The *cis*-regulatory effects of modern human-specific variants

Authors

Carly V. Weiss[1,*], Lana Harshman[2,3,*], Fumitaka Inoue[2,3,4], Hunter B. Fraser[1], Dmitri A. Petrov[1,†], Nadav Ahituv[2,3,†], David Gokhman[1,†]

[1] Department of Biology, Stanford University, Stanford, CA 94305, USA

[2] Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, 94158, USA.

[3] Institute for Human Genetics, University of California San Francisco, San Francisco, CA, 94158, USA.

[4] Present address: Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto, 606-8501, Japan

[*] Equal contributors

[†] Corresponding authors

Abstract

The Neanderthal and Denisovan genomes enabled the discovery of sequences that differ between modern and archaic humans, the majority of which are noncoding. However, our understanding of the regulatory consequences of these differences remains limited, in part due to the decay of regulatory marks in ancient samples. Here, we used a massively parallel reporter assay in embryonic stem cells, neural progenitor cells and bone osteoblasts to investigate the regulatory effects of the 14,042 single-nucleotide modern human-specific variants. Overall, 1,791 (13%) of sequences containing these variants showed active regulatory activity, and 407 (23%) of these drove differential expression between human groups. Differentially active sequences were associated with divergent transcription factor binding motifs, and with genes enriched for vocal tract and brain anatomy and function. This work provides insight into the regulatory function of variants that emerged along the modern human lineage and the recent evolution of human gene expression.

Introduction

The fossil record allows us to directly compare skeletons between modern humans and their closest extinct relatives, the Neanderthal and the Denisovan. From this we can make inferences not only about skeletal differences, but also about other systems, such as the brain. These approaches have uncovered a myriad of traits that distinguish modern from archaic humans. For example, our face is flat with smaller jaws, our development is slower, our pelvises are narrower, our limbs tend to be slenderer, and our brain differs in its substructure proportions[1–3] (especially the cerebellum[4]). Despite our considerable base of knowledge of how modern humans differ from archaic humans at the phenotypic level, we know very little about the genetic changes that have given rise to these phenotypic differences.

The Neanderthal and the Denisovan genomes provide a unique insight into the genetic underpinnings of recent human phenotypic evolution. The vast majority of genetic changes that separate modern and archaic humans are found outside protein-coding regions, and some of these likely affect gene expression[5]. Such regulatory changes may have a sizeable impact on human evolution, as alterations in gene regulation are thought to underlie most of the phenotypic differences between closely related groups[6–9]. Indeed, there is mounting evidence that many of the noncoding variants that emerged in modern humans have altered gene expression in *cis*, shaped phenotypes, and have been under selection[5,10–18]. Fixed variants, in particular, could potentially underlie phenotypes specific to modern humans, and some of these variants might have been driven to fixation by positive selection.

Unfortunately, our ability to infer the regulatory function of noncoding variants is currently limited[19]. In archaic humans, incomplete information on gene regulation is further exacerbated by the lack of RNA molecules and epigenetic marks in these degraded samples[5]. We have previously used patterns of cytosine degradation in ancient samples to reconstruct whole-genome archaic DNA methylation maps[12,20,21]. However, despite various approaches to extract regulatory information from ancient genomes[5,13,21–26], our understanding of gene regulation in archaic humans remains minimal, with most archaic regulatory information being currently inaccessible[5]. Additionally, whereas expression quantitative locus (eQTL) mapping can be used to identify variants that drive differential expression between individuals, it can only be applied to loci that are variable within the present-day human population. Therefore, fixed noncoding variants are of particular interest in the study of human evolution, but are also particularly difficult to characterize.

Massively parallel reporter assays (MPRAs) provide the ability to interrogate the regulatory effects of thousands of variants *en masse*[27]. By cloning a candidate regulatory sequence downstream to a short transcribable sequence-based barcode, thousands of sequences and variants can be tested for regulatory activity in parallel. Thus, MPRA is an effective high-throughput tool to identify variants underlying divergent regulation, especially in organisms where experimental options are limited[28–31]. Here, we conducted a lentivirus-based MPRA (lentiMPRA[32]) on the 14,042 fixed or nearly fixed single-nucleotide variants that emerged along the modern human lineage. We generated a library of both the derived (modern human) and ancestral (archaic human and ape) sequences of each locus and expressed them in three human cell types: embryonic stem cells (ESCs), neural progenitor cells (NPCs), and primary fetal

osteoblasts. By comparing the transcriptional activities of each pair of sequences, we generated a comprehensive catalog providing a map of sequences capable of promoting expression, and those that alter gene expression. We found that 1,791 (13%) of the sequence pairs promote expression and that 407 (23%) of these active sequences drive differential expression between the modern and archaic alleles. These differentially active sequences are associated with differential transcription factor binding affinity and are enriched for genes that affect the vocal tract and brain. This work provides a genome-wide catalog of the *cis*-regulatory effects of genetic variants unique to modern humans, allowing us to systematically interrogate recent human gene regulatory evolution.

Results

LentiMPRA design and validation

To define a set of variants that likely emerged and reached fixation or near fixation along the modern human lineage, we took all the single-nucleotide variants where modern humans differ from archaic humans and great apes (based on three Neanderthal genomes[33–35], one Denisovan genome[36], and 114 chimpanzee, bonobo, and gorilla genomes[37]). We excluded any polymorphic sites within modern humans (in either the 1000 Genomes Project[38] or in dbSNP[39]), or within archaic humans and great apes[33–37] (see Methods). The resulting set of 14,042 variants comprises those changes that likely emerged and reached fixation or near fixation along the modern human lineage (**Supplementary File 1**). The vast majority of these variants are intergenic (**Supplementary Fig. 1a**). By definition, this list does not include variants that introgressed from archaic humans into modern humans and spread to detectable frequencies. We refer to the

derived version of each sequence as the *modern human sequence* and the ancestral version as the *archaic human sequence*.

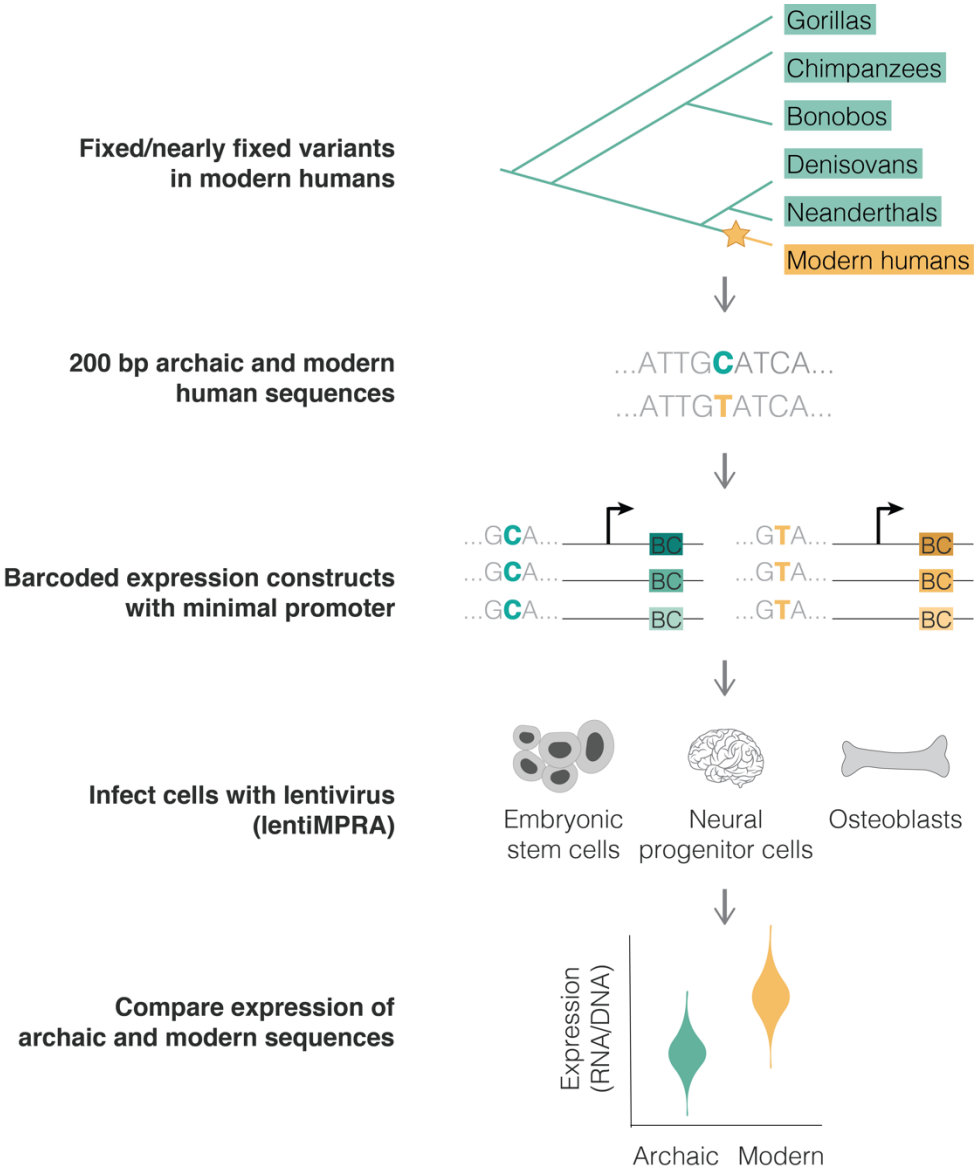
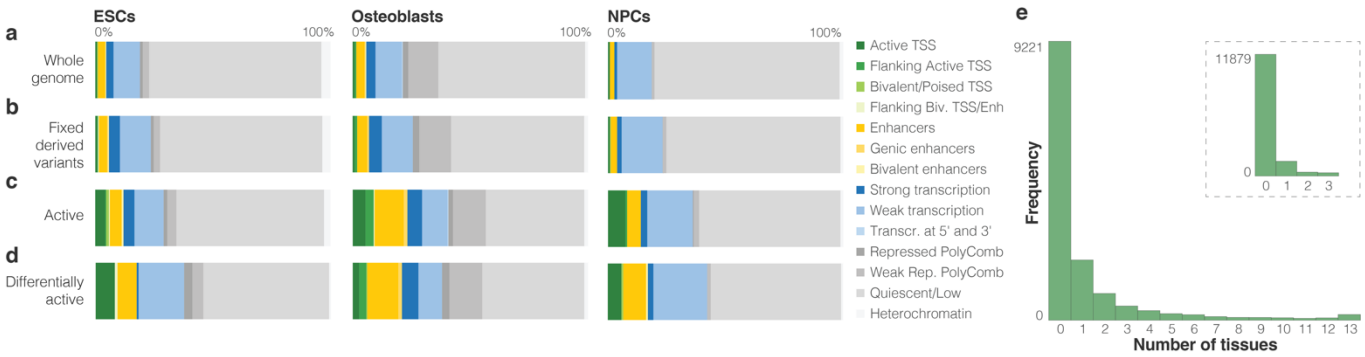


Figure 1. Using lentiMPRA to identify variants driving differential expression in modern humans. We analyzed variants that likely emerged and reached fixation or near fixation along the modern human lineage (yellow) and that were not polymorphic in any other ape or archaic genome (green) (top). The modern and archaic human variants and their surrounding 200 bp were synthesized, cloned into barcoded expression constructs and infected in triplicates into three human cell lines using a chromosomally integrating vector, following the lentiMPRA protocol[32] (see methods). We compared the activity (RNA/DNA) of the modern and archaic human constructs to identify variants promoting differential expression using MPRAnalyze[40] (bottom).



Supplementary Figure 1. Classification of chromHMM annotations for different groups of variants. Relative percentage of bases in each chromHMM[41,42] category throughout the entire genome (**a**), in fixed or nearly fixed modern human-derived variants (**b**), in active sequences (**c**) and in differentially active sequences (**d**), per cell type. See Discussion for cell-type specificity and enhancer enrichment. **e.** Histogram of the number of tissues and number of sequences with TSS- or enhancer-related chromHMM marks for all 14,042 sequences. Tissues and cell types investigated include ESCs, osteoblasts, NPCs, mesenchymal stem cells, monocytes, skin fibroblasts, brain hippocampus, skeletal muscle, heart left ventricle, sigmoid colon, ovary, fetal lung, and liver. Inset shows data for ESC, osteoblast and NPC only.

We synthesized a library composed of 200 base pair (bp) sequences (due to oligonucleotide synthesis length limitations) per each of the 14,042 variants (one sequence for the modern human allele and one for the archaic human allele, **Fig. 1, Supplementary File 1**). Each sequence contained at its center either the modern or archaic human variant. 13,680 out of 14,042 sequence pairs (90%) had a single variant separating the human groups. For the 1,362 sequence pairs containing additional variants within the 200 bp window, we used either the modern-only or archaic-only variants throughout the sequence. We amplified this library of sequences, each along with a minimal promoter and barcode. We then inserted these constructs into the lentiMPRA vector, so that the barcode, which is the readout of activity, is located within the 5'UTR of the reporter gene and is transcribed if the assayed sequence is an active regulatory element[32]. We associated each sequence with multiple barcodes to achieve a high number of independent replicates of expression per sequence, thereby reducing potential site-of-integration effects. 97% of sequences had at least 10 barcodes associated with them, with a median of 96

barcodes per sequence (**Supplementary Fig. 2a**). Furthermore, we used a chromosomally integrating construct rather than an episomal construct due to the improved technical reproducibility and correlation of results from chromosomally integrating constructs with functional genomic signals like transcription factor ChIP-seq and histone acetylation marks[43]. To further reduce lentivirus site-of-integration effects, this vector contained antirepressors on either side and was integrated in multiple independent sites, with each sequence marked by multiple barcodes. (see Discussion for additional lentiMPRA limitations). Importantly, despite the caveat of interrogating sequences outside of their endogenous context, MPRAs were shown to generally replicate the endogenous activity of sequences[43–45].

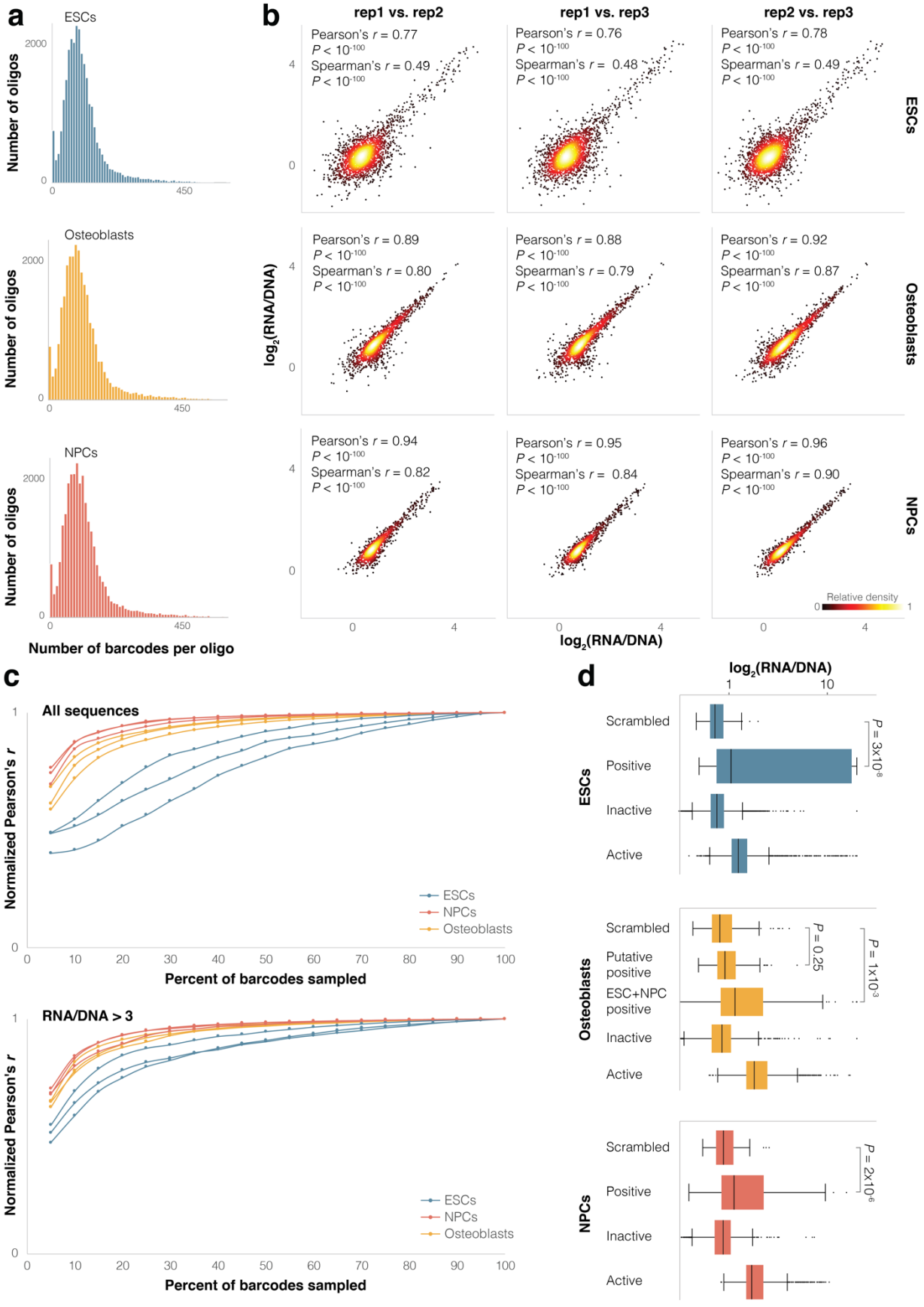
The brain and skeleton have been the focus of evolutionary studies due to their extensive phenotypic divergence among human lineages[3]. Therefore, we chose human cells related to each of these central systems: NPCs and primary fetal osteoblasts. In addition, we used ESCs (line H1, from which the NPCs were derived) to gain insight into early stages of development. Finally, the abundance of previously published regulatory maps for these three cell types[20,41,42] also enables the investigation of the dynamics of evolutionary divergence at different regulatory levels. While these cell types represent diverse systems, further studies are needed in order to characterize the activity of these sequences in other cell types.

We used the library of 14,042 pairs of archaic and modern human sequences, together with positive and negative control sequences, to infect each cell type. As positive controls for ESCs and NPCs, we added a set of 199 sequences with known regulatory capacity from previous MPRAs (**Supplementary File 1**). To our knowledge, there have not been any MPRAs conducted in osteoblasts, so we searched the literature for putative regulatory regions in osteoblasts and

other bone cell types and used these as putative positive controls (**Supplementary File 1**, see Methods.). As negative controls, in all cell types, we randomly chose 100 sequences from the library and scrambled the order of their bases, creating a set of GC-content matching sequences that had not been previously established to drive expression.

We performed three replicates of library infection in each cell type and quantified barcode abundance for each sequence in RNA and DNA (**Fig. 1**). To assess the reproducibility of our lentiMPRA results, we calculated the RNA/DNA ratio (a measure of expression normalized to the number of integrated DNA molecules) for each sequence and compared it across the three replicates per cell type. We saw a strong correlation of RNA/DNA ratios between replicates for all cell types (Pearson's r = 0.76 – 0.96, $P < 10^{-100}$, **Supplementary Fig. 2b**), with the lower correlation scores being in ESC, likely due to our use of lower multiplicity of infection (MOI) in these cells due to their increased sensitivity to lentivirus infection. High barcode and read coverage in MPRA generally provides increased power to detect differences in allelic expression[32,45]. Thus, to determine how variability depended on our barcode counts, we downsampled the number of barcodes per sequence and calculated the RNA/DNA ratio at each step for each of the three replicates. In agreement with previous studies[43], we found that the number of barcodes used in this study is well within the plateau, suggesting that the number of barcodes is not a limiting factor in our experiment (**Supplementary Fig. 2c**). Finally, we assessed the distribution of RNA/DNA ratios across our scrambled sequences and positive controls. The mean RNA/DNA ratio of the scrambled sequences was lower than that of the positive control sequences in ESCs and NPCs ($P = 2.7\text{x}10^{-8}$ for ESCs and $P = 1.8\text{x}10^{-6}$ for NPCs, *t*-test, see Methods, **Supplementary Fig. 2d**), but not in osteoblasts ($P = 0.25$). This is unlikely

due to a problem with the osteoblasts, as the osteoblast-related controls show similar expression in all three cell types. Moreover, ESC and NPC positive controls are active in osteoblasts ($P = 1.1x10^{-3}$). The correlation between replicates was also similar between osteoblasts and the other two cell types (**Supplementary Fig. 2b**). Thus, the lack of activity of the osteoblast putative positive controls is likely because, in contrast to the ESC and NPC confirmed positive controls, the osteoblast putative positive controls were not previously tested in an MPRA, and some of these putative enhancers were identified in mouse and were not validated in human. Overall, these results suggest that the lentiMPRA was technically reproducible and adequately powered to detect expression.



Supplementary Figure 2. Reproducibility of lentiMPRA data. a. Distribution of number of barcodes per each sequence. **b.** Replicate-by-replicate correlation of expression (RNA/DNA). Each point represents an active sequence. **c.** Simulations of barcode down-sampling showing Pearson's correlation of expression (RNA/DNA) between replicates. Upper panel shows all sequences and lower panel shows sequences with higher expression (RNA/DNA > 3). Pearson's *r* values are normalized to maximum Pearson's *r* observed for each pair of replicates. **d.** Box plots of scrambled, positive control, inactive and active sequences. One-sided *t*-test *P*-values are shown. Boxes show interquartile range (IQR), black line within box shows median, whiskers show 1.5xIQR from box borders, points show outliers.

Characterization of active regulatory sequences

We first examined which of the assayed sequences are able to drive expression. To do so, we utilized MPRAnalyze[40], which uses a model for each of the RNA and DNA counts, estimates transcription rate and then identifies sequences driving significant expression. We also added an additional stringency filter whereby a sequence is only considered expressed if it had an RNA/DNA ratio significantly higher than that of the scrambled sequences (FDR < 0.05). We found that in ESCs, 8% (1,183) of sequence pairs drove expression in at least one of the alleles, 6% (814) in osteoblasts, and 4% (602) in NPCs (FDR < 0.05, **Supplementary File 1**, **Supplementary Fig. 2d**, see Methods). Hereinafter, we refer to these sequences as *active* sequences. Overall, 13% (1,791) of archaic and modern human sequence pairs were active in at least one cell type, 4% (586) in at least two cell types, and 2% (222) in all three cell types (overlap of 75-fold higher than expected, $P < 10^{-100}$, Super Exact test[46], **Fig. 2a**).

Some of these sequences may show activity in the lentiMPRA experiment, but not in their endogenous genomic context. To test whether activity in our lentiMPRA reflects true biological function, we investigated whether our active sequences had expected regulatory characteristics in the modern human genome. Active regulatory sequences in the genome tend to bear active chromatin marks. Therefore, we examined whether active sequences in lentiMPRA tend to be enriched for markers of active chromatin in their endogenous context. We first tested overlap

with five histone modification marks and one histone variant associated with active chromatin (H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K9ac, and H2A.Z), as well as with two histone modification marks associated with repressed chromatin (H3K9me3 and H3K27me3, see Methods)[42]. We found that on average, active sequences were 1.6-2.7-fold more likely than inactive sequences to have active chromatin marks, depending on cell type. Also, these sequences tended to show relatively fewer repressive marks compared to active marks (**Fig. 2b-d, Supplementary File 2**). These trends get stronger when looking at more highly active sequences. For example, while only 18% of inactive sequences in ESCs overlap H3K4me2 peaks, 70% of active sequences with an RNA/DNA ratio ≥ 3 in ESCs overlap H3K4me2 peaks ($FDR = 4.4\text{x}10^{-16}$, Fisher's exact test, **Fig. 2b-d, Supplementary File 2**). To further test the functional characteristics of active sequences, we analyzed chromHMM annotation[41,42], which uses chromatin signatures to subdivide the genome into functional regions. 2,163 of the 14,042 sequences (15%) overlapped promoter or enhancer chromHMM annotations in at least one of the three cell types. Additional 2,658 sequences (19%) overlapped such marks in other cell types not included in this study. Compared to inactive sequences, we found that active sequences are enriched for promoter and enhancer marks (FDR < 0.05 in each of the cell types for overlap with *Active TSS* and *Enhancers*, **Fig. 2e, Supplementary Fig. 1**, **Supplementary File 1-2)**. We also found that compared to inactive sequences, active sequences are 6-32% closer to GTEx[47] eQTLs, depending on cell type (FDR < 0.05, *t*-test). Active sequences are also 1.2-1.3x closer to transcription start sites (TSSs), with 32-39% of them located within 10 kb of a TSS, depending on cell type (FDR < 0.05, *t*-test, **Supplementary File 2**).

Active genomic regions often show reduced DNA methylation levels compared to inactive regions[48]. To further test if the activity we detected in the lentiMPRA reflects true biological

function, we tested whether the active sequences in the lentiMPRA tend to be hypomethylated in their endogenous genomic context. To do so, we used our previously published modern and archaic human DNA methylation maps[12,20,21]. Because the DNA methylation maps originate from skeletal samples, we compared them to the osteoblast lentiMPRA data. We found that active sequences are significantly hypomethylated compared to inactive sequences ($P = 5.5x10^{-13}$, *t*-test, **Fig. 2f**) and that their activity level (RNA/DNA ratio) is negatively correlated with methylation levels ($6.0x10^{-9}$, Pearson's $r = -0.24$).

Finally, compared to inactive sequences, active sequences show slightly higher sequence conservation in primates, indicating a potential functional role (PhyloP, -0.05 on average for inactive, -0.04 for active, FDR = $1.1x10^{-3}$, *t*-test) with more highly active sequences showing higher conservation levels (e.g., 0.24 for active sequences with RNA/DNA ratio ≥ 4, **Supplementary Fig. 3a**, **Supplementary File 2**). In summary, we found that sequences that are capable of driving expression tend to overlap active chromatin marks, are depleted of repressive chromatin marks, closer to TSSs and eQTLs, and have higher sequence conservation, giving us confidence that the MPRA provides us with biologically meaningful results.

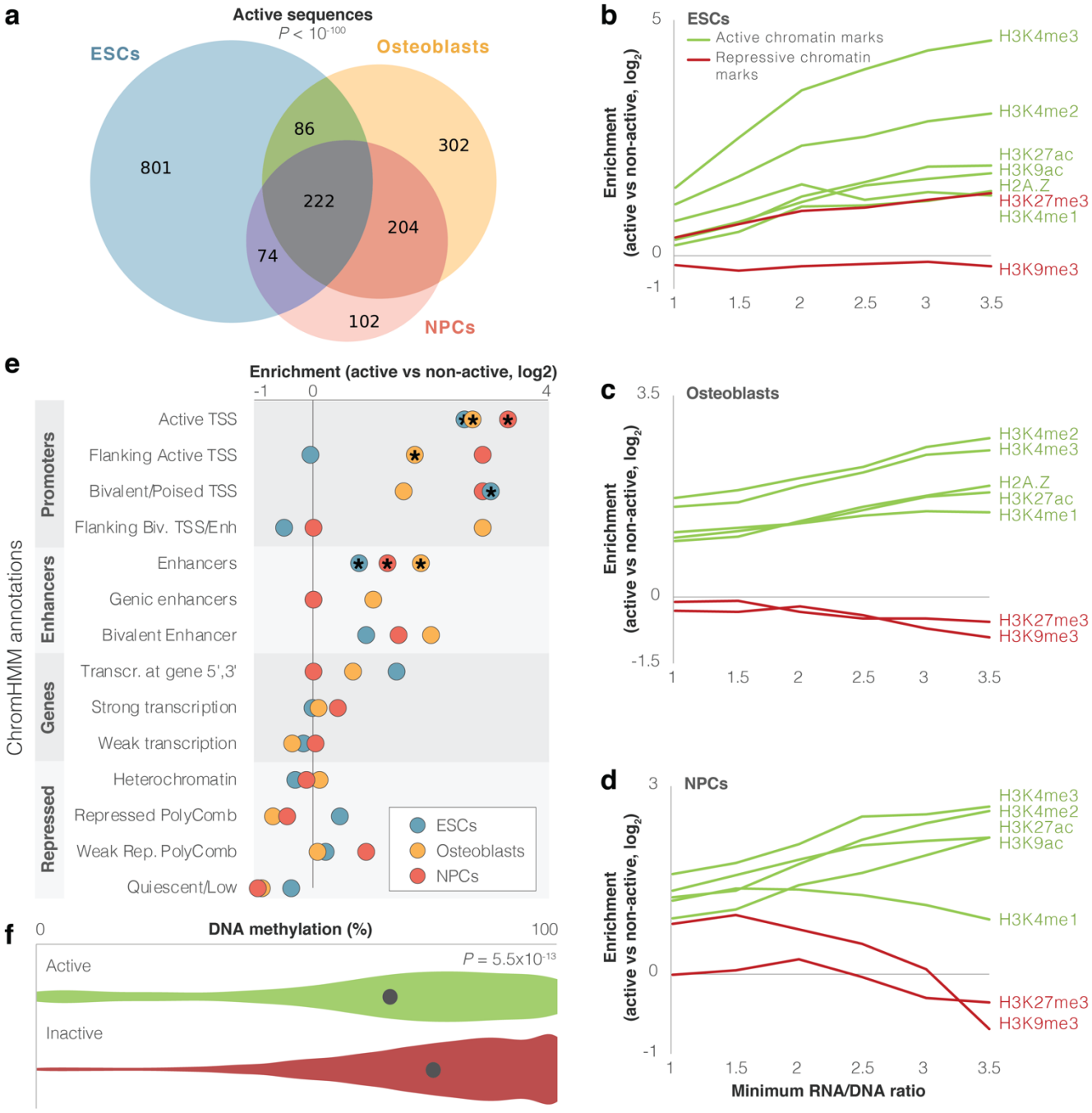


Figure 2. Identification of modern human sequences promoting expression in lentiMPRA. a. Overlap between cell types of active sequences. Super Exact test *P*-value is shown for the overlap of the three groups. **b-d.** Enrichment levels of active and repressive histone modification marks within active sequences. Enrichment is computed compared to inactive sequences. The enrichment of H3K27me3 in ESCs possibly reflects the presence of this mark in bivalent genes, which become active in later stages of development[49]. For confidence intervals see Supplementary Table 2. **e.** Enrichment of differentially active sequences in various chromatin-based genomic annotations. Missing circles reflect no differentially active sequences in that category. Stars mark significant enrichments (FDR < 0.05). **f.** Violin plots of DNA methylation levels for active (green) vs. inactive (red) sequences in osteoblasts. Methylation levels per sequence were computed as the mean methylation across all modern and

archaic human bone methylation samples. The circle marks mean methylation across all sequences in each group. *t*-test *P*-value is shown.

Differentially active sequences between modern and archaic humans

We next set out to identify modern and archaic human sequences driving differential expression. We used MPRAnalyze[40] to compare expression driven by the modern and archaic sequences. Out of the active sequence pairs in each cell type, 110 (9%) in ESCs drive significantly differential expression between modern and archaic humans, 243 (30%) in osteoblasts, and 153 (25%) in NPCs (FDR $\leq$ 0.05, see Methods, **Fig. 3a-c, Supplementary Fig. 2**, see Discussion for cell-type differences). We refer to these sequence pairs hereinafter as *differentially active* sequences. Overall, we see significant overlap between cell types in differentially active sequences: 407 sequences (23% of active sequences) were differentially active in at least one cell type, 89 (5%) in at least two cell types, and 10 (0.6%) in all three cell types (8-fold higher than expected compared to active sequences, $P = 5\text{x}10^{-7}$, Super Exact test[46], **Fig. 3d**).

As expected from such closely related organisms, and similar to other MPRAs that compared nucleotide variants (see Discussion), including one that compared human and chimp sequences[30], most sequences drove modest magnitudes of expression difference; of the 407 differentially active sequences, the median fold-change was 1.2x, and only five sequences had a fold-change greater than 2x (**Fig. 3a-c**). We refer to differentially active sequences where modern human expression is higher/lower than archaic human expression as up/downregulating sequences, respectively. In ESCs and NPCs, sequences were equally likely to be up- or downregulating (51% and 52% of differentially active sequences were downregulating, $P = 0.92$ and $P = 0.63$,

respectively, Binomial test), while in osteoblasts downregulation was observed slightly more often (59%, $P = 6.9\text{x}10^{-3}$). We identified 109 sequence pairs that were differentially active in more than one cell type. Out of these 109, we found that 107 show the same direction of differential activity across cell types ($P = 9.2\text{x}10^{-30}$, Binomial test), and we also observed a high correlation between the magnitudes of differential activity (Pearson's $r = 0.82$, $P = 1.6\text{x}10^{-27}$). That differentially active sequences from one cell type are predictive of differential activity in other cell types, even of cell types as disparate as those used here, suggests that these sequences are likely to be differentially active in other cell types not assayed in this lentiMPRA.

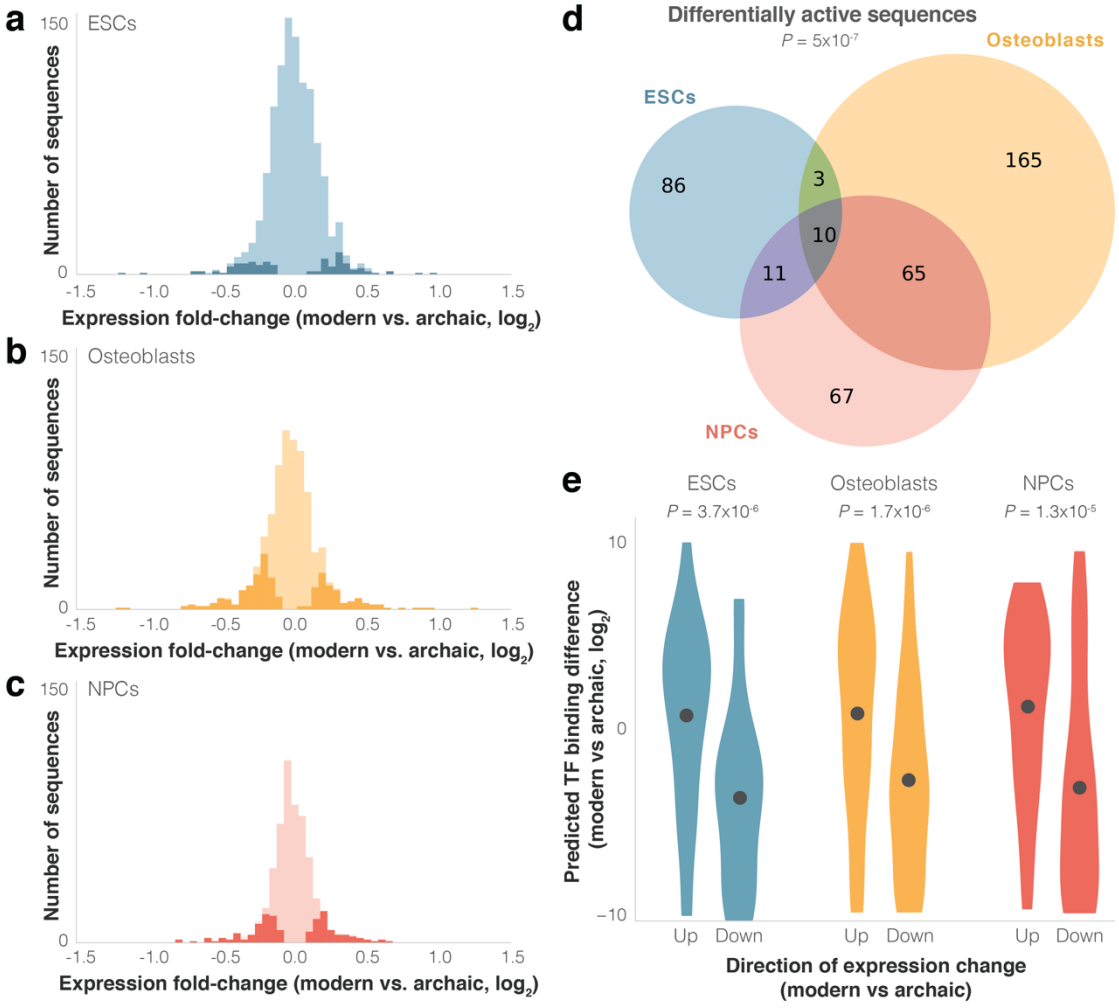
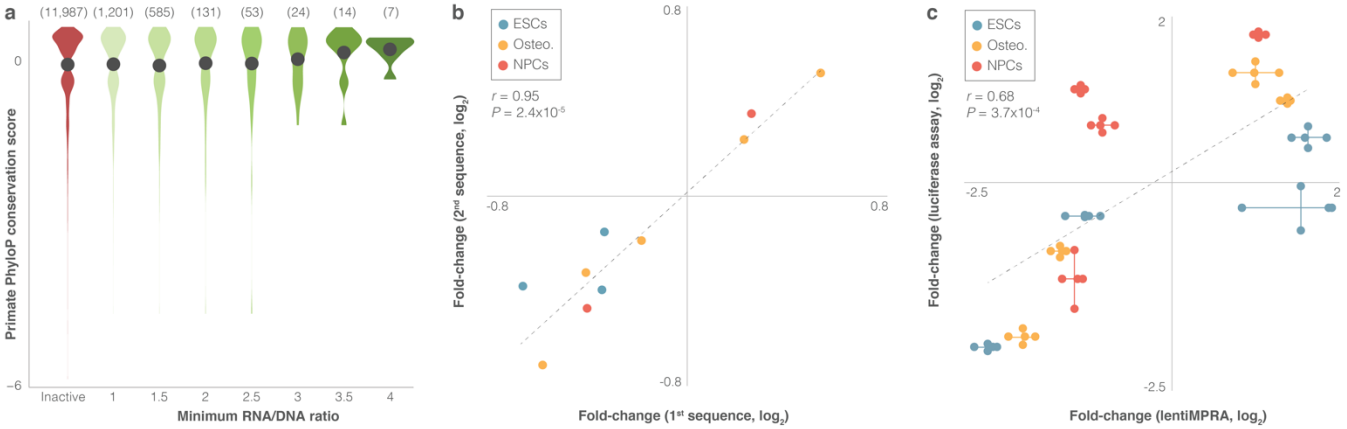


Figure 3. Differential activity of derived modern human sequences. a-c. Distributions of expression fold-changes (RNA/DNA) of active (light) and differentially active (dark) sequences in each cell type. **d.** Overlap of differentially active sequences between cell types. Super Exact test *P*-value is presented for the overlap of the three groups compared to active sequences. In the 10 sequences that were differentially active across all three cells types, the direction of fold-change was identical across all cell types ($P = 1.9x10^{-3}$, Binomial test). **e.** Violin plots of predicted TF binding score difference between modern and archaic sequences. Positive scores represent increased binding in the modern sequence. Points show mean.

To further test the replicability of these results, we examined the relationship between pairs of overlapping differentially active sequences (i.e., variants that are < 200bp apart and thus appear in more than one sequence, three overlapping pairs in ESCs, five in osteoblasts, and two in NPCs). We found that the direction of expression change is identical in all pairs of overlapping sequences ($P = 2.0x10^{-3}$, binomial test), and that the magnitude of their expression change is highly correlated (Pearson's $r = 0.95$, $2.4x10^{-5}$, **Supplementary Fig. 3b**). To validate these results with an orthogonal method, we tested four differentially active sequences from each cell type in a luciferase reporter assay and found that the direction and magnitude of differential expression tended to replicate the lentiMPRA results (9 out of 12 sequences, Pearson's $r = 0.67$, $P = 3.7x10^{-4}$, **Supplementary Fig. 3c, Supplementary File 1**). These results suggest that the lentiMPRA was both technically reproducible across cell types and assays and also indicative of true biological signal.

Finally, we examined the endogenous genomic locations of differentially active sequences, focusing on promoters and enhancers. Between 33-45% of these sequences are within 10 kb of a TSS (depending on cell type, **Supplementary File 1**). Analyzing chromHMM[41,42], we found that between 20-25% of the differentially active sequences are within promoter or enhancer regions (**Supplementary File 1**). To test if differentially active sequences are enriched within regulatory elements, we compared the proportion overlapping chromHMM promoters and enhancers in

differentially active sequences to that proportion in the other active sequences. We found that differentially active sequences are over-represented within putative enhancer regions in NPCs (2.2-fold, FDR = 0.03, Fisher's exact test, **Supplementary Fig. 1c-d**). These results support a model of rapid enhancer evolution in modern humans, as previously reported for other mammals[50] (see Discussion).



Supplementary Figure 3. Differential expression is replicated across overlapping sequences and in a reporter assay validation. a. Primate PhyloP conservation scores in inactive sequences and active sequences with increasingly higher RNA/DNA ratios (maximum RNA/DNA across the three cell types). Dots signify mean conservation per bin. Numbers in parentheses show number of sequences per bin. **b.** Expression fold-change of overlapping pairs of sequences. Pearson's *r* and *P*-value are presented. **c.** Expression fold-change of lentiMPRA vs luciferase assay. Each pair of points connected by a vertical line represents two replicates in the luciferase assay. Each triplet of points connected by a horizontal line represents three lentiMPRA replicates. Pearson's *r* and *P*-value are presented.

Molecular mechanisms underlying differential activity

Next, we sought to understand what regulatory mechanisms might be associated with differential activity. Changes in expression are often linked to changes in regulatory marks. For example, increased DNA methylation tends to be associated with reduced activity[48]. We therefore tested methylation levels in each pair of sequences and examined if the human group with the lower sequence activity tends to show higher methylation levels. Here too, because the DNA

methylation maps originate from bone samples[12,20,21], we compared them to the osteoblast lentiMPRA data. We found that upregulating sequences indeed have a slight but significant tendency to be hypomethylated in modern compared to archaic humans, and that downregulating sequences tend to be hypermethylated in modern compared to archaic humans (on average -2% methylation in upregulating sequences, and +1% methylation in downregulating sequences, in the modern compared to the archaic genomes, $P = 0.028$, paired *t*-test, **Supplementary Fig. 4a**). This trend is slightly more pronounced when looking at the most differentially regulating sequences. For example, the top ten most downregulating sequences show on average +8% methylation in modern compared to archaic humans, whereas the top ten most upregulating sequences show -7% methylation in modern compared to archaic humans. We also examined promoter regions (5 kb upstream to 1 kb downstream of a TSS), where the association between methylation and reduced activity is known to be stronger compared to the rest of the genome[48]. Indeed, we found that upregulating promoter sequences have +5% methylation on average in the modern compared to the archaic genomes, while downregulating promoter sequences have -8% methylation ($P = 0.034$, paired *t*-test; **Supplementary Fig. 4b**). This trend is more pronounced in CpG-poor promoters, where the link between methylation and expression is known to be stronger[51–53] (-15% methylation in upregulating sequences, and +15% methylation in downregulating promoter sequences in modern compared to archaic humans; $P = 6\text{x}10^{-3}$, paired *t*-test; **Supplementary Fig. 4c**).

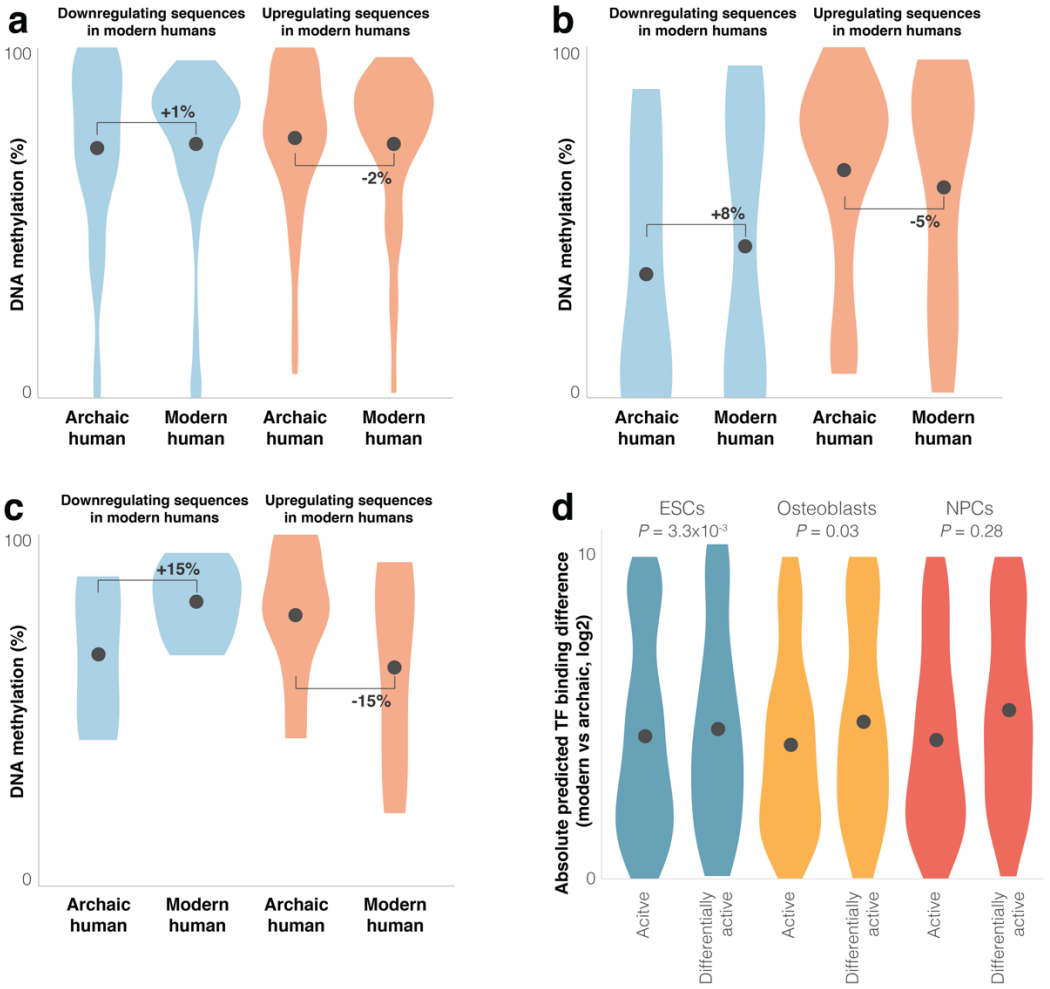
We conjectured that some of the differential activity in these loci might have been driven by alterations in transcription factor (TF) binding. To investigate this, we compared predicted TF binding affinity to the modern and archaic sequences using FIMO[54]. We found that: (1) compared to other active sequences, the difference in predicted binding between the modern and

archaic human alleles tends to be larger for differentially active sequences (combined across cell types: 4.3x, $P = 0.02$, *t*-test, **Supplementary Fig. 4d**); (2) the directionality of differential expression tends to match the directionality of differential binding, i.e., upregulating sequences tend to have stronger predicted binding for the modern human sequence, whereas downregulating sequences tend to have stronger predicted binding for the archaic sequence ($P = 3.7\text{x}10^{-6}$ for ESCs, $P = 1.7\text{x}10^{-6}$ for osteoblasts, and $P = 1.3\text{x}10^{-5}$ for NPCs, binomial test, **Fig. 3e,** see Methods); and (3) the magnitude of expression difference is correlated with the magnitude of predicted binding difference (Pearson's $r = 0.43$ and $P = 1.2\text{x}10^{-3}$ for ESCs, Pearson's $r = 0.23$ and $P = 0.02$ for osteoblasts, and Pearson's $r = 0.35$ and $P = 2.4\text{x}10^{-3}$ for NPCs, **Supplementary Fig. 5a-c** and **Supplementary File 3**). These results support the notion that alterations in TF binding played a role in shaping some of the expression differences between modern and archaic humans.

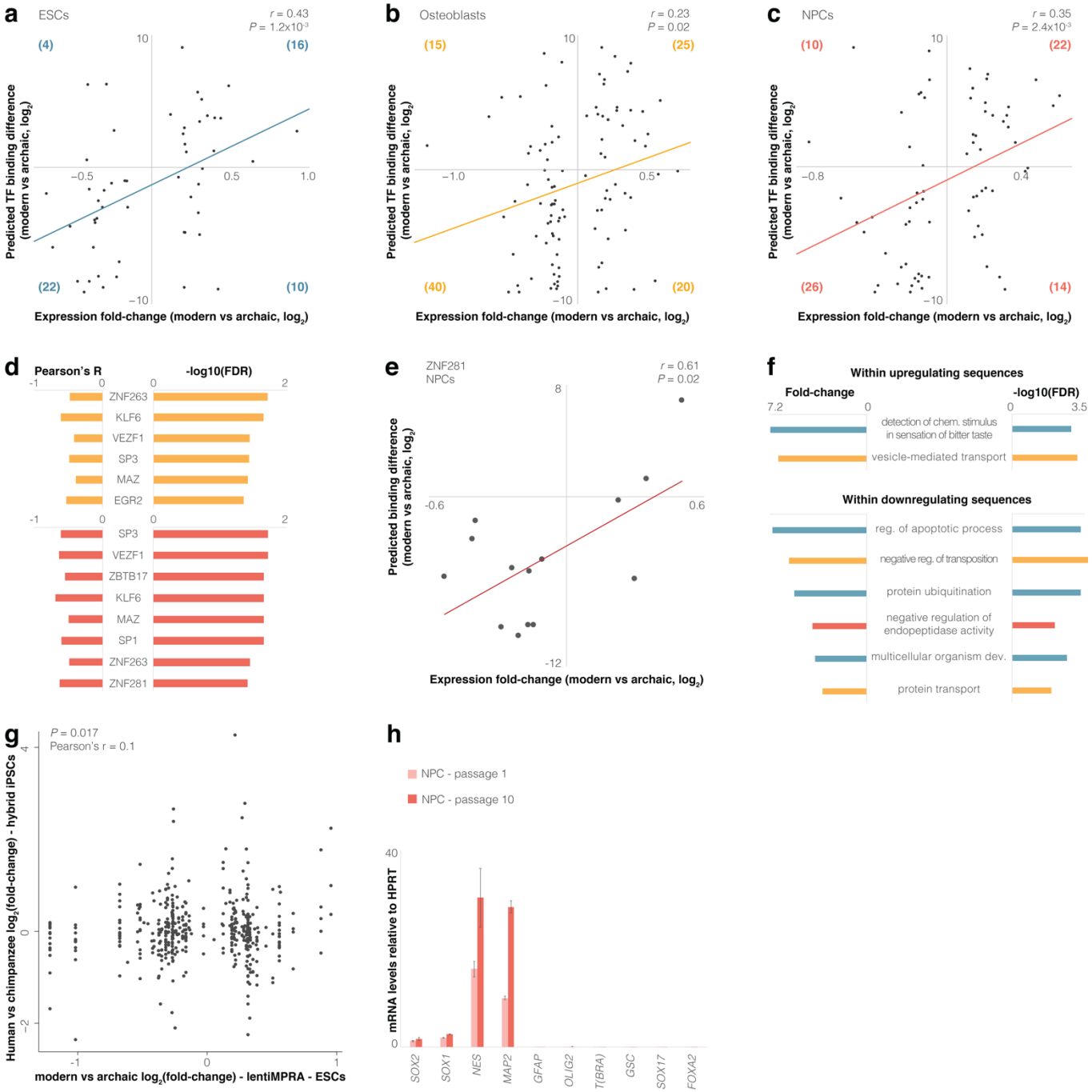
To identify the TFs that primarily drove these observations, we investigated which motif changes are most predictive of expression changes. For each TF and the sequences it is predicted to differentially bind, we examined the correlation between binding and expression fold-change (either positive or negative). We found that changes to the motifs of 14 TFs were predictive of expression changes (**Supplementary Fig. 5d, Supplementary File 3**). All of these TFs had a positive correlation between changes in their predicted binding affinity and changes in expression of their bound sequences, reflective of their known capability to promote transcription[55–63]. Of note, the use of a minimal promoter with basal activity in the MPRA design means that transcriptional repression is less likely to be detected, and therefore, further

investigation is required in order to identify potential repressive activity in these sequences (see Discussion).

Next, we sought to explore if some motif changes are particularly over-represented within differentially active sequences, suggestive of a more central role in shaping modern human regulatory evolution. To control for sequence composition biases, we used active sequences as a background to search for motif enrichment within differentially active sequences. We found that ZNF281, an inhibitor of neuronal differentiation[64], is significantly enriched: out of 153 differentially active sequences in NPCs, 14 are predicted to be bound by ZNF281 (4.6-fold, FDR = 0.04, **Supplementary File 3**). Notably, ZNF281 is also one of the TFs whose predicted differential binding is most tightly linked with differential expression (**Supplementary Fig. 5d,e**). Overall, these data support a model whereby variants in ZNF281 motifs might have modulated ZNF281 binding in NPCs, thereby contributing to neural expression differences between modern and archaic humans.



Supplementary Figure 4. Differential activity is associated with differential DNA methylation and TF binding. a,b. Violin plots of DNA methylation levels in modern and archaic human bone methylation samples, for differentially active (a), promoter differentially active (**b**), and CpG-poor promoter differentially active (**c**) sequences in osteoblasts. Promoter sequences are sequences between 5 kb upstream to 1 kb downstream of a TSS. CpG-poor promoter sequences were defined as the bottom 50% promoter sequences. **d.** Violin plots of absolute predicted TF binding score difference between modern and archaic sequences. Points show mean.



Supplementary Figure 5. Predicted TF binding is correlated with differential activity. a-c. Expression fold-change vs predicted TF binding fold-change for each sequence. Positive scores represent increased binding in the modern sequence. Parentheses show number of points in each quadrant with a score difference > 0. **d.** Pearson's correlation between differential expression and predicted differential binding affinity. Only significant TFs (FDR <= 0.05, Supplementary File 3) are shown for osteoblasts (yellow) and NPCs (red). **e.** Expression fold-change vs predicted TF binding fold-change for ZNF281 in NPCs. Pearson's *r* and *P*-value are shown. **f.** Enriched Gene Ontology terms for ESCs (blue), osteoblasts (yellow) and NPCs (red). **g.** Expression fold-change of differentially

active sequences compared to the *cis*-regulatory expression fold-change between human and chimpanzee of genes associated with these sequences. *cis*-regulatory expression changes were taken from hybrid human-chimpanzee induced pluripotent stem cells (iPSCs)[65]. **h.** RT-qPCR validation of NPCs at passage 1 (pink) and passage 10 (red). Expression levels are normalized to *HPRT* expression.

Potential phenotypic consequences of differential expression

In an attempt to assess the functional effects of the differential transcriptional activity we detected, we first sought to link each sequence to the gene(s) it might regulate in its endogenous genomic location. While most regulatory sequences are known to affect their closest gene[66,67], some exert their function through interactions with more distal genes, often reflected in chromatin conformation capture assays, such as Hi-C interactions[68], or eQTL associations[68,69]. To predict the genes linked to each sequence we combined data from four sources: (1) proximity to transcription start sites; (2) proximity to eQTLs[47]; (3) proximity to putative enhancers[70]; and (4) spatial interaction with promoters using Hi-C data[69] (see Methods). Using these data, we generated for each cell type a list of genes potentially regulated by each sequence. Overall, 1,341 out of the 1,791 active sequences (75%) were linked to at least one putative target gene (**Supplementary File 1**).

To study the potential functional effects of differentially active sequences, we analyzed functions associated with their linked genes. To control for confounders such as cell type-specific regulation, gene length, and GC content, we compared differentially active sequences to other active sequences (instead of the genomic background), which minimizes inherent biases in the active sequences. First, we tested Gene Ontology terms and found an enrichment of the following terms within downregulating sequences: *vesicle-mediated transport* (6.6-fold, FDR = $1.9\text{x}10^{-3}$, in osteoblasts), *regulation of apoptotic process* (6.0-fold, FDR = $1.9\text{x}10^{-3}$, in ESCs),

protein ubiquitination (4.7-fold, FDR = $1.9x10^{-3}$, in ESCs), multicellular organism development (3.3-fold, FDR = 0.01, in ESCs), and *protein transport* (3.3-fold, FDR = 0.02, in osteoblasts, **Supplementary Fig. 5f**, **Supplementary File 4**). No enriched terms were found within upregulating sequences. To obtain a more detailed picture of phenotypic function, we ran Gene ORGANizer, a tool that uses monogenic disorders to link genes to the organs they affect[71]. We analyzed the genes linked to differentially active sequences and found that for genes linked to sequences driving upregulation, the most enriched body parts belong to the vocal tract, i.e., the vocal cords (5.0-fold, FDR = $1.3x10^{-3}$), voice box (larynx, 3.8-fold, FDR = $4.8x10^{-3}$), and pharynx (3.3-fold, FDR = $9.5x10^{-3}$, all within ESCs, **Fig. 4a**). Interestingly, we have previously reported that the most extensive DNA methylation changes in modern compared to archaic humans arose in genes affecting the vocal cords and voice box[12]. Conversely, within sequences driving downregulation, the enriched body part is the cerebellum (3.0-fold, FDR = $9.2x10^{-3}$, in NPCs, **Fig. 4a, Supplementary File 4**). This is in line with previous reports of cerebellar anatomy differences between modern humans and Neanderthals[1–3], including results suggesting that the biggest differences in brain anatomy are in the cerebellum[4]. These data also provide leads into the functional divergence of organs, like the voice box, that are not preserved in the fossil record.

Next, we delved into individual phenotypes associated with the differentially active sequences. To this end, we used the Human Phenotype Ontology (HPO) database[72], a curated database of genes and the phenotypes they underlie in monogenic disorders. HPO covers a broad range of phenotypes related to anatomy, physiology, and behavior. We found that enriched phenotypes were involved in speech, heart morphology testicular descent, and kidney function (FDR < 0.05,

Fig. 4b, Supplementary File 4). These results reveal body parts and phenotypes that were particularly associated with gene expression changes between modern and archaic humans, and could be new candidates for phenotypes under selection.

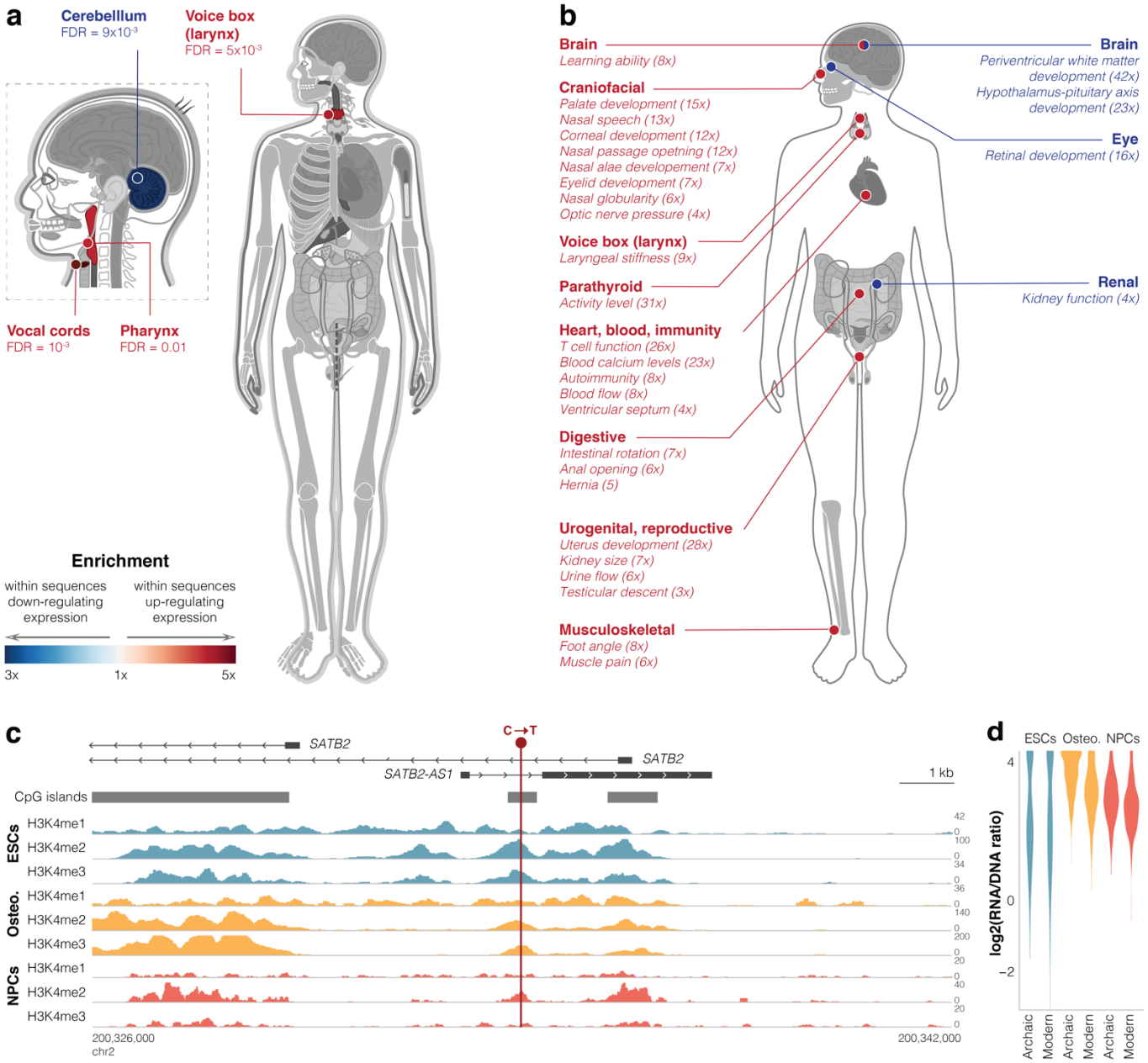


Figure 4. Differentially active sequences are linked to genes affecting the vocal tract and brain. a. Gene ORGANizer enrichment map showing body parts that are significantly over-represented within genes linked to differentially active sequences (FDR < 0.05). Organs are colored according to the enrichment scale. See Supplementary File 4 for cell types. **b.** HPO phenotypes significantly enriched (FDR < 0.05) within differentially

active sequences. Fold-enrichment is shown in parentheses. See Supplementary File 4 for cell types. **c.** CpG islands and read density of active histone modification marks[42] around the differentially active sequence in *SATB2* (GRCh37 genome version). **d.** Violin plots of archaic vs. modern activity of the differentially active sequence in *SATB2*.

Downregulation of *SATB2* potentially underlies brain and skeletal differences

This catalog of *cis*-regulatory changes allows us to explore specific sequence changes that potentially underlie divergent phenotypes observed from fossils. To use the most robust data from lentiMPRA, we examined the ten sequences that are differentially active across all three cell types, and looked at their linked genes. To investigate the phenotypes that are potentially linked to these genes, we looked for those genes whose phenotypes can be compared to the fossil record (i.e., skeletal phenotypes). The only gene that fit these criteria was *SATB2*, a regulator of brain and skeletal phenotypes[73]. First, we analyzed its linked variant (C to T transition), which is at a position that is relatively conserved in vertebrates (GRCh38: 199,469,203 on chromosome 2, PhyloP score = 0.996). This position is found within a CpG island between two alternative TSSs of *SATB2* (**Fig. 4c**). It is also found in the first intron of *SATB2-AS1*, an antisense lncRNA which upregulates SATB2 protein levels[74]. To determine if this position lies within a regulatory region, we investigated it for chromatin marks in modern humans. We found that it overlaps a DNase I-hypersensitive site[75] and shows many peaks of active histone modification marks in all three cell types (**Fig. 4c, Supplementary File 1**). Indeed, this sequence drives high expression in all three cell types (fourth, eighth, and 19th percentile among active sequences, in ESCs, osteoblasts, and NPCs, respectively, FDR < 10^{-5} across all). Although this sequence shows hallmarks of activity in modern humans, compared to the archaic sequence the modern human sequence is downregulating in all three cell types (-9% in ESCs, FDR = 6.8×10^{-4}, -27% in osteoblasts, FDR

= $2.2x10^{-42}$, and -12% in NPCs, FDR = $1.1x10^{-7}$, **Fig. 4d**). These results suggest that the ancestral version of this sequence possibly promoted even higher expression in archaic humans.

SATB2 encodes a transcription factor expressed in developing bone and brain. Its activity promotes bone formation, jaw patterning, cortical upper layer neuron specification, and tumorigenesis[73]. Genome-wide association studies (GWAS) show that common variants near and within *SATB2* are mainly associated with brain and bone phenotypes, such as reaction time, anxiety, mathematical abilities, schizophrenia, autism, bone density, and facial morphology[76,77]. Heterozygous loss-of-function mutations in *SATB2* result in the *SATB2*-associated syndrome, which primarily affects neurological and craniofacial phenotypes. This includes speech delay, behavioral anomalies (e.g., jovial personality, aggressive outbursts, and hyperactivity), autistic tendencies, small jaws, dental abnormalities, and morphological changes to the palate[78]. Additionally, reduced functional levels of SATB2 due to heterozygous loss-of-function have been shown to be the cause of these phenotypes in both human[73,78–80] and mouse[81–83]. Because these phenotypes are driven by changes to functional SATB2 levels[73], we conjectured that the differential expression of *SATB2* predicted from lentiMPRA might be linked to divergent modern human phenotypes. Thus, we examined whether the phenotypes *SATB2* affects are divergent between archaic and modern humans (e.g., if modern human jaw size is different than the jaw size of archaic humans). We focused on phenotypes available for examination from the fossil record, primarily skeletal differences between modern humans and Neanderthals. From HPO, we generated a list of 17 phenotypes known to be affected by *SATB2* and found that 88% (15) of them are divergent between these human groups (**Supplementary File 5**). These include the length of the skull, size of the jaws, and length of the dental arch. Next, based on *SATB*2

downregulation in modern humans predicted from lentiMPRA, we examined whether the direction of a phenotypic change between patients and healthy individuals matches the direction of phenotypic change between modern and archaic humans. For example, given that *SATB2*-associated syndrome patients have smaller jaws, we tested if modern human jaws are smaller compared to archaic humans. If *SATB2* expression is not in fact related to phenotypic divergence, there is a 50% likelihood for a given phenotype to match the fossil record. Yet, we observed a match in direction in 80% of the phenotypes (12 out of 15, **Supplementary File 5**). This includes smaller jaws, flatter face, and higher forehead in modern compared to archaic humans. Overall, the observed number of phenotypes that are both divergent and match in their direction of change is 2.3-fold higher than expected by chance ($P = 1.3\text{x}10^{-4}$, hypergeometric test, **Supplementary File 5,** see Methods). Together, these data support a model whereby the C→T substitution in the putative promoter of *SATB2*, which emerged and reached fixation in modern humans, possibly reduced the expression of *SATB2* and possibly affected brain and craniofacial phenotypes. However, further evidence is required to elucidate the potential role of this variant in modern human evolution.

Discussion

Identifying noncoding sequence changes underlying human traits is one of the biggest challenges in genetics. This is particularly difficult in ancient samples, where regulatory information is scarce[5,21]. Here, we use an MPRA-based framework to study how sequence changes shaped human gene regulation. By comparing modern to archaic sequences, we investigated the regulatory potential of each of the 14,042 single-nucleotide variants that emerged and reached

fixation or near fixation in modern humans. We found an association between divergent TF motifs and the sequences driving expression changes, suggesting that changes to TF binding might have played a central role in shaping divergent modern human expression. Our results also suggest that genes affecting the vocal tract and cerebellum might have been particularly affected by these expression changes, which is in line with previous comparisons based on the fossil record[1–4] and DNA methylation[12]. More importantly, these results provide candidate sequence changes underlying these evolutionary trends.

LentiMPRA is designed for linking DNA sequence changes to expression changes *en masse*. Notably, it has limitations that could influence our results, mainly by potentially generating false negatives. First, our lentiMPRA library inserts were limited to ~200bp in length, due to oligonucleotide synthesis technical restrictions, which may be insufficient to detect the activity of longer enhancer sequences[43]. Second, some minimally active sequences may not be expressed at a high enough level to pass our limit of detection. At the same time, some minimally active sequences may not be biologically significant. Third, some sequences may regulate expression post-transcriptionally, which lentiMPRA is not designed to detect. Fourth, since test sequences are randomly integrated into the genome, sequences that are dependent on their endogenous genomic environments (e.g., on nearby TF binding sites) might show reduced activity when inserted in new locations, while others might show activity that they otherwise would not have. Our design partially addresses this through the use of antirepressors and multiple independent integrations, which are intended to dilute location-specific effects. Additionally, all biases are expected to similarly affect the modern and archaic human versions of each sequence[43]. Fifth, transcriptional repression is less likely to be detected due to the low basal activity of the minimal promoter used. Sixth, the level of sequence activity may depend on more than one variant

(including non-fixed variants, which we have not tested here). In the cases of non-fixed variants, the extent of differential activity could vary between individuals. At the same time, in the 10% of sequences that include more than one fixed variant, it is generally impossible to determine which of the variants drives the differential activity (with the exception of cases with more than two variants where the tiled sequences include a different combination of these variants).

Finally, differences in the *trans* environment of a cell could have an effect on the ability of a sequence to exert its *cis*-regulatory effect, resulting in cell-type-specific cis-regulatory effects, as we observed in our data. The *trans* environment of the same cell type might also differ between two organisms. However, the majority of the *cis*-regulatory changes we observed would be expected to be present in archaic human cells as well, considering that such conservation has been observed between substantially more divergent organisms (e.g., human-chimpanzee[30] and human-mouse[84]). In other words, while *trans*-regulatory changes play a key role in species divergence, the *trans* environments of the same cell type in two closely related organisms tend to affect *cis*-regulation similarly. Despite these caveats, MPRAs have been repeatedly shown to be able to replicate the activity of sequences in their endogenous context[43–45].

Importantly, when genomes from additional modern human individuals are sequenced and new variants mapped, it might become clear that some of the variants we analyzed have not reached fixation. However, regardless of whether they are completely fixed or not, these variants represent derived substitutions that likely emerged in modern humans and spread to considerable frequency. Further investigation is required to determine when they emerged, how rapidly they spread, and whether their effect was neutral or adaptive.

As expected, we observed differences in activity and differential activity between cell types[28,45,84]. Although some of this variation is likely biological (i.e., cell type-specific gene regulation), it is difficult to determine what proportion of it is due to biological versus technical factors (e.g., differences in lentivirus preparation, infection rate, or cell growth, see Methods). Importantly, these differences are expected to result in false negatives rather than false positives. In other words, some of the sequences that appear as active or differentially active in one cell type might actually be active or differentially active in additional cell types (including cell types that were not tested in this study). Thus, we largely refrained from comparisons between cell types and the overlap observed in Fig. 2a and Fig. 3a should not be used to define such similarities. Rather, these diagrams should be used to examine the replicability of our results. Despite these caveats and limitations, lentiMPRA is a powerful high-throughput tool to characterize the regulatory activity of derived variants, and indeed has become a common assay to study the capability of sequences to promote expression[19].

With this method, we found that 1,791 (13%) of the 14,042 sequence pairs can promote expression in at least one of the three cell types tested, and that 405 (23%) of these active sequences show differential activity between modern and archaic humans (average fold-change: 1.24x, standard deviation: 0.18, **Fig. 2**, **Supplementary File 1**). Interpreting these results in light of previous MPRAs is challenging, not only because of key differences in statistical power and experimental design (e.g., sequence length), but also because of differing variant selection processes for each MPRA. With the exception of highly repetitive regions, which were removed from our library for technical reasons, the sequences we selected included all known modern human-derived fixed or nearly fixed variants (see Methods). Conversely, previous reporter assays and MPRAs on human intra- or inter-species variation used biased sets of variants by

selecting sequences with putative regulatory function (e.g., eQTLs[28], TF binding sites[16], ChIP-seq peaks[29], or TSSs[84]) and/or regions showing particularly rapid evolution (e.g., human accelerated regions[30,31,85,86]). In line with the fact that our data was not pre-filtered for putative regulatory regions, the proportion of active sequences we observed tends to be slightly lower than these previous studies. However, the magnitude of differential activity, as well as the fraction of differentially active sequences out of the active sequences was similar to previous studies[16,28–31,84–86]. At the same time, we were capable of measuring regulatory activity in regions that would otherwise be excluded by filtering for a specific set of marks. Thus, future MPRAs on unfiltered sets of variants will enable the comparison of the patterns we observed to patterns within modern humans, between more deeply divergent clades, and of non-fixed modern-archaic differences.

Our results also suggest that differentially active sequences are over-represented within putative enhancers in NPCs (**Supplementary Fig. 1c-d**, **Supplementary File 1**). Enhancers have been suggested to be an ideal substrate for evolution because of their tissue-specificity and temporal modularity[87]. Indeed, previous studies of introgression between archaic and modern humans suggested that enhancers are some of the most divergent regions between modern and archaic humans[11,25,88]. In line with the enrichment we observed in NPCs, brain-related putative enhancers show particularly low introgression, perhaps suggesting that the modern human sequences in these regions were adaptive[25,88]. To fully characterize the underlying mechanisms of differential activity in enhancers, it is important to disentangle the various factors and confounders that might contribute to this enrichment. There are several alternative explanations for the enrichment we observe, namely that variants within enhancers could be more likely to alter expression

compared to other active sequences, or they could be particularly detectable in lentiMPRA. This could be tested using saturation mutagenesis MPRAs[45] to compare the effect of random mutations in enhancer and non-enhancer modern human-derived active sequences.

Our results suggest that differentially active sequences are not randomly distributed across the genome, but rather tend to be linked to genes affecting particular body parts and phenotypes. The most pronounced enrichment was in the vocal tract, i.e., the vocal cords, larynx, and pharynx. This was evident in the Gene ORGANizer analysis, where these organs are over-represented by up to 5-fold, as well as in the HPO phenotype analysis, where some of the most enriched phenotypes are *nasal speech*, *palate development*, *nasal passage opening*, and *laryngeal stiffness* (**Fig. 4b**, **Supplementary File 4**). Overall, 53 of the 407 differentially active sequences were linked to genes which are known to affect one or more vocal tract phenotypes. Previous reports have also suggested that the vocal tract went through particularly extensive regulatory changes between modern and archaic humans[12], as well as between humans and chimpanzees[65,89]. Intriguingly, the anatomy of the vocal tract differs between humans and chimpanzees, and has been suggested to affect human phonetic range[90]. Comparing the anatomy of archaic and modern human larynges is currently impossible because the soft tissues of the larynx rapidly decay postmortem. However, together with these previous reports[12,65,89], our results enable the study of vocal tract evolution from a genetic point of view and suggest that genes influencing the modern human vocal tract have possibly gone through regulatory changes that are not shared by archaic humans.

We also identified an enrichment of brain-related phenotypes, particularly those affecting the size of the cerebellum (**Fig. 4**, **Supplementary File 4**). The cerebellum is involved in motor control and perception, as well as more complex functions such as cognitive processing, emotional regulation, language, and working memory[91]. Interestingly, the cerebellum has been described as the most morphologically divergent brain region between modern and archaic humans[1,4]. Evidence of divergent brain and cerebellar evolution can also be found at the regulatory level. Studies of Neanderthal alleles introduced into modern humans through introgression provide a clue as to the functional effects of divergent loci between archaic and modern humans. These works have shown that many of the introgressed sequences were likely negatively selected, with the strongest effect in regulatory regions[11,25], particularly in brain enhancers[88]. Studies of introgressed sequences have also shown that the cerebellum is one of the regions with the most divergent expression between Neanderthal and modern human alleles[10]. Together with our results, these data collectively suggest that sequences separating archaic and modern humans are particularly linked to functions of the brain, and especially the cerebellum.

Functional information on archaic human genomes is particularly challenging to obtain because of the postmortem decay of RNA and epigenetic marks in ancient samples. MPRA not only provides a new avenue to identify differential regulation in archaic samples, but also reveals the sequence changes underlying these differences. Here, we present a catalog providing regulatory insight into the sequence changes that separate modern from archaic humans. This resource will hopefully help assign functional context to various signatures of sequence divergence, such as selective sweeps and introgression deserts, and facilitate the study of modern human evolution through the lens of gene regulation.

Methods

Code and data availability

Code is available for download on Github: https://github.com/weiss19/AH-v-MH. Data was deposited in GEO under accession number: GSE152404.

Selection of fixed, derived variants and design of DNA oligonucleotides

We selected the variants for our lentiMPRA in the following manner. As a basis, we used the list of 321,820 modern human-derived single nucleotide changes reported to differ between modern humans and the Altai Neanderthal genome[33]. We then filtered this list to include only positions where the Vindija Neanderthal[34] and Denisovan sequences[36] both match the Altai Neanderthal variant, and are also not polymorphic in any of the four ape species examined (61 *Pan troglodytes,* 10 *Pan paniscus,* 15 *Gorilla beringei*, and 28 *Gorilla gorilla*)[37]. Next, we excluded loci which had any observed variation within modern humans in dbSNP, as annotated by Prüfer et al.[33] or in the 1000 Genomes project (phase 3)[38]. Finally, for technical limitations in downstream synthesis and cloning, we excluded variants at which the surrounding 200 base pairs (bp) had >25% repetitive elements as defined by RepeatMasker[92]. The resulting list contained 14,297 sequences and was used to design the initial set of DNA fragments. Upon completion of the lentiMPRA, another high-coverage Neanderthal genome (the Chagyrskaya Neanderthal) was published[35], and we subsequently also filtered out loci at which the Chagyrskaya Neanderthal genome did not match the ancestral sequence, bringing the final list of analyzed loci to 14,042 (28,082 archaic and modern sequences, **Supplementary File 1**).

We designed DNA fragments (oligonucleotides, hereinafter oligos) centered on each variant, including the 99 bp upstream and 100 bp downstream of each variant (200 bp total). For each variant we designed two fragments, one with the ancestral (archaic human and ape) sequence and one with the derived (modern human) sequence. For cases where two or more variants would be included in the same oligo, we used either derived-only (modern human) or ancestral-only (archaic human and ape) variants throughout the oligo. The average variants per oligo out of the 14,042 oligos was 1.1, with 12,680 containing one variant, 1,259 containing two, 96 containing three and seven containing four. We also included 100 negative control fragments, created by randomly picking 100 of the designed DNA fragments and scrambling their sequence (**Supplementary File 1**). Lastly, we incorporated 299 positive control fragments[30,85,93–101] (i.e., expected to drive expression; **Supplementary File 1**). As the library was infected into three cell types (see later), we designed positive controls for each of the cell types. For human embryonic stem cells (ESCs) and human neural progenitor cells (NPCs), we used sequences which were previously shown to drive expression in MPRA in each of these cell types (**Supplementary File 1**). For fetal osteoblast cells (Hobs), we used putative and confirmed enhancers from mouse and human (**Supplementary File 1**). 15 bp adapter sequences for downstream cloning were added to the 5' (5'-AGGACCGGATCAACT) and 3' (5'-CATTGCGTGAACCGA) ends of each fragment, bringing the total length of each fragment to 230 bp. We synthesized each fragment as an oligonucleotide through Agilent Technologies, twice independently to minimize synthesis errors (**Supplementary File 1**).

Production of the plasmid lentiMPRA library and barcode association sequencing

The plasmid lentiMPRA library was generated as described in Gordon et al. [32]. In brief, the two independently synthesized Agilent Technology oligo pools were amplified separately via a 5-cycle PCR using a different pairs of primers for each pool (forward primers, 5BC-AG-f01.1 and 5BC-AG-f01.2; reverse primers, 5BC-AG-r01.1 and 5BC-AG-r01.2; **Supplementary File 1**), adding a minimal promoter (mP) downstream of the test sequence. A second round of 5-cycle PCR was performed with the same primers for both pools (5BC-AG-f02 and 5BC-AG-r02; **Supplementary File 1**) to add a 15-bp random barcode downstream of the mP. The two pools were then combined at a 1:1 ratio and cloned into a doubled digested (AgeI/SbfI) pLS-SceI vector (Addgene, 137725) with NEBuilder HiFi Master Mix (NEB). The resulting plasmid lentiMPRA library was electroporated into 10-beta competent cells (NEB) using a Gemini X2 electroporation system (BTX) [2kv, 25uF, 200Ω] and allowed to grow up overnight on twelve 15cm 100 mg/mL carbenicillin LB agar plates. Colonies were pooled and midiprepped (Qiagen). We collected approximately 6 million colonies, such that ~200 barcodes were associated with each oligo on average. To determine the sequences of the random barcodes and which oligos they were associated with, we first amplified a fragment containing the oligo, mP and barcode from each plasmid in the lentiMPRA library using primers that contain Illumina flow cell adapters (P7-pLSmp-ass-gfp and P5-pLSmP-ass-i#, **Supplementary File 1**). We sequenced these amplified sequences with a NextSeq 150PE kit using custom primers (R1, pLSmP-ass-seq-R1; R2 (index read), pLSmP-ass-seq-ind1; R3, pLSmP-ass-seq-R2, **Supplementary File 1**) to obtain approximately 150M total reads. We later did a second round of barcode association sequencing of these fragments to obtain approximately 76M additional reads, for a combined total of 225,592,667 reads. To associate barcodes with oligos, we first mapped read pairs (R1

and R3) to the original list of 28,993 oligos using bowtie2 (--very-sensitive)[102]. Next, we filtered out pairs of reads that (1) did not map to the same oligo, (2) did not have at least one of the reads in the pair with a mapping quality of ≥ 6, or (3) did not have the "proper pair" SAM designation. We linked each pair of reads with the read covering its barcode (R2) and saved only those barcode reads having at least a quality score of 30 across all 15 bases in the R2 read. We removed any barcodes associated with more than a single unique oligo (i.e., "promiscuous" barcodes), as well as any barcodes where we did not see evidence of its oligo association at least three times. We then created a list of barcode-oligo associations – this final list comprised 3,495,698 unique barcodes spanning 28,678 oligos (98.9% of the original list of 14,297 variant sequence pairs, 100 negative sequences and 299 positive control sequences), which we refer to as the barcode-oligo association list.

Cell culture and differentiation

Human fetal osteoblasts were purchased from Cell Applications Inc. (406K-05f, tested negative for mycoplasma) and were maintained in osteoblast Growth Medium (Cell Applications Inc.). For passaging, cells were washed with 1x PBS, dissociated with Trypsin/EDTA (Cell Applications Inc.), and plated at approximately 5,000 cells/cm^2. H1-ESCs (embryonic stem cells, ESCs, WiCell WA-01, RRID:CVCL_9771, identity authenticated via STR profiling, and tested negative for mycoplasma) were cultured on Matrigel (Corning) in mTeSR1 media (STEMCELL Technologies) and medium was changed daily. For passaging, cells were dissociated using StemPro Accutase (Thermo Fisher Scientific), washed and re-plated on Matrigel-coated dishes at a dilution of 1:5 to 1:10 in mTeSR1 media supplemented with 10 μM Y-27632 (Selleck Chemicals). ESCs were differentiated into neural progenitor cells (NPCs) by dual-Smad

inhibition as previously described (Chambers et al., 2009; Inoue et al., 2019). Briefly, ESCs were cultured in mTeSR1 media until the cells became 80% confluent and then the media was replaced with neural differentiation media consisting of: KnockOut DMEM (Life Technologies) supplemented with KnockOut Serum Replacement (Life Technologies), 2 mM L-glutamine, 1x MEM-NEAA (Life Technologies), 1x beta-mercaptoethanol (Life Technologies), 200 ng/mL Recombinant mouse Noggin (R&D systems), and 10 μM SB431542 (EMD Millipore). On day 4 of differentiation, the neural differentiation media was gradually replaced by N2 media [DMEM/F12 (Thermo Fisher Scientific) supplemented with N2 (Thermo Fisher Scientific)] every 2 days (3:1 ratio on day 6, 1:1 on day 8 and 1:3 on day 10) while maintaining 200 ng/mL Noggin and 10 μM SB431542. On day 12, cells were dissociated into single-cell using TrypLE Express (Thermo Fisher Scientific) and cultured in N2B27 media [1:1 mixture of N2 media and Neurobasal media (Thermo Fisher Scientific) with B27 (Thermo Fisher Scientific)] supplemented with 20 ng/mL bFGF (R&D systems) and 20 ng/mL EGF (Millipore sigma)] on Matrigel-coated dish. NPCs were maintained in N2B27 with bFGF and EGF for a month and used for the following experiments at passage 15.

NPCs were validated through RT-qPCR at passage 1 (after one week of culturing in N2B27 media supplemented with bFGF and EGF) and at passage 10. RT-qPCR primers were designed for neural marker genes: *SOX1/2*, *NES* (*NESTIN*), *MAP2*; glial marker genes: *GFAP*, *OLIG2*; mesoderm marker genes: *T*(*BRA*), *GSC*; and endoderm marker genes: *SOX17*, *FOXA2* (**Supplementary File 1**). Expression of each marker was compared to *HPRT* expression (**Supplemental fig. 5h**). Additionally, validation via RNA-seq at passage 1 was performed. Results can be found in Figure 7A and 7D of Inoue, et al.[94] (data in GEO under accession number: GSE115046).

Cell line infection with lentiMPRA library, RNA- and DNA-seq and read processing

Lentivirus was produced and packaged with the plasmid lentiMPRA library in twelve 15cm dishes of HEK293T cells using the Lenti-Pac HIV expression packaging kit, following the manufacturer's protocol (GeneCopoeia). Additional lentivirus was produced as needed in batches of ten 15cm dishes. Lentivirus containing the lentiMPRA library (referred to hereafter as lentivirus) was filtered through a 0.45μm PES filter system (Thermo Scientific) and concentrated with Lenti-X concentrator (Takara Bio). Titration reactions using varying amounts of lentivirus were conducted on each cell type to determine the best volume to add, based on an optimal number of viral particles per cell, as described in Gordon et al.[32]. Lentiviral infection, DNA/RNA extraction, and barcode sequencing were all performed as described in Gordon et al.[32]. Briefly, each replicate consisted of approximately 9.6 million cells each of ESC and osteoblast, and 20 million cells of NPC. ESC and osteoblast cells were seeded into four 10cm dishes per replicate (with approximately 2.4 million cells in each dish), while NPCs were seeded into five 10cm dishes per replicate (with approximately 4 million cells per dish). Additional cells were used for NPCs due to decreased efficiency of DNA/RNA extraction in NPCs). Three replicates were

performed per cell type. Cells were infected with the lentiMPRA library at a MOI of 50 for NPCs and osteoblasts, and a MOI of 10 for ESCs. We used a lower MOI for ESC because the cells are very sensitive to infection and a MOI higher than 10 would result in cell death. For ESC and osteoblasts, cell media was changed to include 8ug/mL polybrene before the addition of the lentiMPRA library to increase infection efficiency. The media was replaced with growth media without polybrene approximately 24 hours after infection. Infected cells were grown for three days before combining the plates of each replicate for extraction of RNA and DNA via the Qiagen AllPrep mini kit (Qiagen). We subsequently purified mRNA from the RNA using the Oligotex mRNA prep kit (Qiagen) and synthesized cDNA from the resulting mRNA with SuperScript II RT (Invitrogen), using a primer containing a unique molecular identifier (UMI) (P7-pLSmp-ass16UMI-gfp, **Supplementary File 1**). DNA fragments were amplified from both the isolated DNA and generated cDNA, keeping each replicate and DNA type separate, with 3-cycle PCR using primers that include adapters necessary for sequencing (P7-pLSmp-ass16UMI-gfp and P5-pLSmP-5bc-i#, **Supplementary File 1**). These primers also contained a sample index for demultiplexing and a UMI for consolidating replicate molecules (see later). A second round of PCR was performed to amplify the library for sequencing using primers targeting the adapters (P5, P7, **Supplementary File 1**). The fragments were purified and further sequenced with six runs of NextSeq 15PE with 10-cycle dual index reads, using custom primers (R1, pLSmP-ass-seq-ind1; R2 (read for UMI), pLSmP-UMI-seq; R3, pLSmP-bc-seq; R4 (read for sample index), pLSmP-5bc-seq-R2, **Supplementary File 1**). Later, an additional two runs of 15PE of only the ESC samples were performed due to lower lentivirus infection efficiency in this cell type. Each samples' R1 and R3 reads (containing the barcode) were mapped with bowtie2 [[102]] (--very-sensitive) to the barcode-oligo association list. Next, we applied several quality filters on the

resulting alignments. We first filtered out read pairs that didn't map as proper pairs, and then ensured the mapped sequence completely matched the known barcode sequence by requiring that both R1 and R3 reads have CIGAR stings = 15M, MD flags = 15 and a mapping quality of at least 20. Next, we consolidated read abundance per barcode by selecting only reads with unique UMIs, the result being abundance counts for each barcode, across each replicate library of each cell type for both RNA and DNA.

Data was deposited in GEO under accession number: GSE152404.

Measurement of expression and differential expression

We used the R package MPRAnalyze[40] (version 1.3.1, https://github.com/YosefLab/MPRAnalyze) to analyze lentiMPRA data. To determine which oligos were capable of promoting expression, we modeled replicate information into both the RNA and DNA models of MPRAnalyze's quantification framework (rnaDesign = ~ replicate and dnaDesign = ~ replicate) and extracted alpha, the transcription rate, for each oligo. MPRAnalyze used the expression of our 100 scrambled oligos as a baseline against which to measure the level of expression of each tested oligo. We corrected the mean absolute deviation (MAD) score-based *P*-values from MPRAnalyze for multiple testing across tested oligos, including positive controls and excluding scrambled sequences, using the Benjamini-Hochberg method, thus generating an MAD score-based expression false discovery rate (FDR) for each oligo. For each variant and for each cell type, we looked at both the archaic and modern sequence oligos and assigned an oligo as potentially capable of driving expression if it had an FDR ≤ 0.05 in at least one sequence, and at least 10 barcodes in both sequences (**Supplementary File 1**). This left 2,097 sequences in ESCs, 1,059 in osteoblasts, and 664 in NPCs. Next, we applied a second test for activity, to

account for potential overestimation of active sequences in ESCs due to the lower lentiviral infection efficiency in these cells. We aggregated UMI-normalized read abundances across all barcodes of each oligo, across all replicates in a given cell type, and calculated a simple ratio of expression as RNA abundance normalized to DNA abundance (RNA/DNA ratio). Next, similarly to Kwasnieski et al.[103], we determined an RNA/DNA ratio threshold per cell type. This was done by first removing scrambled sequences that show RNA/DNA ratios >2 standard deviations away from the average RNA/DNA ratio of all of the scrambled sequences, as these likely represent oligos that are, by chance, capable of driving some expression. This left 95 scrambled sequences in ESCs, 94 in osteoblasts and 97 in NPCs. Then, we used the distribution of RNA/DNA ratios of the remaining scrambled sequences to assign an FDR for each of the non-scrambled oligos. FDR was calculated as the fraction of scrambled sequences that showed an RNA/DNA ratio as high or higher than each non-scrambled oligo. Only oligos that passed both tests described above (FDR ≤ 0.05 in each test) were considered as "active" (i.e., capable of driving expression). This resulted in 1,183 sequences in ESCs, 814 in osteoblasts and 602 in NPCs.

To measure differential expression between archaic and modern sequences, we used MPRAnalyze's comparative framework. In essence, this tool uses a barcode's RNA reads as an indicator of expression level and normalizes this to the DNA reads as a measure of the number of genomic insertions of that barcode (i.e., the number of fragments from which RNA can be transcribed). MPRAnalyze uses information across all the barcodes for both alleles of a given sequence, as well as information across all replicates. For the terms of the model, we included replicate information in the RNA, DNA and reduced (null) models, allele information in the RNA and DNA models, and barcode information only in the DNA model (rnaDesign = ~

replicate + allele, dnaDesign = ~ replicate + barcode + allele, reducedDesign = ~ replicate). We extracted *P*-values and the differential expression estimate (fold-change of the modern relative to archaic sequence). Then, we corrected the *P*-values of the set of active oligos (see above) for multiple testing with the Benjamini-Hochberg method to generate an FDR for each sequence. We set a cutoff of FDR ≤ 0.05 to call a sequence capable of driving differential expression. From this we generated, for each cell type, a list of sequences with differential expression between the archaic and modern alleles (**Supplementary File 1**).

Luciferase validation assays

Each assayed oligo was synthesized by Twist Biosciences and cloned into the pLS-mP-Luc vector (Addgene 106253) upstream of the luciferase gene. Lentivirus was generated independently for each vector using techniques as described for MPRA (see above), with the omission of the filtering and concentration step, which was replaced with the collection of the entirety of the cell culture media for use in subsequent infections. In addition, pLS-SV40-mP-Rluc (Addgene 106292), to adjust for infection efficiency, was added at a 1:3 ratio to the assayed vector for a total of 4ug for lentivirus production. We infected each cell type individually with each viral prep. The amount of lentivirus added was based on titrations in which varying amounts of a subset of viral preps were added to each cell type and cell death was observed 3 days post infection; the virus volume that produced between 30-50% death was used for subsequent experiments. Approximately 20,000 cells were plated in 96-well plates and grown for 24-48 hours (~70% confluent) before the addition of lentivirus. For osteoblasts and ESCs, 8ug/mL polybrene was added to the culture media at the same time as the addition of the lentivirus. The media was changed 24 hours after infection and cells were grown for an

additional 48 hours. The cells were then washed with PBS and lysed. Firefly and renilla luciferase expression were measured using the Dual-Luciferase Reporter Assay System (Promega) on the GloMax plate reader (Promega). Each oligo was tested using two biological replicates on different days and each biological replicate consisted of three technical replicates. Activity of a given oligo was calculated by normalizing the firefly luciferase activity to the renilla luciferase. We then calculated the $\log_2$ fold change (LFC) between the modern and archaic alleles as $\log_2$(modern / archaic). A full list of oligos tested and their LFC can be found in **Supplementary File 1**.

We found that the mean difference in fold-change between replicates was 3-fold lower for the differentially active vs other active sequences (0.22 vs 0.60), and that the variance of these differences was 9-fold lower for differentially active sequences compared to other active sequences (0.09 vs 0.83, **Supplementary File 1**), suggesting that differentially active sequences reflect a true biological signal.

Predicting target genes

To connect the surrounding locus of each variant to genes it potentially regulates, we combined four data sources. For each locus, we generated four types of gene lists, based on four largely complementary approaches: (1) overlap with known expression quantitative trait loci (eQTLs); (2) spatial interaction with promoters; (3) proximity to putative enhancers; and (4) proximity to a transcription start site (TSS, **Supplementary File 1**). Each data source was obtained and incorporated into each type of list as described below:

1) *Proximity to known eQTLs*

eQTLs are genetic variants between individuals shown to be associated with expression differences. We reasoned that the target genes of the sequence surrounding a variant are potentially similar to the target genes of nearby eQTLs. We downloaded eQTLs and their associated genes from GTEx[47] (www.gtexportal.org, v8 on August 26, 2019) and overlapped the locations of each eQTL with our list of sequences. We linked the target genes of any eQTLs within +/-1 kb to each variant. We used all tissue types reported by GTEx, for each cell type in the lentiMPRA. 9,503 out of the 14,042 loci were found within +/- 1 kb of an eQTL, with 83,777 eQTLS overall overlapping them.

2) Spatial interaction with a promoter via Hi-C data

High-throughput chromosome conformation capture (Hi-C) techniques map spatial interactions between segments of DNA. We reasoned that if a variant is found within or near a region that was shown to interact physically with a promoter, that variant could be in a region involved in regulating that promoter. We downloaded promoter capture Hi-C data from Jung et al.[69], containing a list of all the significant interactions between promoters and other segments of the genome across 27 tissue and cell types. We overlapped our variants with the locations of interacting genomic fragments to find interactions within +/-10 kb of each variant. We then linked each variant with the promoters that each interacting fragment was shown to contact. We repeated this process twice: once to obtain a cell type-specific list, and once to obtain a generic list. For the cell type-specific (stringent) list of locus-gene links, we included only those interactions observed in cell types corresponding to the cell lines used in our lentiMPRA: ESCs, NPCs and mesenchymal stem cells as an approximation for osteoblasts (given that osteoblast Hi-C data is not publicly available to the best of our knowledge, and that osteoblasts differentiate

from MSCs). For the generic (non-stringent) list, we used interactions across any of the 27 tissue and cell types analyzed by Jung et al.[69]. 4,688 out of the 14,042 loci overlapped at least one region that interacts with a promoter.

3) Putative enhancers

Lastly, we checked which of our variants were in previously reported putative enhancers. To this end, we downloaded the GeneHancer database[70] V4_12 and searched for putative enhancers within +/- 10kb of each of our variants, linking each variant to the target genes of each putative enhancer within that distance. GeneHancer provides "elite" or "non-elite" status to their defined enhancer-target gene connections depending on the strength of the evidence supporting each connection. Using this information, we repeated the process twice: once for the elite status and once for all annotations. 5,017 out of the 14,042 loci overlapped at least one putative enhancer

4) Promoters

Promoters were defined as the region 5kb upstream to 1kb downstream of GENCODE[104] v29 GRCh38 TSSs. If a variant fell within this region, we linked it to that TSS's gene. Each variant was assigned to all the promoters it fell within. 1,466 out of the 14,042 loci were found within a promoter.

Overall, 11,207 out of the 14,042 loci were linked to at least one putative target gene, with a median of four target genes per locus. 2,830 of the remaining loci were linked to their closest TSS, regardless of distance. The last 5 without hg38 coordinates for their closest TSS were not linked to a gene. Importantly, these links do not necessarily mean that these target genes are

regulated by these loci, but rather they serve as a list of potential target genes for the loci showing a regulatory function through lentiMPRA.

DNA methylation in active and differentially active sequences

The four highest resolution DNA methylation maps for modern and archaic bone samples were taken from Gokhman et al. 2014 [ref [20]] and Gokhman et al. 2020 [ref [12]]. Promoter sequences were defined as sequences within 5 kb upstream to 1 kb downstream of a TSS. CpG-poor promoter sequences were defined as promoter sequences ranking at the bottom half based on their CpG density. Enhancer sequences were defined as sequences annotated in chromHMM as putative enhancers (i.e., enhancers, genic enhancers, and bivalent enhancer) in osteoblast cells. In putative enhancer sequences we found a slightly weaker link between methylation and activity compared to promoter sequences, with 3% hypermethylation of downregulating sequences and 5% hypomethylation of upregulating sequences. Perhaps in accordance with the much weaker link between enhancer methylation and activity[48], this trend is not significant despite having similar statistical power to the promoter analysis ($P = 0.12$, paired t-test). To test whether our results might have been affected by CpG density, we compared CpG density in differentially active compared to non-differentially active sequences, and in upregulating compared to downregulating sequences. We found no significant difference in CpG density between these groups (P-values > 0.05, t-test).

The hypermethylation of downregulating sequences in modern compared to archaic humans, and the hypomethylation of upregulating sequences in modern compared to archaic humans is also observed to some extent when testing these sequences in NPCs, but not in ESCs. For example, the top 10 upregulating sequences are hypomethylated by 7% on average in modern compared to

archaic humans, top 10 downregulating sequences are hypermethylated by 13% in modern compared to archaic humans. This is in line with previous observations that differentially methylated regions tend to be shared across tissues[105].

Differential transcription factor binding sites

We predicted differences in binding of human transcription factors caused by each of our variants as follows. First, we downloaded the entire set of publicly available human transcription factor binding motifs (7,705 motifs, 6,608 publicly available) from the Catalogue of Inferred Sequence Binding Preferences (CIS-BP) database (http://cisbp.ccbr.utoronto.ca/), and filtered them to include only motifs labeled as *directly determined* (i.e., we filtered out inferred motifs), resulting in 4,351 motifs. Next, to enrich our mapping result for matches covering the variant location, we trimmed each of our oligo sequences containing a single variant to +/- 30 bp around the variant (the length of the longest motif). We did not trim oligos containing >1 variant. We used FIMO[54] to map each remaining motif to both the archaic and modern alleles of each trimmed sequence (or untrimmed, for sequences with >1 variant). A background model was generated using fasta-get-markov using the trimmed (or untrimmed, if >1 variant) sequences. For each motif mapping to both the archaic and modern alleles at the same strand and location, we required that at least one allele had a *q*-value (as supplied by FIMO) ≤ 0.05). Then, we found cases where the FIMO predicted binding score of a motif differed between the archaic and modern alleles. FIMO uses a *P*-value cutoff of 10^{-4} for reporting predicted binding. Therefore, some sequence pairs have a reported score for only one of the alleles. To assign these sequence pairs with a score difference, we used a conservative approach where we assigned the unscored allele with this lowest score reported for that motif, representing a score that is closest to a *P*-value of 10^{-4}. Because the unreported score could be anywhere below the lowest reported score,

but could not have been above it, this results in a conservative underestimation of the score difference. Finally, we linked each motif to the transcription factor (TF) it is most confidently associated with in CIS-BP, thereby generating lists of TFs that showed differential predicted binding for each sequence. For cases in which multiple unique motifs corresponded to the same TF, we used the motif with the largest score difference between alleles. TF enrichment analyses were done on all predicted differential TF binding sites for TFs with a minimum of 10 predicted differential sites. TFs that are not expressed in the cell types we examined in this study (FPKM < 1) were removed from the analyses. For TF expression in ESCs, we used ENCODE RNA-seq data for H1-hESC[75]. For osteoblast expression, data[106] was downloaded from GEO under accession number: GSE57925. For NPC expression, data[107] was downloaded from GEO under accession number: GSE115407. Fisher's exact test was used to compute enrichment of a TF among differentially active sequences compared to other active sequences. *P*-values were FDR-adjusted across all three cell lines combined.

To further test the enrichment of ZNF281, we examined various cutoffs of the number of predicted bound motifs, ranging from 5 to a maximum of 14 (the number of motifs predicted to be differentially bound by ZNF281) in steps of 1. We found that with the exception of the cutoffs of 5 and 6 (where ZNF281 is only slightly above the significance threshold: FDR = 0.058 and FDR = 0.053, respectively), ZNF281 is the only significant TF across all of these cutoffs (FDR ≤ 0.05). We repeated the same test for FPKM cutoffs, ranging from 0.5 to 3 in steps of 0.5, and found that ZNF281 is the only significantly enriched TF (FDR ≤ 0.05) across all of these cutoffs. For the predicted binding vs. expression correlation analysis, a cutoff of 10 sites per TF was used. *P*-values were computed using Pearson's correlation.

Overlapping loci with genomic features

The following datasets were used for the overlap analyses: GENCODE v28 GRCh38 human genome TSSs[108], GTEx v8 eQTLs[47], and broad peaks for the following histone modification marks: H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, and H3K27me, and the histone variant H2A.Z from the Roadmap Project for ESCs, ESC-derived NPCs, and osteoblasts[42]. We overlapped each of these datasets with the lists of inactive and active sequences, and computed enrichment *P*-values using a Fisher's exact test. We repeated this for various RNA/DNA cutoffs (1, 1.5, 2, 2.5, 3 and 3.5). Sex chromosomes were removed from the analyses. *P*-values were FDR-adjusted using the Benjamini-Hochberg procedure.

Sequence conservation within primates was taken from the Altai Neanderthal genome annotation, which used the PhyloP metric[33].

Human-chimpanzee *cis*-regulatory expression changes

We investigated the expression of genes associated with differentially active sequences by analyzing human and chimp RNA-seq data. As the expression changes we report are driven by *cis*-regulatory changes, we used our recently generated RNA-seq data from human-chimp hybrid cells[65] (GEO accession numbers: GSE146481 and GSE144825). In these hybrid cells, the human and chimpanzee chromosomes are found within the same nuclear environment and are exposed to the same trans factors (e.g., transcription factors). Therefore, any differential expression observed between the human and chimpanzee alleles within these hybrid cells is attributed to *cis*-regulatory changes. These cells are hybrid human-chimpanzee induced pluripotent stem cells (iPSCs), and we therefore investigated whether genes associated with upregulating sequences in our ESC lentiMPRA data tend to be upregulated in the hybrid iPSCs, and vice versa. It is

important to note that differential expression between humans and chimpanzees reflects ~12 million years of evolution (i.e., changes that emerged along the human as well as along the chimpanzee lineages since their split from their common ancestor ~6 million years ago). However, our lentiMPRA data was done on sequences that changed along the modern human lineage (~550-765 thousand years). Therefore, the human-chimpanzee differences span an evolutionary time that is ~20-fold longer than the modern human lineage, and the effect of modern-derived variants on gene expression between humans and chimpanzees is expected to be largely diluted by the many other changes that accumulated along the rest of this time. Indeed, we observe a very slight, but significant correlation between differential expression observed in the lentiMPRA data and differential expression observed in the human-chimp hybrid data (P = 0.017, Pearson's r = 0.1, **Supplementary Fig. 5g**).

Phenotype enrichment analyses

Body part enrichment analyses were conducted using Gene ORGANizer v13. The analyses were conducted on sequences driving increased expression, sequences driving decreased expression, and all differentially active sequences. This was done in each of the three cell types. We conducted these analyses using various $\log_2$(fold-change) thresholds: 0, 0.5, and 0.75, on the non-stringent locus-gene associations, and using a cutoff of 5 genes per term. Analyses were done against the active sequences as background, and using the ORGANizer tool with the *confident* option. *P*-values were FDR-adjusted using the Benjamini-Hochberg procedure across all three cell lines combined. For osteoblasts, non-skeletal organs were removed from the analyses. For NPCs, non-neuronal organs were removed.

For the HPO analyses, we used HPO[72] build 1268 (08 November, 2019), analyzing gene lists identical to the Gene ORGANizer analyses, with the exception of using a cutoff of 3 genes per

term, because fewer genes are linked to HPO terms than to Gene ORGANizer terms. Lists of phenotypes from HPO were generated for each variant through its linked genes. Hypergeometric test *P*-values were computed per phenotype and FDR-adjusted. Similarly to the Gene ORGANizer analysis, we removed non-skeletal phenotypes from the osteoblast results, and non-neuronal phenotypes from the NPC results.

Gene Ontology, Gene ORGANizer and HPO analyses were also done on the full set of genes linked to the 14,042 fixed variants using the same parameters described above (**Supplementary Table 7**). Importantly, unlike the analyses of differentially active sequences, which can be compared against a non-differentially active sequences background to control for potential biases, the full set of sequences cannot be compared against a background set. Therefore, these results may be affected by different confounders such as GC content, the ability to call SNPs, DNA degradation patterns, and it is still to be determined to what extent these results reflect true evolutionary trends.

SATB2 phenotypic analysis was done as previously described in Gokhman et al[14]. In short, we used HPO[72] build 1268 (08 November, 2019) to link phenotypes to *SATB2*. In addition, we conducted a literature search to expand gene-phenotype links to include studies that did not appear on HPO (**Supplementary File 5**). We used only skeletal directional phenotypes, i.e., phenotypes that could be described on a scale (e.g., smaller/larger hands), as these could be examined against the fossil record. This resulted in 34 phenotypes that are the result of *SATB2* heterozygous loss-of-function (LOF) (**Supplementary File 5**). Phenotypes that are included in another phenotype (e.g., *Prominent nasal bridge* and *Prominent nose*) were merged, and contradicting phenotypes (e.g., *Broad nose* and *Thin/small nose*) were removed. This resulted in

a final list of 17 phenotypes (**Supplementary File 5**). Given that the mechanism underlying these phenotypes is a decrease in the dosage of SATB2, and that *SATB2* is possibly downregulated in modern humans, we sought to investigate if similar phenotypes exist between modern human patients with *SATB2* heterozygous LOF and archaic humans. For each phenotype, we determined if it is divergent between the modern and archaic humans based on previously published annotation[14]. Then, for remaining divergent phenotypes, we tested if the direction between patients and healthy individuals matches the direction between modern and archaic humans. The significance of directionality match was computed using a binomial test, with a random probability of success $p = 0.5$. To compute the significance of the overall number of phenotypes that are divergent and match in direction, we compared the overall number of annotated divergent phenotypes to the number of divergent phenotypes associated with *SATB2* using a hypergeometric test. Out of a total of 696 annotated phenotypes between modern and archaic humans[14], 434 are annotated as divergent, and the direction of 50% of them (217 phenotypes) is expected to match by chance.

Acknowledgements

We would like to thank Tal Ashuach (MPRAnalyze), Terence Capellini, Evelyn Jagoda, Martin Kircher, and the Fraser, Petrov and McCoy labs for helpful feedback. D.G. was supported by the Human Frontier, Rothschild and Zuckerman fellowships. This work was supported in part by the National Human Genome Research Institute grant 1UM1HG009408 (N.A.), the National Institute of Mental Health grants 1R01MH109907 (N.A.) and 1U01MH116438 (N.A.), the Uehara Memorial Foundation (F.I.) and the Stanford Center for Computational, Evolutionary and Human Genomics (CEHG).

References

1. Neubauer, S., Hublin, J. J. & Gunz, P. The evolution of modern human brain shape. *Sci. Adv.* (2018). doi:10.1126/sciadv.aao5961
2. Gunz, P. *et al.* Neandertal Introgression Sheds Light on Modern Human Endocranial Globularity. *Curr. Biol.* (2019). doi:10.1016/j.cub.2018.10.065
3. Aiello, L. & Dean, C. *An Introduction to Human Evolutionary Anatomy*. (Elsevier, 2002).
4. Kochiyama, T. *et al.* Reconstructing the Neanderthal brain using computational anatomy. *Sci. Rep.* (2018). doi:10.1038/s41598-018-24331-0
5. Yan, S. M. & McCoy, R. C. Archaic hominin genomics provides a window into gene expression evolution. *Curr. Opin. Genet. Dev.* **62**, 44–49 (2020).
6. Britten, R. J. & Davidson, E. H. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q. Rev. Biol.* (1971). doi:10.1086/406830
7. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
8. Enard, D., Messer, P. W. & Petrov, D. A. Genome-wide signals of positive selection in human evolution. *Genome Res.* (2014). doi:10.1101/gr.164822.113
9. Fraser, H. B. Gene expression drives local adaptation in humans. *Genome Res.* **23**, 1089–1096 (2013).
10. McCoy, R. C., Wakefield, J. & Akey, J. M. Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression. *Cell* **168**, 916-927.e12 (2017).
11. Petr, M., Pääbo, S., Kelso, J. & Vernot, B. Limits of long-term selection against Neandertal introgression. *Proc. Natl. Acad. Sci. U. S. A.* (2019). doi:10.1073/pnas.1814338116
12. Gokhman, D. *et al.* Differential DNA methylation of vocal and facial anatomy genes in modern humans. *Nat. Commun.* **11**, 1189 (2020).
13. Colbran, L. L. *et al.* Inferred divergent gene regulation in archaic hominins reveals potential phenotypic differences. *Nat. Ecol. Evol.* (2019). doi:10.1038/s41559-019-0996-x
14. Gokhman, D. *et al.* Reconstructing Denisovan Anatomy Using DNA Methylation Maps. *Cell* **179**, 180-192.e10 (2019).

15. Dannemann, M. & Racimo, F. Something old, something borrowed: admixture and adaptation in human evolution. *Curr. Opin. Genet. Dev.* **53**, 1–8 (2018).
16. Weyer, S. & Pääbo, S. Functional analyses of transcription factor binding sites that differ between present-day and archaic humans. *Mol. Biol. Evol.* (2016). doi:10.1093/molbev/msv215
17. Vespasiani, D. M., Jacobs, G. S., Brucato, N., Cox, M. P. & Romero, I. G. Denisovan introgression has shaped the immune system of present-day Papuans. *bioRxiv* 2020.07.09.196444 (2020). doi:10.1101/2020.07.09.196444
18. Grogan, K. E. & Perry, G. H. Studying human and nonhuman primate evolutionary biology with powerful in vitro and in vivo functional genomics tools. *Evolutionary Anthropology* (2020). doi:10.1002/evan.21825
19. Chatterjee, S. & Ahituv, N. Gene Regulatory Elements, Major Drivers of Human Disease. *Annu. Rev. Genomics Hum. Genet.* (2017). doi:10.1146/annurev-genom-091416-035537
20. Gokhman, D. *et al.* Reconstructing the DNA methylation maps of the Neandertal and the Denisovan. *Science* **344**, 523–7 (2014).
21. Gokhman, D., Meshorer, E. & Carmel, L. Epigenetics: It's Getting Old. Past Meets Future in Paleoepigenetics. *Trends Ecol. Evol.* **31**, 290–300 (2016).
22. Barker, H. R., Parkkila, S. & Tolvanen, M. E. E. Evolution is in the details: Regulatory differences in modern human and Neanderthal. *bioRxiv* (2020). doi:doi.org/10.1101/2020.09.04.282749
23. Batyrev, D., Lapid, E., Carmel, L. & Meshorer, E. Predicted Archaic 3D Genome Organization Reveals Genes Related to Head and Spinal Cord Separating Modern from Archaic Humans. *Cells* (2019). doi:10.3390/cells9010048
24. Pedersen, J. S. *et al.* Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome Res.* **24**, 454–466 (2014).
25. Silvert, M., Quintana-Murci, L. & Rotival, M. Impact and Evolutionary Determinants of Neanderthal Introgression on Transcriptional and Post-Transcriptional Regulation. *Am. J. Hum. Genet.* (2019). doi:10.1016/j.ajhg.2019.04.016
26. Moriano, J. & Boeckx, C. Modern human changes in regulatory regions implicated in cortical development. *BMC Genomics* **21**, 304 (2020).
27. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays.

Genomics (2015). doi:10.1016/j.ygeno.2015.06.005

28. Tewhey, R. *et al.* Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* (2016). doi:10.1016/j.cell.2016.04.027
29. Klein, J. C., Keith, A., Agarwal, V., Durham, T. & Shendure, J. Functional characterization of enhancer evolution in the primate lineage. *Genome Biol.* (2018). doi:10.1186/s13059-018-1473-6
30. Ryu, H. *et al.* Massively parallel dissection of human accelerated regions in human and chimpanzee neural progenitors. *bioRxiv* (2018). doi:10.1101/256313
31. Uebbing, S. *et al.* Massively parallel discovery of human-specific substitutions that alter enhancer activity. *Proc. Natl. Acad. Sci. U. S. A.* (2021). doi:10.1073/pnas.2007049118
32. Gordon, M. G. *et al.* lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nat. Protoc.* (2020). doi:10.1038/s41596-020-0333-5
33. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–9 (2014).
34. Prüfer, K. *et al.* A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science (80-.).* **358**, 655–658 (2017).
35. Mafessoni, F. *et al.* A high-coverage Neandertal genome from Chagyrskaya Cave. *Proc. Natl. Acad. Sci.* (2020). doi:10.1073/pnas.2004944117
36. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–6 (2012).
37. De Manuel, M. *et al.* Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science (80-.).* (2016). doi:10.1126/science.aag2602
38. Auton, A. *et al.* A global reference for human genetic variation. *Nature* (2015). doi:10.1038/nature15393
39. Sherry, S. T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* (2001). doi:10.1093/nar/29.1.308
40. Ashuach, T. *et al.* MPRAnalyze: Statistical framework for massively parallel reporter assays. *Genome Biol.* (2019). doi:10.1186/s13059-019-1787-z
41. Ernst, J. & Kellis, M. ChromHMM: Automating chromatin-state discovery and characterization. *Nature Methods* (2012). doi:10.1038/nmeth.1906

42. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–329 (2015).
43. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* (2017). doi:10.1101/gr.212092.116
44. Klein, J. C. *et al.* A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat. Methods* (2020). doi:10.1038/s41592-020-0965-y
45. Kircher, M. *et al.* Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* (2019). doi:10.1038/s41467-019-11526-w
46. Wang, M., Zhao, Y. & Zhang, B. Efficient Test and Visualization of Multi-Set Intersections. *Sci. Rep.* (2015). doi:10.1038/srep16923
47. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (80-.).* **348**, 648–660 (2015).
48. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–92 (2012).
49. Blanco, E., González-Ramírez, M., Alcaine-Colet, A., Aranda, S. & Di Croce, L. The Bivalent Genome: Characterization, Structure, and Regulation. *Trends in Genetics* (2020). doi:10.1016/j.tig.2019.11.004
50. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* (2015). doi:10.1016/j.cell.2015.01.006
51. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* (2009). doi:10.1038/nature08514
52. Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* (2011). doi:10.1038/nature10716
53. Schlesinger, F., Smith, A. D., Gingeras, T. R., Hannon, G. J. & Hodges, E. De novo DNA demethylation and noncoding transcription define active intergenic regulatory elements. *Genome Res.* (2013). doi:10.1101/gr.157271.113
54. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* (2011). doi:10.1093/bioinformatics/btr064
55. Suske, G. NF-Y and SP transcription factors — New insights in a long-standing liaison. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* (2017).

doi:10.1016/j.bbagrm.2016.08.011

56. Frey-Jakobs, S. *et al.* ZNF341 controls STAT3 expression and thereby immunocompetence. *Sci. Immunol.* (2018). doi:10.1126/sciimmunol.aat4941
57. Bruderer, M., Alini, M. & Stoddart, M. J. Role of HOXA9 and VEZF1 in endothelial biology. *Journal of Vascular Research* (2013). doi:10.1159/000353287
58. Frietze, S., Lan, X., Jin, V. X. & Farnham, P. J. Genomic targets of the KRAB and SCAN domain-containing zinc finger protein 263. *J. Biol. Chem.* (2010). doi:10.1074/jbc.M109.063032
59. Song, J. *et al.* Transcriptional regulation by zinc-finger proteins Sp1 and MAZ involves interactions with the same cis-elements. *International journal of molecular medicine* (2003). doi:10.3892/ijmm.11.5.547
60. Zhu, C., Chen, G., Zhao, Y., Gao, X. M. & Wang, J. Regulation of the development and function of B cells by ZBTB transcription factors. *Frontiers in Immunology* (2018). doi:10.3389/fimmu.2018.00580
61. Ji, W., Mu, Q., Liu, X. Y., Cao, X. C. & Yu, Y. ZNF281-miR-543 Feedback Loop Regulates Transforming Growth Factor-β-Induced Breast Cancer Metastasis. *Mol. Ther. - Nucleic Acids* (2020). doi:10.1016/j.omtn.2020.05.020
62. Morita, K. *et al.* Emerging roles of Egr2 and Egr3 in the control of systemic autoimmunity. *Rheumatol. (United Kingdom)* (2016). doi:10.1093/rheumatology/kew342
63. Syafruddin, S. E., Mohtar, M. A., Nazarie, W. F. W. M. & Low, T. Y. Two sides of the same coin: The roles of KLF6 in physiology and pathophysiology. *Biomolecules* (2020). doi:10.3390/biom10101378
64. Pieraccioli, M. *et al.* ZNF281 inhibits neuronal differentiation and is a prognostic marker for neuroblastoma. *Proc. Natl. Acad. Sci. U. S. A.* (2018). doi:10.1073/pnas.1801435115
65. Gokhman, D. *et al.* Human-chimpanzee fused cells reveal cis-regulation underlying skeletal evolution. *Nat. Genet.* **in press**, (2021).
66. Gasperini, M. *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* (2019). doi:10.1016/j.cell.2018.11.029
67. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics* (2019). doi:10.1038/s41588-019-0538-0

68. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews Genetics* (2020). doi:10.1038/s41576-019-0209-0

69. Jung, I. *et al.* A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat. Genet.* (2019). doi:10.1038/s41588-019-0494-8

70. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford).* (2017). doi:10.1093/database/bax028

71. Gokhman, D. *et al.* Gene ORGANizer: Linking genes to the organs they affect. *Nucleic Acids Res.* **45**, W138–W145 (2017).

72. Köhler, S. *et al.* The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* **42**, (2014).

73. Zarate, Y. A. & Fish, J. L. SATB2-associated syndrome: Mechanisms, phenotype, and practical recommendations. *American Journal of Medical Genetics, Part A* (2017). doi:10.1002/ajmg.a.38022

74. Liu, S. H. *et al.* A novel antisense long non-coding RNA SATB2-AS1 overexpresses in osteosarcoma and increases cell proliferation and growth. *Mol. Cell. Biochem.* (2017). doi:10.1007/s11010-017-2953-9

75. Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

76. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gky1120

77. Claes, P. *et al.* Modeling 3D Facial Shape from DNA. *PLoS Genet.* **10**, e1004224 (2014).

78. Zarate, Y. A., Kaylor, J. & Fish, J. SATB2-Associated Syndrome. in (eds. Adam, M. P. et al.) (1993).

79. Gigek, C. O. *et al.* A molecular model for neurodevelopmental disorders. *Transl. Psychiatry* (2015). doi:10.1038/tp.2015.56

80. Qian, Y. *et al.* Paternal Low-Level Mosaicism-Caused SATB2-Associated Syndrome. *Front. Genet.* **10**, 630 (2019).

81. Li, Y. *et al.* Satb2 Ablation Impairs Hippocampus-Based Long-Term Spatial Memory and Short-Term Working Memory and Immediate Early Genes (IEGs)-Mediated Hippocampal

Synaptic Plasticity. *Mol. Neurobiol.* (2017). doi:10.1007/s12035-017-0531-5

82. Zhang, Q., Huang, Y., Zhang, L., Ding, Y. Q. & Song, N. N. Loss of satb2 in the cortex and hippocampus leads to abnormal behaviors in mice. *Front. Mol. Neurosci.* (2019). doi:10.3389/fnmol.2019.00033
83. Dobreva, G. *et al.* SATB2 Is a Multifunctional Determinant of Craniofacial Patterning and Osteoblast Differentiation. *Cell* (2006). doi:10.1016/j.cell.2006.05.012
84. Mattioli, K. *et al.* Cis and trans effects differentially contribute to the evolution of promoters and enhancers. *Genome Biol.* **21**, 210 (2020).
85. Prabhakar, S. *et al.* Human-specific gain of function in a developmental enhancer. *Science (80-.).* (2008). doi:10.1126/science.1159974
86. 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. B Biol. Sci.* (2013). doi:10.1098/rstb.2013.0025
87. True, J. R. & Carroll, S. B. Gene co-option in physiological and morphological evolution. *Annual Review of Cell and Developmental Biology* (2002). doi:10.1146/annurev.cellbio.18.020402.140619
88. Telis, N., Aguilar, R. & Harris, K. Selection against archaic hominin genetic variation in regulatory regions. *Nat Ecol Evol* (2020). doi:10.1038/s41559-020-01284-0
89. Prescott, S. L. *et al.* Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. *Cell* **163**, 68–84 (2015).
90. Lieberman, P. The Evolution of Human Speech: Its Anatomical and Neural Bases. *Curr. Anthropol.* **48**, 39–66 (2007).
91. Mariën, P. *et al.* Consensus paper: Language and the cerebellum: An ongoing enigma. *Cerebellum* (2014). doi:10.1007/s12311-013-0540-5
92. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-3.0. *RepeatMasker Open-3.0* (1996).
93. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser - A database of tissue-specific human enhancers. *Nucleic Acids Res.* (2007). doi:10.1093/nar/gkl822
94. Inoue, F., Kreimer, A., Ashuach, T., Ahituv, N. & Yosef, N. Identification and Massively Parallel Characterization of Regulatory Elements Driving Neural Induction. *Cell Stem*

Cell (2019). doi:10.1016/j.stem.2019.09.010

95. Hojo, H., Ohba, S., He, X., Lai, L. P. & McMahon, A. P. Sp7/Osterix Is Restricted to Bone-Forming Vertebrates where It Acts as a Dlx Co-factor in Osteoblast Specification. *Dev. Cell* (2016). doi:10.1016/j.devcel.2016.04.002
96. Meyer, M. B., Benkusky, N. A., Onal, M. & Pike, J. W. Selective regulation of Mmp13 by 1,25(OH)2D3, PTH, and Osterix through distal enhancers. *Journal of Steroid Biochemistry and Molecular Biology* (2016). doi:10.1016/j.jsbmb.2015.09.001
97. Khalid, A. B. *et al.* GATA4 represses RANKL in osteoblasts via multiple long-range enhancers to regulate osteoclast differentiation. *Bone* (2018). doi:10.1016/j.bone.2018.07.014
98. Khalid, A. B. *et al.* GATA4 Directly Regulates Runx2 Expression and Osteoblast Differentiation . *JBMR Plus* (2018). doi:10.1002/jbm4.10027
99. Loots, G. G. *et al.* Genomic deletion of a long-range bone enhancer misregulates sclerostin in Van Buchem disease. *Genome Res.* (2005). doi:10.1101/gr.3437105
100. Fukami, M., Kato, F., Tajima, T., Yokoya, S. & Ogata, T. Transactivation function of an ~800-bp evolutionarily conserved sequence at the SHOX 3′ region: Implication for the downstream enhancer. *American Journal of Human Genetics* (2006). doi:10.1086/499254
101. Kawane, T. *et al.* Runx2 is required for the proliferation of osteoblast progenitors and induces proliferation by regulating Fgfr2 and Fgfr3. *Sci. Rep.* (2018). doi:10.1038/s41598-018-31853-0
102. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* (2012). doi:10.1038/nmeth.1923
103. Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* (2014). doi:10.1101/gr.173518.114
104. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* (2012). doi:10.1101/gr.135350.111
105. Hernando-Herraez, I. *et al.* Dynamics of DNA Methylation in Recent Human and Great Ape Evolution. *PLOS Genet* **9**, e1003763 (2013).
106. Moriarity, B. S. *et al.* A Sleeping Beauty forward genetic screen identifies new genes and pathways driving osteosarcoma development and metastasis. *Nat. Genet.* **47**, 615–24

(2015).

107. Lu, L. *et al.* Robust Hi-C Maps of Enhancer-Promoter Interactions Reveal the Function of Non-coding Genome in Neural Development and Diseases. *Mol. Cell* (2020). doi:10.1016/j.molcel.2020.06.007

108. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gky955