	1.00E-05	3.00E-05	
chr4	11691180	11691330	9.58861	13.94589	1.00E-05	0	
chr14	85124220	85124670	9.51808	39.25349	1.00E-05	0	
chr9	469645	470020	9.47839	6.50365	1.00E-05	0.00016	phastBias
chr17	74908980	74909130	9.44122	10.63759	1.00E-05	0	
chr1	194307560	194308050	9.43164	6.82423	1.00E-05	0.00011	
chr10	4320380	4320750	9.40734	5.46565	1.00E-05	0.00047	
chr3	52354240	52354470	9.27965	4.11415	1.00E-05	0.00206	
chr4	81707620	81707890	9.13279	11.37266	1.00E-05	0	
chr15	81822500	81822970	9.1023	4.94647	1.00E-05	0.00083	
chr10	128956040	128956455	8.96495	22.56732	1.00E-05	0	
chr5	158635020	158635355	8.96068	12.91105	1.00E-05	0	
chr3	89536540	89536690	8.93921	5.15004	1.00E-05	0.00067	
chr9	16729200	16729570	8.89653	33.6907	1.00E-05	0	
chr1	71184680	71185270	8.87476	5.84678	1.00E-05	0.00031	
chr4	4794040	4794190	8.87132	8.50001	1.00E-05	2.00E-05	
chr15	77849020	77849790	8.82479	18.84373	1.00E-05	0	
chr4	33264460	33264830	8.80736	5.40637	1.00E-05	5.00E-04	
chr12	109692000	109692270	8.80467	9.67135	1.00E-05	1.00E-05	
chr3	86986760	86986990	8.78181	23.44433	1.00E-05	0	
chr1	97463380	97463550	8.77642	12.21853	1.00E-05	0	
chr16	64803220	64803370	8.77145	5.87161	1.00E-05	0.00031	
chr2	135118260	135118550	8.76123	5.40104	1.00E-05	0.00051	
chr18	3214860	3215270	8.72089	15.48977	1.00E-05	0	phastBias
chr2	238224260	238224490	8.63958	4.67525	2.00E-05	0.00111	
chr1	91286480	91286630	8.633	8.94717	2.00E-05	1.00E-05	
chr2	145234920	145235190	8.62405	17.68046	2.00E-05	0	
chr8	32065060	32065395	8.58367	20.02507	2.00E-05	0	
chr16	85399600	85400330	8.55987	18.41768	2.00E-05	0	phastBias
chr15	91009460	91009795	8.53265	7.09292	2.00E-05	8.00E-05	
chr15	34030260	34030735	8.42475	9.01799	2.00E-05	1.00E-05	
chr11	107364840	107365090	8.34704	4.77665	2.00E-05	0.001	
chr8	76636560	76636835	8.34346	31.15276	2.00E-05	0	
chr3	55577180	55577630	8.31572	15.04355	2.00E-05	0	
chr3	101569620	101569950	8.31264	4.82423	2.00E-05	0.00095	
chr1	3285380	3285535	8.27572	6.51963	2.00E-05	0.00015	

chr2	151644940	151645290	8.24587	6.19818	2.00E-05	0.00022	
chr1	42166820	42167010	8.20137	4.07487	3.00E-05	0.00215	
chr1	244200840	244200990	8.17905	14.83063	3.00E-05	0	
chr2	144483820	144483970	8.16721	21.02716	3.00E-05	0	
chr10	123064660	123064910	8.15363	15.16718	3.00E-05	0	
chr11	114099820	114100210	8.13029	55.03235	3.00E-05	0	
chr6	164984980	164985290	8.09328	12.02406	3.00E-05	0	
chr19	31154040	31154330	8.0629	10.68109	3.00E-05	0	
chr2	208003340	208003890	8.05686	6.49751	3.00E-05	0.00016	phastBias
chr10	123914220	123914600	8.00411	13.68355	3.00E-05	0	phastBias
chr4	16813760	16814170	7.98167	9.27382	3.00E-05	1.00E-05	
chr18	62504900	62505110	7.9812	4.80949	3.00E-05	0.00096	
chr5	4570760	4571010	7.9356	20.29314	3.00E-05	0	
chr18	68137920	68138170	7.92901	5.79134	3.00E-05	0.00033	phastBias
chr6	11607620	11608050	7.91972	8.29282	3.00E-05	2.00E-05	
chr7	154412000	154412150	7.91129	6.10499	3.00E-05	0.00024	
chr1	107292620	107292910	7.90481	7.65388	4.00E-05	5.00E-05	
chr4	85395340	85395570	7.87939	23.42346	4.00E-05	0	
chr11	17786920	17787230	7.83507	10.87022	4.00E-05	0	
chr6	46012640	46012930	7.83135	4.65864	4.00E-05	0.00114	
chr1	101490800	101491270	7.83121	6.82833	4.00E-05	0.00011	
chr3	193801080	193801330	7.81137	8.806	4.00E-05	1.00E-05	
chr10	70010900	70011355	7.75879	51.22538	4.00E-05	0	
chr5	117901160	117901310	7.75023	5.20538	4.00E-05	0.00063	
chr14	40133125	40133715	7.73307	9.47614	4.00E-05	1.00E-05	
chr18	44068825	44069010	7.72838	6.40099	4.00E-05	0.00017	
chr7	138969080	138969270	7.69353	31.14558	4.00E-05	0	
chr9	603160	603470	7.67708	9.51843	4.00E-05	1.00E-05	phastBias
chr16	17812480	17812630	7.66327	6.76044	5.00E-05	0.00012	
chr1	203336280	203336530	7.64177	10.96814	5.00E-05	0	
chr8	19613780	19614220	7.62523	10.90291	5.00E-05	0	
chr18	22657985	22658150	7.61584	23.9339	5.00E-05	0	
chr15	96952460	96953100	7.60359	4.44952	5.00E-05	0.00143	phastBias
chr1	238497460	238497710	7.59611	7.2316	5.00E-05	7.00E-05	
chr1	183746320	183746470	7.59091	5.52819	5.00E-05	0.00044	
chr11	113936620	113937250	7.58612	8.82828	5.00E-05	1.00E-05	
chr11	131852600	131852790	7.56905	5.23425	5.00E-05	0.00061	
chr16	78471280	78471570	7.56446	13.48144	5.00E-05	0	
chr18	73257060	73257390	7.5167	8.62283	5.00E-05	2.00E-05	
chr18	45735100	45735390	7.49692	19.90667	5.00E-05	0	
chr17	42215400	42215995	7.49374	4.06484	5.00E-05	0.00218	
chr7	15582640	15582790	7.48242	8.30386	5.00E-05	2.00E-05	
chr10	122945740	122945950	7.46987	12.62244	6.00E-05	0	
chr6	91190720	91190910	7.46445	4.21264	6.00E-05	0.00185	
chr18	38430800	38430950	7.43318	4.54008	6.00E-05	0.00129	
chr13	107001720	107002090	7.42956	20.68312	6.00E-05	0	
chr4	54809800	54810210	7.41874	10.26178	6.00E-05	0	
chr15	33306860	33307070	7.38567	5.52905	6.00E-05	0.00044	phastBias
chr2	64067160	64067310	7.34946	6.63902	6.00E-05	0.00013	
chr5	128988040	128988310	7.34811	4.08359	6.00E-05	0.00213	

chr8	26087760	26088020	7.34424	13.49294	6.00E-05	0	
chr1	176663120	176663435	7.34015	14.11764	6.00E-05	0	
chr1	2928620	2929080	7.33732	15.4797	6.00E-05	0	
chr8	89128940	89129090	7.33218	24.27049	6.00E-05	0	
chr7	11765940	11766250	7.31519	6.81804	7.00E-05	0.00011	
chr14	33180340	33180630	7.28184	4.37031	7.00E-05	0.00156	
chr7	26765820	26765970	7.27854	20.10153	7.00E-05	0	
chr7	69944180	69944510	7.26717	6.85065	7.00E-05	0.00011	
chr1	246230840	246231550	7.25262	4.754	7.00E-05	0.00102	phastBias
chr3	61268940	61269090	7.24787	6.90751	7.00E-05	1.00E-04	
chr1	200438420	200438690	7.24403	15.70494	7.00E-05	0	
chr16	80155720	80155870	7.23488	18.06394	7.00E-05	0	
chr3	127168440	127168630	7.23142	17.42399	7.00E-05	0	
chr16	59587760	59587930	7.2306	8.30513	7.00E-05	2.00E-05	
chr4	130202780	130202930	7.22606	7.26416	7.00E-05	7.00E-05	
chr15	70306080	70306390	7.215	14.28578	7.00E-05	0	
chr5	158229000	158229150	7.21157	21.19217	7.00E-05	0	
chr14	99502300	99502660	7.20301	23.59046	7.00E-05	0	
chr9	1743600	1743750	7.20055	14.87247	7.00E-05	0	phastBias
chr9	20804580	20804730	7.19762	11.07936	7.00E-05	0	
chr15	38170560	38170950	7.19738	16.99665	7.00E-05	0	
chr1	223059740	223060440	7.19615	14.59616	7.00E-05	0	
chr3	65869200	65869630	7.18423	5.59107	8.00E-05	0.00041	
chr12	43150800	43151235	7.14745	15.68766	8.00E-05	0	
chrX	33152220	33152610	7.13501	11.17207	8.00E-05	0	
chrX	17065460	17065630	7.13103	8.54149	8.00E-05	2.00E-05	
chr8	77282900	77283130	7.12941	14.4694	8.00E-05	0	
chr7	69187880	69188250	7.12031	6.37981	8.00E-05	0.00018	
chr18	22714860	22715210	7.09872	17.76958	8.00E-05	0	
chr15	75746245	75746770	7.09162	8.85071	8.00E-05	1.00E-05	
chr8	72639380	72639530	7.07661	6.97782	8.00E-05	9.00E-05	
chr6	90861800	90862070	7.07399	6.53653	8.00E-05	0.00015	
chr12	99515380	99515710	7.0706	11.3903	8.00E-05	0	
chrX	22410280	22410760	7.05818	16.72301	9.00E-05	0	
chr18	23259560	23259950	7.03998	15.81859	9.00E-05	0	
chr8	130368840	130368990	7.03963	4.87359	9.00E-05	9.00E-04	
chr1	49082760	49082910	7.03696	7.84365	9.00E-05	4.00E-05	
chr6	98946600	98946750	7.03378	7.52245	9.00E-05	5.00E-05	
chr7	138564200	138564350	7.03229	6.201	9.00E-05	0.00021	
chr13	67009340	67009630	7.01356	9.63075	9.00E-05	1.00E-05	
chrX	138284080	138284310	6.99982	5.75194	9.00E-05	0.00035	
chr5	122812460	122812690	6.99694	4.62626	9.00E-05	0.00118	
chr2	26792905	26793190	6.99527	17.02905	9.00E-05	0	
chr11	103924820	103925070	6.99303	6.31144	9.00E-05	0.00019	
chr3	175569100	175569890	6.96996	4.65922	9.00E-05	0.00113	
chr14	98099540	98099730	6.96368	30.67556	1.00E-04	0	
chr9	13755620	13755990	6.96162	15.76978	1.00E-04	0	phastBias
chr14	90040900	90041090	6.95291	5.99893	1.00E-04	0.00027	
chr11	125063060	125063290	6.9353	18.39724	1.00E-04	0	
chr9	126225240	126225430	6.92746	4.74366	1.00E-04	0.00103	

chr8	96356960	96357130	6.92596	4.89145	1.00E-04	0.00088	
chr5	28215200	28215750	6.91972	25.622	1.00E-04	0	
chr5	157753260	157753530	6.89591	15.78608	1.00E-04	0	
chr10	3324160	3324450	6.89517	7.42248	1.00E-04	6.00E-05	
chr21	17443380	17443790	6.88296	51.00045	1.00E-04	0	
chr14	48143340	48143880	6.86022	45.55564	0.00011	0	
chr4	183128080	183128490	6.85904	25.61924	0.00011	0	
chr4	17022180	17022330	6.85636	12.0283	0.00011	0	
chr6	16431400	16431670	6.85318	22.97086	0.00011	0	
chr11	86299520	86299880	6.84477	6.16557	0.00011	0.00022	
chr12	65297600	65297870	6.84028	7.76565	0.00011	4.00E-05	
chr2	208001560	208002050	6.83986	6.0199	0.00011	0.00026	
chr10	4764160	4764570	6.8085	5.09574	0.00011	0.00071	
chr1	19668300	19668935	6.80736	7.44345	0.00011	6.00E-05	
chr13	101126060	101126210	6.80355	11.8689	0.00011	0	phastBias
chr9	3646960	3647110	6.79457	11.63061	0.00011	0	
chr2	12061180	12061650	6.79043	21.24816	0.00011	0	
chr12	78240800	78241080	6.78709	19.62179	0.00011	0	
chr2	145868640	145868810	6.76842	6.32686	0.00012	0.00019	
chr1	232170400	232170630	6.75444	8.55839	0.00012	2.00E-05	
chr16	82661420	82661700	6.74521	16.69747	0.00012	0	
chr12	97553225	97553530	6.74312	12.9683	0.00012	0	
chr1	107291285	107291490	6.74136	4.2427	0.00012	0.00179	
chr10	119354320	119354570	6.73242	5.11509	0.00012	0.00069	
chr6	98990580	98990730	6.7283	7.89063	0.00012	4.00E-05	
chr5	168313300	168313470	6.72558	11.17172	0.00012	0	
chr3	12940580	12940875	6.72149	17.3563	0.00012	0	
chrX	18050260	18050450	6.71955	7.12773	0.00012	8.00E-05	
chr14	74600580	74600770	6.71889	13.51546	0.00012	0	
chr12	128346080	128346410	6.71882	7.85558	0.00012	4.00E-05	
chr6	40489120	40489330	6.71689	17.53398	0.00012	0	
chr1	209630840	209631070	6.7149	14.47003	0.00012	0	
chr7	31232560	31232875	6.70694	12.14865	0.00012	0	
chr5	5539725	5540070	6.69344	7.00774	0.00013	9.00E-05	
chr9	82394840	82395230	6.68891	21.39883	0.00013	0	
chr1	216356780	216357170	6.67748	48.84537	0.00013	0	
chr9	102495040	102495410	6.67642	28.59623	0.00013	0	
chr5	134570320	134570750	6.66112	14.92614	0.00013	0	
chr9	109622100	109622540	6.62697	19.25658	0.00014	0	
chr10	108787280	108787430	6.62253	12.4163	0.00014	0	
chrX	153220145	153220335	6.62028	8.01089	0.00014	3.00E-05	
chr21	33220160	33220450	6.59476	4.05326	0.00014	0.00221	
chr6	93472080	93472790	6.56107	50.37847	0.00015	0	
chr3	2701980	2702310	6.56032	4.4762	0.00015	0.00139	phastBias
chr9	2623240	2623780	6.55366	4.05271	0.00015	0.00221	phastBias
chr7	31512940	31513090	6.53055	12.11563	0.00015	0	
chr18	53359420	53359710	6.52864	24.16927	0.00015	0	
chr4	145793560	145793990	6.52274	38.27262	0.00015	0	
chr17	71800180	71800755	6.51875	23.87087	0.00015	0	
chr2	8549680	8550270	6.49304	7.3796	0.00016	6.00E-05	

chr2	106054240	106054550	6.47396	12.39828	0.00016	0	
chr1	19688140	19688655	6.4688	16.75899	0.00016	0	
chrX	21816980	21817370	6.44854	5.75492	0.00016	0.00035	
chr13	48224440	48224670	6.44691	5.08229	0.00016	0.00072	
chr13	110900260	110900755	6.4322	11.40595	0.00017	0	phastBias
chr18	64447520	64447670	6.43133	5.56192	0.00017	0.00043	
chr14	60999240	60999450	6.4263	4.54208	0.00017	0.00129	
chr2	225510500	225510790	6.42194	13.06154	0.00017	0	
chr18	19402440	19402590	6.41675	4.78007	0.00017	0.00099	
chr2	21980620	21981350	6.403	48.34143	0.00017	0	
chr4	143114420	143114570	6.40249	5.8235	0.00017	0.00032	
chr9	14531525	14531730	6.40195	19.12015	0.00017	0	
chr12	130548545	130548710	6.39576	15.06851	0.00017	0	
chr1	202251720	202251990	6.38367	6.01938	0.00018	0.00026	
chr14	69172700	69173160	6.38063	7.17701	0.00018	8.00E-05	
chr7	8418920	8419230	6.36997	14.63057	0.00018	0	
chr8	50969280	50969870	6.3659	8.41177	0.00018	2.00E-05	
chr6	1626280	1626610	6.36572	6.49008	0.00018	0.00016	phastBias
chrX	133309720	133309950	6.35656	6.74584	0.00018	0.00012	
chr12	102864780	102865030	6.35529	7.01287	0.00018	9.00E-05	
chr10	77809800	77810195	6.35415	23.36604	0.00018	0	
chr3	97454780	97454930	6.35269	20.76596	0.00018	0	
chr15	102010900	102011355	6.34955	4.95738	0.00018	0.00082	
chr5	88753780	88754150	6.34801	22.48894	0.00018	0	
chr2	22822540	22822770	6.34109	14.16527	0.00018	0	
chr6	140783380	140783530	6.33662	4.56659	0.00019	0.00126	
chr18	56578960	56579470	6.33601	4.46689	0.00019	0.0014	
chr2	215442400	215442850	6.31618	23.26651	0.00019	0	
chr7	152999040	152999310	6.31405	6.98168	0.00019	9.00E-05	phastBias
chr10	114345620	114346070	6.28283	29.17486	2.00E-04	0	
chr20	11108780	11109390	6.275	20.19699	2.00E-04	0	
chr10	73113680	73113870	6.27092	10.60549	2.00E-04	0	
chr11	87015480	87015850	6.26881	4.40699	2.00E-04	0.00149	
chr4	128260880	128261070	6.26792	12.67184	2.00E-04	0	
chr7	119916880	119917090	6.25303	6.05647	2.00E-04	0.00025	
chr4	182728980	182729210	6.24889	5.2548	2.00E-04	0.00059	
chr4	182271860	182272010	6.24793	15.24122	2.00E-04	0	
chr7	69075265	69075610	6.24329	10.98945	2.00E-04	0	
chr11	31893460	31893930	6.2249	4.59325	0.00021	0.00122	
chr4	14771660	14771910	6.22427	29.31265	0.00021	0	
chr3	41381360	41381510	6.20582	7.20889	0.00021	7.00E-05	
chr1	57184940	57185130	6.205	14.72981	0.00021	0	
chr8	92020860	92021350	6.2	12.75931	0.00021	0	
chr8	25837740	25837890	6.19765	7.83672	0.00022	4.00E-05	
chr6	45868860	45869175	6.19732	8.70022	0.00022	2.00E-05	
chr4	105341940	105342490	6.19646	50.12171	0.00022	0	
chr12	92788500	92788870	6.19578	6.96279	0.00022	1.00E-04	
chr5	5582205	5582470	6.19302	8.61688	0.00022	2.00E-05	
chr10	48557700	48558350	6.19004	7.16783	0.00022	8.00E-05	
chr9	2159560	2160390	6.1869	88.19634	0.00022	0	phastBias

chr6	14431640	14431950	6.18352	19.71739	0.00022	0	
chr10	8794120	8794270	6.17884	21.58621	0.00022	0	
chr7	23146520	23147230	6.17637	3.97341	0.00022	0.00241	
chr11	103151280	103151430	6.16621	7.5626	0.00022	5.00E-05	
chr3	6969840	6970070	6.16382	7.71973	0.00022	4.00E-05	
chr9	272805	273130	6.14444	8.10975	0.00023	3.00E-05	phastBias
chr10	116675140	116675470	6.13586	14.36484	0.00023	0	
chr11	15561260	15561470	6.13551	7.53585	0.00023	5.00E-05	
chr1	113811840	113811990	6.13197	7.15106	0.00023	8.00E-05	
chr15	35814080	35814650	6.12692	32.6043	0.00023	0	
chr3	143201620	143201790	6.12194	4.32063	0.00023	0.00164	
chr14	33526920	33527130	6.10268	17.02481	0.00024	0	
chr4	99567320	99567490	6.09322	4.65701	0.00024	0.00114	phastBias
chr10	109431580	109431970	6.09262	10.21058	0.00024	0	
chr5	132803540	132803770	6.09	10.89924	0.00024	0	
chr5	5374160	5374520	6.08973	11.46633	0.00024	0	
chr6	73208800	73208950	6.08624	19.11502	0.00024	0	
chr2	119600160	119600580	6.08533	26.22568	0.00024	0	phastBias
chr8	10234780	10235030	6.08518	4.68835	0.00024	0.0011	
chr7	134501240	134501650	6.08217	8.52102	0.00024	2.00E-05	
chr7	18534740	18535050	6.07807	32.51562	0.00024	0	
chr8	110650340	110650590	6.07723	12.93978	0.00024	0	
chr1	245851420	245851950	6.06931	20.68844	0.00025	0	phastBias
chr18	57368400	57368850	6.06905	9.07093	0.00025	1.00E-05	
chr4	179441860	179442010	6.06829	4.79847	0.00025	0.00097	
chr3	35721780	35722050	6.06411	9.91103	0.00025	0	
chr1	54725380	54725670	6.0606	22.53293	0.00025	0	
chr7	127176020	127176170	6.0563	17.11584	0.00025	0	
chr2	36654940	36655220	6.05288	5.6118	0.00025	4.00E-04	
chr3	157212480	157213050	6.04976	7.46145	0.00025	6.00E-05	
chr17	6063620	6063810	6.04395	7.95604	0.00025	3.00E-05	
chr9	2241880	2242215	6.04278	5.67766	0.00025	0.00038	phastBias
chr15	67687200	67687490	6.03718	10.49884	0.00026	0	
chr2	195351180	195351330	6.0324	5.61642	0.00026	4.00E-04	
chr10	78874300	78874535	6.03135	11.42221	0.00026	0	
chr3	34354220	34354450	6.02749	18.91686	0.00026	0	
chr3	185957140	185957395	6.02326	21.21842	0.00026	0	phastBias
chr13	100395325	100395650	6.02158	31.46687	0.00026	0	
chr15	94433300	94433450	6.01707	4.6842	0.00026	0.0011	
chr12	18694240	18694390	6.0145	11.0403	0.00026	0	
chr2	218770160	218770410	6.01304	9.69315	0.00026	1.00E-05	
chr16	65043720	65043870	6.00717	8.37	0.00026	2.00E-05	
chr4	155151360	155151550	6.00557	6.11433	0.00026	0.00024	
chr3	19442100	19442350	5.99662	4.52394	0.00027	0.00132	
chr5	11888460	11888650	5.98699	4.66801	0.00027	0.00112	
chr1	87718160	87718430	5.97907	5.19454	0.00027	0.00063	
chr5	124726000	124726350	5.96612	7.98792	0.00028	3.00E-05	
chr4	20195240	20195950	5.96027	10.47781	0.00028	0	
chr2	84461280	84461510	5.95386	8.72871	0.00028	1.00E-05	
chr8	10981620	10981970	5.93776	6.00035	0.00028	0.00027	

chr8	18438200	18438710	5.93752	7.23511	0.00028	7.00E-05	
chr17	73686725	73686970	5.92761	9.32337	0.00029	1.00E-05	
chr1	214549280	214549570	5.92437	7.44667	0.00029	6.00E-05	
chr14	32346360	32346890	5.91992	7.71196	0.00029	4.00E-05	phastBias
chr10	77499545	77499875	5.91506	15.94968	0.00029	0	
chr5	81518480	81518990	5.91298	17.56043	0.00029	0	
chr10	24642360	24642590	5.9117	4.29826	0.00029	0.00168	
chr4	48081980	48082790	5.90987	7.20808	0.00029	7.00E-05	
chr10	123072640	123072930	5.90486	21.52758	0.00029	0	
chr22	28145200	28145355	5.9024	6.64351	3.00E-04	0.00013	
chr8	6366400	6366690	5.89776	15.88817	3.00E-04	0	
chr6	100694180	100694410	5.89351	7.64273	3.00E-04	5.00E-05	
chr19	31398500	31398650	5.89072	19.60622	3.00E-04	0	
chr2	21676640	21676870	5.88006	14.22221	3.00E-04	0	
chr2	53341860	53342050	5.87653	17.6954	3.00E-04	0	
chr18	39998340	39998530	5.87439	5.37806	3.00E-04	0.00052	
chr8	144561860	144562030	5.8738	23.96702	3.00E-04	0	
chr3	114378340	114378770	5.87325	7.53557	3.00E-04	5.00E-05	
chr7	94051000	94051390	5.87276	4.25027	3.00E-04	0.00178	
chr6	1385280	1386010	5.87257	10.07699	0.00031	0	phastBias
chr7	126631180	126631530	5.87188	7.71638	0.00031	4.00E-05	
chr9	119139565	119140010	5.86879	8.90089	0.00031	1.00E-05	
chr2	22551960	22552550	5.86673	24.98196	0.00031	0	
chr6	46922380	46922900	5.86526	7.66083	0.00031	5.00E-05	
chr4	93686580	93686730	5.86253	6.82773	0.00031	0.00011	
chr10	117873800	117874030	5.85923	13.25868	0.00031	0	
chr11	12351420	12351650	5.85525	6.2578	0.00031	2.00E-04	
chr5	151462960	151463170	5.85344	7.28231	0.00031	7.00E-05	
chr11	16191520	16191870	5.85149	20.13669	0.00031	0	
chr20	9455560	9456030	5.8445	9.25426	0.00031	1.00E-05	
chr18	22622580	22622930	5.83472	13.01643	0.00032	0	
chr2	60280860	60281290	5.83384	4.94119	0.00032	0.00083	
chr3	1237640	1237970	5.83338	6.60493	0.00032	0.00014	
chr10	14144800	14145010	5.83307	4.31049	0.00032	0.00166	
chr21	42059585	42059850	5.82907	4.02319	0.00032	0.00228	
chr8	38237780	38238200	5.82643	9.30675	0.00032	1.00E-05	
chrX	31850280	31850670	5.81778	9.55228	0.00032	1.00E-05	
chr8	58779560	58779890	5.81635	5.12254	0.00032	0.00069	
chr3	178785140	178785410	5.80596	5.20236	0.00033	0.00063	
chr11	30952360	30952610	5.80389	8.98512	0.00033	1.00E-05	
chr15	36587220	36587370	5.80316	9.23805	0.00033	1.00E-05	
chr15	61591340	61591590	5.80129	10.05083	0.00033	0	
chr19	34857520	34858080	5.80067	4.27275	0.00033	0.00173	
chr7	25802220	25802490	5.79735	5.74454	0.00033	0.00035	
chr9	74296120	74296600	5.78627	4.44737	0.00033	0.00143	
chr16	73086940	73087560	5.77803	34.88889	0.00034	0	phastBias
chr9	98803060	98803360	5.77685	8.15385	0.00034	3.00E-05	
chr10	123268680	123268930	5.7652	9.98692	0.00034	0	
chr17	46222880	46223110	5.75896	15.42062	0.00034	0	
chr12	1157465	1157770	5.75777	17.62297	0.00035	0	

chr5	74434620	74434870	5.75232	4.53447	0.00035	0.0013	
chr14	78184040	78184190	5.75086	6.09491	0.00035	0.00024	
chr6	152451620	152451870	5.74938	8.1691	0.00035	3.00E-05	
chr4	38948180	38948430	5.74863	9.68613	0.00035	1.00E-05	
chr15	37190760	37191210	5.74845	44.24177	0.00035	0	
chr3	67458320	67458470	5.74515	7.87013	0.00035	4.00E-05	
chr7	39447320	39447510	5.74468	13.69174	0.00035	0	
chr16	77613220	77613560	5.74351	11.68105	0.00035	0	
chr3	147239500	147239770	5.73663	6.68317	0.00035	0.00013	
chr9	16830880	16831150	5.73563	13.66304	0.00035	0	
chr1	79659180	79659440	5.73273	9.80635	0.00035	0	
chr4	23979680	23979930	5.72649	6.99298	0.00036	9.00E-05	
chr5	54752080	54752400	5.71787	6.27182	0.00036	2.00E-04	
chr7	33483600	33483750	5.71203	7.83977	0.00036	4.00E-05	
chr9	16964160	16964530	5.71092	5.68389	0.00036	0.00037	phastBias
chr9	129200200	129200350	5.7106	16.10676	0.00036	0	
chr4	138683860	138684010	5.7058	7.2848	0.00036	7.00E-05	
chr11	31623260	31623710	5.70046	49.91765	0.00037	0	
chr7	39018140	39018550	5.69836	27.91519	0.00037	0	
chr1	237207560	237207935	5.68625	7.72336	0.00037	4.00E-05	
chr3	7961680	7961930	5.68434	4.42005	0.00037	0.00147	
chr15	96389100	96389790	5.68375	28.8499	0.00037	0	
chr6	98946320	98946550	5.681	26.4023	0.00037	0	
chr12	102029520	102029750	5.67838	7.34222	0.00038	6.00E-05	
chr7	13563400	13563650	5.67296	7.24688	0.00038	7.00E-05	
chr18	43142480	43143030	5.67177	30.71191	0.00038	0	
chr3	114137840	114138090	5.66986	16.71227	0.00038	0	
chr2	161453020	161453410	5.66778	8.18752	0.00038	3.00E-05	
chr2	119036605	119036870	5.66685	4.12129	0.00038	0.00205	
chr18	2240380	2240610	5.66416	6.26486	0.00038	2.00E-04	
chr4	105345660	105346230	5.65984	72.12672	0.00038	0	
chr16	82816280	82817030	5.65837	13.50274	0.00038	0	
chr4	95855420	95855810	5.65761	14.42909	0.00038	0	
chr3	67637740	67638010	5.65732	13.35934	0.00038	0	
chr6	155988260	155988410	5.65716	10.04728	0.00038	0	
chr9	18448240	18448410	5.6537	6.48793	0.00039	0.00016	
chr3	29482120	29482455	5.65109	25.62184	0.00039	0	
chr1	175173420	175173770	5.64911	8.43545	0.00039	2.00E-05	
chr14	78365365	78365515	5.64193	6.935	0.00039	1.00E-04	
chr8	116714600	116715110	5.64029	37.6073	0.00039	0	
chr1	209989200	209989690	5.63995	24.82846	0.00039	0	
chr6	5167780	5168070	5.6337	5.98228	0.00039	0.00027	phastBias
chr2	227362040	227362190	5.632	6.30885	4.00E-04	0.00019	
chr3	60787840	60787990	5.62918	6.39765	4.00E-04	0.00017	
chr5	110456700	110456890	5.62593	9.82798	4.00E-04	0	
chr2	108602605	108603115	5.61576	12.05958	4.00E-04	0	
chr13	105442800	105442950	5.6143	11.92769	4.00E-04	0	
chr18	27836740	27836890	5.60644	7.01153	0.00041	9.00E-05	
chr9	134891940	134892170	5.60494	14.77874	0.00041	0	
chr1	108225700	108226070	5.60412	4.75986	0.00041	0.00102	

chr9	18496700	18497010	5.5988	8.24914	0.00041	2.00E-05
chr5	102276780	102277170	5.59869	5.6554	0.00041	0.00039
chr6	11228800	11229030	5.59782	4.24771	0.00041	0.00178
chr10	78010380	78010950	5.59614	39.83546	0.00041	0
chr2	151478040	151478710	5.59397	5.73174	0.00041	0.00035
chr8	94090760	94091090	5.59287	19.86644	0.00041	0
chr3	63953580	63953820	5.59202	14.20165	0.00041	0
chr14	70102660	70103430	5.59184	4.73817	0.00041	0.00104
chr16	27811380	27811550	5.59052	5.33722	0.00041	0.00054
chr11	113985980	113986695	5.58923	21.27426	0.00041	0
chr11	88908980	88909310	5.5836	15.02096	0.00042	0
chr21	36341100	36341590	5.57646	4.38096	0.00042	0.00154
chr1	88261640	88261890	5.57563	6.83069	0.00042	0.00011
chr8	40068200	40068610	5.57503	8.04841	0.00042	3.00E-05
chr16	87098800	87099370	5.5737	9.13728	0.00042	1.00E-05
chr1	183820000	183820490	5.56956	8.49409	0.00042	2.00E-05
chr9	82283020	82283350	5.5682	27.19005	0.00042	0
chr9	125043820	125044490	5.56463	26.22455	0.00042	0
chr8	93115000	93115270	5.56262	13.35181	0.00043	0
chr20	51759220	51759370	5.55841	5.03357	0.00043	0.00075
chr6	140887760	140888180	5.5506	5.57062	0.00043	0.00042
chr2	105489020	105489380	5.54952	5.03477	0.00043	0.00075
chr8	135421560	135421710	5.54927	10.98151	0.00043	0
chr4	55948740	55948890	5.54779	7.41379	0.00043	6.00E-05
chr12	108066120	108066350	5.54402	8.32031	0.00043	2.00E-05
chr7	127295120	127295330	5.53866	5.84444	0.00044	0.00031
chr4	16758780	16759210	5.53601	22.72875	0.00044	0
chrX	41510360	41510510	5.53302	4.03715	0.00044	0.00224
chr8	145693700	145694130	5.52326	25.34977	0.00044	0
chr6	33692040	33692330	5.52176	7.91759	0.00044	3.00E-05
chr4	72188240	72188450	5.51702	9.73126	0.00045	1.00E-05
chr11	115915700	115915910	5.51019	25.6847	0.00045	0
chr4	183228820	183229250	5.50769	4.60664	0.00045	0.0012
chr9	8784020	8784350	5.50202	31.80054	0.00045	0
chr1	81011420	81011570	5.50039	4.27683	0.00046	0.00172
chr11	42103800	42104190	5.49201	7.23318	0.00046	7.00E-05
chr21	45502400	45502940	5.48498	8.37157	0.00046	2.00E-05
chr3	107724960	107725555	5.48251	6.83176	0.00046	0.00011
chr15	94988820	94989050	5.48223	9.77964	0.00046	0
chr14	64705360	64705795	5.47917	4.94137	0.00047	0.00083
chr2	169630245	169630560	5.47689	4.72066	0.00047	0.00106
chr2	108677920	108678150	5.47092	7.65536	0.00047	5.00E-05
chr7	146351220	146351510	5.46985	17.63155	0.00047	0
chr3	71577620	71577830	5.46939	4.71356	0.00047	0.00107
chr18	53999705	53999890	5.46753	23.81228	0.00047	0
chr20	21922580	21922890	5.46638	12.52929	0.00047	0
chr3	168594220	168594500	5.46289	14.50059	0.00047	0
chr5	33616100	33616250	5.46062	6.43562	0.00048	0.00017
chr12	97752840	97752990	5.45377	9.69628	0.00048	1.00E-05
chr10	132615380	132615630	5.45148	4.93645	0.00048	0.00084

chr17	9877400	9877710	5.45119	6.08079	0.00048	0.00024	
chr18	31165125	31165670	5.45015	38.77336	0.00048	0	
chr2	217728640	217729030	5.44919	10.82164	0.00048	0	
chr8	57125060	57125410	5.44735	9.41833	0.00048	1.00E-05	
chr17	28004700	28004850	5.44521	4.09121	0.00048	0.00211	
chr2	124996400	124996550	5.44479	9.83856	0.00048	0	
chr20	51981200	51981370	5.44452	10.49546	0.00048	0	
chr8	62006300	62006670	5.4415	7.85534	0.00049	4.00E-05	
chr10	28552440	28552755	5.43854	11.46425	0.00049	0	
chr6	51504840	51505250	5.43725	35.31501	0.00049	0	
chr14	55113560	55113750	5.4351	8.43798	0.00049	2.00E-05	
chr6	70980385	70980615	5.43408	6.17329	0.00049	0.00022	
chr7	18332260	18332410	5.43098	6.83724	0.00049	0.00011	
chr6	77595560	77595820	5.4291	22.75157	0.00049	0	
chr2	114647840	114648480	5.42496	18.41802	0.00049	0	phastBias
chr4	102082560	102082810	5.42295	12.02182	5.00E-04	0	
chr12	79559080	79559230	5.42288	5.08174	5.00E-04	0.00072	
chr9	16138580	16139590	5.41997	12.30692	5.00E-04	0	
chr12	110977600	110977870	5.41714	4.74677	5.00E-04	0.00103	
chr1	109260560	109260850	5.41429	24.11853	5.00E-04	0	
chr13	67462900	67463330	5.41268	12.6813	5.00E-04	0	
chr3	39136440	39136870	5.41093	9.58441	5.00E-04	1.00E-05	
chr5	142005180	142005550	5.40763	7.00887	5.00E-04	9.00E-05	
chr2	23264260	23264470	5.40714	10.41467	5.00E-04	0	
chr4	137813360	137813630	5.40638	13.82638	5.00E-04	0	
chr11	19902800	19902990	5.39467	5.43409	0.00051	0.00049	
chr18	35249100	35249370	5.39399	19.79967	0.00051	0	
chr8	145698180	145698795	5.39106	12.7732	0.00051	0	
chr4	183717100	183717635	5.39045	6.36991	0.00051	0.00018	phastBias
chr5	24623160	24623450	5.3874	16.00023	0.00051	0	
chr3	114304600	114305350	5.38675	52.78358	0.00051	0	

Table B.3 Transgenic mouse assay results. Results for all 18 transgenic mouse assays, including published and novel. The results column contains the sequence ID in the vista enhancer browser, and the numbers in parentheses indicate how many embryos were positive for staining at that particular region.

chr	start	stop	PMID if published	results
chr1	209989050	209989824	PMID:18836445	Human_element932_positive_limb[9/9]:branchialarch[9/9]:eye[6/9]:nose[7/9]
chr2	236773458	236774696	PMID:18772437	Human_element521_positive_limb[6/6]:branchialarch[6/6]:eye[4/6]:ear[6/6]
chr2	238221820	238225273	PMID: 22138689	Human_element1951_positive_somite[9/11]:heart[11/11]
chr4	105345575	105346895	PMID: 23253453	Human_element260_positive_midbrain(mesencephalon)[6/8]
chr5	157752608	157754317		Human_element1735_negative
chr5	158227696	158229500	PMID:23375746	Human_element1137_negative
chr7	127174386	127177546		Human_element1308_positive_midbrain(mesencephalon)[7/7]:forebrain[7/7]
chr8	27752440	27753351		Human_element1741_positive_hindbrain(rhombencephalon)[4/6]:limb[6/6]:eye[5/6]:nose[5/6]:facialmesenchyme[5/6]
chr8	32063683	32067187	PMID:23375746	Human_element1719_negative
chr8	32064662	32066126		Human_element1742_negative
chr9	2240936	2242833		Human_element1743_positive_neuraltube[4/4]:hindbrain(rhombencephalon)[4/4]:midbrain(mesencephalon)[4/4]:forebrain[4/4]:trigeminalV(ganglion,cranial)[4/4]:cranialnerve[4/4]
chr9	82224085	82226757	PMID:23375746	Human_element1078_positive_neuraltube[7/8]:hindbrain(rhombencephalon)[7/8]:midbrain(mesencephalon)[7/8]:forebrain[7/8]:cranialnerve[7/8]
chr9	129198400	129200739		Human_element186_positive_midbrain(mesencephalon)[6/8]
chr11	16190865	16192453		Human_element1720_positive_branchialarch[5/8]
chr11	31622822	31624118		Human_element565_positive_midbrain(mesencephalon)[8/10]:forebrain[8/10]
chr12	102863674	102866150	PIMD:24159046	Mouse_element694_positive_limb[6/10]:branchialarch[6/10]:nose[7/10]:tail[7/10]:facialmesenchyme[7/10]:other[7/10]

chr15	37240805	37242498		Human_element181_positive_midbrain(mesencephalon)[4/4]
chr18	22657391	22658788	PMID:23375746	Human_element1104_negative

Table B.4 Luciferase data. This table contains computed p values for each experiment. Tests for allelic differences were two sided, and others were one sided. All tests were t tests.

#	cell type	allele diff	Human > control	Chimp > control
1	imr90	0.008470509	0.003768851	0.3379964
2	sknmc	0.151543137	0.06015201	0.5180263
3	imr90	0.01464019	0.9178468	0.9999885
3	sknmc	0.144598153	0.2833337	0.9862089
4	sknmc	0.760708679	0.1115235	0.1119847
5	sknmc	0.357990842	0.009119022	0.06230156
6	sknmc	0.460078893	0.2482261	0.09797473
7	imr90	0.912083125	0.7629743	0.6919462
8	imr90	0.395070846	0.09678877	0.03159584
8	sknmc	0.180841419	0.3724764	0.05890304
9	sknmc	0.525926478	0.1062244	0.3230781
10	imr90	0.578749003	0.1138591	0.1118162
10	sknmc	0.091772528	0.7989634	0.05580223
11	sknmc	0.770086866	0.9091693	0.8756497
12	imr90	0.139393889	0.3825551	0.04039827
12	sknmc	0.004029742	0.01747009	0.001415629
13	sknmc	0.078363482	0.2746814	0.9755962
14	imr90	0.350270611	0.1953485	0.6115706
14	sknmc	0.061485734	0.9591503	0.1011531
15	sknmc	0.090275717	0.9070156	0.1462604
16	imr90	0.657217251	0.3078073	0.126206
16	sknmc	0.126676566	0.004794537	0.007884615
17	sknmc	0.012989243	0.5029828	0.9993407
18	imr90	0.931721174	0.4403896	0.4850841
18	sknmc	0.399813735	0.9534786	0.9312041
19	sknmc	0.32197722	6.31E-05	0.008039423
20	sknmc	0.007848733	0.9998017	0.7268415
21	sknmc	0.133778516	0.9994674	0.9841761
22	imr90	0.260237768	0.9970872	0.8870729
22	sknmc	0.467189406	0.6701458	0.2915965
23	sknmc	0.67770751	0.9089011	0.8134739
24	sknmc	0.218799951	0.004455866	0.03014861
25	imr90	0.378459242	0.1116381	0.0127046
25	sknmc	0.132336448	0.03924705	0.1326494
26	sknmc	0.029236463	0.008692464	0.02253292
26	imr90	0.61360849	0.02195662	0.3913769
26	sknmc	0.441676485	0.006907962	0.002690956

27	imr90	0.411448242	0.9652618	0.9914308
28	sknmc	0.191372613	0.9282409	0.4229041
29	imr90	0.856602474	0.1440336	0.1575513
29	sknmc	0.004769836	0.5759711	0.001849556
30	imr90	0.17835091	0.2363692	0.8671196
30	sknmc	0.081365018	0.05005254	1.53E-09
31	imr90	0.50936479	0.154245	0.2607579
31	sknmc	0.192377757	0.1206292	0.04788078
32	imr90	0.072813686	0.3427268	0.9901729
32	sknmc	0.597401406	0.8052173	0.7544005
33	sknmc	0.905599441	0.5506777	0.5019928
34	sknmc	0.198637906	0.1335978	0.7113559
35	sknmc	0.211466635	0.9996813	0.9987167
36	imr90	0.097048706	0.528099	0.9831434
36	sknmc	0.248091792	0.2411794	0.827505
37	imr90	0.622853773	0.9258588	0.8053394
37	sknmc	0.622486223	0.05023015	0.05549676
38	imr90	0.702728852	0.1938564	0.1418484
38	sknmc	0.034704661	0.1262765	0.005463785
39	sknmc	0.018548005	0.9484238	0.04235603

Table B.5 GO Results. The top 30 GO results from the WebGestault server for both conserved DHS and haDHS target genes.

conserved DHS or haDHS?	term	raw P value	adjusted P value
consDHS	single-organism process	4.13E-16	1.59E-12
consDHS	single-multicellular organism process	1.22E-13	2.35E-10
consDHS	multicellular organismal process	7.04E-13	9.03E-10
consDHS	anatomical structure development	1.06E-12	1.02E-09
consDHS	anatomical structure morphogenesis	3.41E-11	2.19E-08
consDHS	system development	3.41E-11	2.19E-08
consDHS	signaling	5.72E-11	2.75E-08
consDHS	single organism signaling	5.72E-11	2.75E-08
consDHS	cell communication	1.52E-10	6.50E-08
consDHS	multicellular organismal development	3.42E-10	1.32E-07
consDHS	organ morphogenesis	4.11E-10	1.44E-07
consDHS	anatomical structure formation involved in morphogenesis	8.74E-10	2.80E-07
consDHS	developmental process	1.19E-09	3.52E-07
consDHS	cellular component organization or biogenesis	4.82E-09	1.32E-06
consDHS	cellular component organization	7.73E-09	1.98E-06
consDHS	cell development	2.21E-08	5.31E-06
consDHS	organ development	4.01E-08	9.07E-06
consDHS	neuron differentiation	7.19E-08	1.54E-05
consDHS	cellular component organization or biogenesis at cellular level	7.77E-08	1.57E-05
consDHS	cellular component organization at cellular level	8.84E-08	1.66E-05
consDHS	system process	9.08E-08	1.66E-05
consDHS	cellular component morphogenesis	1.25E-07	2.19E-05
consDHS	generation of neurons	2.73E-07	4.57E-05
consDHS	cell morphogenesis	3.01E-07	4.82E-05
consDHS	heart development	3.59E-07	5.30E-05
consDHS	neuron development	3.72E-07	5.30E-05
consDHS	regulation of multicellular organismal process	3.69E-07	5.30E-05
consDHS	signal transduction	4.14E-07	5.69E-05
consDHS	cellular response to stimulus	5.75E-07	7.37E-05
consDHS	regulation of multicellular organismal development	5.59E-07	7.37E-05
haDHS	anatomical structure development	0.0001	0.0023
haDHS	multicellular organismal process	4.67E-05	0.0023
haDHS	organ morphogenesis	5.88E-05	0.0023

haDHS	regulation of multicellular organismal development	0.0001	0.0023
haDHS	regulation of multicellular organismal process	0.0001	0.0023
haDHS	single-multicellular organism process	4.17E-05	0.0023
haDHS	single-organism process	2.55E-05	0.0023
haDHS	system development	6.74E-05	0.0023
haDHS	cell development	0.0002	0.0041
haDHS	multicellular organismal development	0.0003	0.0051
haDHS	organ development	0.0003	0.0051
haDHS	anatomical structure morphogenesis	0.0006	0.0074
haDHS	positive regulation of biological process	0.0006	0.0074
haDHS	positive regulation of cellular process	0.0006	0.0074
haDHS	system process	0.0005	0.0074
haDHS	developmental process	0.0007	0.0077
haDHS	enzyme linked receptor protein signaling pathway	0.0007	0.0077
haDHS	positive regulation of transcription from RNA polymerase II promoter	0.0009	0.0088
haDHS	regulation of developmental process	0.0009	0.0088
haDHS	embryo development	0.001	0.0089
haDHS	generation of neurons	0.001	0.0089
haDHS	negative regulation of biosynthetic process	0.0012	0.01
haDHS	negative regulation of gene expression	0.0014	0.01
haDHS	nervous system development	0.0014	0.01
haDHS	neuron differentiation	0.0014	0.01
haDHS	regulation of cell communication	0.0014	0.01
haDHS	negative regulation of cellular macromolecule biosynthetic process	0.0016	0.0103
haDHS	negative regulation of transcription, DNA-dependent	0.0016	0.0103
haDHS	neurogenesis	0.0016	0.0103
haDHS	cell communication	0.0018	0.0105

Table B.6 Second set of haDHS. This set of haDHS was derived only from DHS that are proximal to enhancer histone modifications from the same cell type.

chr	start	stop	lnL human acc	lnL primate cons	p value human acc	p value primate cons
chr3	11642000	11642350	24.30836	14.84102	0	0
chr15	34030420	34030570	14.42594	5.99008	0	0.00027
chr2	237776620	237776830	14.19327	12.19665	0	0
chr10	80423620	80423770	12.94383	14.679	0	0
chr9	2622740	2623190	12.86197	5.23353	0	0.00061
chr2	147187020	147187170	12.49586	14.82778	0	0
chr9	101866180	101866470	12.16231	15.22475	0	0
chr9	3968980	3969210	11.43568	8.60616	0	2.00E-05
chr1	116854460	116854950	11.36872	6.41801	0	0.00017
chr6	1287420	1287850	11.30355	13.72748	0	0
chr12	12735920	12736170	11.14543	4.84065	0	0.00093
chr10	33735040	33735190	10.92966	16.49732	0	0
chr15	37242100	37242250	10.82079	8.11558	0	3.00E-05
chr12	68286920	68287070	10.46006	12.79841	0	0
chr13	109926560	109926830	10.3645	4.40847	0	0.00149
chr14	37608260	37608410	10.29167	5.40972	0	5.00E-04
chrX	133703420	133703570	10.27414	4.93573	0	0.00084
chr7	134380100	134380290	10.12252	4.14886	0	0.00198
chr10	4320440	4320650	9.97598	7.87197	0	4.00E-05
chr1	5409720	5409870	9.93022	4.71312	0	0.00107
chr10	3628700	3628850	9.89106	5.85455	0	0.00031
chr3	193662120	193662350	9.86452	5.96396	0	0.00028
chr15	81822500	81822650	9.69953	4.24982	1.00E-05	0.00178
chr8	37378320	37378610	9.5391	28.49988	1.00E-05	0
chr6	11607660	11608030	9.52703	9.07266	1.00E-05	1.00E-05
chr2	238224280	238224470	9.49018	4.50125	1.00E-05	0.00135
chr12	97553320	97553530	8.96517	12.14936	1.00E-05	0
chr6	46012660	46012930	8.83843	5.83469	1.00E-05	0.00032
chr4	154347700	154347930	8.81391	6.35395	1.00E-05	0.00018
chr6	139184540	139184690	8.78172	7.72135	1.00E-05	4.00E-05
chr5	77287045	77287210	8.73636	5.2085	1.00E-05	0.00062
chr10	114345820	114346070	8.54588	16.04802	2.00E-05	0
chr11	31893720	31893930	8.48951	3.95819	2.00E-05	0.00245
chr2	135118340	135118530	8.24279	10.52949	2.00E-05	0
chr6	22363400	22363710	8.23948	7.89663	2.00E-05	4.00E-05
chr6	38428060	38428230	8.21109	6.94749	3.00E-05	1.00E-04
chr10	14144800	14144950	8.15501	4.87244	3.00E-05	9.00E-04
chr2	79152960	79153110	8.10795	3.93897	3.00E-05	0.0025
chr5	4570760	4570970	8.04697	15.27359	3.00E-05	0
chr11	48098900	48099110	8.02147	9.7893	3.00E-05	0
chr9	25311020	25311170	7.8399	13.01141	4.00E-05	0
chr20	60204500	60204670	7.82551	16.30437	4.00E-05	0
chr7	154412000	154412150	7.80664	6.115	4.00E-05	0.00024

chr2	53341860	53342010	7.7528	18.34111	4.00E-05	0
chr1	180313080	180313230	7.70431	5.53684	4.00E-05	0.00044
chr18	3262880	3263110	7.7027	3.99091	4.00E-05	0.00236
chr4	54809800	54810190	7.66395	10.14267	5.00E-05	0
chr3	67652720	67652970	7.64351	6.89885	5.00E-05	1.00E-04
chr4	13145100	13145330	7.63646	26.20278	5.00E-05	0
chr5	117901160	117901310	7.58411	5.07918	5.00E-05	0.00072
chr1	107291300	107291490	7.57172	5.25504	5.00E-05	0.00059
chr14	33180380	33180590	7.54552	5.22703	5.00E-05	0.00061
chr12	97931360	97931510	7.53316	20.75868	5.00E-05	0
chr1	2828020	2828170	7.47038	19.55583	6.00E-05	0
chr17	31866200	31866350	7.45474	4.29204	6.00E-05	0.0017
chr9	27599200	27599370	7.33141	9.89794	6.00E-05	0
chr8	92901340	92901490	7.30157	5.60384	7.00E-05	0.00041
chr7	26765820	26765970	7.30083	20.13501	7.00E-05	0
chr3	61268940	61269090	7.29822	6.86409	7.00E-05	0.00011
chr1	68925000	68925150	7.25755	5.50758	7.00E-05	0.00045
chr4	130202780	130202930	7.24356	7.23708	7.00E-05	7.00E-05
chr10	78874325	78874475	7.19569	6.48941	7.00E-05	0.00016
chr6	1626280	1626550	7.16484	5.87361	8.00E-05	3.00E-04
chr1	19688220	19688655	7.15007	12.49442	8.00E-05	0
chr11	121666020	121666190	7.12819	22.61846	8.00E-05	0
chr5	88753800	88754130	7.12002	22.37645	8.00E-05	0
chr18	42879020	42879190	7.07281	7.2288	8.00E-05	7.00E-05
chr10	77357520	77357810	7.04549	39.38224	9.00E-05	0
chr3	2141600	2141930	6.9342	5.4099	1.00E-04	5.00E-04
chr5	5374200	5374430	6.89755	17.96218	1.00E-04	0
chr2	84461280	84461490	6.84945	10.32984	0.00011	0
chr5	91951880	91952030	6.84944	6.46503	0.00011	0.00016
chr4	17022180	17022330	6.83512	12.01809	0.00011	0
chr5	40502940	40503090	6.80483	11.93433	0.00011	0
chr8	109280460	109280650	6.78643	6.18171	0.00011	0.00022
chr20	19255320	19255530	6.78008	7.59925	0.00012	5.00E-05
chr3	114304940	114305350	6.77086	25.40282	0.00012	0
chr8	110650400	110650570	6.77071	8.42827	0.00012	2.00E-05
chr4	105345800	105346050	6.75726	27.65063	0.00012	0
chr9	3646960	3647110	6.75678	11.67129	0.00012	0
chr22	27554060	27554210	6.74166	15.81295	0.00012	0
chr8	62006300	62006610	6.7277	8.86859	0.00012	1.00E-05
chr14	60999240	60999450	6.7132	4.3934	0.00012	0.00152
chr5	168313300	168313450	6.67911	8.78195	0.00013	1.00E-05
chr10	128956040	128956250	6.66906	10.35667	0.00013	0
chr11	20047000	20047310	6.66147	6.95847	0.00013	1.00E-04
chr7	152999080	152999310	6.65303	4.93916	0.00013	0.00084
chr12	80081620	80081995	6.64893	10.56509	0.00013	0
chr14	51750080	51750370	6.63283	4.2104	0.00014	0.00185
chr10	108787280	108787430	6.62117	12.31351	0.00014	0
chr17	77967320	77967690	6.61526	11.2166	0.00014	0
chr18	38919080	38919230	6.60155	5.45349	0.00014	0.00048
chr10	34264660	34264810	6.59703	11.51274	0.00014	0

chr9	14516800	14516950	6.58451	20.28333	0.00014	0
chr2	111880700	111881130	6.58173	5.41499	0.00014	5.00E-04
chr3	34354240	34354450	6.55512	18.17932	0.00015	0
chr4	58631220	58631370	6.54384	5.87925	0.00015	3.00E-04
chr1	232060000	232060250	6.54048	14.16113	0.00015	0
chr18	53359420	53359710	6.51429	23.96639	0.00015	0
chr8	25372580	25372730	6.50688	18.63481	0.00015	0
chr14	85124260	85124410	6.5025	10.81401	0.00016	0
chr13	107592900	107593050	6.49239	9.56453	0.00016	1.00E-05
chr2	23777040	23777310	6.48864	21.45281	0.00016	0
chr1	223060080	223060230	6.46392	4.42215	0.00016	0.00147
chr3	56002380	56002590	6.45463	4.10994	0.00016	0.00207
chr5	73943020	73943270	6.45011	19.03901	0.00016	0
chr8	108080960	108081190	6.43384	7.83087	0.00017	4.00E-05
chr14	37374900	37375090	6.43187	11.66182	0.00017	0
chrX	31850460	31850610	6.42825	9.64093	0.00017	1.00E-05
chr2	215442560	215442790	6.39627	12.25992	0.00017	0
chr3	44151640	44151790	6.36092	14.33487	0.00018	0
chr2	60729340	60729490	6.3577	7.02964	0.00018	9.00E-05
chr5	57423280	57423670	6.35241	21.17436	0.00018	0
chr13	39224700	39224850	6.30636	11.52159	0.00019	0
chr20	20077260	20077410	6.29055	16.65363	0.00019	0
chr3	21314860	21315070	6.26529	5.56443	2.00E-04	0.00042
chr5	167014840	167015075	6.23073	21.91232	0.00021	0
chr2	152631660	152631890	6.20418	4.08674	0.00021	0.00213
chr4	182271860	182272010	6.19343	15.73423	0.00022	0
chr10	25101280	25101430	6.18006	10.03361	0.00022	0
chr5	164651480	164651630	6.1759	4.06152	0.00022	0.00219
chr8	19613860	19614390	6.16883	13.30211	0.00022	0
chr2	31298360	31298570	6.14783	7.44714	0.00023	6.00E-05
chr4	85395340	85395490	6.14722	17.7362	0.00023	0
chr2	119067540	119068370	6.13463	43.83978	0.00023	0
chr4	17070740	17070890	6.10607	4.04955	0.00024	0.00221
chr16	82816280	82816690	6.09751	12.49037	0.00024	0
chr2	195351180	195351330	6.09747	5.57971	0.00024	0.00042
chr11	19902800	19902950	6.08322	4.56041	0.00024	0.00126
chr4	179441860	179442010	6.06725	4.80797	0.00025	0.00096
chr4	93686580	93686730	6.03487	6.7327	0.00026	0.00012
chr20	9237220	9237370	6.03436	8.76524	0.00026	1.00E-05
chr12	43150860	43151235	6.00962	16.30504	0.00026	0
chr11	14075760	14076030	6.0095	8.22563	0.00026	2.00E-05
chr7	127176020	127176170	6.00681	17.14928	0.00026	0
chr5	4848000	4848150	6.00183	21.56987	0.00027	0
chr1	97463380	97463530	5.99481	9.60902	0.00027	1.00E-05
chr2	21676640	21676870	5.98741	14.26821	0.00027	0
chr2	12061220	12061450	5.98212	13.226	0.00027	0
chr14	49440220	49440470	5.96738	4.35346	0.00028	0.00159
chrX	139438160	139438310	5.96515	14.43431	0.00028	0
chr13	81415700	81415850	5.94231	19.85944	0.00028	0
chr13	34611920	34612250	5.94077	12.72445	0.00028	0

chr10	125538300	125538450	5.9392	9.9867	0.00028	0
chr2	10111320	10111530	5.92847	5.73903	0.00029	0.00035
chr11	15695040	15695430	5.90422	8.38267	0.00029	2.00E-05
chr4	72188260	72188410	5.8993	5.48547	3.00E-04	0.00046
chr1	72750080	72750290	5.88993	13.2161	3.00E-04	0
chr5	81518500	81518990	5.88905	16.79786	3.00E-04	0
chr8	38238020	38238350	5.88312	4.80881	3.00E-04	0.00096
chr2	205569560	205569710	5.87761	19.14996	3.00E-04	0
chr16	65043720	65043870	5.8631	8.40992	0.00031	2.00E-05
chr18	72842940	72843090	5.86292	7.54175	0.00031	5.00E-05
chr6	73208800	73208950	5.85589	18.35164	0.00031	0
chr6	170031565	170031715	5.84437	5.5457	0.00031	0.00043
chr20	11176580	11176890	5.83295	7.80525	0.00032	4.00E-05
chr6	77595580	77595790	5.82274	18.0739	0.00032	0
chr7	17572760	17573050	5.81946	5.6345	0.00032	0.00039

Appendix C

SUPPLEMENT FOR CHAPTER 3: ARCHAIC HOMININ ADMIXTURE FACILITATED ADAPTATION TO OUT-OF-AFRICA ENVIRONMENTS

This appendix contains material from Gittelman et al

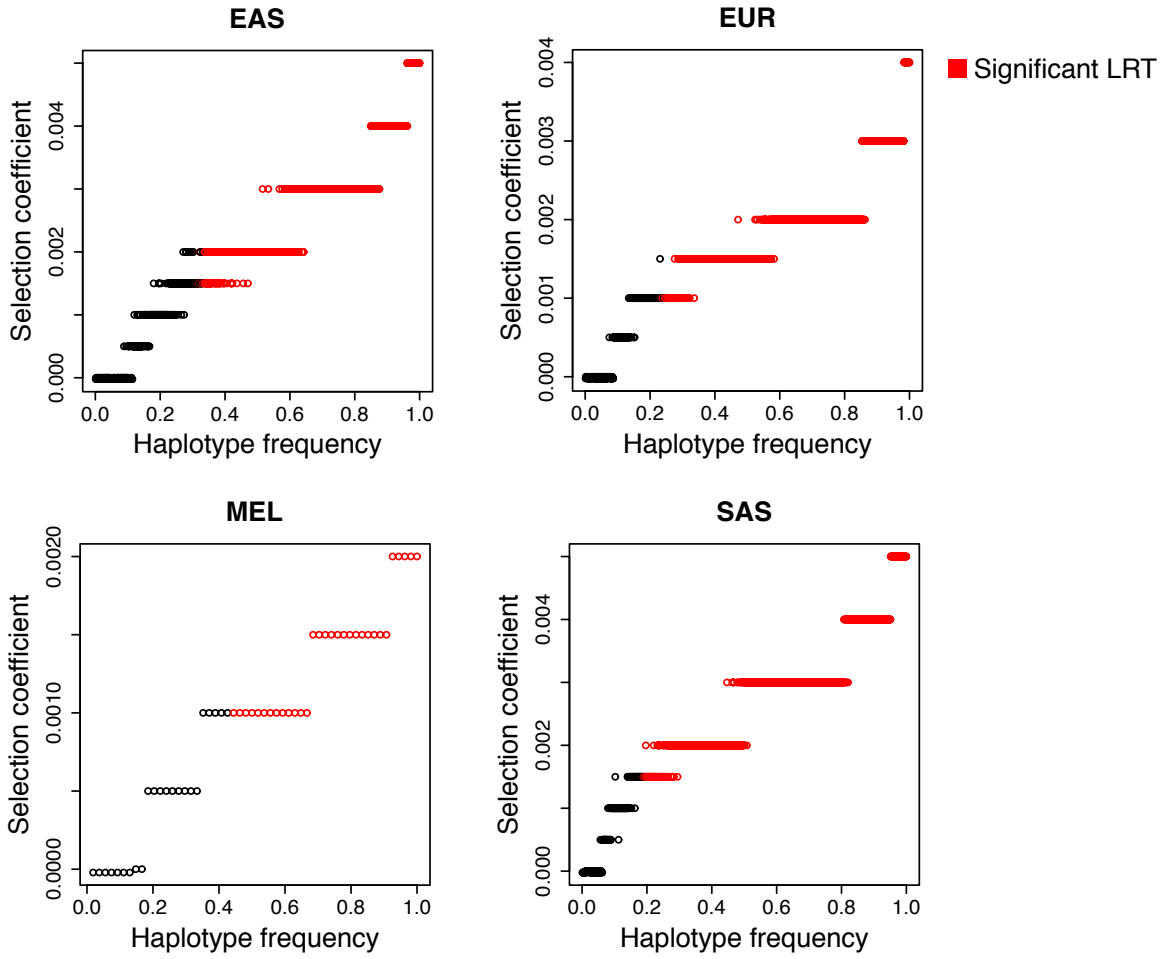

Fig. C.1. Estimating the strength of positive selection. The maximum likelihood selection coefficient is depicted for all frequencies of introgressed haplotypes in each of the four populations studied. Frequencies in which the maximum likelihood selection coefficient is a significantly better fit than the mildly deleterious model are shown in red.

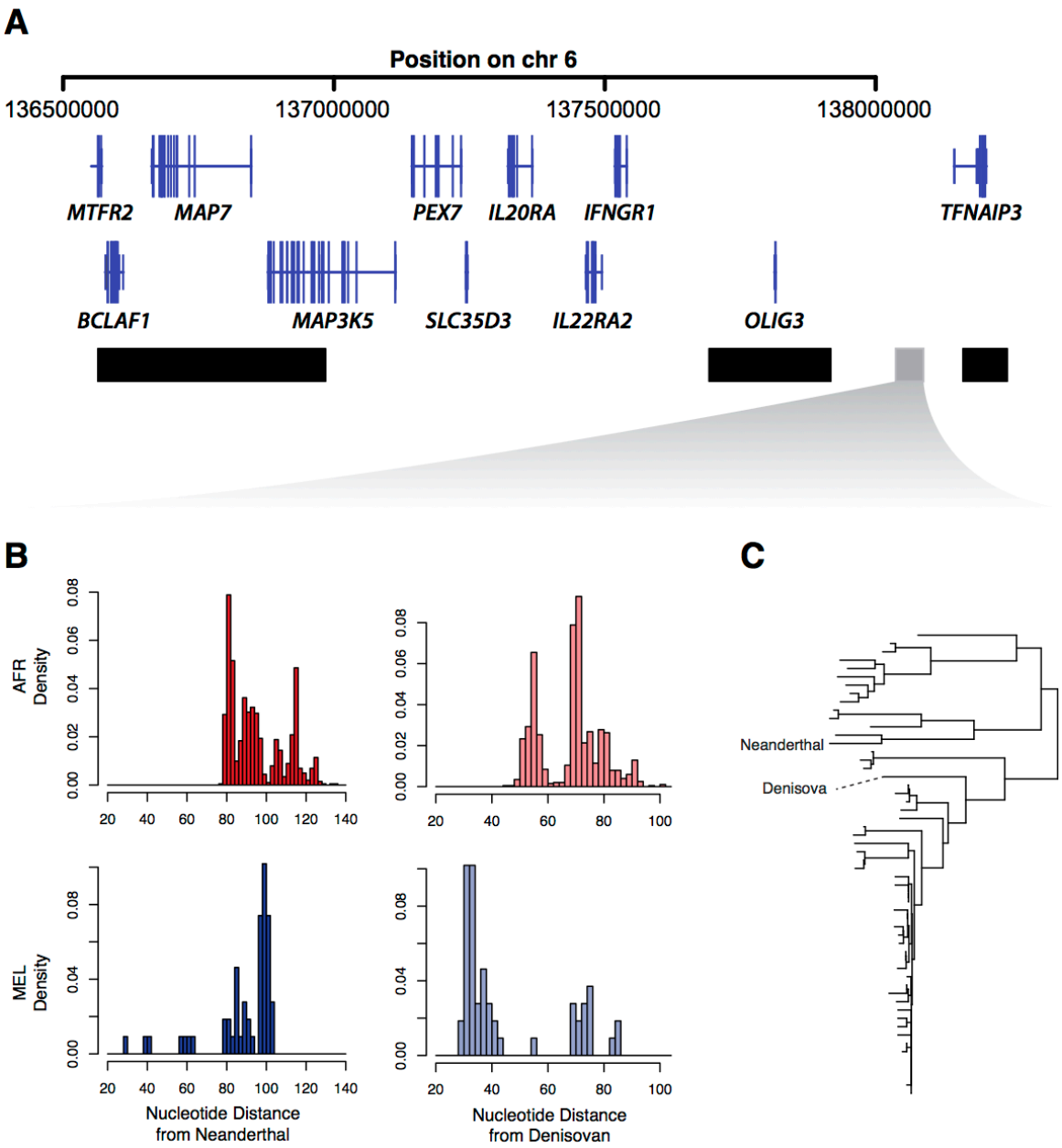

Figure C.2. A region with both Neandertal and Denisovan ancestry. A. A schematic of the region harboring 4 high frequency Melanesian regions that segregate both Neandertal and Denisovan sequence. The bars indicate the distinct regions, and the grey bar indicates the region that is further characterized in the next panels. **B.** The distribution of absolute genetic distances from Neandertal (left column) and Denisovan (right column) are shown for Africans (top row) and Melanesians (bottom row). **C.** A neighbor-joining tree constructed with sequence from Melanesians, Densiovan, and Neandertal is shown.

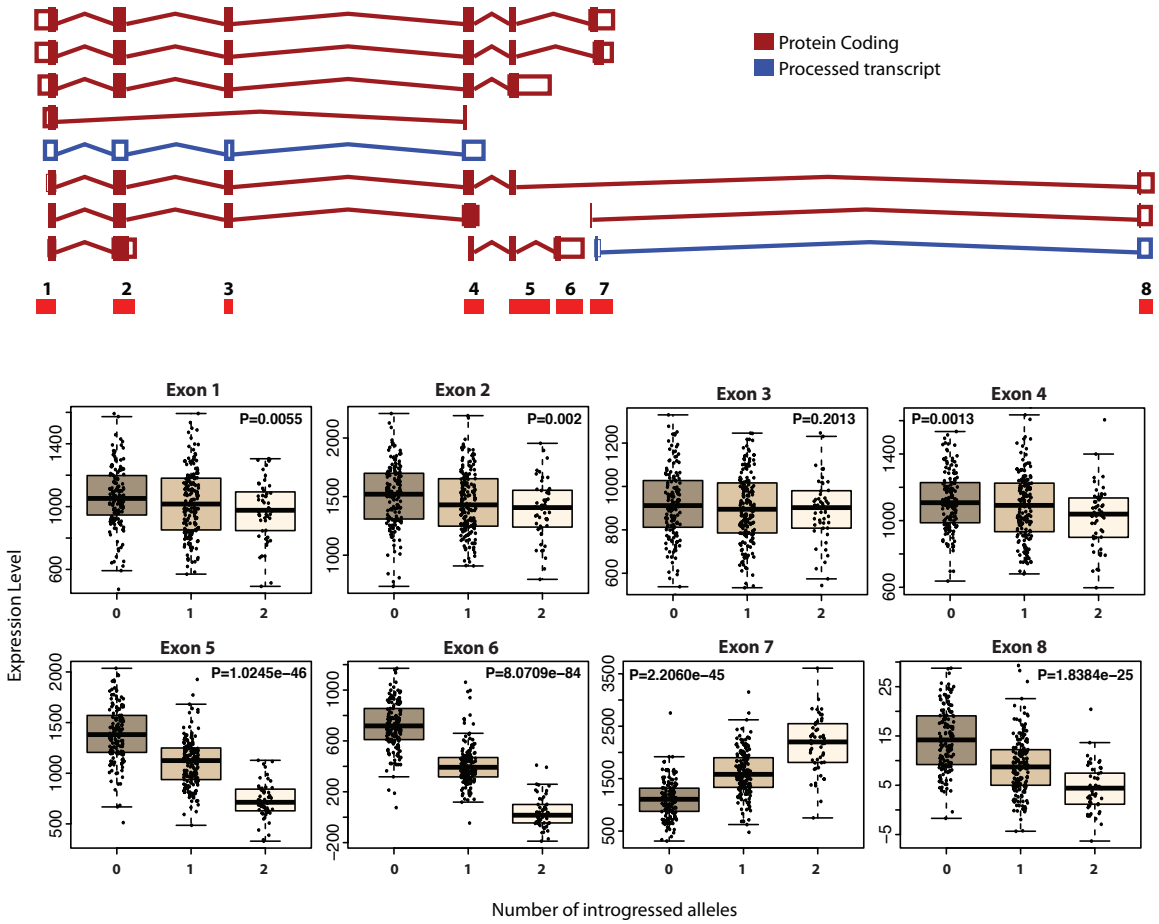

Figure C.3. Neandertal haplotype association with *OAS1* exon expression. A. Gene expression for each exon of *OAS1* is shown stratified by the number of Neandertal alleles each sample has. Data is from Geuvadis project LCLs. **B.** Schematics of observed *OAS1* isoforms are shown. Boxes indicate which exons are included in each transcript.

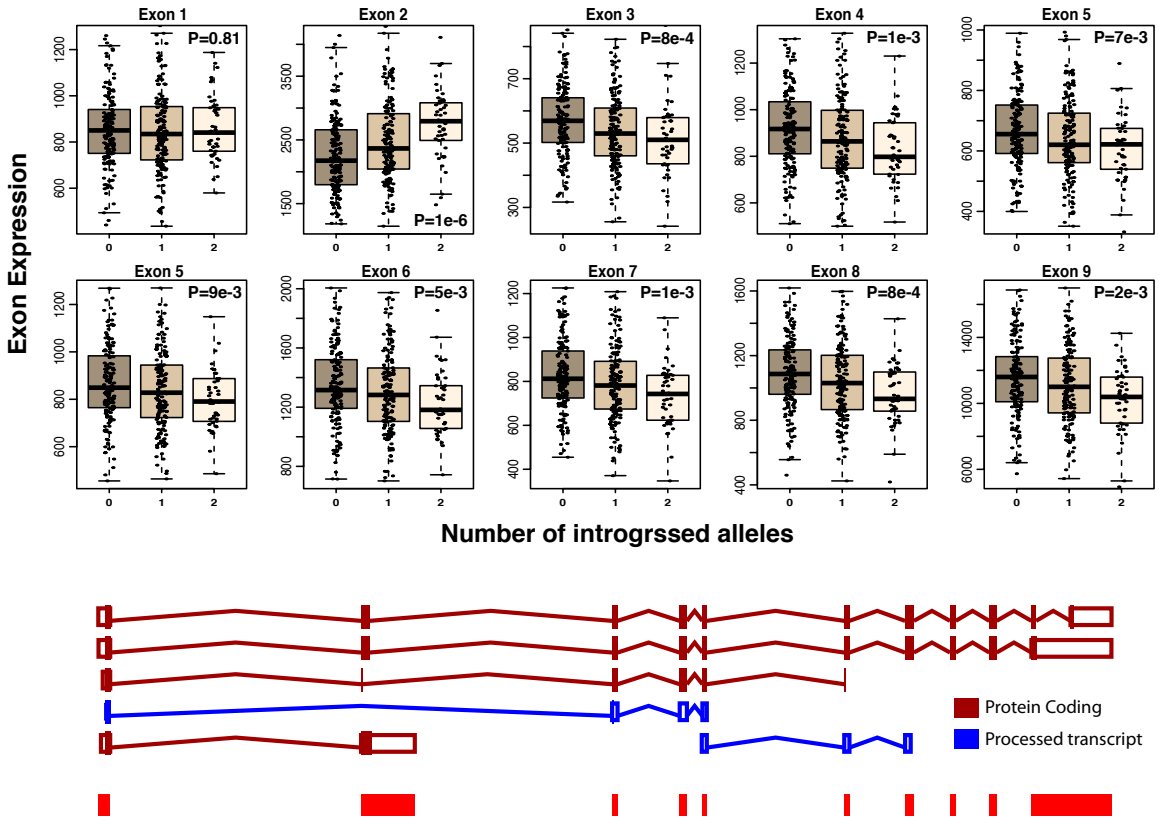

Fig. C.4. Neandertal haplotype association with *OAS2* exon expression. A. Gene expression for each exon of *OAS2* is shown stratified by the number of Neandertal alleles each sample has. Data is from Geuvadis project LCLs. **B.** Schematics of observed *OAS2* isoforms are shown. Boxes indicate which exons are included in each transcript.

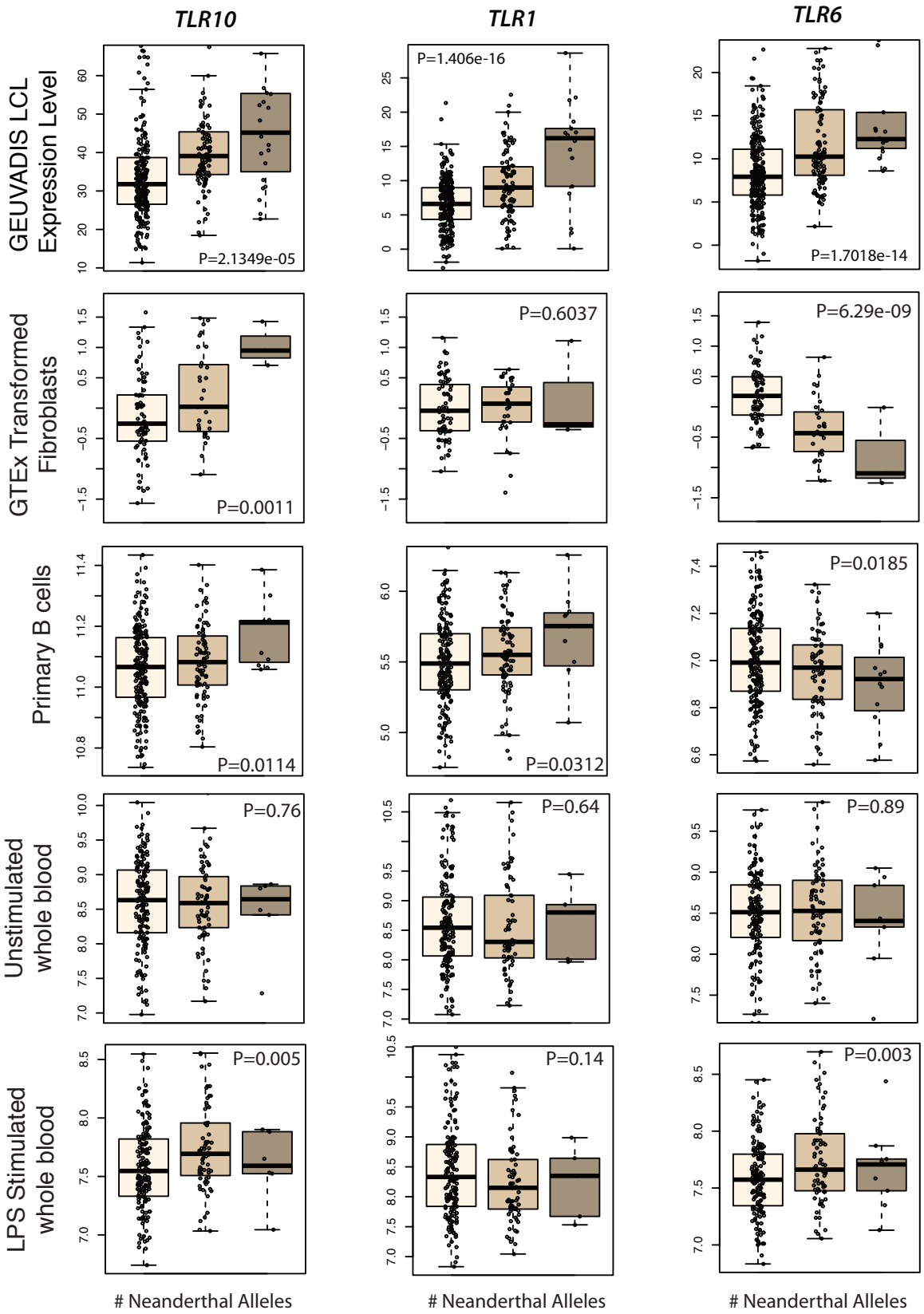
TLR10
TLR1
TLR6
GEUVADIS LCL Expression Level
GTEx Transformed Fibroblasts
Primary B cells
Unstimulated whole blood
LPS Stimulated whole blood
P=2.1349e-05
P=1.406e-16
P=1.7018e-14
P=0.0011
P=0.6037
P=6.29e-09
P=0.0114
P=0.0312
P=0.0185
P=0.76
P=0.64
P=0.89
P=0.005
P=0.14
P=0.003
# Neanderthal Alleles
# Neanderthal Alleles
# Neanderthal Alleles

Fig. C.5. Neandertal haplotype association with *TLR1/6/10* expression in multiple cell types. Gene expression for *TLR10*/1/6 is shown stratified by the number of Neandertal alleles each sample has. Rows consist of data from a single cell type. P values are indicated for each plot.

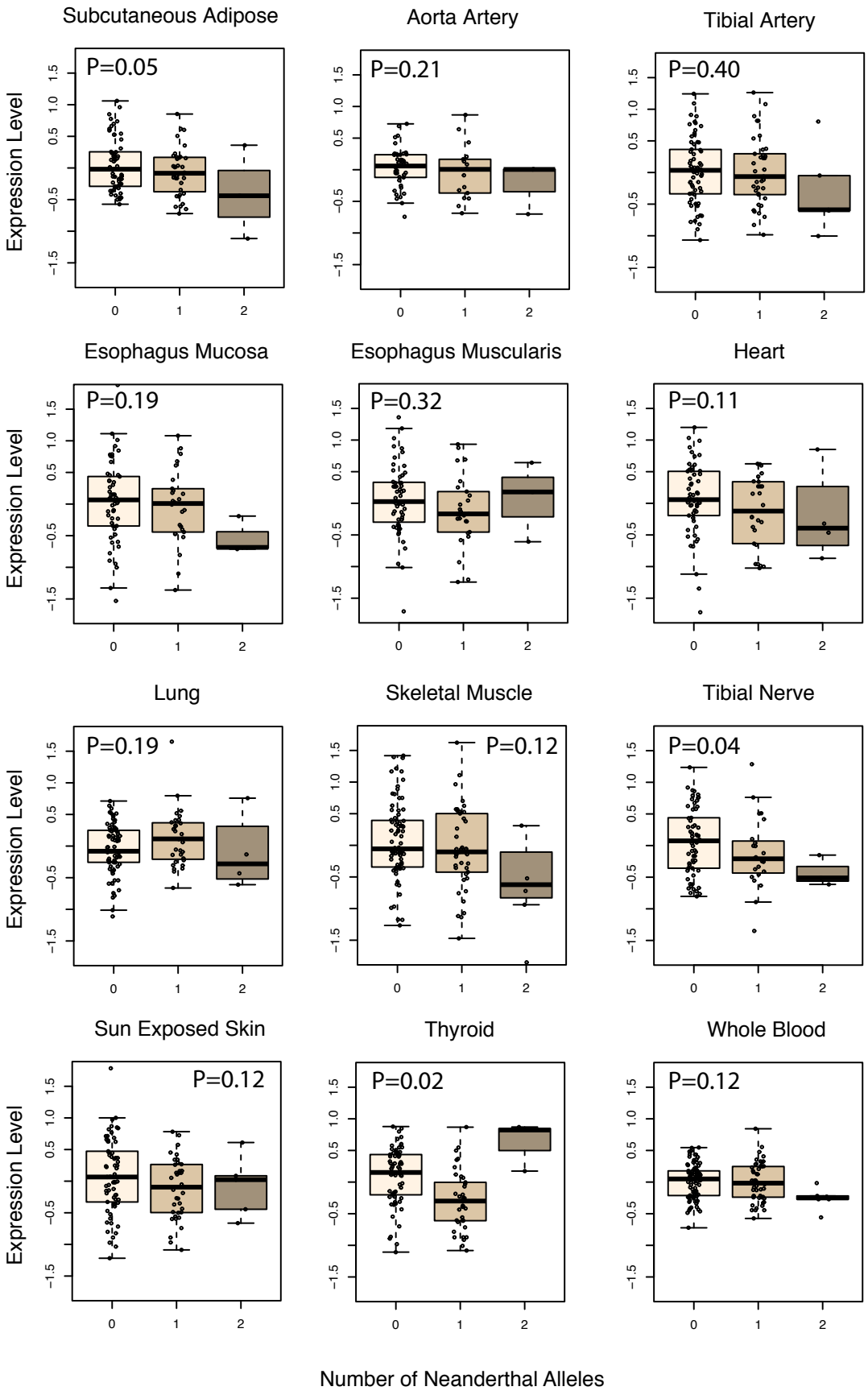
Subcutaneous Adipose
P=0.05
Aorta Artery
P=0.21
Tibial Artery
P=0.40
Esophagus Mucosa
P=0.19
Esophagus Muscularis
P=0.32
Heart
P=0.11
Lung
P=0.19
Skeletal Muscle
P=0.12
Tibial Nerve
P=0.04
Sun Exposed Skin
P=0.12
Thyroid
P=0.02
Whole Blood
P=0.12
Expression Level
Number of Neanderthal Alleles

Fig. C.6. Neandertal haplotype association with *TLR6* expression in all additional GTEx cell types. Gene expression for *TLR*6 is shown in additional GTEx cell types that didn't have a significant association (FDR $\leq$ 0.05). Data is stratified by the number of Neandertal alleles each sample has. Each plot P values are indicated for each plot.

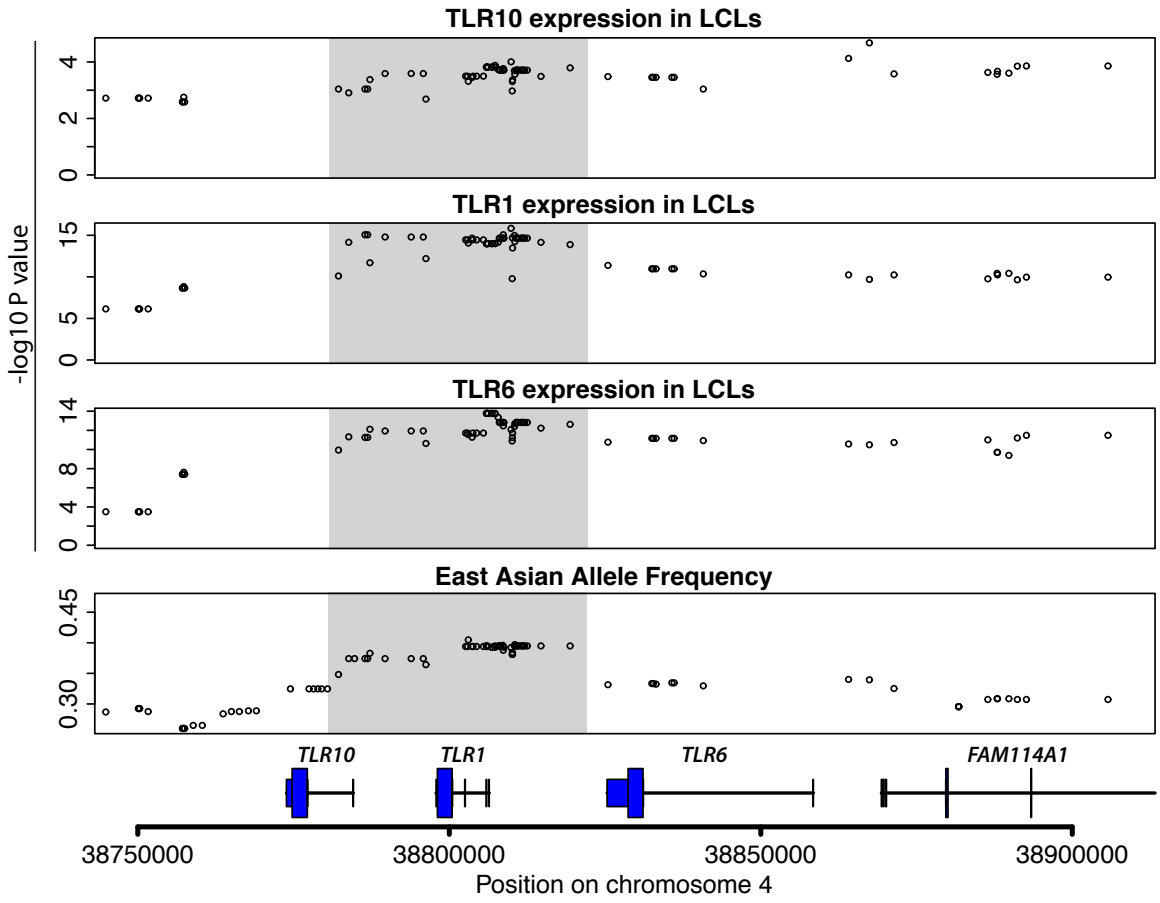

Fig. C.7. Fine-mapping the TLR Neandertal haplotype *TLR1/6/10* eQTL. The –log10(p value) of association with gene expression for *TLR1/6/10* is shown for all variants on the TLR haplotype. On the bottom, allele frequency in East Asians of each variant is shown. They grey box highlights the region of maximal p value and allele frequency.

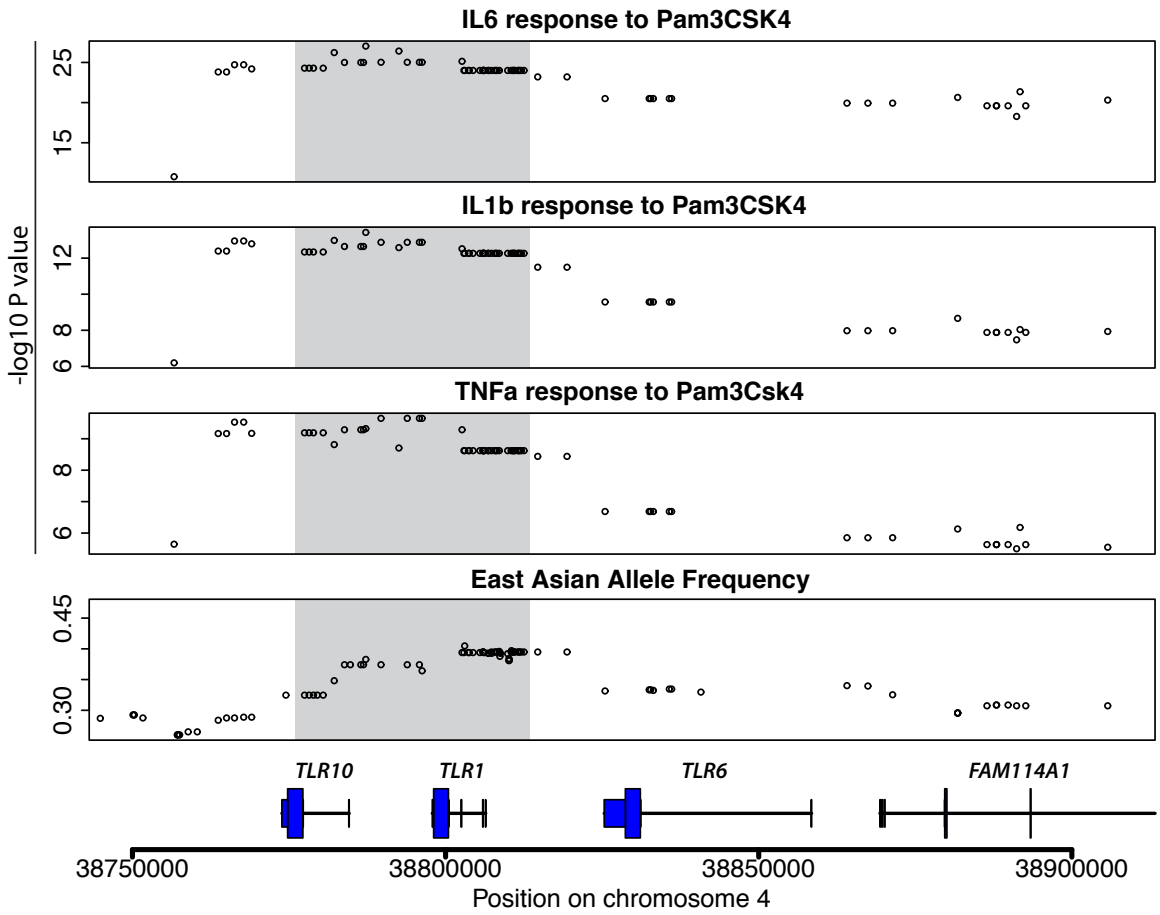

Fig. C.8. Fine-mapping the TLR Neandertal haplotype PAM3CSK4 response association. The –log10(p value) of association with interleukin response to *TLR1* stimulation by PAM3CSK4 is shown for all variants on the TLR haplotype. On the bottom, allele frequency in East Asians of each variant is shown. They grey box highlights the region of maximal p value and allele frequency.

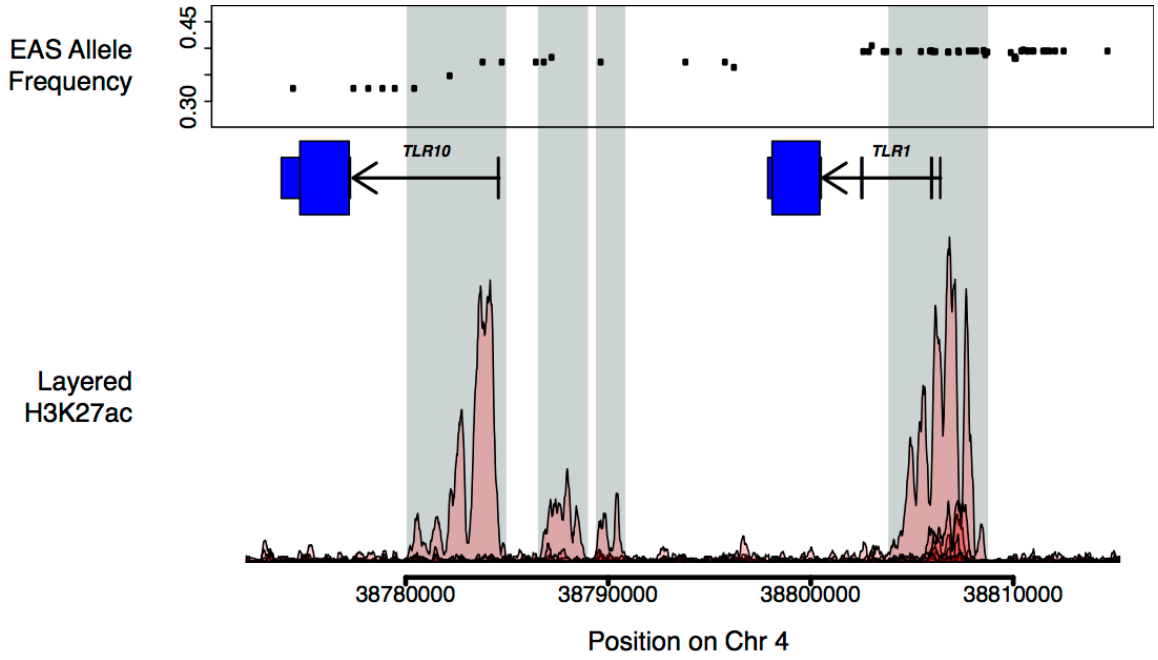

Fig. C.9. Regulatory regions within the TLR Neandertal haplotype. The region within the *TLR1/6/10* haplotype of maximal trait associations and allele frequency in East Asians is shown. Pink peaks represent H3K27ac signal from the Layered H3k27ac track in the UCSC genome browser. Grey bars indicate regions of high H3K27ac signal.

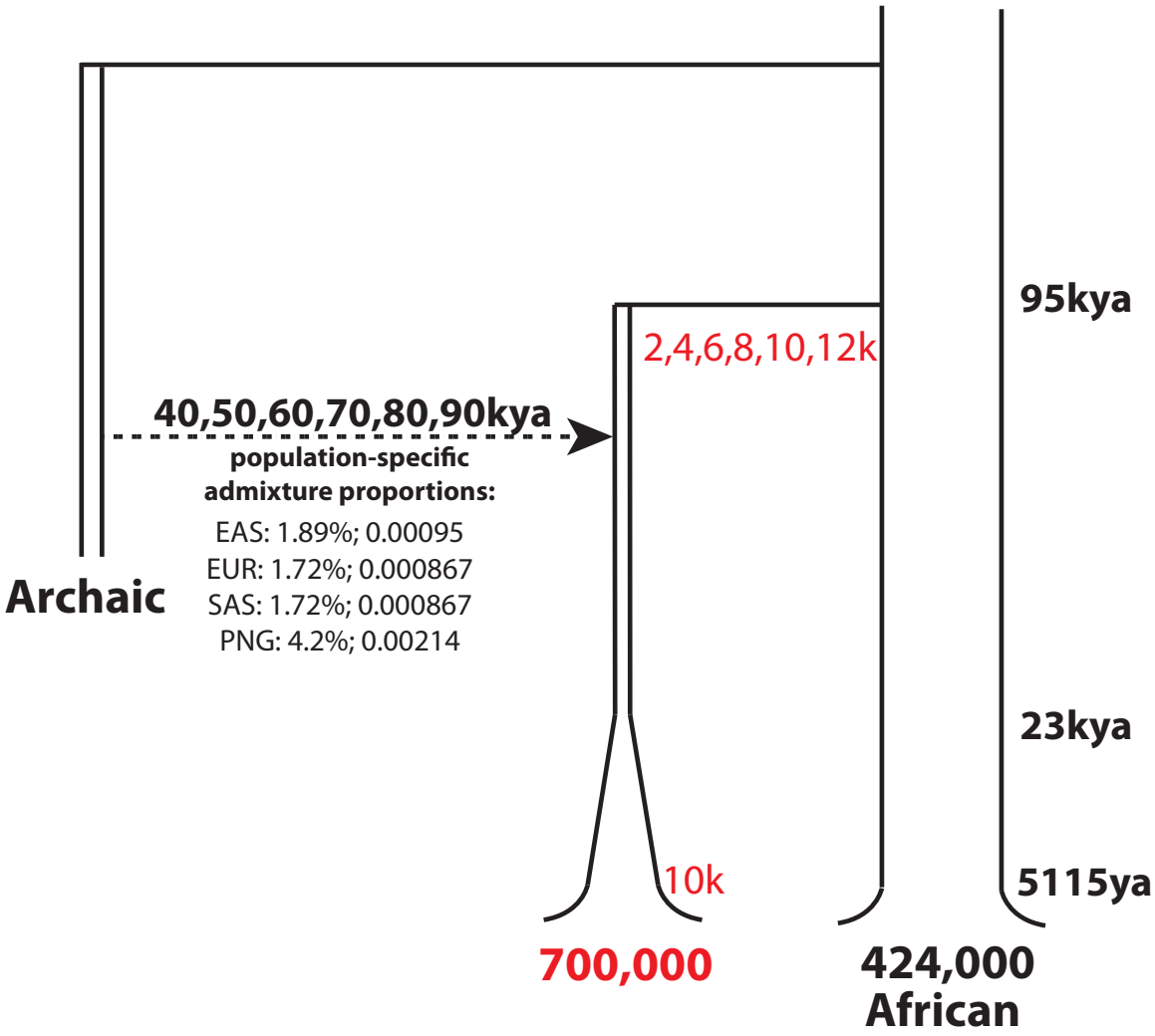

Fig. C.10. Demographic model for simulations. The base demographic model for the simulation framework is depicted with the grid of parameters used. Text in red indicates effective population sizes in the introgressed population.

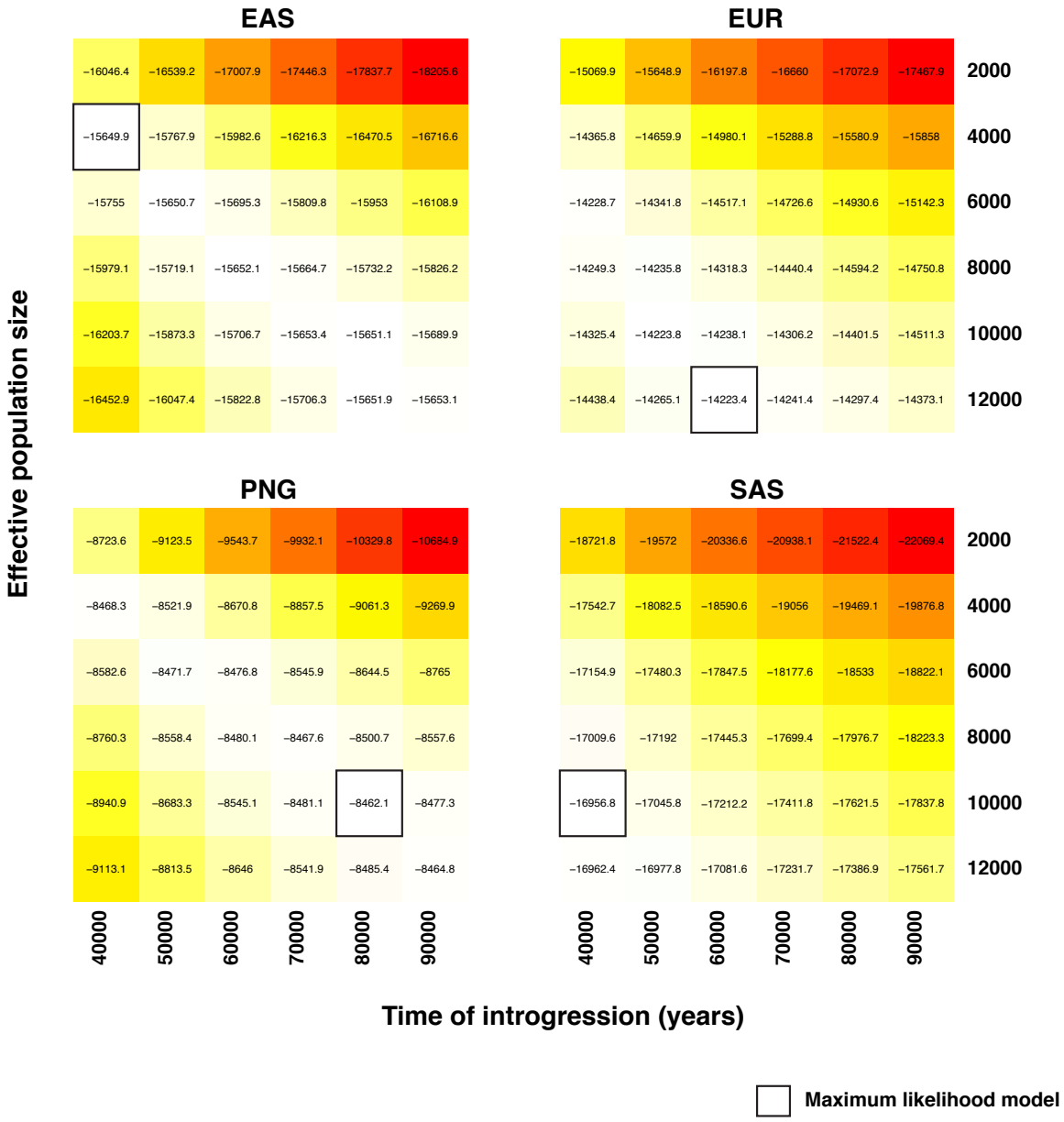

Fig. C.11. Demographic model likelihoods. Likelihoods for each of the 36 demographic models tested is depicted for each population. Color corresponds to the likelihoods, with more likely models shown in lighter colors. The maximum likelihood model used in subsequent simulations and FDR is highlighted with a black box.

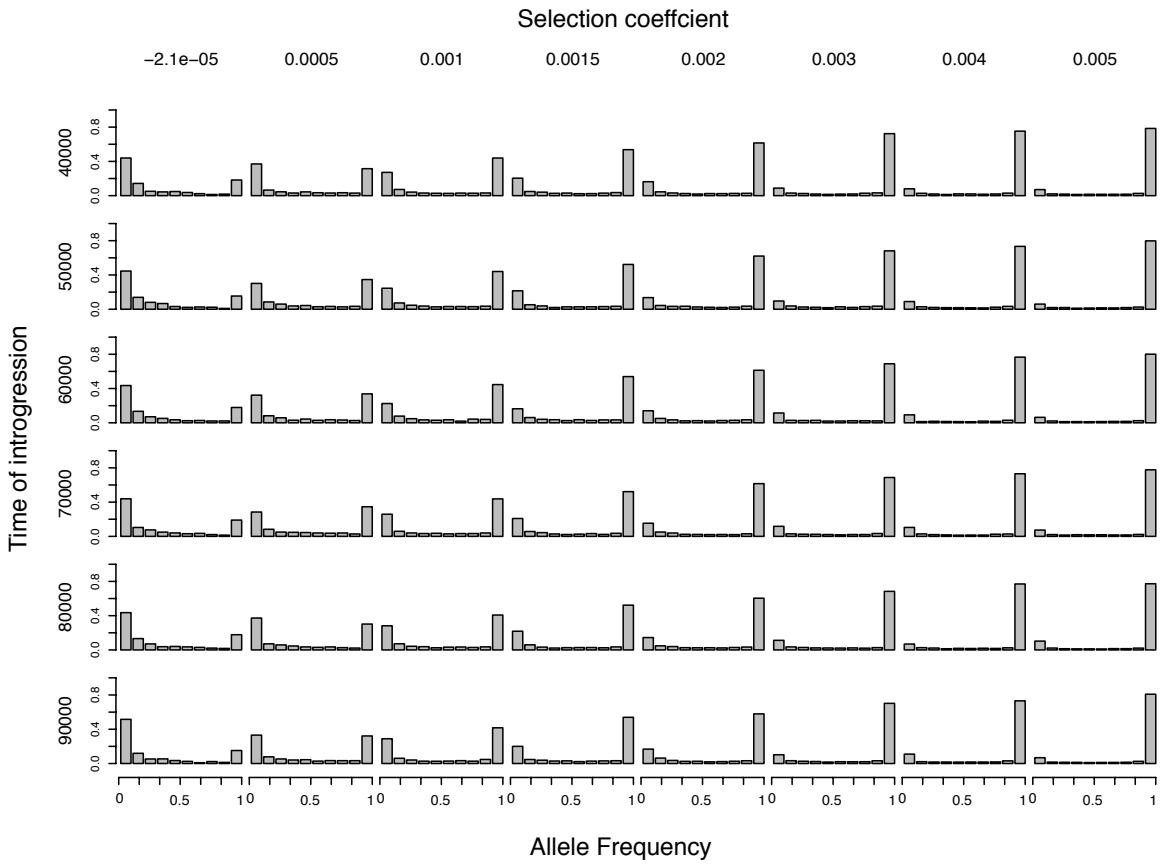

Fig. C.12. Frequency distributions of the selected allele at the time of introgression. Frequency distributions at the time of introgression for alleles that arose in the archaic lineage are depicted for different selection coefficients.

Table C.1. Adaptively introgressed loci. This table contains information about the regions identified as adaptively introgressed at the 50% FDR in each population.

EAS			
Chr	**Start**	**Stop**	**frequency**
chr1	164560944	164595665	0.42659
chr1	208552587	208627757	0.45933
chr1	212347149	212805881	0.38492
chr1	215932956	215967857	0.38393
chr1	232529815	232673154	0.45933
chr10	7006868	7043252	0.43155
chr10	7074204	7156076	0.44444
chr10	21257682	21276783	0.37698
chr11	98938409	99279895	0.42113
chr11	120097174	120178414	0.57143
chr12	52759502	52805755	0.474205
chr12	52931746	52943144	0.499505
chr12	114188041	114244795	0.39137
chr12	129892165	129931969	0.40873
chr13	108420509	108467914	0.41766
chr14	58320130	58444752	0.3869
chr15	28160571	28177529	0.62202
chr16	76994846	77221577	0.3874
chr18	51521835	51531031	0.51587
chr19	56656315	56693894	0.38393
chr2	150040223	150049794	0.56746
chr2	239724967	239763994	0.39385
chr20	62067608	62095558	0.47917
chr20	62195671	62229244	0.54167
chr22	20755526	20930065	0.38095
chr3	33036764	33084884	0.435515
chr3	50200348	50414404	0.61905
chr3	191259248	191313260	0.46329
chr4	38309802	38484762	0.41766
chr4	38570177	38905731	0.39484
chr4	167098529	167559248	0.4375
chr5	39393624	39437163	0.40476
chr6	39093550	39129242	0.41071
chr6	167189068	167234330	0.375
chr8	5723060	5774871	0.51835
chr8	103601704	103693753	0.42163
chr9	90765291	90890583	0.4504
chr9	96895107	96994063	0.413195

chr9	112790035	112855172	0.41171
chr9	112875127	112965275	0.44643
chr9	113094933	113236127	0.40873
chr9	115788475	115798998	0.391865
chr9	129327668	129409062	0.40377
chr9	129555992	129593549	0.38889

EUR

Chr	Start	Stop	Frequency
chr1	34249159	34371256	0.27833
chr1	169018901	169078990	0.332505
chr1	170233051	170543775	0.36282
chr1	193783042	193938169	0.275845
chr1	216737860	216894402	0.27336
chr10	6457665	6503894	0.284295
chr10	133872132	133920011	0.32505
chr11	11569019	11590971	0.35089
chr12	52931746	52943144	0.50596
chr12	54195032	54235868	0.47316
chr12	90768548	90784728	0.34493
chr12	113350795	113474299	0.36382
chr12	133337351	133449837	0.27932
chr13	20726920	20733403	0.37177
chr15	85535260	85694548	0.294235
chr15	85867872	86013059	0.384195
chr15	86093925	86319706	0.297715
chr16	77979107	78273131	0.28032
chr18	1945193	1965174	0.36879
chr18	51521835	51531031	0.30219
chr18	60106280	60240333	0.27833
chr19	33554150	33700078	0.64811
chr19	33724557	33760114	0.40457
chr2	68348010	68524491	0.36282
chr2	150040223	150049794	0.38966
chr2	154829586	155011132	0.30716
chr2	159672625	160058450	0.29622
chr2	238850458	239036423	0.47614
chr20	4130254	4160952	0.33499
chr20	15074448	15145194	0.32803
chr3	2182146	2201147	0.30616
chr3	11021892	11039202	0.39463
chr3	188439935	188448320	0.334
chr4	181149180	181185720	0.31412

chr5	24572260	24583085	0.53579
chr5	101635198	101643293	0.29324
chr6	44130869	44141717	0.389165
chr6	52139615	52183126	0.44732
chr7	13573003	13631076	0.347415
chr7	136607897	136620040	0.36581
chr8	13773103	13891785	0.37177
chr8	13955787	14164646	0.3499
chr8	14175805	14394081	0.29622
chr9	16720121	16833797	0.6829
chr9	112790035	112855172	0.29026

MEL

Chr	Start	Stop	Frequency
chr1	49512225	49545650	0.59259
chr1	86693866	86938383	0.55556
chr1	89641240	89665347	0.57407
chr1	208760708	208922357	0.77778
chr1	208924350	208995044	0.685185
chr1	209093258	209107029	0.53704
chr18	60056542	60240333	0.62963
chr18	70967371	71024040	0.55556
chr18	71049647	71092638	0.59259
chr18	71156813	71215957	0.7037
chr2	3815476	3874363	0.59259
chr2	162998940	163348463	0.55556
chr2	241983021	242393658	0.62963
chr2	242502039	242604376	0.55556
chr22	20755526	20788590	0.62037
chr5	55987313	56264124	0.57407
chr5	56278198	56293149	0.703705
chr6	136563964	136984706	0.55556
chr6	137692158	137917432	0.59259
chr6	138037107	138088266	0.77778
chr6	138160925	138243220	0.59259
chr9	114722659	114766818	0.53704

SAS

Chr	Start	Stop	Frequency
chr1	33643982	33751945	0.27096
chr1	57259057	57270607	0.26074
chr10	121365413	121873278	0.30266

chr10	125687979	125790354	0.24847
chr11	11494384	11530287	0.24438
chr11	11569019	11590971	0.41002
chr11	107092780	107257389	0.39264
chr12	52583916	52588644	0.28681
chr12	52931746	52943144	0.26687
chr12	90768548	90784728	0.25153
chr12	113341597	113474299	0.28425
chr15	28160571	28177529	0.28528
chr16	2010477	2058040	0.27198
chr16	76904775	76985391	0.29448
chr18	1945193	1965174	0.24335
chr19	33554150	33700078	0.26176
chr2	105877716	106020468	0.256645
chr2	150040223	150049794	0.5317
chr22	39290233	39302174	0.35123
chr3	21900762	22100476	0.260225
chr3	45864731	46474902	0.29755
chr3	100514847	100546733	0.27914
chr3	188415647	188448320	0.27914
chr4	37337745	37367110	0.24131
chr4	37501255	37534576	0.28732
chr5	24572260	24583085	0.47546
chr5	55987313	56335027	0.2955
chr5	57742572	57788035	0.24642
chr6	44130869	44141717	0.25511
chr6	52138225	52183126	0.41258
chr6	137308579	137541230	0.26278
chr8	13955787	14023668	0.34305
chr9	32135872	32174572	0.2546
chr9	92683272	92786683	0.31902
chr9	96895107	97025979	0.34458
chr9	112790035	112855172	0.34049
chr9	114397251	114572327	0.24131
chr9	119140656	119464644	0.3318

Table C.2. Gene Ontology enrichments in genes near adaptively introgressed loci. In this table I present the top 10 gene ontology terms for genes surrounding adaptively introgressed loci.

Category	ID	Adjusted P value
cellular response to cytokine stimulus	GO:0071345	adjP=0.0008
defense response	GO:0006952	adjP=0.0008
response to cytokine stimulus	GO:0034097	adjP=0.0017
cytokine-mediated signaling pathway	GO:0019221	adjP=0.0055
inflammatory response	GO:0006954	adjP=0.0055
innate immune response	GO:0045087	adjP=0.0183
positive regulation of defense response	GO:0031349	adjP=0.0196
cellular response to organic substance	GO:0071310	adjP=0.0518
response to virus	GO:0009615	adjP=0.0518
chemotaxis	GO:0006935	adjP=0.0708

Table C.3. Introgressed OAS coding variants. This table lists protein-coding variants that segregate between the introgressed and modern human haplotypes at the *OAS1/2/3* locus.

chromosome	base	RS number	annotation	gene
chr12	113357193	rs10774671	intron,splice-3	OAS1
chr12	113357209	rs1131476	intron,missense	OAS1
chr12	113357442	rs2660	missense,untranslated-3	OAS1
chr12	113376388	rs1859330	missense	OAS3
chr12	113425154	rs1293767	intron,missense	OAS2
chr12	113448288	rs15895	stop-loss,untranslated-3	OAS2