The cis-regulatory effects of modern human-specific variants

Authors

Carly V. Weiss¹,*, Lana Harshman²,³,*”, Fumitaka Inoue²,³,⁴, Hunter B. Fraser¹, Dmitri A. Petrov¹,†,
†, Nadav Ahituv²,³,†, David Gokhman¹,†

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\textsuperscript{1–3} (especially the cerebellum\textsuperscript{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\textsuperscript{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\textsuperscript{6–9}. Indeed, there is mounting evidence that many of the noncoding variants that emerged in modern humans have altered gene expression in \textit{cis}, shaped phenotypes, and have been under selection\textsuperscript{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\(^1\). 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 with regulatory potential 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\textsuperscript{33–35}, one Denisovan genome\textsuperscript{36}, and 114 chimpanzee, bonobo, and gorilla genomes\textsuperscript{37}). We excluded any polymorphic sites within modern humans (in either the 1000 Genomes Project\textsuperscript{38} or in dbSNP\textsuperscript{39}), or within archaic humans and great apes\textsuperscript{33–37} (see Methods). The resulting set of 14,042 pairs of sequences comprises those changes that likely emerged and reached fixation or near fixation along the modern human lineage (Supplementary File 1). 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 variant 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 (see methods). We compared the activity (RNA/DNA) of the modern and archaic human constructs to identify variants promoting differential expression using MPRA

We synthesized a library composed of 200 base pair (bp) sequences per each of the 14,042 pairs of sequences (Fig. 1, Supplementary File 1). Each sequence contained at its center either the modern or archaic human variant. 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. We associated each sequence with multiple barcodes to achieve a high number of independent replicates of expression per sequence. 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. To 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. 97% of sequences had at least 10 barcodes associated with them, with a median of 96 barcodes per sequence (Supplementary Fig. 1a).

The brain and skeleton have been the focus of evolutionary studies due to their extensive phenotypic divergence within human lineages. Therefore, we chose human cells related to each of these central systems: NPCs and primary fetal osteoblasts. In addition, we used ESCs due to their globally active transcription, which provides a general view of gene expression. 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 ($R = 0.76 - 0.96$, $P < 10^{-100}$, Supplementary Fig. 1b). High barcode and read coverage in MPRA generally provides increased power to detect differences in allelic expression$^{32,43}$. 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$^{41}$, 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. 1c). 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 ($P = 2.7 \times 10^{-8}$ for ESCs and $P = 1.8 \times 10^{-6}$ for NPCs, $t$-test, see Methods, Supplementary Fig. 1d), 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 ($P = 1.1 \times 10^{-3}$). The correlation between replicates was also similar between osteoblasts and the other
two cell types (Supplementary Fig. 1b). Thus, the lack of activity of the osteoblast putative positive controls is likely because, in contrast to the ESC and NPC positive controls, the osteoblast putative positive controls were not previously tested in an MPRA and are not driving as much expression as the true positive controls in the ESCs and NPCs. Overall, these results suggest that the lentiMPRA was technically reproducible and adequately powered to detect expression.

Supplementary Figure 1. 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. Pearson’s R and P-values are shown. 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 vs. positive control sequences. One-sided t-test P-values are shown. ESC RNA/DNA values greater than 20 were set to 20. Black lines show the median.

Characterization of active regulatory sequences

We first examined which of the assayed sequences are able to drive expression. To do so, we utilized MPRAnalyze, 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, 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, Fig. 2a).
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 sequences tend to bear active chromatin marks in their endogenous genomic environment. Therefore, we examined whether active sequences in lentiMPRA tend to be enriched for markers of active chromatin. 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). We found that on average, active sequences were 1.7x more likely than inactive sequences to have an active chromatin mark. At the same time, these sequences tended to be depleted of repressive marks (Fig. 2b-d, Supplementary File 2). This trend gets stronger when looking at more highly active sequences. For example, while only 18% of inactive sequences overlap an H3K4me2 peak, 70% of active sequences with an RNA/DNA ratio > 3 in ESCs overlap H3K4me2 peaks ($FDR = 1.9 \times 10^{-21}$, hypergeometric test, Fig. 2b-d, Supplementary File 2). To further test the functional characteristics of active sequences, we analyzed chromHMM annotation, which uses chromatin signatures to subdivide the genome into functional regions. Compared to inactive sequences, we found that active sequences are enriched for promoter and enhancer marks (Supplementary Fig. 2a-c, Fig. 2e, Supplementary File 1-2). We also found that compared to inactive sequences, active sequences are 6-32% closer to GTEx 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).
Finally, compared to inactive sequences, active sequences show higher sequence conservation in primates, indicating a functional role (PhyloP, FDR = 1.1x10^{-3}, t-test, Supplementary File 2). In summary, we found that sequences that are capable of driving expression tend to overlap active chromatin marks, be depleted of repressive chromatin marks, be 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. **e.** Enrichment of differentially active sequences in various chromatin-based genomic annotations. Missing circles reflect no differentially active sequences in that category. Asterisks mark significant enrichments (FDR < 0.05).

**Supplementary Figure 2. Classification of chromHMM annotations for different groups of variants.** Relative percentage of bases in each chromHMM 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.

**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 ≤ 0.05, see Methods, **Fig. 3a-c, Supplementary File 1**, 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, $P = 5 \times 10^{-7}$, Super Exact test$^{44}$, **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\textsuperscript{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%, $P = 6.9 \times 10^{-3}$). For sequences that are differentially active in more than one cell type, we found an exceptional agreement between the direction of differential activity, with the regulation of 107 out of 109 pairs going in the same direction ($P = 9.2 \times 10^{-30}$, Binomial test), and a high correlation between the magnitudes of differential activity (Pearson’s $R = 0.82$, $P = 1.6 \times 10^{-27}$). That differentially activity 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. In the 10 sequences that were differentially active across all three cell types, the direction of fold-change was identical across all cell types \( (P = 1.9 \times 10^{-3}, \text{Binomial test}) \). 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 \( (P = 2.0 \times 10^{-3}, \text{binomial test}) \), and that the magnitude of their expression change is
highly correlated (Pearson’s R = 0.95, 2.4x10^{-5}, Supplementary Fig. 3a). 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, \( P = 3.7 \times 10^{-4} \), Supplementary Fig. 3b, 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. Out of these 407 sequences, 34\% (138) are within 10 kb of a TSS. Analyzing chromHMM\(^{46}\), we found that 8\% (32) of the sequences are within or flanking an active TSS, and 17\% (70) are within a putative enhancer. To test if differentially active sequences are enriched within specific chromHMM annotations, we compared the proportion overlapping each annotation to that proportion in the other active sequences. We found that sequences driving differential expression are over-represented within putative enhancer regions (\( P = 3.6 \times 10^{-3} \), Fisher’s exact test, see Discussion). This is mainly driven by enrichment within ESCs and NPCs (2.2x and 2.54x, \( P = 5.1 \times 10^{-3} \), and \( P = 2.0 \times 10^{-6} \), respectively, Supplementary Fig. 2c-d, Supplementary File 3, see Methods). These results support a model of rapid enhancer evolution in modern humans, as previously reported for other mammals\(^{48}\).
Molecular mechanisms underlying differential activity

Next, we sought to understand what regulatory mechanisms could be associated with differential activity. Changes in expression are often linked to changes in regulatory marks, such as DNA methylation. Thus, we checked the overlap between differentially active sequences and differential methylation in modern versus archaic humans\textsuperscript{12,20,21}. As hypermethylation of regulatory elements is often associated with transcriptional downregulation\textsuperscript{49}, we tested if higher activity levels in the lentiMPRA are associated with hypomethylation and vice versa. Because the DNA methylation maps originate from bone samples, we compared them to the osteoblast lentiMPRA data. We found that upregulating sequences indeed have a slight but significant tendency to be hypomethylated and that downregulating sequences tend to be hypermethylated in modern compared to archaic humans (-1% difference for upregulating sequences, and +2% for downregulating sequences, $P = 0.028$, paired $t$-test). This trend strengthens when looking at the most differentially regulating sequences. For example, the top ten most downregulating
sequences show on average +3.2% methylation in modern humans, whereas the top ten most
upregulating sequences show -7.4% methylation. We also found that differentially active
sequences tend to overlap previously reported differentially methylated regions more often than
expected by chance \( (P = 0.01, \text{overlap within 10 kb, hypergeometric test}) \).

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. 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. 4a**); (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 \( (P =
3.7x10^{-6} \text{ for ESCs}, P = 1.7x10^{-6} \text{ for osteoblasts, and } P = 1.3x10^{-5} \text{ for NPCs, binomial test, Fig. 3e, see Methods}); \) and (3) the magnitude of expression difference is correlated with the
magnitude of predicted binding difference (Pearson’s R = 0.43 and \( P = 1.2x10^{-3} \text{ for ESCs,}
Pearson’s R = 0.23 and \( P = 0.02 \text{ for osteoblasts, and Pearson’s R = 0.35 and } P = 2.4x10^{-3} \text{ for}
NPCs, **Supplementary Fig. 4b-d** and **Supplementary File 4**). 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. We found that changes to the motifs of 11 TFs in osteoblasts and 13 in NPCs were predictive of expression changes, i.e., higher affinity to the modern human sequences was predictive of higher expression of that sequence (Supplementary Fig. 4e, Supplementary File 4). 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 an activator function.

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 in differentially active sequences. We found two enriched motifs: ZNF281 (4.6-fold, FDR = 0.016, in NPCs) and SP3 (2.5-fold, FDR = 0.046, in NPCs, Supplementary File 4). ZNF281, among other functions, is an inhibitor of neuronal differentiation. SP3 is a ubiquitously expressed TF involved in a wide variety of processes including cell differentiation, proliferation, and synaptic gene regulation. Both TFs bind GC-rich sequences with overlapping binding specificities, evident also through the overlap between them in predicted differential binding (Supplementary File 4). Out of 153 differentially active sequences in NPCs, 23 (15%) are predicted to be differentially bound by SP3 (2.5-fold, FDR = 0.046), and 14 by ZNF281 (4.6-fold, FDR = 0.016). Intriguingly, ZNF281 and SP3 are also two of the TFs whose predicted differential binding is most tightly linked with differential expression (Supplementary Fig. 4e-g). Overall, these data support a model whereby variants in
ZNF281 and SP3 motifs might have modulated their binding in NPCs, thereby shaping neural expression differences between modern and archaic humans.

Supplementary Figure 4. Predicted TF binding is correlated with expression changes. a. Violin plots of absolute predicted TF binding score difference between modern and archaic sequences. Points show mean. b-d. Expression fold-change vs predicted TF binding fold-change. Positive scores represent increased binding in the modern sequence. Parentheses show number of points in each quadrant with a score difference > 0. e. Pearson’s correlation between differential expression and predicted differential binding affinity. Only significant TFs (FDR <= 0.05, Supplementary File 4) are shown for osteoblasts (yellow) and NPCs (red). f-g. Expression fold-change vs predicted TF binding fold-change for SP3 (e) and ZNF281 (f) in NPCs. Pearson’s correlation R and P-value are shown.

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\textsuperscript{55,56}, some exert their function through interactions with more distal genes, often reflected in
chromatin conformation capture assays, such as Hi-C interactions\textsuperscript{57}, or eQTL associations\textsuperscript{57,58}. To predict the genes linked to each sequence we combined data from four sources: (1) proximity to transcription start sites; (2) proximity to eQTLs\textsuperscript{47}; (3) proximity to putative enhancers\textsuperscript{59}; and (4) spatial interaction with promoters using Hi-C data\textsuperscript{58} (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 sequences driving differential expression, 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 terms \textit{G-protein coupled receptor signaling pathway} (2.9-fold, FDR = 0.019, in ESCs), and \textit{regulation of transcription, DNA-templated} (1.8-fold, FDR = 3.3x10\textsuperscript{-3}, in downregulating sequences across all cells combined, Supplementary File 4). 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\textsuperscript{60}. We analyzed the genes linked to differentially active sequences and found that for genes linked to sequences driving up-regulation, the most enriched body parts belong to the vocal tract, i.e., the vocal cords (5.0-fold, FDR = 1.1x10\textsuperscript{-3}), voice box (larynx, 3.8-fold, FDR = 4.2x10\textsuperscript{-3}), and pharynx (3.3-fold, FDR = 8.3x10\textsuperscript{-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
tract\textsuperscript{1,2}. Most subcranial, appendicular and axial skeletal parts are also over-represented within upregulating sequences. Conversely, within sequences driving downregulation, the enriched body parts are the cerebellum (3.0-fold, FDR = 3.9x10\textsuperscript{-3}, in NPCs) and urethra (2.3-fold, FDR = 0.04, in ESCs, \textbf{Fig. 4a, Supplementary File 5}). This is in line with previous reports of cerebellar anatomy differences between modern humans and Neanderthals\textsuperscript{1-3}, including results suggesting that the biggest differences in brain anatomy are in the cerebellum\textsuperscript{4}. These data also provide leads into the functional divergence of organs, like the urethra, 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\textsuperscript{61}, 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. Within upregulating sequences, the enriched phenotypes were involved in learning ability, craniofacial morphology, speech, hormonal activity, heart anatomy, testicular descent, osteoarthritis, muscle tone, and spine curvature (FDR < 0.05, \textbf{Fig. 4b}). Within downregulating sequences, we detected an enrichment of phenotypes affecting cerebellar size, intellectual ability, seizures, kidney function, and walking (FDR < 0.05, \textbf{Fig. 4b, Supplementary File 5}). 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). Results for the top enriched organs are labeled. These, as well as all remaining enriched organs, are colored according to the enrichment scale. For full results see Supplementary File 5.

b. HPO phenotypes significantly enriched within differentially active sequences. Fold enrichment and FDR are shown in parentheses. C. CpG islands and read density of active histone modification marks around the differentially active sequence in SATB2 (GRCh37 genome version)
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. First, we analyzed its linked variant (C to T transition), which is at a position that is highly 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 IncRNA which upregulates SATB2 protein levels. 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 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 = 6.8x10^-4, -27% in osteoblasts, FDR = 2.2x10^-42, and -12% in NPCs, FDR = 1.1x10^-7). 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. Genome-wide association studies (GWAS) show that common variants near *SATB2* are mainly associated with brain and bone phenotypes, such as reaction time, anxiety, mathematical abilities, schizophrenia, autism, and bone density. 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. 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 and mouse. Because these phenotypes are driven by changes to functional SATB2 levels, 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 6*). These include the length of the skull, size of the jaws, and length of the dental arch. Next, based on *SATB2* 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 6). 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 ($P = 1.3 \times 10^{-4}$, hypergeometric test, Supplementary File 6, see Methods).

Together, these data support a model whereby the C$\rightarrow$T substitution in the promoter of SATB2,
which emerged and reached fixation in modern humans, likely 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\textsuperscript{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 how each
of the 14,042 single-nucleotide variants that emerged and reached fixation or near fixation in
modern humans have affected expression. 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\textsuperscript{1–4} and DNA methylation\textsuperscript{12}. More importantly, these results provide candidate sequence changes underlying these evolutionary trends.

LentiMPRA is designed for linking DNA sequence changes to expression changes\textit{ 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\textsuperscript{41}. Second, some minimally active sequences may not be expressed at a high enough level to pass our limit of detection. 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 insulators 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\textsuperscript{41}. Fifth, transcriptional repression is less likely to be detected due to the low basal activity of the minimal promoter used. Finally, differences in the\textit{ trans} environment of a cell could have an effect on the ability of a sequence to exert its\textit{ cis}-regulatory effects. However, previous data comparing substantially more divergent organisms (e.g., human-chimpanzee\textsuperscript{30} and human-mouse\textsuperscript{72}) show that\textit{ cis}-regulatory changes exert their effects similarly across different
trans environments, suggesting the potential effect of such trans differences is likely minimal between modern and archaic humans.

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. 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. 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. 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: 27
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\textsuperscript{28}, TF binding sites\textsuperscript{16}, ChIP-seq peaks\textsuperscript{29}, or TSSs\textsuperscript{72}) and/or regions showing particularly rapid evolution (e.g., human accelerated regions\textsuperscript{30,31,73,74}). In line with the fact that our data was not pre-filtered for putative regulatory regions, the proportion of active and differentially active sequences we observed tends to be slightly lower than these previous studies, but the magnitude of differential expression in our active sequences was similar\textsuperscript{16,28–31,72–74}. 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.

Emergence of new regulatory elements is a common phenomenon in mammalian genomes\textsuperscript{48,75}. Indeed, some instances of differential activity in our dataset could potentially be explained by the emergence of entirely new regulatory sequences: in 22-34% (depending on cell type) of differentially active sequences, only the derived sequence promotes activity (**Supplementary File 1**). Some of these cases could potentially be the result of a new TF binding site. This is
supported by the observation that 20 of these sequences show significant predicted TF binding only in the derived allele (Supplementary File 4).

Our results also suggest that sequences driving differential expression are over-represented within putative enhancers, especially in ESCs and NPCs (Supplementary Fig. 2c-d, Supplementary File 3). Enhancers have been suggested to be an ideal substrate for evolution because of their tissue-specificity and temporal modularity. 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. 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. 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 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 both with Gene ORGANizer, where these organs are over-represented by up
to 5-fold, and also by HPO phenotype analysis, where the most enriched phenotype was nasal speech (a phenotype that occurs when there is an abnormal closure of the soft palate against the walls of the pharynx, 13-fold, FDR = 3x10^{-4}, Fig. 4b, Supplementary File 5). 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\(^\text{12}\), as well as between humans and chimpanzees\(^\text{78}\). Intriguingly, the anatomy of the vocal tract differs between humans and chimpanzees, and has been suggested to affect human phonetic range\(^\text{79}\). 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\(^\text{12,78}\), 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 5). 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\(^\text{80}\). Interestingly, the cerebellum has been described as the most morphologically divergent brain region between modern and archaic humans\(^\text{1,4}\). 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\textsuperscript{11,25}, particularly in brain
enhancers\textsuperscript{77}. 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\textsuperscript{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

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\textsuperscript{33}. We then filtered this list to include only positions where the Vindija Neanderthal\textsuperscript{34}, and Denisovan\textsuperscript{36} match the Altai Neanderthal variant (or have no information), and are also not polymorphic in any of the four ape species examined (61 \textit{Pan troglodytes}, 10 \textit{Pan paniscus}, 15 \textit{Gorilla beringei}, and 28 \textit{Gorilla gorilla})\textsuperscript{37}. Next, we excluded any loci which had any observed variation within modern humans in dbSNP, as annotated by Prüfer et al.\textsuperscript{33} or in the 1000 Genomes project (phase 3)\textsuperscript{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\textsuperscript{81}. 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\textsuperscript{35}, and we subsequently decided to also filter 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 ancestral and derived 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 sequence and one with the derived sequence. For cases where two or more variants would be included in the same oligo, we used either derived-only or ancestral-only 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 (i.e., expected to drive expression). 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’-CATTGCCTGAACCGA) 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)\textsuperscript{82}. 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) 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/cm². H1-ESCs (embryonic stem cells, ESCs, WiCell WA-01, RRID:CVCL_9771) 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 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.

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.\textsuperscript{32}. Lentiviral infection, DNA/RNA extraction, and barcode sequencing were all performed as described in Gordon et al.\textsuperscript{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 decreased efficiency of DNA/RNA extraction in NPCs. Three replicates were
performed per cell type. Cells were infected with the lentiMPRA library at a multiplicity of infection (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 [82] (--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 (reviewer access token: avujcogkdixizij).

Measurement of expression and differential expression

We used the R package MPRAnalyze\(^4\) (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 ancestral and derived sequence oligos and assigned an oligo as potentially capable of driving expression if it had an FDR \(\leq 0.05\) in at least one sequence, and at least 10 barcodes in both sequences (Supplementary File 1). 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, 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 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).

To measure differential expression between ancestral and derived 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 derived relative to ancestral 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 ancestral and derived 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 log2 fold change (LFC) between the ancestral and derived alleles as log2(derived / ancestral). 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 GTEx47 (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.\textsuperscript{58}, 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 list of sequence-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 list, we used interactions across any of the 27 tissue and cell types analyzed by Jung et al.\textsuperscript{58}. 4,688 out of the 14,042 loci overlapped at least one region that interacts with a promoter.
Lastly, we checked which of our variants were in previously reported putative enhancers. To this end, we downloaded the GeneHancer database\textsuperscript{59} 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\textsuperscript{83} 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.
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 to map each remaining motif to both the ancestral and derived alleles of each trimmed sequence (or untrimmed, for sequences with >1 variant). For each motif mapping to both the ancestral and derived alleles at the same strand and location, we required that at least one allele had a q-value (as supplied by FIMO) \( \leq 0.05 \). For motifs where only one of the ancestral or derived alleles had a reported mapping from FIMO for a given oligo, we searched all the mappings of that motif for the lowest mapping score, and substituted that score for the missing allele’s score. Then, we found cases where the FIMO predicted binding score of a motif differed between the ancestral and derived alleles. 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. This was done for each cell type, as well as for the union of all differentially active sequences across all cell types.
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. A threshold of 10 sites per TF was used for the analysis of correlation between predicted binding and expression.

Overlapping loci with genomic features

The following datasets were used for the overlap analyses: GENCODE v28 GRCh38 human genome TSSs, GTEx v8 eQTLs, 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. We overlapped each of these datasets with the lists of non-active and active sequences, and computed enrichment P-values using a hypergeometric 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.

Phenotype enrichment analysis

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, as well as on the union of all three cell types. We repeated the analyses using various log2(fold-change) thresholds: 0, 0.25, 0.5, 0.75 and 1, and on both the stringent and non-stringent locus-gene associations. Analyses were done against the Gene ORGANizer genomic background using the
ORGANizer tool with the *confident* option. *P*-values were FDR-adjusted using the Benjamini-Hochberg procedure. For osteoblasts, non-skeletal organs were removed from the analyses. For NPCs, non-neuronal organs were removed.

For the HPO analyses, we used HPO\textsuperscript{61} build 1268 (08 November, 2019), analyzing gene lists identical to the Gene ORGANizer analyses. 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 (GO) analyses on the same groups of genes described above revealed two enriched terms: GO:0007186, G-protein coupled receptor signaling pathway (2.9-fold within ESCs, up- and downregulating sequences, log\textsubscript{2} Fold-change cutoff: 0.25, FDR = 0.018), and GO:0006355, regulation of transcription, DNA-templated (1.8-fold within the union of cell types, downregulating sequences, log\textsubscript{2} Fold-change cutoff: 0.25, FDR = 3.3x10\textsuperscript{-3}).

*SATB2* phenotypic analysis was done as previously described in Gokhman et al\textsuperscript{14}. In short, we used HPO\textsuperscript{61} 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 6). 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 6). 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 6). 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 (MPRAalyze), 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


doi:10.1093/molbev/msv215
doi:doi.org/10.1101/2020.09.04.282749
27. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays.


68. Qian, Y. *et al.* Paternal Low-Level Mosaicism-Caused SATB2-Associated Syndrome.


