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.

52 Unfortunately, our ability to infer the regulatory function of noncoding variants is currently
53 limited[19]. In archaic humans, incomplete information on gene regulation is further exacerbated
54 by the lack of RNA molecules and epigenetic marks in these degraded samples[5]. We have
55 previously used patterns of cytosine degradation in ancient samples to reconstruct whole-genome
56 archaic DNA methylation maps[12,20,21]. However, despite various approaches to extract regulatory
57 information from ancient genomes[5,13,21–26], our understanding of gene regulation in archaic
58 humans remains minimal, with most archaic regulatory information being currently
59 inaccessible[5]. Additionally, whereas expression quantitative locus (eQTL) mapping can be used
60 to identify variants that drive differential expression between individuals, it can only be applied
61 to loci that are variable within the present-day human population. Therefore, fixed noncoding
62 variants are of particular interest in the study of human evolution, but are also particularly
63 difficult to characterize.
64
65 Massively parallel reporter assays (MPRAs) provide the ability to interrogate the regulatory
66 effects of thousands of variants *en masse*[27]. By cloning a candidate regulatory sequence
67 downstream to a short transcribable sequence-based barcode, thousands of sequences and
68 variants can be tested for regulatory activity in parallel. Thus, MPRA is an effective high-
69 throughput tool to identify variants underlying divergent regulation, especially in organisms
70 where experimental options are limited[28–31]. Here, we conducted a lentivirus-based MPRA
71 (lentiMPRA[32]) on the 14,042 fixed or nearly fixed single-nucleotide variants that emerged along
72 the modern human lineage. We generated a library of both the derived (modern human) and
73 ancestral (archaic human and ape) sequences of each locus and expressed them in three human
74 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[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 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*.

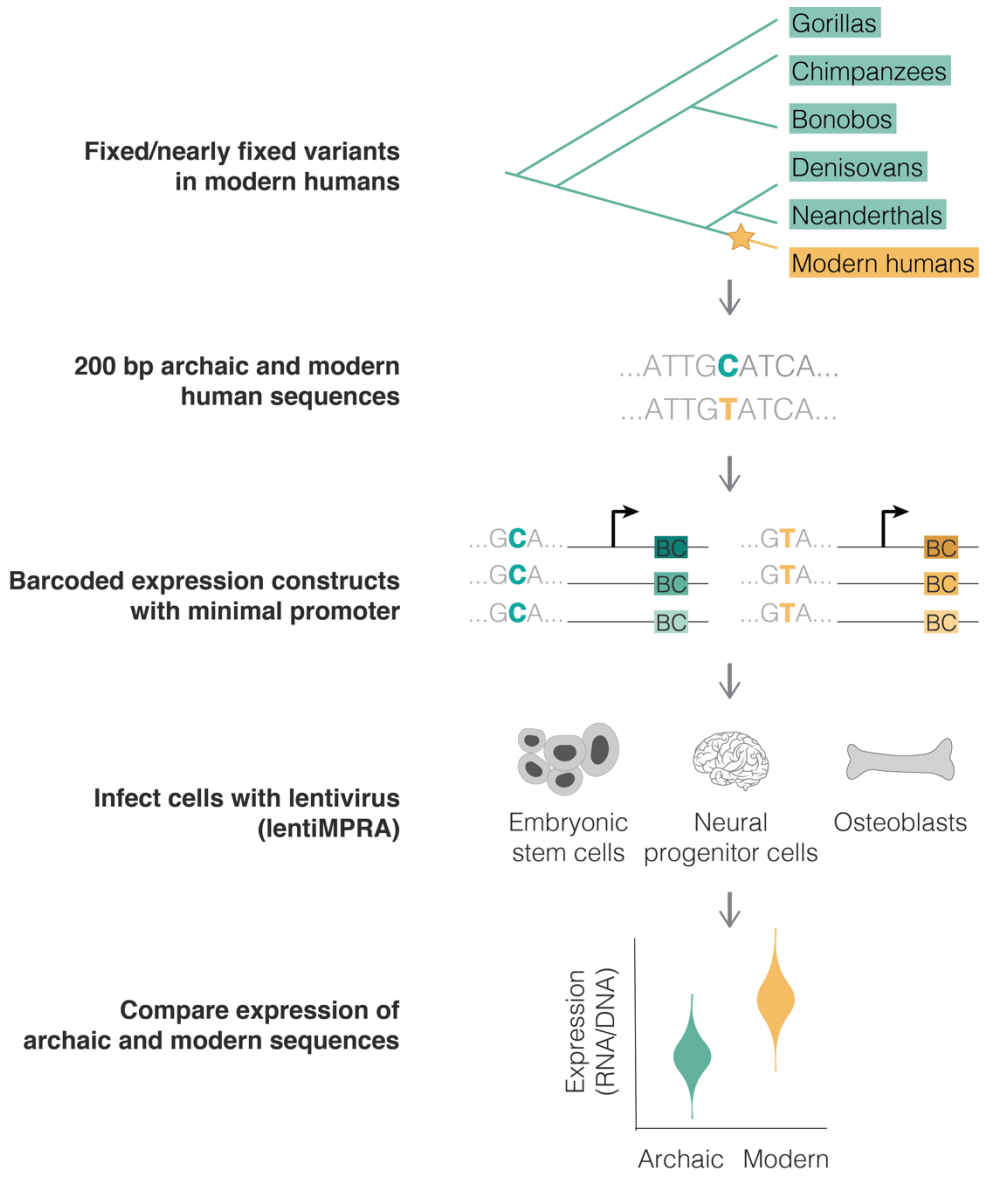


Figure 1. Using lentiMPRA to identify variants driving differential expression in modern humans. We analyzed variants that likely emerged and reached fixation or near fixation along the modern human lineage (yellow) and that were not polymorphic in any other ape or archaic genome (green) (top). The modern 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[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).

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

109 minimal promoter and barcode. We then inserted these constructs into the lentiMPRA vector, so
110 that the barcode, which is the readout of activity, is located within the 5'UTR of the reporter
111 gene and is transcribed if the assayed sequence is an active regulatory element[32]. We associated
112 each sequence with multiple barcodes to achieve a high number of independent replicates of
113 expression per sequence. Furthermore, we used a chromosomally integrating construct rather
114 than an episomal construct due to the improved technical reproducibility and correlation of
115 results from chromosomally integrating constructs with functional genomic signals like
116 transcription factor ChIP-seq and histone acetylation marks[41]. To reduce lentivirus site-of-
117 integration effects, this vector contained antirepressors on either side and was integrated in
118 multiple independent sites, with each sequence marked by multiple barcodes. 97% of sequences
119 had at least 10 barcodes associated with them, with a median of 96 barcodes per sequence
120 (**Supplementary Fig. 1a**).
121
122 The brain and skeleton have been the focus of evolutionary studies due to their extensive
123 phenotypic divergence within human lineages[3]. Therefore, we chose human cells related to each
124 of these central systems: NPCs and primary fetal osteoblasts. In addition, we used ESCs due to
125 their globally active transcription, which provides a general view of gene expression[42]. We used
126 the library of 14,042 pairs of archaic and modern human sequences, together with positive and
127 negative control sequences, to infect each cell type. As positive controls for ESCs and NPCs, we
128 added a set of 199 sequences with known regulatory capacity from previous MPRAs
129 (**Supplementary File 1**). To our knowledge, there have not been any MPRAs conducted in
130 osteoblasts, so we searched the literature for putative regulatory regions in osteoblasts and other
131 bone cell types and used these as putative positive controls (**Supplementary File 1**, see

132 Methods.). As negative controls, in all cell types, we randomly chose 100 sequences from the
133 library and scrambled the order of their bases, creating a set of GC-content matching sequences
134 that had not been previously established to drive expression.
135
136 We performed three replicates of library infection in each cell type and quantified barcode
137 abundance for each sequence in RNA and DNA (**Fig. 1**). To assess the reproducibility of our
138 lentiMPRA results, we calculated the RNA/DNA ratio (a measure of expression normalized to
139 the number of integrated DNA molecules) for each sequence and compared it across the three
140 replicates per cell type. We saw a strong correlation of RNA/DNA ratios between replicates for
141 all cell types (R = 0.76 – 0.96, $P < 10^{-100}$, **Supplementary Fig. 1b**). High barcode and read
142 coverage in MPRA generally provides increased power to detect differences in allelic
143 expression[32,43]. Thus, to determine how variability depended on our barcode counts, we
144 downsampled the number of barcodes per sequence and calculated the RNA/DNA ratio at each
145 step for each of the three replicates. In agreement with previous studies[41], we found that the
146 number of barcodes used in this study is well within the plateau, suggesting that the number of
147 barcodes is not a limiting factor in our experiment (**Supplementary Fig. 1c**). Finally, we
148 assessed the distribution of RNA/DNA ratios across our scrambled sequences and positive
149 controls. The mean RNA/DNA ratio of the scrambled sequences was lower than that of the
150 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,
151 *t*-test, see Methods, **Supplementary Fig. 1d**), but not in osteoblasts ($P = 0.25$). This is unlikely
152 due to a problem with the osteoblasts, as the osteoblast-related controls show similar expression
153 in all three cell types. Moreover, ESC and NPC positive controls are active in osteoblasts ($P =$
154 $1.1\text{x}10^{-3}$). The correlation between replicates was also similar between osteoblasts and the other

155 two cell types (**Supplementary Fig. 1b**). Thus, the lack of activity of the osteoblast putative
156 positive controls is likely because, in contrast to the ESC and NPC positive controls, the
157 osteoblast putative positive controls were not previously tested in an MPRA and are not driving
158 as much expression as the true positive controls in the ESCs and NPCs. Overall, these results
159 suggest that the lentiMPRA was technically reproducible and adequately powered to detect
160 expression.

161

162 **Supplementary Figure 1. Reproducibility of lentiMPRA data. a.** Distribution of number of barcodes per each
163 sequence. **b.** Replicate-by-replicate correlation of expression (RNA/DNA). Each point represents an active
164 sequence. Pearson's R and *P*-values are shown. **c.** Simulations of barcode down-sampling showing Pearson's
165 correlation of expression (RNA/DNA) between replicates. Upper panel shows all sequences and lower panel shows
166 sequences with higher expression (RNA/DNA > 3). Pearson's R values are normalized to maximum Pearson's R
167 observed for each pair of replicates. **d.** Box plots of scrambled vs. positive control sequences. One-sided *t*-test *P*-
168 values are shown. ESC RNA/DNA values greater than 20 were set to 20. Black lines show the median.

169

170 Characterization of active regulatory sequences

171 We first examined which of the assayed sequences are able to drive expression. To do so, we
172 utilized MPRAnalyze[40], which uses a model for each of the RNA and DNA counts, estimates
173 transcription rate and then identifies sequences driving significant expression. We also added an
174 additional stringency filter whereby a sequence is only considered expressed if it had an
175 RNA/DNA ratio significantly higher than that of the scrambled sequences (FDR < 0.05). We
176 found that in ESCs, 8% (1,183) of sequence pairs drove expression in at least one of the alleles,
177 6% (814) in osteoblasts, and 4% (602) in NPCs (FDR < 0.05, see Methods). Hereinafter, we
178 refer to these sequences as *active* sequences. Overall, 13% (1,791) of archaic and modern human
179 sequence pairs were active in at least one cell type, 4% (586) in at least two cell types, and 2%
180 (222) in all three cell types (overlap of 75-fold higher than expected, $P < 10^{-100}$, Super Exact
181 test[44], **Fig. 2a**).

182

183 To test whether activity in our lentiMPRA reflects true biological function, we investigated
184 whether our active sequences had expected regulatory characteristics in the modern human
185 genome. Active sequences tend to bear active chromatin marks in their endogenous genomic
186 environment. Therefore, we examined whether active sequences in lentiMPRA tend to be
187 enriched for markers of active chromatin. We first tested overlap with five histone modification
188 marks and one histone variant associated with active chromatin (H3K27ac, H3K4me1,
189 H3K4me2, H3K4me3, H3K9ac, and H2A.Z), as well as with two histone modification marks
190 associated with repressed chromatin (H3K9me3 and H3K27me3, see Methods)[45]. We found that
191 on average, active sequences were 1.7x more likely than inactive sequences to have an active
192 chromatin mark. At the same time, these sequences tended to be depleted of repressive marks
193 (**Fig. 2b-d, Supplementary File 2**). This trend gets stronger when looking at more highly active
194 sequences. For example, while only 18% of inactive sequences overlap an H3K4me2 peak, 70%
195 of active sequences with an RNA/DNA ratio > 3 in ESCs overlap H3K4me2 peaks (*FDR* =
196 1.9×10^{-21}, hypergeometric test, **Fig. 2b-d, Supplementary File 2**). To further test the functional
197 characteristics of active sequences, we analyzed chromHMM annotation[45,46], which uses
198 chromatin signatures to subdivide the genome into functional regions. Compared to inactive
199 sequences, we found that active sequences are enriched for promoter and enhancer marks
200 (**Supplementary Fig. 2a-c**, **Fig. 2e, Supplementary File 1-2)**. We also found that compared to
201 inactive sequences, active sequences are 6-32% closer to GTEx[47] eQTLs, depending on cell type
202 (FDR < 0.05, *t*-test). Active sequences are also 1.2-1.3x closer to transcription start sites (TSSs),
203 with 32-39% of them located within 10 kb of a TSS, depending on cell type (FDR < 0.05, *t*-test,
204 **Supplementary File 2**).

205

206 Finally, compared to inactive sequences, active sequences show higher sequence conservation in

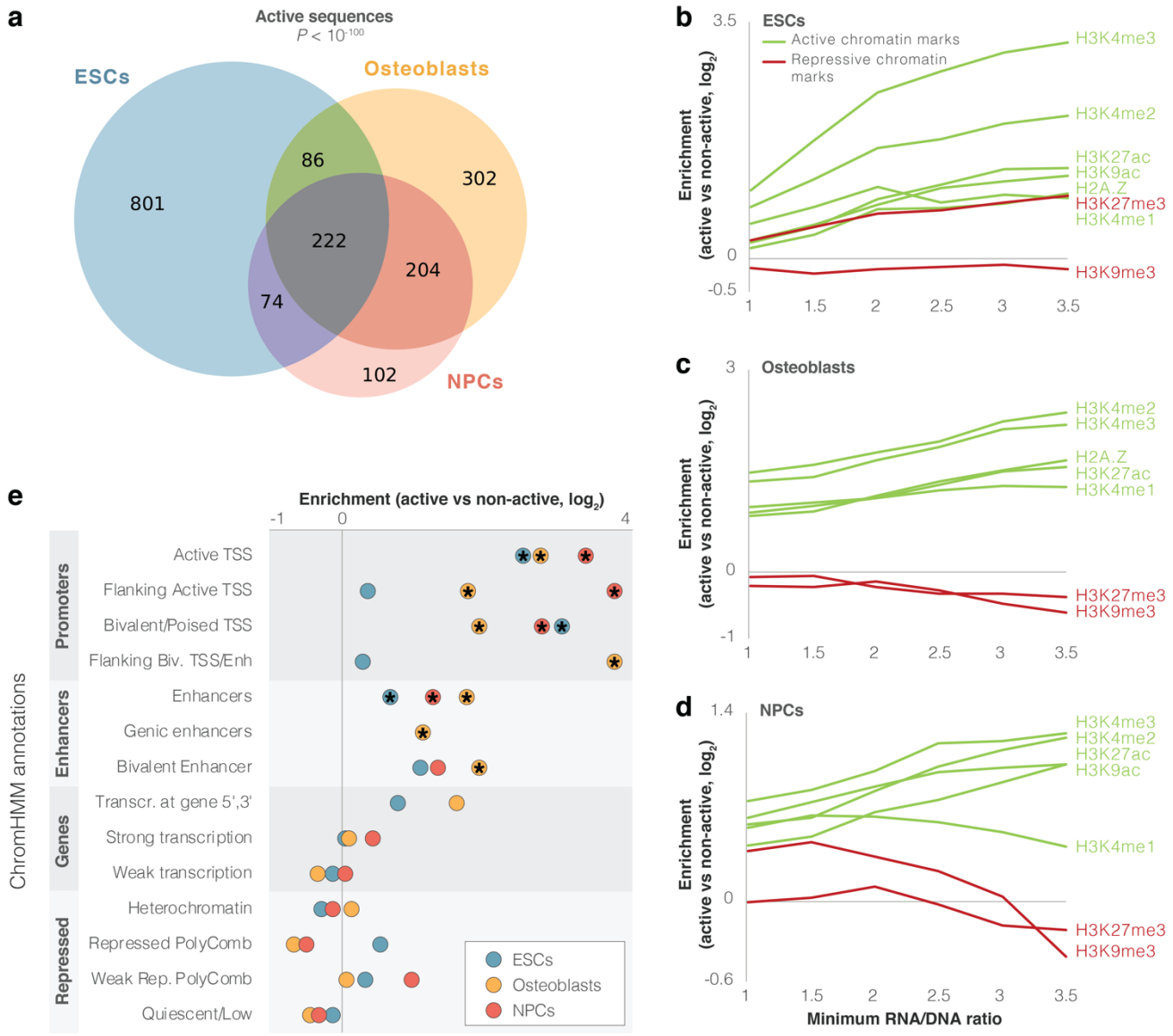
207 primates, indicating a functional role (PhyloP, FDR = 1.1×10^{-3}, *t*-test, **Supplementary File 2**). In

208 summary, we found that sequences that are capable of driving expression tend to overlap active

209 chromatin marks, be depleted of repressive chromatin marks, be closer to TSSs and eQTLs, and

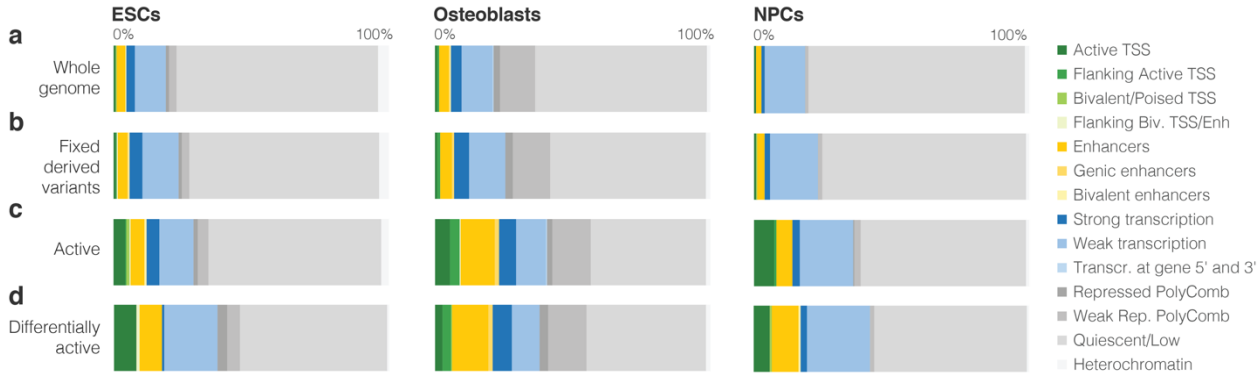
210 have higher sequence conservation, giving us confidence that the MPRA provides us with

211 biologically meaningful results.



212

213 **Figure 2. Identification of modern human sequences promoting expression in lentiMPRA. a.** Overlap between
214 cell types of active sequences. Super Exact test *P*-value is shown for the overlap of the three groups. **b-d.**
215 Enrichment levels of active and repressive histone modification marks within active sequences. Enrichment is
216 computed compared to inactive sequences. **e.** Enrichment of differentially active sequences in various chromatin-
217 based genomic annotations. Missing circles reflect no differentially active sequences in that category. Asterisks
218 mark significant enrichments (FDR < 0.05).
219



220

221 **Supplementary Figure 2. Classification of chromHMM annotations for different groups of variants.** Relative
222 percentage of bases in each chromHMM category throughout the entire genome (**a**), in fixed or nearly fixed modern
223 human-derived variants (**b**), in active sequences (**c**) and in differentially active sequences (**d**), per cell type.
224

225 ## Differentially active sequences between modern and archaic humans

226 We next set out to identify modern and archaic human sequences driving differential expression.
227 We used MPRAnalyze[40] to compare expression driven by the modern and archaic sequences.
228 Out of the active sequence pairs in each cell type, 110 (9%) in ESCs drive significantly
229 differential expression between modern and archaic humans, 243 (30%) in osteoblasts, and 153
230 (25%) in NPCs (FDR ≤ 0.05, see Methods, **Fig. 3a-c, Supplementary File 1**, see Discussion for
231 cell-type differences). We refer to these sequence pairs hereinafter as *differentially active*
232 sequences. Overall, we see significant overlap between cell types in differentially active
233 sequences: 407 sequences (23% of active sequences) were differentially active in at least one cell
234 type, 89 (5%) in at least two cell types, and 10 (0.6%) in all three cell types (8-fold higher than
235 expected, $P = 5\text{x}10^{-7}$, Super Exact test[44], **Fig. 3d**).
236

237 As expected from such closely related organisms, and similar to other MPRAs that compared
238 nucleotide variants (see Discussion) including one that compared human and chimp sequences[30],
239 most sequences drove modest magnitudes of expression difference; of the 407 differentially
240 active sequences, the median fold-change was 1.2x, and only five sequences had a fold-change
241 greater than 2x (**Fig. 3a-c**). We refer to differentially active sequences where modern human
242 expression is higher/lower than archaic human expression as up/downregulating sequences,
243 respectively. In ESCs and NPCs, sequences were equally likely to be up- or downregulating
244 (51% and 52% of differentially active sequences were downregulating, $P = 0.92$ and $P = 0.63$,
245 respectively, Binomial test), while in osteoblasts downregulation was observed slightly more
246 often (59%, $P = 6.9\text{x}10^{-3}$). For sequences that are differentially active in more than one cell type,
247 we found an exceptional agreement between the direction of differential activity, with the
248 regulation of 107 out of 109 pairs going in the same direction ($P = 9.2\text{x}10^{-30}$, Binomial test), and
249 a high correlation between the magnitudes of differential activity (Pearson's R = 0.82, $P =$
250 $1.6\text{x}10^{-27}$). That differentially activity sequences from one cell type are predictive of differential
251 activity in other cell types, even of cell types as disparate as those used here, suggests that these
252 sequences are likely to be differentially active in other cell types not assayed in this lentiMPRA.
253

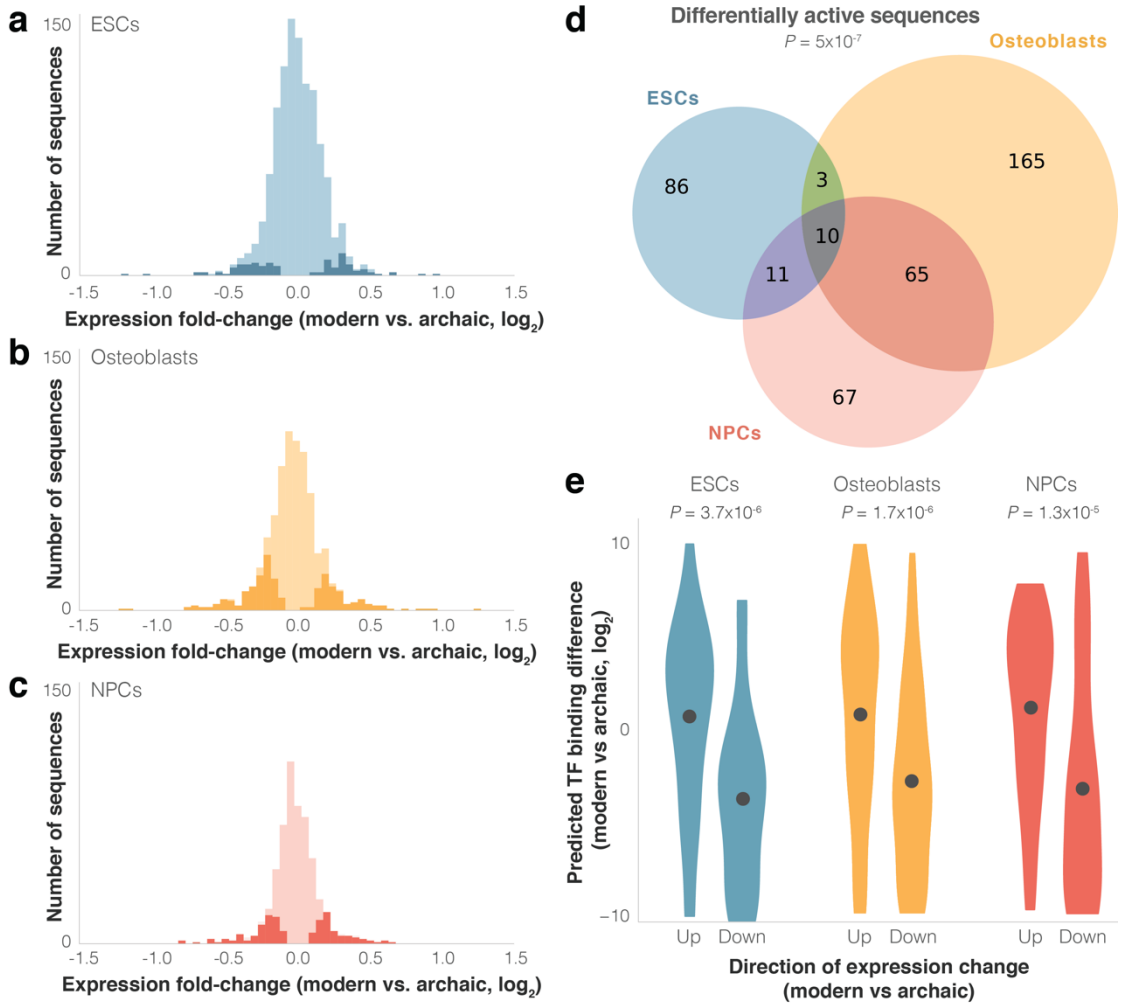
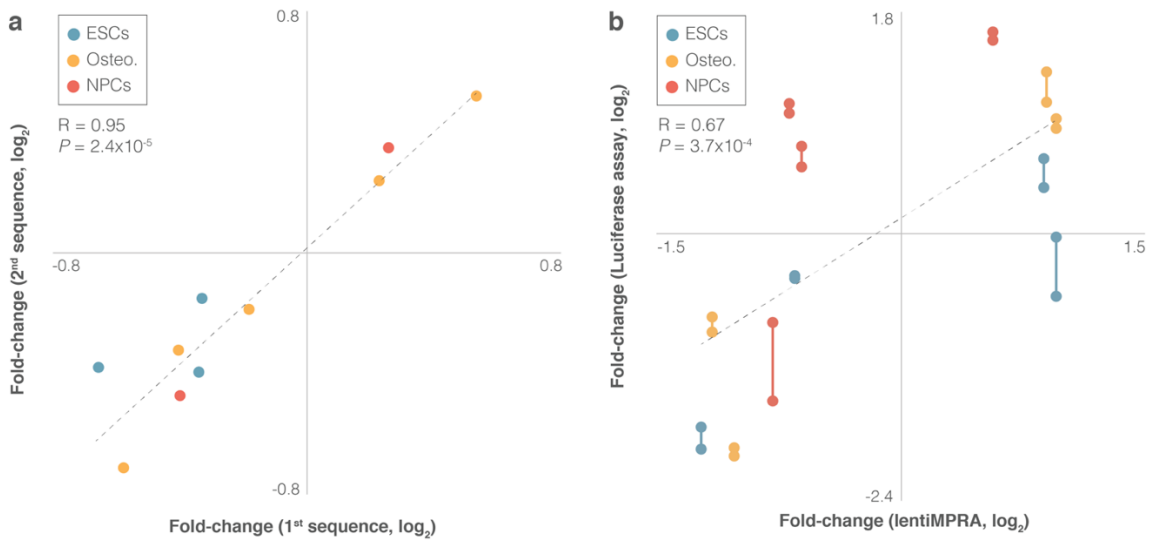


Figure 3. Differential activity of derived modern human sequences. a-c. Distributions of expression fold-changes (RNA/DNA) of active (light) and differentially active (dark) sequences in each cell type. **d.** Overlap of differentially active sequences between cell types. Super Exact test *P*-value is presented for the overlap of the three groups. In the 10 sequences that were differentially active across all three cells types, the direction of fold-change was identical across all cell types ($P = 1.9\text{x}10^{-3}$, 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 ($P = 2.0\text{x}10^{-3}$, binomial test), and that the magnitude of their expression change is

268 highly correlated (Pearson's R = 0.95, 2.4×10^{-5}, **Supplementary Fig. 3a**). To validate these
269 results with an orthogonal method, we tested four differentially active sequences from each cell
270 type in a luciferase reporter assay and found that the direction and magnitude of differential
271 expression tended to replicate the lentiMPRA results (9 out of 12 sequences, Pearson's R = 0.67,
272 $P = 3.7 \times 10^{-4}$, **Supplementary Fig. 3b, Supplementary File 1**). These results suggest that the
273 lentiMPRA was both technically reproducible across cell types and assays and also indicative of
274 true biological signal.
275
276 Finally, we examined the endogenous genomic locations of differentially active sequences,
277 focusing on promoters and enhancers. Out of these 407 sequences, 34% (138) are within 10 kb of
278 a TSS. Analyzing chromHMM[46], we found that 8% (32) of the sequences are within or flanking
279 an active TSS, and 17% (70) are within a putative enhancer. To test if differentially active
280 sequences are enriched within specific chromHMM annotations, we compared the proportion
281 overlapping each annotation to that proportion in the other active sequences. We found that
282 sequences driving differential expression are over-represented within putative enhancer regions
283 ($P = 3.6 \times 10^{-3}$, Fisher's exact test, see Discussion). This is mainly driven by enrichment within
284 ESCs and NPCs (2.2x and 2.54x, $P = 5.1 \times 10^{-3}$, and $P = 2.0 \times 10^{-6}$, respectively, **Supplementary**
285 **Fig. 2c-d**, **Supplementary File 3**, see Methods). These results support a model of rapid enhancer
286 evolution in modern humans, as previously reported for other mammals[48].
287



Supplementary Figure 3. Differential expression is replicated across overlapping sequences and in a reporter assay validation. a. Expression fold-change of overlapping pairs of sequences. Pearson's R and *P*-value are presented. **b.** Expression fold-change of lentiMPRA vs luciferase assay. Each pair of points connected by a line represent two replicates. Pearson's R and *P*-value are presented.

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[12,20,21]. As hypermethylation of regulatory elements is often associated with transcriptional downregulation[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

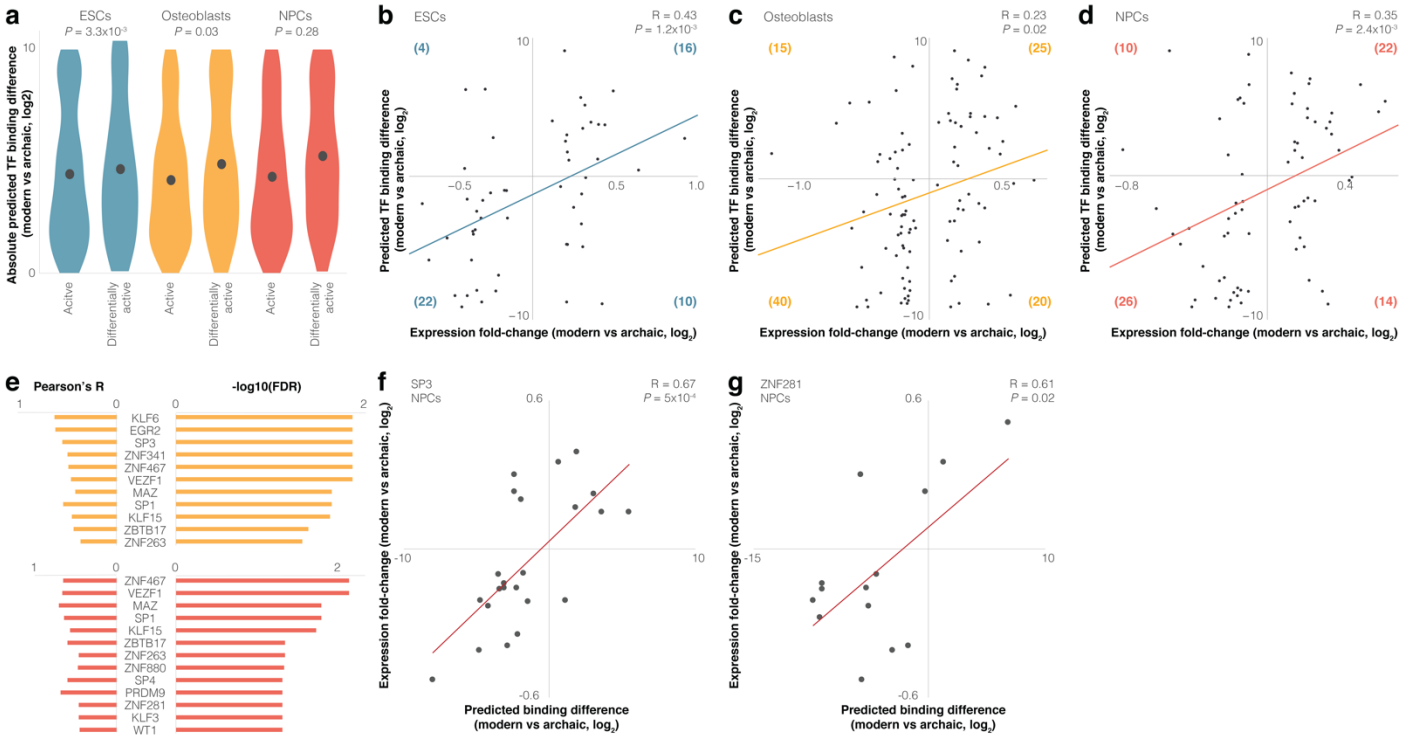
307 sequences show on average +3.2% methylation in modern humans, whereas the top ten most
308 upregulating sequences show -7.4% methylation. We also found that differentially active
309 sequences tend to overlap previously reported differentially methylated regions[12] more often than
310 expected by chance ($P = 0.01$, overlap within 10 kb, hypergeometric test).
311
312 We conjectured that some of the differential activity in these loci might have been driven by
313 alterations in transcription factor (TF) binding. To investigate this, we compared predicted TF
314 binding affinity to the modern and archaic sequences using FIMO[50]. We found that: (1)
315 compared to other active sequences, the difference in predicted binding between the modern and
316 archaic human alleles tends to be larger for differentially active sequences (combined across cell
317 types: 4.3x, $P = 0.02$, *t*-test, **Supplementary Fig. 4a**); (2) the directionality of differential
318 expression tends to match the directionality of differential binding, i.e., upregulating sequences
319 tend to have stronger predicted binding for the modern human sequence, whereas
320 downregulating sequences tend to have stronger predicted binding for the archaic sequence ($P =$
321 $3.7\text{x}10^{-6}$ for ESCs, $P = 1.7\text{x}10^{-6}$ for osteoblasts, and $P = 1.3\text{x}10^{-5}$ for NPCs, binomial test, **Fig.**
322 **3e,** see Methods); and (3) the magnitude of expression difference is correlated with the
323 magnitude of predicted binding difference (Pearson's R = 0.43 and $P = 1.2\text{x}10^{-3}$ for ESCs,
324 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
325 NPCs, **Supplementary Fig. 4b-d** and **Supplementary File 4**). These results support the notion
326 that alterations in TF binding played a role in shaping some of the expression differences
327 between modern and archaic humans.
328

329 To identify the TFs that primarily drove these observations, we investigated which motif changes
330 are most predictive of expression changes. For each TF and the sequences it is predicted to
331 differentially bind, we examined the correlation between binding and expression fold-change.
332 We found that changes to the motifs of 11 TFs in osteoblasts and 13 in NPCs were predictive of
333 expression changes, i.e., higher affinity to the modern human sequences was predictive of higher
334 expression of that sequence (**Supplementary Fig. 4e, Supplementary File 4**). All of these TFs
335 had a positive correlation between changes in their predicted binding affinity and changes in
336 expression of their bound sequences, reflective of an activator function.
337
338 Next, we sought to explore if some motif changes are particularly over-represented within
339 differentially active sequences, suggestive of a more central role in shaping modern human
340 regulatory evolution. To control for sequence composition biases, we used active sequences as a
341 background to search for motif enrichment in differentially active sequences. We found two
342 enriched motifs: ZNF281 (4.6-fold, FDR = 0.016, in NPCs) and SP3 (2.5-fold, FDR = 0.046, in
343 NPCs, **Supplementary File 4**). ZNF281, among other functions, is an inhibitor of neuronal
344 differentiation[51]. SP3 is a ubiquitously expressed TF involved in a wide variety of processes
345 including cell differentiation, proliferation, and synaptic gene regulation[52]. Both TFs bind GC-
346 rich sequences with overlapping binding specificities[53,54], evident also through the overlap
347 between them in predicted differential binding (**Supplementary File 4)**. Out of 153 differentially
348 active sequences in NPCs, 23 (15%) are predicted to be differentially bound by SP3 (2.5-fold,
349 FDR = 0.046), and 14 by ZNF281 (4.6-fold, FDR = 0.016). Intriguingly, ZNF281 and SP3 are
350 also two of the TFs whose predicted differential binding is most tightly linked with differential
351 expression (**Supplementary Fig. 4e-g**). Overall, these data support a model whereby variants in

352 ZNF281 and SP3 motifs might have modulated their binding in NPCs, thereby shaping neural

353 expression differences between modern and archaic humans.

354



355

356 **Supplementary Figure 4. Predicted TF binding is correlated with expression changes. a.** Violin plots of
357 absolute predicted TF binding score difference between modern and archaic sequences. Points show mean. **b-d.**
358 Expression fold-change vs predicted TF binding fold-change. Positive scores represent increased binding in the
359 modern sequence. Parentheses show number of points in each quadrant with a score difference > 0. **e.** Pearson's
360 correlation between differential expression and predicted differential binding affinity. Only significant TFs (FDR <=
361 0.05, Supplementary File 4) are shown for osteoblasts (yellow) and NPCs (red). **f,g.** Expression fold-change vs
362 predicted TF binding fold-change for SP3 (e) and ZNF281 (f) in NPCs. Pearson's correlation R and *P*-value are
363 shown.

364

365 Phenotypic consequences of differential expression

366 In an attempt to assess the functional effects of the differential transcriptional activity we

367 detected, we first sought to link each sequence to the gene(s) it might regulate in its endogenous

368 genomic location. While most regulatory sequences are known to affect their closest gene[55,56],

369 some exert their function through interactions with more distal genes, often reflected in

370 chromatin conformation capture assays, such as Hi-C interactions[57], or eQTL associations[57,58].
371 To predict the genes linked to each sequence we combined data from four sources: (1) proximity
372 to transcription start sites; (2) proximity to eQTLs[47]; (3) proximity to putative enhancers[59]; and
373 (4) spatial interaction with promoters using Hi-C data[58] (see Methods). Using these data, we
374 generated for each cell type a list of genes potentially regulated by each sequence. Overall, 1,341
375 out of the 1,791 active sequences (75%) were linked to at least one putative target gene
376 (**Supplementary File 1**).
377
378 To study the potential functional effects of sequences driving differential expression, we
379 analyzed functions associated with their linked genes. To control for confounders such as cell
380 type-specific regulation, gene length, and GC content, we compared differentially active
381 sequences to other active sequences (instead of the genomic background), which minimizes
382 inherent biases in the active sequences. First, we tested Gene Ontology terms and found an
383 enrichment of the terms *G-protein coupled receptor signaling pathway* (2.9-fold, FDR = 0.019,
384 in ESCs), and *regulation of transcription, DNA-templated* (1.8-fold, FDR = 3.3×10^{-3}, in
385 downregulating sequences across all cells combined, **Supplementary File 4**). To obtain a more
386 detailed picture of phenotypic function, we ran Gene ORGANizer, a tool that uses monogenic
387 disorders to link genes to the organs they affect[60]. We analyzed the genes linked to differentially
388 active sequences and found that for genes linked to sequences driving up-regulation, the most
389 enriched body parts belong to the vocal tract, i.e., the vocal cords (5.0-fold, FDR = 1.1×10^{-3}),
390 voice box (larynx, 3.8-fold, FDR = 4.2×10^{-3}), and pharynx (3.3-fold, FDR = 8.3×10^{-3}, all within
391 ESCs, **Fig. 4a**). Interestingly, we have previously reported that the most extensive DNA
392 methylation changes in modern compared to archaic humans arose in genes affecting the vocal

393 tract[12]. Most subcranial, appendicular and axial skeletal parts are also over-represented within
394 upregulating sequences. Conversely, within sequences driving downregulation, the enriched
395 body parts are the cerebellum (3.0-fold, FDR = 3.9×10^{-3}, in NPCs) and urethra (2.3-fold, FDR =
396 0.04, in ESCs, **Fig. 4a, Supplementary File 5**). This is in line with previous reports of cerebellar
397 anatomy differences between modern humans and Neanderthals[1–3], including results suggesting
398 that the biggest differences in brain anatomy are in the cerebellum[4]. These data also provide
399 leads into the functional divergence of organs, like the urethra, that are not preserved in the fossil
400 record.
401
402 Next, we delved into individual phenotypes associated with the differentially active sequences.
403 To this end, we used the Human Phenotype Ontology (HPO) database[61], a curated database of
404 genes and the phenotypes they underlie in monogenic disorders. HPO covers a broad range of
405 phenotypes related to anatomy, physiology, and behavior. Within upregulating sequences, the
406 enriched phenotypes were involved in learning ability, craniofacial morphology, speech,
407 hormonal activity, heart anatomy, testicular descent, osteoarthritis, muscle tone, and spine
408 curvature (FDR < 0.05, **Fig. 4b**). Within downregulating sequences, we detected an enrichment
409 of phenotypes affecting cerebellar size, intellectual ability, seizures, kidney function, and
410 walking (FDR < 0.05, **Fig. 4b, Supplementary File 5**). These results reveal body parts and
411 phenotypes that were particularly associated with gene expression changes between modern and
412 archaic humans, and could be new candidates for phenotypes under selection.
413

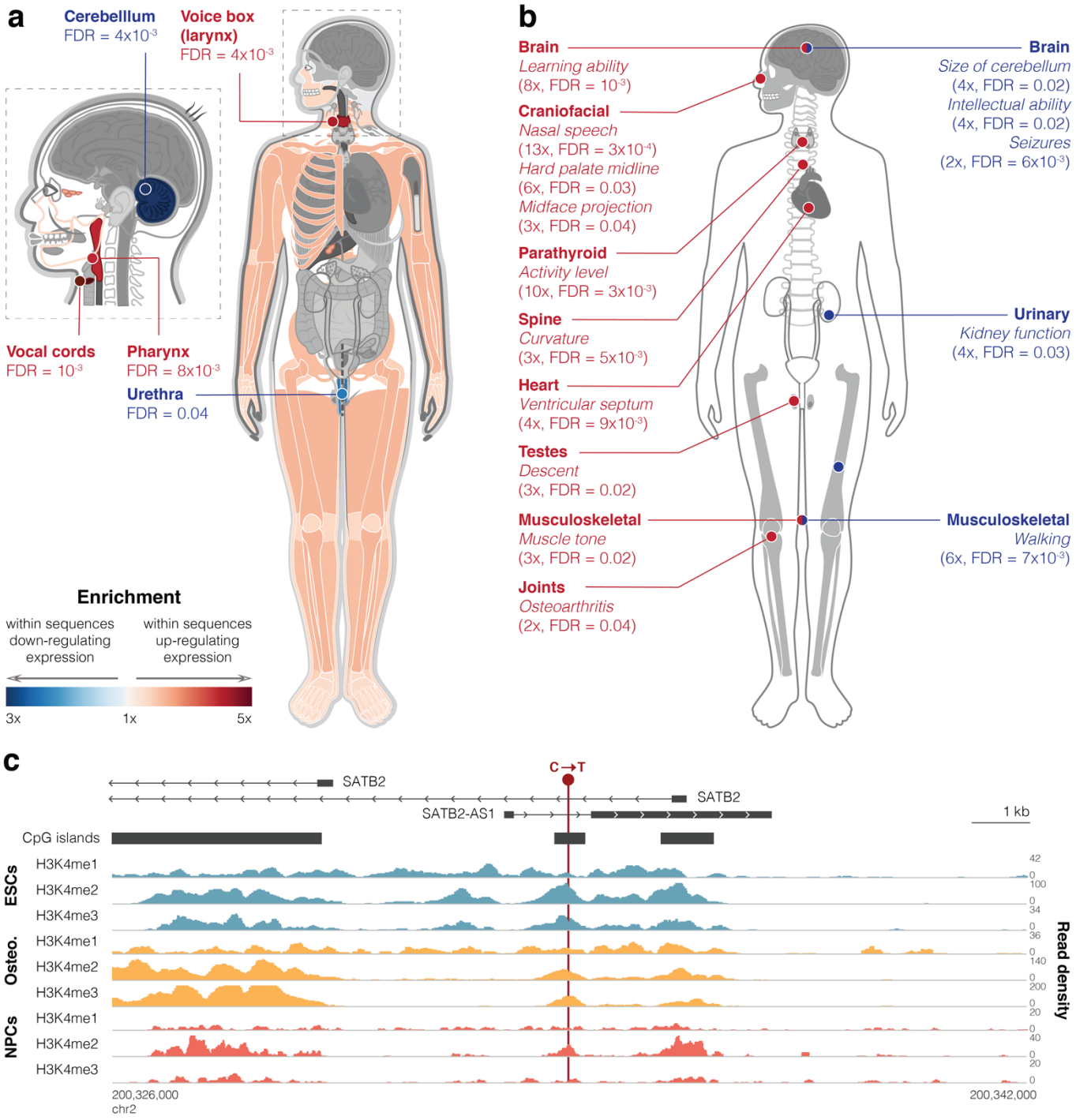


Figure 4. Differentially active sequences are linked to genes affecting the vocal tract and brain. a. Gene ORGANizer enrichment map showing body parts that are significantly over-represented within genes linked to differentially active sequences (FDR < 0.05). Results for the top enriched organs are labeled. 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[45] around the differentially active sequence in *SATB2* (GRCh37 genome version).

422

423 ## Downregulation of *SATB2* potentially underlies brain and skeletal differences

424 This catalog of *cis*-regulatory changes allows us to explore specific sequence changes that
425 potentially underlie divergent phenotypes observed from fossils. To use the most robust data
426 from lentiMPRA, we examined the ten sequences that are differentially active across all three
427 cell types, and looked at their linked genes. To investigate the phenotypes that are potentially
428 linked to these genes, we looked for those genes whose phenotypes can be compared to the fossil
429 record (i.e., skeletal phenotypes). The only gene that fit these criteria was *SATB2*, a regulator of
430 brain and skeletal phenotypes[62]. First, we analyzed its linked variant (C to T transition), which is
431 at a position that is highly conserved in vertebrates (GRCh38: 199,469,203 on chromosome 2,
432 PhyloP score = 0.996). This position is found within a CpG island between two alternative TSSs
433 of *SATB2* (**Fig. 4c**). It is also found in the first intron of *SATB2-AS1*, an antisense lncRNA which
434 upregulates SATB2 protein levels[63]. To determine if this position lies within a regulatory region,
435 we investigated it for chromatin marks in modern humans. We found that it overlaps a DNase I-
436 hypersensitive site[64] and shows many peaks of active histone modification marks in all three cell
437 types (**Fig. 4c, Supplementary File** 1). Indeed, this sequence drives high expression in all three
438 cell types (fourth, eighth, and 19th percentile among active sequences, in ESCs, osteoblasts, and
439 NPCs, respectively, FDR < 10^{-5} across all). Although this sequence shows hallmarks of activity
440 in modern humans, compared to the archaic sequence the modern human sequence is
441 downregulating in all three cell types (-9% in ESCs, FDR = 6.8×10^{-4}, -27% in osteoblasts, FDR
442 = 2.2×10^{-42}, and -12% in NPCs, FDR = 1.1×10^{-7}). These results suggest that the ancestral version
443 of this sequence possibly promoted even higher expression in archaic humans.

444

445 *SATB2* encodes a transcription factor expressed in developing bone and brain. Its activity
446 promotes bone formation, jaw patterning, cortical upper layer neuron specification, and
447 tumorigenesis[62]. Genome-wide association studies (GWAS) show that common variants near
448 *SATB2* are mainly associated with brain and bone phenotypes, such as reaction time, anxiety,
449 mathematical abilities, schizophrenia, autism, and bone density[65]. Heterozygous loss-of-function
450 mutations in *SATB2* result in the *SATB2*-associated syndrome, which primarily affects
451 neurological and craniofacial phenotypes. This includes speech delay, behavioral anomalies
452 (e.g., jovial personality, aggressive outbursts, and hyperactivity), autistic tendencies, small jaws,
453 dental abnormalities, and morphological changes to the palate[66]. Additionally, reduced functional
454 levels of SATB2 due to heterozygous loss-of-function have been shown to be the cause of these
455 phenotypes in both human[62,66–68] and mouse[69–71]. Because these phenotypes are driven by
456 changes to functional SATB2 levels[62], we conjectured that the differential expression of *SATB2*
457 predicted from lentiMPRA might be linked to divergent modern human phenotypes. Thus, we
458 examined whether the phenotypes *SATB2* affects are divergent between archaic and modern
459 humans (e.g., if modern human jaw size is different than the jaw size of archaic humans). We
460 focused on phenotypes available for examination from the fossil record, primarily skeletal
461 differences between modern humans and Neanderthals. From HPO, we generated a list of 17
462 phenotypes known to be affected by *SATB2* and found that 88% (15) of them are divergent
463 between these human groups (**Supplementary File 6**). These include the length of the skull, size
464 of the jaws, and length of the dental arch. Next, based on *SATB*2 downregulation in modern
465 humans predicted from lentiMPRA, we examined whether the direction of a phenotypic change
466 between patients and healthy individuals matches the direction of phenotypic change between
467 modern and archaic humans. For example, given that *SATB2*-associated syndrome patients have

468 smaller jaws, we tested if modern human jaws are smaller compared to archaic humans. If
469 *SATB2* expression is not in fact related to phenotypic divergence, there is a 50% likelihood for a
470 given phenotype to match the fossil record. Yet, we observed a match in direction in 80% of the
471 phenotypes (12 out of 15, **Supplementary File 6**). This includes smaller jaws, flatter face, and
472 higher forehead in modern compared to archaic humans. Overall, the observed number of
473 phenotypes that are both divergent and match in their direction of change is 2.3-fold higher than
474 expected by chance ($P = 1.3 \times 10^{-4}$, hypergeometric test, **Supplementary File 6,** see Methods).
475 Together, these data support a model whereby the C→T substitution in the promoter of *SATB2*,
476 which emerged and reached fixation in modern humans, likely reduced the expression of *SATB2*
477 and possibly affected brain and craniofacial phenotypes. However, further evidence is required to
478 elucidate the potential role of this variant in modern human evolution.
479

480 ## Discussion

481 Identifying noncoding sequence changes underlying human traits is one of the biggest challenges
482 in genetics. This is particularly difficult in ancient samples, where regulatory information is
483 scarce[5,21]. Here, we use an MPRA-based framework to study how sequence changes shaped
484 human gene regulation. By comparing modern to archaic sequences, we investigated how each
485 of the 14,042 single-nucleotide variants that emerged and reached fixation or near fixation in
486 modern humans have affected expression. We found an association between divergent TF motifs
487 and the sequences driving expression changes, suggesting that changes to TF binding might have
488 played a central role in shaping divergent modern human expression. Our results also suggest
489 that genes affecting the vocal tract and cerebellum might have been particularly affected by these

490 expression changes, which is in line with previous comparisons based on the fossil record[1–4] and
491 DNA methylation[12]. More importantly, these results provide candidate sequence changes
492 underlying these evolutionary trends.

493

494 LentiMPRA is designed for linking DNA sequence changes to expression changes *en masse*.
495 Notably, it has limitations that could influence our results, mainly by potentially generating false
496 negatives. First, our lentiMPRA library inserts were limited to ~200bp in length, due to
497 oligonucleotide synthesis technical restrictions, which may be insufficient to detect the activity
498 of longer enhancer sequences[41]. Second, some minimally active sequences may not be expressed
499 at a high enough level to pass our limit of detection. Third, some sequences may regulate
500 expression post-transcriptionally, which lentiMPRA is not designed to detect. Fourth, since test
501 sequences are randomly integrated into the genome, sequences that are dependent on their
502 endogenous genomic environments (e.g., on nearby TF binding sites) might show reduced
503 activity when inserted in new locations, while others might show activity that they otherwise
504 would not have. Our design partially addresses this through the use of insulators and multiple
505 independent integrations, which are intended to dilute location-specific effects. Additionally, all
506 biases are expected to similarly affect the modern and archaic human versions of each
507 sequence[41]. Fifth, transcriptional repression is less likely to be detected due to the low basal
508 activity of the minimal promoter used. Finally, differences in the *trans* environment of a cell
509 could have an effect on the ability of a sequence to exert its *cis*-regulatory effects. However,
510 previous data comparing substantially more divergent organisms (e.g., human-chimpanzee[30] and
511 human-mouse[72]) show that *cis*-regulatory changes exert their effects similarly across different

512 *trans* environments, suggesting the potential effect of such *trans* differences is likely minimal
513 between modern and archaic humans.

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

520 As expected, we observed differences in activity and differential activity between cell
521 types[28,43,72]. Although some of this variation is likely biological (i.e. cell type-specific gene
522 regulation), it is difficult to determine what proportion of it is due to biological versus technical
523 factors (e.g., differences in lentivirus preparation, infection rate, or cell growth, see Methods).
524 Importantly, these differences are expected to result in false negatives rather than false positives.
525 In other words, some of the sequences that appear as active or differentially active in one cell
526 type might actually be active or differentially active in additional cell types (including cell types
527 that were not tested in this study). Thus, we largely refrained from comparisons between cell
528 types. Despite these caveats and limitations, lentiMPRA is a powerful high-throughput tool to
529 characterize the regulatory activity of derived variants, and indeed has become a common assay
530 to study the capability of sequences to promote expression[19].

531 With this method, we found that 1,791 (13%) of the 14,042 sequence pairs can promote
532 expression in at least one of the three cell types tested, and that 405 (23%) of these active
533 sequences show differential activity between modern and archaic humans (average fold-change:

534 1.24x, standard deviation: 0.18, **Fig. 2**, **Supplementary File 1**). Interpreting these results in light
535 of previous MPRAs is challenging, not only because of key differences in statistical power and
536 experimental design (e.g., sequence length), but also because of differing variant selection
537 processes for each MPRA. With the exception of highly repetitive regions, which were removed
538 from our library for technical reasons, the sequences we selected included all known modern
539 human-derived fixed or nearly fixed variants (see Methods). Conversely, previous reporter
540 assays and MPRAs on human intra- or inter-species variation used biased sets of variants by
541 selecting sequences with putative regulatory function (e.g., eQTLs[28], TF binding sites[16], ChIP-
542 seq peaks[29], or TSSs[72]) and/or regions showing particularly rapid evolution (e.g., human
543 accelerated regions[30,31,73,74]). In line with the fact that our data was not pre-filtered for putative
544 regulatory regions, the proportion of active and differentially active sequences we observed tends
545 to be slightly lower than these previous studies, but the magnitude of differential expression in
546 our active sequences was similar[16,28–31,72–74]. At the same time, we were capable of measuring
547 regulatory activity in regions that would otherwise be excluded by filtering for a specific set of
548 marks. Thus, future MPRAs on unfiltered sets of variants will enable the comparison of the
549 patterns we observed to patterns within modern humans, between more deeply divergent clades,
550 and of non-fixed modern-archaic differences.

551 Emergence of new regulatory elements is a common phenomenon in mammalian genomes[48,75].
552 Indeed, some instances of differential activity in our dataset could potentially be explained by the
553 emergence of entirely new regulatory sequences: in 22-34% (depending on cell type) of
554 differentially active sequences, only the derived sequence promotes activity (**Supplementary**
555 **File 1**). Some of these cases could potentially be the result of a new TF binding site. This is

556 supported by the observation that 20 of these sequences show significant predicted TF binding
557 only in the derived allele (**Supplementary File 4**).
558
559 Our results also suggest that sequences driving differential expression are over-represented
560 within putative enhancers, especially in ESCs and NPCs (**Supplementary Fig. 2c-d**,
561 **Supplementary File 3**). Enhancers have been suggested to be an ideal substrate for evolution
562 because of their tissue-specificity and temporal modularity[76]. Indeed, previous studies of
563 introgression between archaic and modern humans suggested that enhancers are some of the
564 most divergent regions between modern and archaic humans[11,25,77]. In line with the enrichment
565 we observed in NPCs, brain-related putative enhancers show particularly low introgression,
566 perhaps suggesting that the modern human sequences in these regions were adaptive[25,77]. To
567 fully characterize the underlying mechanisms of differential activity in enhancers, it is important
568 to disentangle the various factors and confounders that might contribute to this enrichment.
569 There are several alternative explanations for the enrichment we observe, namely that variants
570 within enhancers could be more likely to alter expression compared to other active sequences, or
571 they could be particularly detectable in lentiMPRA. This could be tested using saturation
572 mutagenesis MPRAs[43] to compare the effect of random mutations in enhancer and non-enhancer
573 modern human-derived active sequences.
574
575 Our results suggest that differentially active sequences are not randomly distributed across the
576 genome, but rather tend to be linked to genes affecting particular body parts and phenotypes. The
577 most pronounced enrichment was in the vocal tract, i.e., the vocal cords, larynx, and pharynx.
578 This was evident in both with Gene ORGANizer, where these organs are over-represented by up

579 to 5-fold, and also by HPO phenotype analysis, where the most enriched phenotype was nasal
580 speech (a phenotype that occurs when there is an abnormal closure of the soft palate against the
581 walls of the pharynx, 13-fold, FDR = $3x10^{-4}$, **Fig. 4b**, **Supplementary File 5**). Overall, 53 of the
582 407 differentially active sequences were linked to genes which are known to affect one or more
583 vocal tract phenotypes. Previous reports have also suggested that the vocal tract went through
584 particularly extensive regulatory changes between modern and archaic humans[12], as well as
585 between humans and chimpanzees[78]. Intriguingly, the anatomy of the vocal tract differs between
586 humans and chimpanzees, and has been suggested to affect human phonetic range[79]. Comparing
587 the anatomy of archaic and modern human larynges is currently impossible because the soft
588 tissues of the larynx rapidly decay postmortem. However, together with these previous
589 reports[12,78], our results enable the study of vocal tract evolution from a genetic point of view and
590 suggest that genes influencing the modern human vocal tract have possibly gone through
591 regulatory changes that are not shared by archaic humans.
592
593 We also identified an enrichment of brain-related phenotypes, particularly those affecting the
594 size of the cerebellum (**Fig. 4**, **Supplementary File 5**). The cerebellum is involved in motor
595 control and perception, as well as more complex functions such as cognitive processing,
596 emotional regulation, language, and working memory[80]. Interestingly, the cerebellum has been
597 described as the most morphologically divergent brain region between modern and archaic
598 humans[1,4]. Evidence of divergent brain and cerebellar evolution can also be found at the
599 regulatory level. Studies of Neanderthal alleles introduced into modern humans through
600 introgression provide a clue as to the functional effects of divergent loci between archaic and
601 modern humans. These works have shown that many of the introgressed sequences were likely

602 negatively selected, with the strongest effect in regulatory regions[11,25], particularly in brain
603 enhancers[77]. Studies of introgressed sequences have also shown that the cerebellum is one of the
604 regions with the most divergent expression between Neanderthal and modern human alleles[10].
605 Together with our results, these data collectively suggest that sequences separating archaic and
606 modern humans are particularly linked to functions of the brain, and especially the cerebellum.
607
608 Functional information on archaic human genomes is particularly challenging to obtain because
609 of the postmortem decay of RNA and epigenetic marks in ancient samples. MPRA not only
610 provides a new avenue to identify differential regulation in archaic samples, but also reveals the
611 sequence changes underlying these differences. Here, we present a catalog providing regulatory
612 insight into the sequence changes that separate modern from archaic humans. This resource will
613 hopefully help assign functional context to various signatures of sequence divergence, such as
614 selective sweeps and introgression deserts, and facilitate the study of modern human evolution
615 through the lens of gene regulation.

616 # Methods

617 ## Selection of fixed, derived variants and design of DNA oligonucleotides

618 We selected the variants for our lentiMPRA in the following manner. As a basis, we used the list
619 of 321,820 modern human-derived single nucleotide changes reported to differ between modern
620 humans and the Altai Neanderthal genome[33]. We then filtered this list to include only positions
621 where the Vindija Neanderthal[34], and Denisovan[36] match the Altai Neanderthal variant (or have
622 no information), and are also not polymorphic in any of the four ape species examined (61 *Pan*
623 *troglodytes,* 10 *Pan paniscus,* 15 *Gorilla beringei*, and 28 *Gorilla gorilla*)[37]. Next, we excluded
624 any loci which had any observed variation within modern humans in dbSNP, as annotated by
625 Prüfer et al.[33] or in the 1000 Genomes project (phase 3)[38]. Finally, for technical limitations in
626 downstream synthesis and cloning, we excluded variants at which the surrounding 200 base pairs
627 (bp) had >25% repetitive elements as defined by RepeatMasker[81]. The resulting list contained
628 14,297 sequences and was used to design the initial set of DNA fragments. Upon completion of
629 the lentiMPRA, another high-coverage Neanderthal genome (the Chagyrskaya Neanderthal) was
630 published[35], and we subsequently decided to also filter out loci at which the Chagyrskaya
631 Neanderthal genome did not match the ancestral sequence, bringing the final list of analyzed loci
632 to 14,042 (28,082 ancestral and derived sequences, Supplementary File 1).
633 We designed DNA fragments (oligonucleotides, hereinafter oligos) centered on each variant,
634 including the 99 bp upstream and 100 bp downstream of each variant (200 bp total). For each
635 variant we designed two fragments, one with the ancestral sequence and one with the derived
636 sequence. For cases where two or more variants would be included in the same oligo, we used
637 either derived-only or ancestral-only variants throughout the oligo. The average variants per

638 oligo out of the 14,042 oligos was 1.1, with 12,680 containing one variant, 1,259 containing two,
639 96 containing three and seven containing four. We also included 100 negative control fragments,
640 created by randomly picking 100 of the designed DNA fragments and scrambling their sequence
641 (Supplementary File 1). Lastly, we incorporated 299 positive control fragments (i.e., expected to
642 drive expression). As the library was infected into three cell types (see later), we designed
643 positive controls for each of the cell types. For human embryonic stem cells (ESCs) and human
644 neural progenitor cells (NPCs), we used sequences which were previously shown to drive
645 expression in MPRA in each of these cell types (Supplementary File 1). For fetal osteoblast cells
646 (Hobs), we used putative and confirmed enhancers from mouse and human (Supplementary File
647 1). 15 bp adapter sequences for downstream cloning were added to the 5' (5'-
648 AGGACCGGATCAACT) and 3' (5'-CATTGCGTGAACCGA) ends of each fragment, bringing
649 the total length of each fragment to 230 bp. We synthesized each fragment as an oligonucleotide
650 through Agilent Technologies, twice independently to minimize synthesis errors (Supplementary
651 File 1).
652

653 ## Production of the plasmid lentiMPRA library and barcode association sequencing

654 The plasmid lentiMPRA library was generated as described in Gordon et al. [32]. In brief, the two
655 independently synthesized Agilent Technology oligo pools were amplified separately via a 5-
656 cycle PCR using a different pairs of primers for each pool (forward primers, 5BC-AG-f01.1 and
657 5BC-AG-f01.2; reverse primers, 5BC-AG-r01.1 and 5BC-AG-r01.2; Supplementary File 1),
658 adding a minimal promoter (mP) downstream of the test sequence. A second round of 5-cycle
659 PCR was performed with the same primers for both pools (5BC-AG-f02 and 5BC-AG-r02;
660 Supplementary File 1) to add a 15-bp random barcode downstream of the mP. The two pools

661 were then combined at a 1:1 ratio and cloned into a doubled digested (AgeI/SbfI) pLS-SceI
662 vector (Addgene, 137725) with NEBuilder HiFi Master Mix (NEB). The resulting plasmid
663 lentiMPRA library was electroporated into 10-beta competent cells (NEB) using a Gemini X2
664 electroporation system (BTX) [2kv, 25uF, 200Ω] and allowed to grow up overnight on twelve
665 15cm 100 mg/mL carbenicillin LB agar plates. Colonies were pooled and midiprepped (Qiagen).
666 We collected approximately 6 million colonies, such that ~200 barcodes were associated with
667 each oligo on average. To determine the sequences of the random barcodes and which oligos
668 they were associated with, we first amplified a fragment containing the oligo, mP and barcode
669 from each plasmid in the lentiMPRA library using primers that contain Illumina flow cell
670 adapters (P7-pLSmp-ass-gfp and P5-pLSmP-ass-i#, Supplementary File 1). We sequenced these
671 amplified sequences with a NextSeq 150PE kit using custom primers (R1, pLSmP-ass-seq-R1;
672 R2 (index read), pLSmP-ass-seq-ind1; R3, pLSmP-ass-seq-R2, Supplementary File 1) to obtain
673 approximately 150M total reads. We later did a second round of barcode association sequencing
674 of these fragments to obtain approximately 76M additional reads, for a combined total of
675 225,592,667 reads. To associate barcodes with oligos, we first mapped read pairs (R1 and R3) to
676 the original list of 28,993 oligos using bowtie2 (--very-sensitive)[82]. Next, we filtered out pairs of
677 reads that (1) did not map to the same oligo, (2) did not have at least one of the reads in the pair
678 with a mapping quality of ≥ 6, or (3) did not have the "proper pair" SAM designation. We linked
679 each pair of reads with the read covering its barcode (R2) and saved only those barcode reads
680 having at least a quality score of 30 across all 15 bases in the R2 read. We removed any barcodes
681 associated with more than a single unique oligo (i.e., "promiscuous" barcodes), as well as any
682 barcodes where we did not see evidence of its oligo association at least three times. We then
683 created a list of barcode-oligo associations – this final list comprised 3,495,698 unique barcodes

684 spanning 28,678 oligos (98.9% of the original list of 14,297 variant sequence pairs, 100 negative
685 sequences and 299 positive control sequences), which we refer to as the barcode-oligo
686 association list.
687

688 ## Cell culture and differentiation

689 Human fetal osteoblasts were purchased from Cell Applications Inc. (406K-05f) and were
690 maintained in osteoblast Growth Medium (Cell Applications Inc.). For passaging, cells were
691 washed with 1x PBS, dissociated with Trypsin/EDTA (Cell Applications Inc.), and plated at
692 approximately 5,000 cells/cm^2. H1-ESCs (embryonic stem cells, ESCs, WiCell WA-01,
693 RRID:CVCL_9771) were cultured on Matrigel (Corning) in mTeSR1 media (STEMCELL
694 Technologies) and medium was changed daily. For passaging, cells were dissociated using
695 StemPro Accutase (Thermo Fisher Scientific), washed and re-plated on Matrigel-coated dishes at
696 a dilution of 1:5 to 1:10 in mTeSR1 media supplemented with 10 μM Y-27632 (Selleck
697 Chemicals). ESCs were differentiated into neural progenitor cells (NPCs) by dual-Smad
698 inhibition as previously described (Chambers et al., 2009; Inoue et al., 2019). Briefly, ESCs were
699 cultured in mTeSR1 media until the cells became 80% confluent and then the media was
700 replaced with neural differentiation media consisting of: KnockOut DMEM (Life Technologies)
701 supplemented with KnockOut Serum Replacement (Life Technologies), 2 mM L-glutamine, 1x
702 MEM-NEAA (Life Technologies), 1x beta-mercaptoethanol (Life Technologies), 200 ng/mL
703 Recombinant mouse Noggin (R&D systems), and 10 μM SB431542 (EMD Millipore). On day 4
704 of differentiation, the neural differentiation media was gradually replaced by N2 media
705 [DMEM/F12 (Thermo Fisher Scientific) supplemented with N2 (Thermo Fisher Scientific)]
706 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

707 Noggin and 10 μM SB431542. On day 12, cells were dissociated into single-cell using TrypLE
708 Express (Thermo Fisher Scientific) and cultured in N2B27 media [1:1 mixture of N2 media and
709 Neurobasal media (Thermo Fisher Scientific) with B27 (Thermo Fisher Scientific)]
710 supplemented with 20 ng/mL bFGF (R&D systems) and 20 ng/mL EGF (Millipore sigma)] on
711 Matrigel-coated dish. NPCs were maintained in N2B27 with bFGF and EGF for a month and
712 used for the following experiments at passage 15.

713

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

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

729 performed per cell type. Cells were infected with the lentiMPRA library at a multiplicity of
730 infection (MOI) of 50 for NPCs and osteoblasts, and a MOI of 10 for ESCs. We used a lower
731 MOI for ESC because the cells are very sensitive to infection and a MOI higher than 10 would
732 result in cell death. For ESC and osteoblasts, cell media was changed to include 8ug/mL
733 polybrene before the addition of the lentiMPRA library to increase infection efficiency. The
734 media was replaced with growth media without polybrene approximately 24 hours after
735 infection. Infected cells were grown for three days before combining the plates of each replicate
736 for extraction of RNA and DNA via the Qiagen AllPrep mini kit (Qiagen). We subsequently
737 purified mRNA from the RNA using the Oligotex mRNA prep kit (Qiagen) and synthesized
738 cDNA from the resulting mRNA with SuperScript II RT (Invitrogen), using a primer containing
739 a unique molecular identifier (UMI) (P7-pLSmp-ass16UMI-gfp, Supplementary File 1). DNA
740 fragments were amplified from both the isolated DNA and generated cDNA, keeping each
741 replicate and DNA type separate, with 3-cycle PCR using primers that include adapters
742 necessary for sequencing (P7-pLSmp-ass16UMI-gfp and P5-pLSmP-5bc-i#, Supplementary File
743 1). These primers also contained a sample index for demultiplexing and a UMI for consolidating
744 replicate molecules (see later). A second round of PCR was performed to amplify the library for
745 sequencing using primers targeting the adapters (P5, P7, Supplementary File 1). The fragments
746 were purified and further sequenced with six runs of NextSeq 15PE with 10-cycle dual index
747 reads, using custom primers (R1, pLSmP-ass-seq-ind1; R2 (read for UMI), pLSmP-UMI-seq;
748 R3, pLSmP-bc-seq; R4 (read for sample index), pLSmP-5bc-seq-R2, Supplementary File 1).
749 Later, an additional two runs of 15PE of only the ESC samples were performed due to lower
750 lentivirus infection efficiency in this cell type. Each samples' R1 and R3 reads (containing the
751 barcode) were mapped with bowtie2 [82] (--very-sensitive) to the barcode-oligo association list.

752 Next, we applied several quality filters on the resulting alignments. We first filtered out read
753 pairs that didn't map as proper pairs, and then ensured the mapped sequence completely matched
754 the known barcode sequence by requiring that both R1 and R3 reads have CIGAR stings = 15M,
755 MD flags = 15 and a mapping quality of at least 20. Next, we consolidated read abundance per
756 barcode by selecting only reads with unique UMIs, the result being abundance counts for each
757 barcode, across each replicate library of each cell type for both RNA and DNA.
758 Data was deposited in GEO under accession number: GSE152404 (reviewer access token:
759 avujcogkdxixzij).
760

761 ## Measurement of expression and differential expression

762 We used the R package MPRAnalyze[40] (version 1.3.1,
763 https://github.com/YosefLab/MPRAnalyze) to analyze lentiMPRA data. To determine which
764 oligos were capable of promoting expression, we modeled replicate information into both the
765 RNA and DNA models of MPRAnalyze's quantification framework (rnaDesign = ~ replicate and
766 dnaDesign = ~ replicate) and extracted alpha, the transcription rate, for each oligo. MPRAnalyze
767 used the expression of our 100 scrambled oligos as a baseline against which to measure the level
768 of expression of each tested oligo. We corrected the mean absolute deviation (MAD) score-based
769 *P*-values from MPRAnalyze for multiple testing across tested oligos, including positive controls
770 and excluding scrambled sequences, using the Benjamini-Hochberg method, thus generating an
771 MAD score-based expression false discovery rate (FDR) for each oligo. For each variant and for
772 each cell type, we looked at both the ancestral and derived sequence oligos and assigned an oligo
773 as potentially capable of driving expression if it had an FDR ≤ 0.05 in at least one sequence, and
774 at least 10 barcodes in both sequences (Supplementary File 1). We aggregated UMI-normalized

775 read abundances across all barcodes of each oligo, across all replicates in a given cell type, and
776 calculated a simple ratio of expression as RNA abundance normalized to DNA abundance
777 (RNA/DNA ratio). Next, we determined an RNA/DNA ratio threshold per cell type. This was
778 done by first removing scrambled sequences that show RNA/DNA ratios >2 standard deviations
779 away from the average RNA/DNA ratio of all of the scrambled sequences, as these likely
780 represent oligos that are, by chance, capable of driving expression. This left 95 scrambled
781 sequences in ESCs, 94 in osteoblasts and 97 in NPCs. Then, we used the distribution of
782 RNA/DNA ratios of the remaining scrambled sequences to assign an FDR for each of the non-
783 scrambled oligos. FDR was calculated as the fraction of scrambled sequences that showed an
784 RNA/DNA ratio as high or higher than each non-scrambled oligo. Only oligos that passed both
785 tests described above (FDR ≤ 0.05 in each test) were considered as "active" (i.e., capable of
786 driving expression).
787 To measure differential expression between ancestral and derived sequences, we used
788 MPRAnalyze's comparative framework. In essence, this tool uses a barcode's RNA reads as an
789 indicator of expression level and normalizes this to the DNA reads as a measure of the number of
790 genomic insertions of that barcode (i.e., the number of fragments from which RNA can be
791 transcribed). MPRAnalyze uses information across all the barcodes for both alleles of a given
792 sequence, as well as information across all replicates. For the terms of the model, we included
793 replicate information in the RNA, DNA and reduced (null) models, allele information in the
794 RNA and DNA models, and barcode information only in the DNA model (rnaDesign = ~
795 replicate + allele, dnaDesign = ~ replicate + barcode + allele, reducedDesign = ~ replicate). We
796 extracted *P*-values and the differential expression estimate (fold-change of the derived relative to
797 ancestral sequence). Then, we corrected the *P*-values of the set of active oligos (see above) for

798 multiple testing with the Benjamini-Hochberg method to generate an FDR for each sequence.
799 We set a cutoff of FDR ≤ 0.05 to call a sequence capable of driving differential expression. From
800 this we generated, for each cell type, a list of sequences with differential expression between the
801 ancestral and derived alleles (Supplementary File 1).
802

803 ## Luciferase validation assays

804 Each assayed oligo was synthesized by Twist Biosciences and cloned into the pLS-mP-Luc
805 vector (Addgene 106253) upstream of the luciferase gene. Lentivirus was generated
806 independently for each vector using techniques as described for MPRA (see above), with the
807 omission of the filtering and concentration step, which was replaced with the collection of the
808 entirety of the cell culture media for use in subsequent infections. In addition, pLS-SV40-mP-
809 Rluc (Addgene 106292), to adjust for infection efficiency, was added at a 1:3 ratio to the assayed
810 vector for a total of 4ug for lentivirus production. We infected each cell type individually with
811 each viral prep. The amount of lentivirus added was based on titrations in which varying
812 amounts of a subset of viral preps were added to each cell type and cell death was observed 3
813 days post infection; the virus volume that produced between 30-50% death was used for
814 subsequent experiments. Approximately 20,000 cells were plated in 96-well plates and grown for
815 24-48 hours (~70% confluent) before the addition of lentivirus. For osteoblasts and ESCs,
816 8ug/mL polybrene was added to the culture media at the same time as the addition of the
817 lentivirus. The media was changed 24 hours after infection and cells were grown for an
818 additional 48 hours. The cells were then washed with PBS and lysed. Firefly and renilla
819 luciferase expression were measured using the Dual-Luciferase Reporter Assay System
820 (Promega) on the GloMax plate reader (Promega). Each oligo was tested using two biological

821 replicates on different days and each biological replicate consisted of three technical replicates.
822 Activity of a given oligo was calculated by normalizing the firefly luciferase activity to the
823 renilla luciferase. We then calculated the $\log_2$ fold change (LFC) between the ancestral and
824 derived alleles as $\log_2$(derived / ancestral). A full list of oligos tested and their LFC can be found
825 in **Supplementary File 1**.
826 We found that the mean difference in fold-change between replicates was 3-fold lower for the
827 differentially active vs other active sequences (0.22 vs 0.60), and that the variance of these
828 differences was 9-fold lower for differentially active sequences compared to other active
829 sequences (0.09 vs 0.83, Supplementary File 1), suggesting that differentially active sequences
830 reflect a true biological signal.

831 ## Predicting target genes

832 To connect the surrounding locus of each variant to genes it potentially regulates, we combined
833 four data sources. For each locus, we generated four types of gene lists, based on four largely
834 complementary approaches: (1) overlap with known expression quantitative trait loci (eQTLs);
835 (2) spatial interaction with promoters; (3) proximity to putative enhancers; and (4) proximity to a
836 transcription start site (TSS, Supplementary File 1). Each data source was obtained and
837 incorporated into each type of list as described below:
838

839 *1) Proximity to known eQTLs*

840 eQTLs are genetic variants between individuals shown to be associated with expression
841 differences. We reasoned that the target genes of the sequence surrounding a variant are
842 potentially similar to the target genes of nearby eQTLs. We downloaded eQTLs and their
843 associated genes from GTEx[47] (www.gtexportal.org, v8 on August 26, 2019) and overlapped the

844 locations of each eQTL with our list of sequences. We linked the target genes of any eQTLs
845 within +/-1 kb to each variant. We used all tissue types reported by GTEx, for each cell type in
846 the lentiMPRA. 9,503 out of the 14,042 loci were found within +/- 1 kb of an eQTL, with 83,777
847 eQTLS overall overlapping them.
848
849 *2) Spatial interaction with a promoter via Hi-C data*
850 High-throughput chromosome conformation capture (Hi-C) techniques map spatial interactions
851 between segments of DNA. We reasoned that if a variant is found within or near a region that
852 was shown to interact physically with a promoter, that variant could be in a region involved in
853 regulating that promoter. We downloaded promoter capture Hi-C data from Jung et al.[58],
854 containing a list of all the significant interactions between promoters and other segments of the
855 genome across 27 tissue and cell types. We overlapped our variants with the locations of
856 interacting genomic fragments to find interactions within +/-10 kb of each variant. We then
857 linked each variant with the promoters that each interacting fragment was shown to contact. We
858 repeated this process twice: once to obtain a cell type-specific list, and once to obtain a generic
859 list. For the cell type-specific list of sequence-gene links, we included only those interactions
860 observed in cell types corresponding to the cell lines used in our lentiMPRA: ESCs, NPCs and
861 mesenchymal stem cells as an approximation for osteoblasts (given that osteoblast Hi-C data is
862 not publicly available to the best of our knowledge, and that osteoblasts differentiate from
863 MSCs). For the generic list, we used interactions across any of the 27 tissue and cell types
864 analyzed by Jung et al.[58]. 4,688 out of the 14,042 loci overlapped at least one region that
865 interacts with a promoter.
866

867 *3) Putative enhancers*

868 Lastly, we checked which of our variants were in previously reported putative enhancers. To this
869 end, we downloaded the GeneHancer database[59] V4_12 and searched for putative enhancers
870 within +/- 10kb of each of our variants, linking each variant to the target genes of each putative
871 enhancer within that distance. GeneHancer provides "elite" or "non-elite" status to their defined
872 enhancer-target gene connections depending on the strength of the evidence supporting each
873 connection. Using this information, we repeated the process twice: once for the elite status and
874 once for all annotations. 5,017 out of the 14,042 loci overlapped at least one putative enhancer
875

876 *4) Promoters*

877 Promoters were defined as the region 5kb upstream to 1kb downstream of GENCODE[83] v29
878 GRCh38 TSSs. If a variant fell within this region, we linked it to that TSS's gene. Each variant
879 was assigned to all the promoters it fell within. 1,466 out of the 14,042 loci were found within a
880 promoter.

881 Overall, 11,207 out of the 14,042 loci were linked to at least one putative target gene, with a
882 median of four target genes per locus. 2,830 of the remaining loci were linked to their closest
883 TSS, regardless of distance. The last 5 without hg38 coordinates for their closest TSS were not
884 linked to a gene. Importantly, these links do not necessarily mean that these target genes are
885 regulated by these loci, but rather they serve as a list of potential target genes for the loci
886 showing a regulatory function through lentiMPRA.
887

888 ## Differential transcription factor binding sites

889 We predicted differences in binding of human transcription factors caused by each of our
890 variants as follows. First, we downloaded the entire set of publicly available human transcription
891 factor binding motifs (7,705 motifs, 6,608 publicly available) from the Catalogue of Inferred
892 Sequence Binding Preferences (CIS-BP) database (http://cisbp.ccbr.utoronto.ca/), and filtered
893 them to include only motifs labeled as *directly determined* (i.e., we filtered out inferred motifs),
894 resulting in 4,351 motifs. Next, to enrich our mapping result for matches covering the variant
895 location, we trimmed each of our oligo sequences containing a single variant to +/- 30 bp around
896 the variant (the length of the longest motif). We did not trim oligos containing >1 variant. We
897 used FIMO[50] to map each remaining motif to both the ancestral and derived alleles of each
898 trimmed sequence (or untrimmed, for sequences with >1 variant). For each motif mapping to
899 both the ancestral and derived alleles at the same strand and location, we required that at least
900 one allele had a *q*-value (as supplied by FIMO) ≤ 0.05. For motifs where only one of the
901 ancestral or derived alleles had a reported mapping from FIMO for a given oligo, we searched all
902 the mappings of that motif for the lowest mapping score, and substituted that score for the
903 missing allele's score. Then, we found cases where the FIMO predicted binding score of a motif
904 differed between the ancestral and derived alleles. Finally, we linked each motif to the
905 transcription factor (TF) it is most confidently associated with in CIS-BP, thereby generating
906 lists of TFs that showed differential predicted binding for each sequence. For cases in which
907 multiple unique motifs corresponded to the same TF, we used the motif with the largest score
908 difference between alleles. TF enrichment analyses were done on all predicted differential TF
909 binding sites for TFs with a minimum of 10 predicted differential sites. This was done for each
910 cell type, as well as for the union of all differentially active sequences across all cell types.

911 Fisher's exact test was used to compute enrichment of a TF among differentially active
912 sequences compared to other active sequences. *P*-values were FDR-adjusted. A threshold of 10
913 sites per TF was used for the analysis of correlation between predicted binding and expression.

914 ## Overlapping loci with genomic features

915 The following datasets were used for the overlap analyses: GENCODE v28 GRCh38 human
916 genome TSSs[84], GTEx v8 eQTLs[47], and broad peaks for the following histone modification
917 marks: H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, and H3K27me, and
918 the histone variant H2A.Z from the Roadmap Project for ESCs, ESC-derived NPCs, and
919 osteoblasts[45]. We overlapped each of these datasets with the lists of non-active and active
920 sequences, and computed enrichment *P*-values using a hypergeometric test. We repeated this for
921 various RNA/DNA cutoffs (1, 1.5, 2, 2.5, 3 and 3.5). Sex chromosomes were removed from the
922 analyses. *P*-values were FDR-adjusted using the Benjamini-Hochberg procedure.
923 Sequence conservation within primates was taken from the Altai Neanderthal genome
924 annotation, which used the PhyloP metric[33].
925

926 ## Phenotype enrichment analysis

927 Body part enrichment analyses were conducted using Gene ORGANizer v13. The analyses were
928 conducted on sequences driving increased expression, sequences driving decreased expression,
929 and all differentially active sequences. This was done in each of the three cell types, as well as on
930 the union of all three cell types. We repeated the analyses using various $\log_2$(fold-change)
931 thresholds: 0, 0.25, 0.5, 0.75 and 1, and on both the stringent and non-stringent locus-gene
932 associations. Analyses were done against the Gene ORGANizer[60] genomic background using the

933 ORGANizer tool with the *confident* option. *P*-values were FDR-adjusted using the Benjamini-
934 Hochberg procedure. For osteoblasts, non-skeletal organs were removed from the analyses. For
935 NPCs, non-neuronal organs were removed.
936 For the HPO analyses, we used HPO[61] build 1268 (08 November, 2019), analyzing gene lists
937 identical to the Gene ORGANizer analyses. Lists of phenotypes from HPO were generated for
938 each variant through its linked genes. Hypergeometric test *P*-values were computed per
939 phenotype and FDR-adjusted. Similarly to the Gene ORGANizer analysis, we removed non-
940 skeletal phenotypes from the osteoblast results, and non-neuronal phenotypes from the NPC
941 results.
942 Gene Ontology (GO) analyses on the same groups of genes described above revealed two
943 enriched terms: GO:0007186, G-protein coupled receptor signaling pathway (2.9-fold within
944 ESCs, up- and downregulating sequences, $\log_2$Fold-change cutoff: 0.25, FDR = 0.018), and
945 GO:0006355, regulation of transcription, DNA-templated (1.8-fold within the union of cell
946 types, downregulating sequences, $\log_2$Fold-change cutoff: 0.25, FDR = 3.3×10^{-3}).
947
948 *SATB2* phenotypic analysis was done as previously described in Gokhman et al[14]. In short, we
949 used HPO[61] build 1268 (08 November, 2019) to link phenotypes to *SATB2*. In addition, we
950 conducted a literature search to expand gene-phenotype links to include studies that did not
951 appear on HPO (Supplementary File 6). We used only skeletal directional phenotypes, i.e.,
952 phenotypes that could be described on a scale (e.g., smaller/larger hands), as these could be
953 examined against the fossil record. This resulted in 34 phenotypes that are the result of *SATB2*
954 heterozygous loss-of-function (LOF) (Supplementary File 6). Phenotypes that are included in
955 another phenotype (e.g., *Prominent nasal bridge* and *Prominent nose*) were merged, and

956 contradicting phenotypes (e.g., *Broad nose* and *Thin/small nose*) were removed. This resulted in
957 a final list of 17 phenotypes (Supplementary File 6). Given that the mechanism underlying these
958 phenotypes is a decrease in the dosage of SATB2, and that *SATB2* is possibly downregulated in
959 modern humans, we sought to investigate if similar phenotypes exist between modern human
960 patients with *SATB2* heterozygous LOF and archaic humans. For each phenotype, we determined
961 if it is divergent between the modern and archaic humans based on previously published
962 annotation[14]. Then, for remaining divergent phenotypes, we tested if the direction between
963 patients and healthy individuals matches the direction between modern and archaic humans. The
964 significance of directionality match was computed using a binomial test, with a random
965 probability of success $p = 0.5$. To compute the significance of the overall number of phenotypes
966 that are divergent and match in direction, we compared the overall number of annotated
967 divergent phenotypes to the number of divergent phenotypes associated with *SATB2* using a
968 hypergeometric test. Out of a total of 696 annotated phenotypes between modern and archaic
969 humans[14], 434 are annotated as divergent, and the direction of 50% of them (217 phenotypes) is
970 expected to match by chance.
971

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).

980 ## References

981 1. Neubauer, S., Hublin, J. J. & Gunz, P. The evolution of modern human brain shape. *Sci.*
982 *Adv.* (2018). doi:10.1126/sciadv.aao5961
983 2. Gunz, P. *et al.* Neandertal Introgression Sheds Light on Modern Human Endocranial
984 Globularity. *Curr. Biol.* (2019). doi:10.1016/j.cub.2018.10.065
985 3. Aiello, L. & Dean, C. *An Introduction to Human Evolutionary Anatomy*. (Elsevier, 2002).
986 4. Kochiyama, T. *et al.* Reconstructing the Neanderthal brain using computational anatomy.
987 *Sci. Rep.* (2018). doi:10.1038/s41598-018-24331-0
988 5. Yan, S. M. & McCoy, R. C. Archaic hominin genomics provides a window into gene
989 expression evolution. *Curr. Opin. Genet. Dev.* **62**, 44–49 (2020).
990 6. Britten, R. J. & Davidson, E. H. Repetitive and non-repetitive DNA sequences and a
991 speculation on the origins of evolutionary novelty. *Q. Rev. Biol.* (1971).
992 doi:10.1086/406830
993 7. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science*
994 **188**, 107–116 (1975).
995 8. Enard, D., Messer, P. W. & Petrov, D. A. Genome-wide signals of positive selection in
996 human evolution. *Genome Res.* (2014). doi:10.1101/gr.164822.113
997 9. Fraser, H. B. Gene expression drives local adaptation in humans. *Genome Res.* **23**, 1089–
998 1096 (2013).
999 10. McCoy, R. C., Wakefield, J. & Akey, J. M. Impacts of Neanderthal-Introgressed
1000 Sequences on the Landscape of Human Gene Expression. *Cell* **168**, 916-927.e12 (2017).
1001 11. Petr, M., Pääbo, S., Kelso, J. & Vernot, B. Limits of long-term selection against
1002 Neandertal introgression. *Proc. Natl. Acad. Sci. U. S. A.* (2019).
1003 doi:10.1073/pnas.1814338116
1004 12. Gokhman, D. *et al.* Differential DNA methylation of vocal and facial anatomy genes in
1005 modern humans. *Nat. Commun.* **11**, 1189 (2020).
1006 13. Colbran, L. L. *et al.* Inferred divergent gene regulation in archaic hominins reveals
1007 potential phenotypic differences. *Nat. Ecol. Evol.* (2019). doi:10.1038/s41559-019-0996-x
1008 14. Gokhman, D. *et al.* Reconstructing Denisovan Anatomy Using DNA Methylation Maps.
1009 *Cell* **179**, 180-192.e10 (2019).

1010 15. Dannemann, M. & Racimo, F. Something old, something borrowed: admixture and
1011 adaptation in human evolution. *Curr. Opin. Genet. Dev.* **53**, 1–8 (2018).
1012 16. Weyer, S. & Pääbo, S. Functional analyses of transcription factor binding sites that differ
1013 between present-day and archaic humans. *Mol. Biol. Evol.* (2016).
1014 doi:10.1093/molbev/msv215
1015 17. Vespasiani, D. M., Jacobs, G. S., Brucato, N., Cox, M. P. & Romero, I. G. Denisovan
1016 introgression has shaped the immune system of present-day Papuans. *bioRxiv*
1017 2020.07.09.196444 (2020). doi:10.1101/2020.07.09.196444
1018 18. Grogan, K. E. & Perry, G. H. Studying human and nonhuman primate evolutionary
1019 biology with powerful in vitro and in vivo functional genomics tools. *Evolutionary*
1020 *Anthropology* (2020). doi:10.1002/evan.21825
1021 19. Chatterjee, S. & Ahituv, N. Gene Regulatory Elements, Major Drivers of Human Disease.
1022 *Annu. Rev. Genomics Hum. Genet.* (2017). doi:10.1146/annurev-genom-091416-035537
1023 20. Gokhman, D. *et al.* Reconstructing the DNA methylation maps of the Neandertal and the
1024 Denisovan. *Science* **344**, 523–7 (2014).
1025 21. Gokhman, D., Meshorer, E. & Carmel, L. Epigenetics: It's Getting Old. Past Meets Future
1026 in Paleoepigenetics. *Trends Ecol. Evol.* **31**, 290–300 (2016).
1027 22. Barker, H. R., Parkkila, S. & Tolvanen, M. E. E. Evolution is in the details: Regulatory
1028 differences in modern human and Neanderthal. *bioRxiv* (2020).
1029 doi:doi.org/10.1101/2020.09.04.282749
1030 23. Batyrev, D., Lapid, E., Carmel, L. & Meshorer, E. Predicted Archaic 3D Genome
1031 Organization Reveals Genes Related to Head and Spinal Cord Separating Modern from
1032 Archaic Humans. *Cells* (2019). doi:10.3390/cells9010048
1033 24. Pedersen, J. S. *et al.* Genome-wide nucleosome map and cytosine methylation levels of an
1034 ancient human genome. *Genome Res.* **24**, 454–466 (2014).
1035 25. Silvert, M., Quintana-Murci, L. & Rotival, M. Impact and Evolutionary Determinants of
1036 Neanderthal Introgression on Transcriptional and Post-Transcriptional Regulation. *Am. J.*
1037 *Hum. Genet.* (2019). doi:10.1016/j.ajhg.2019.04.016
1038 26. Moriano, J. & Boeckx, C. Modern human changes in regulatory regions implicated in
1039 cortical development. *BMC Genomics* **21**, 304 (2020).
1040 27. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays.

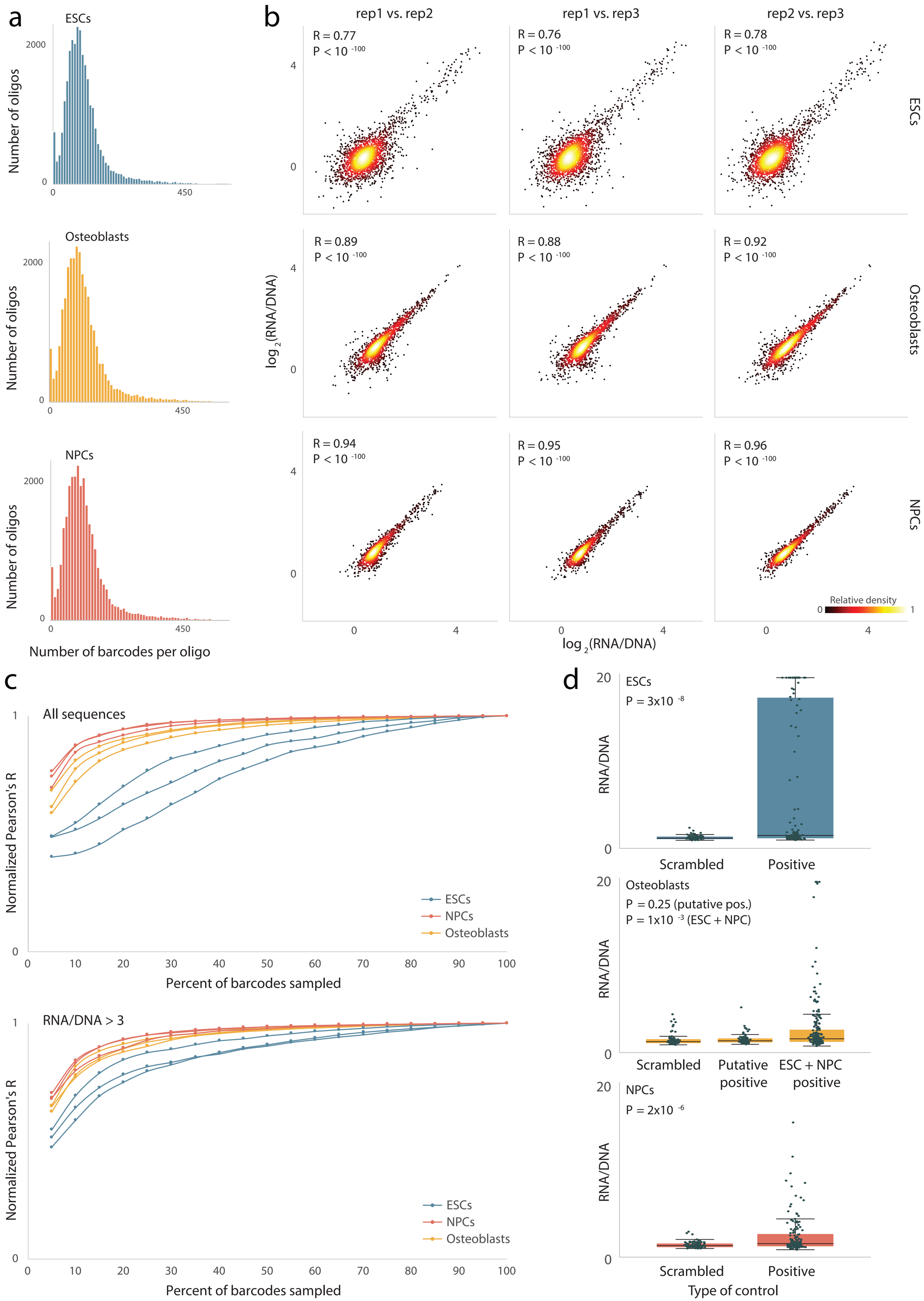
1041 *Genomics* (2015). doi:10.1016/j.ygeno.2015.06.005
1042 28. Tewhey, R. *et al.* Direct identification of hundreds of expression-modulating variants
1043 using a multiplexed reporter assay. *Cell* (2016). doi:10.1016/j.cell.2016.04.027
1044 29. Klein, J. C., Keith, A., Agarwal, V., Durham, T. & Shendure, J. Functional
1045 characterization of enhancer evolution in the primate lineage. *Genome Biol.* (2018).
1046 doi:10.1186/s13059-018-1473-6
1047 30. Ryu, H. *et al.* Massively parallel dissection of human accelerated regions in human and
1048 chimpanzee neural progenitors. *bioRxiv* (2018). doi:10.1101/256313
1049 31. Uebbing, S. *et al.* Massively parallel discovery of human-specific substitutions that alter
1050 neurodevelopmental enhancer activity. *bioRxiv* (2019). doi:10.1101/865519
1051 32. Gordon, M. G. *et al.* lentiMPRA and MPRAflow for high-throughput functional
1052 characterization of gene regulatory elements. *Nat. Protoc.* (2020). doi:10.1038/s41596-
1053 020-0333-5
1054 33. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai
1055 Mountains. *Nature* **505**, 43–9 (2014).
1056 34. Prüfer, K. *et al.* A high-coverage Neandertal genome from Vindija Cave in Croatia.
1057 *Science (80-.).* **358**, 655–658 (2017).
1058 35. Mafessoni, F. *et al.* A high-coverage Neandertal genome from Chagyrskaya Cave. *Proc.*
1059 *Natl. Acad. Sci.* (2020). doi:10.1073/pnas.2004944117
1060 36. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan
1061 individual. *Science* **338**, 222–6 (2012).
1062 37. De Manuel, M. *et al.* Chimpanzee genomic diversity reveals ancient admixture with
1063 bonobos. *Science (80-.).* (2016). doi:10.1126/science.aag2602
1064 38. Auton, A. *et al.* A global reference for human genetic variation. *Nature* (2015).
1065 doi:10.1038/nature15393
1066 39. Sherry, S. T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* (2001).
1067 doi:10.1093/nar/29.1.308
1068 40. Ashuach, T. *et al.* MPRAnalyze: Statistical framework for massively parallel reporter
1069 assays. *Genome Biol.* (2019). doi:10.1186/s13059-019-1787-z
1070 41. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal
1071 versus episomal encoding of enhancer activity. *Genome Res.* (2017).

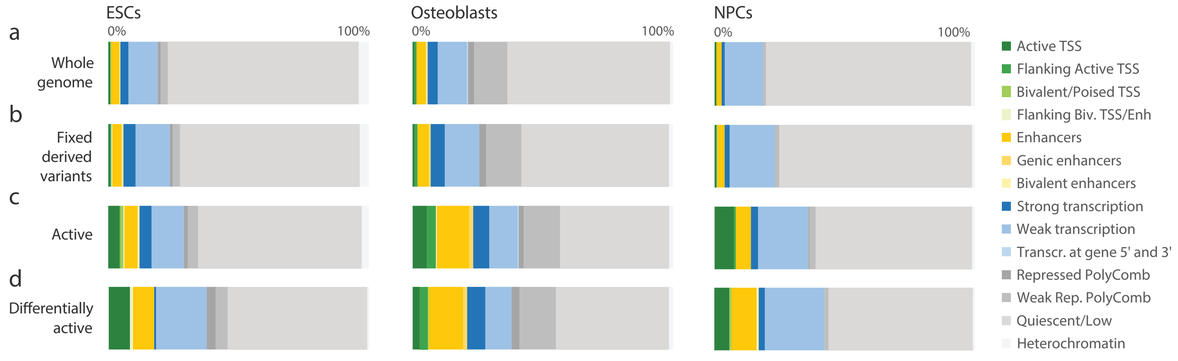
1072 doi:10.1101/gr.212092.116
1073 42. Efroni, S. *et al.* Global Transcription in Pluripotent Embryonic Stem Cells. *Cell Stem Cell*
1074 (2008). doi:10.1016/j.stem.2008.03.021
1075 43. Kircher, M. *et al.* Saturation mutagenesis of twenty disease-associated regulatory elements
1076 at single base-pair resolution. *Nat. Commun.* (2019). doi:10.1038/s41467-019-11526-w
1077 44. Wang, M., Zhao, Y. & Zhang, B. Efficient Test and Visualization of Multi-Set
1078 Intersections. *Sci. Rep.* (2015). doi:10.1038/srep16923
1079 45. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human
1080 epigenomes. *Nature* **518**, 317–329 (2015).
1081 46. Ernst, J. & Kellis, M. ChromHMM: Automating chromatin-state discovery and
1082 characterization. *Nature Methods* (2012). doi:10.1038/nmeth.1906
1083 47. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis:
1084 Multitissue gene regulation in humans. *Science (80-.).* **348**, 648–660 (2015).
1085 48. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* (2015).
1086 doi:10.1016/j.cell.2015.01.006
1087 49. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond.
1088 *Nat. Rev. Genet.* **13**, 484–92 (2012).
1089 50. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given
1090 motif. *Bioinformatics* (2011). doi:10.1093/bioinformatics/btr064
1091 51. Pieraccioli, M. *et al.* ZNF281 inhibits neuronal differentiation and is a prognostic marker
1092 for neuroblastoma. *Proc. Natl. Acad. Sci. U. S. A.* (2018). doi:10.1073/pnas.1801435115
1093 52. Suske, G. NF-Y and SP transcription factors — New insights in a long-standing liaison.
1094 *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* (2017).
1095 doi:10.1016/j.bbagrm.2016.08.011
1096 53. Hahn, S. & Hermeking, H. ZNF281/ZBP-99: A new player in epithelial-mesenchymal
1097 transition, stemness, and cancer. *Journal of Molecular Medicine* (2014).
1098 doi:10.1007/s00109-014-1160-3
1099 54. Yamakawa, H. *et al.* The Transcription Factor Sp3 Cooperates with HDAC2 to Regulate
1100 Synaptic Function and Plasticity in Neurons. *Cell Rep.* (2017).
1101 doi:10.1016/j.celrep.2017.07.044
1102 55. Gasperini, M. *et al.* A Genome-wide Framework for Mapping Gene Regulation via

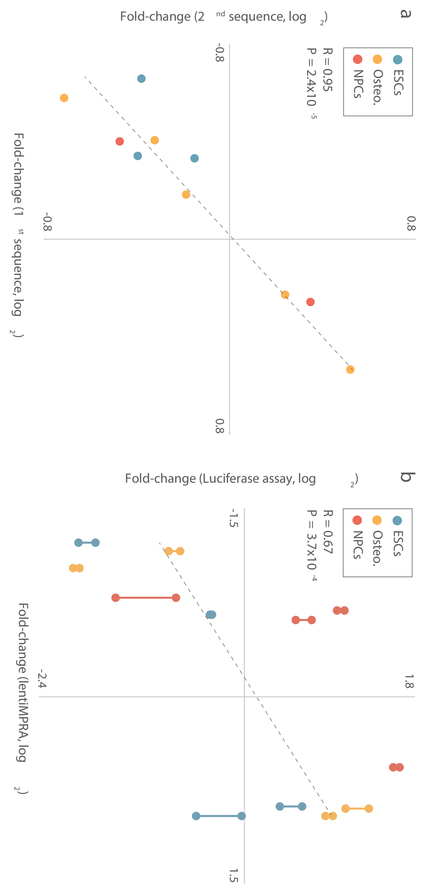
1103 Cellular Genetic Screens. *Cell* (2019). doi:10.1016/j.cell.2018.11.029
1104 56. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from
1105 thousands of CRISPR perturbations. *Nature Genetics* (2019). doi:10.1038/s41588-019-
1106 0538-0
1107 57. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of
1108 validated and target-linked human enhancers. *Nature Reviews Genetics* (2020).
1109 doi:10.1038/s41576-019-0209-0
1110 58. Jung, I. *et al.* A compendium of promoter-centered long-range chromatin interactions in
1111 the human genome. *Nat. Genet.* (2019). doi:10.1038/s41588-019-0494-8
1112 59. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes
1113 in GeneCards. *Database (Oxford).* (2017). doi:10.1093/database/bax028
1114 60. Gokhman, D. *et al.* Gene ORGANizer: Linking genes to the organs they affect. *Nucleic*
1115 *Acids Res.* **45**, W138–W145 (2017).
1116 61. Köhler, S. *et al.* The Human Phenotype Ontology project: Linking molecular biology and
1117 disease through phenotype data. *Nucleic Acids Res.* **42**, (2014).
1118 62. Zarate, Y. A. & Fish, J. L. SATB2-associated syndrome: Mechanisms, phenotype, and
1119 practical recommendations. *American Journal of Medical Genetics, Part A* (2017).
1120 doi:10.1002/ajmg.a.38022
1121 63. Liu, S. H. *et al.* A novel antisense long non-coding RNA SATB2-AS1 overexpresses in
1122 osteosarcoma and increases cell proliferation and growth. *Mol. Cell. Biochem.* (2017).
1123 doi:10.1007/s11010-017-2953-9
1124 64. Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome.
1125 *Nature* **489**, 57–74 (2012).
1126 65. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide
1127 association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*
1128 (2019). doi:10.1093/nar/gky1120
1129 66. Zarate, Y. A., Kaylor, J. & Fish, J. SATB2-Associated Syndrome. in (eds. Adam, M. P. et
1130 al.) (1993).
1131 67. Gigek, C. O. *et al.* A molecular model for neurodevelopmental disorders. *Transl.*
1132 *Psychiatry* (2015). doi:10.1038/tp.2015.56
1133 68. Qian, Y. *et al.* Paternal Low-Level Mosaicism-Caused SATB2-Associated Syndrome.

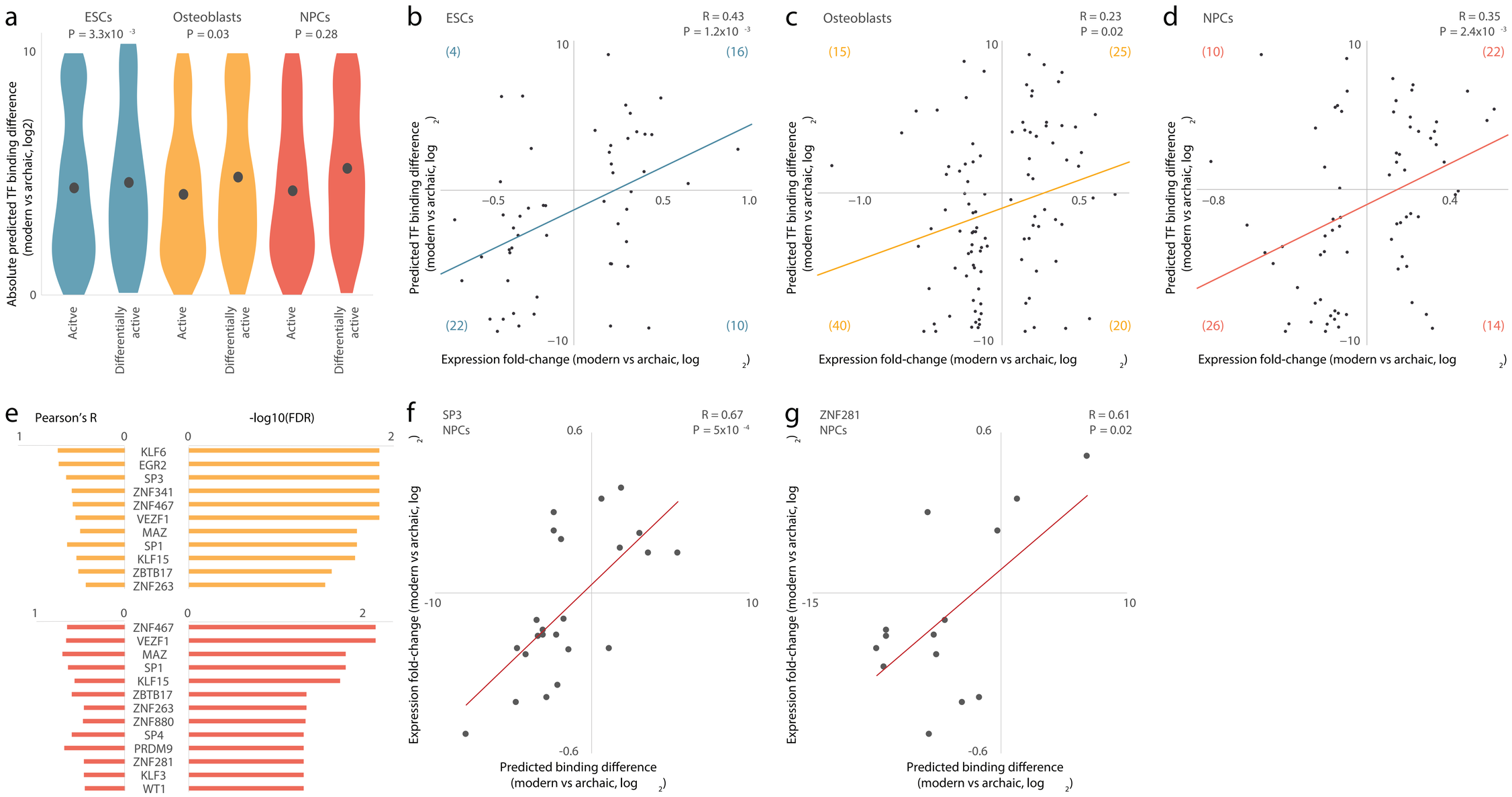
1134 *Front. Genet.* **10**, 630 (2019).
1135 69. Li, Y. *et al.* Satb2 Ablation Impairs Hippocampus-Based Long-Term Spatial Memory and
1136 Short-Term Working Memory and Immediate Early Genes (IEGs)-Mediated Hippocampal
1137 Synaptic Plasticity. *Mol. Neurobiol.* (2017). doi:10.1007/s12035-017-0531-5
1138 70. Zhang, Q., Huang, Y., Zhang, L., Ding, Y. Q. & Song, N. N. Loss of satb2 in the cortex
1139 and hippocampus leads to abnormal behaviors in mice. *Front. Mol. Neurosci.* (2019).
1140 doi:10.3389/fnmol.2019.00033
1141 71. Dobreva, G. *et al.* SATB2 Is a Multifunctional Determinant of Craniofacial Patterning and
1142 Osteoblast Differentiation. *Cell* (2006). doi:10.1016/j.cell.2006.05.012
1143 72. Mattioli, K. *et al.* Cis and trans effects differentially contribute to the evolution of
1144 promoters and enhancers. *Genome Biol.* **21**, 210 (2020).
1145 73. Prabhakar, S. *et al.* Human-specific gain of function in a developmental enhancer. *Science*
1146 *(80-.).* (2008). doi:10.1126/science.1159974
1147 74. Capra, J. A., Erwin, G. D., Mckinsey, G., Rubenstein, J. L. R. & Pollard, K. S. Many
1148 human accelerated regions are developmental enhancers. *Philos. Trans. R. Soc. B Biol.*
1149 *Sci.* (2013). doi:10.1098/rstb.2013.0025
1150 75. Villar, D., Flicek, P. & Odom, D. T. Evolution of transcription factor binding in
1151 metazoans-mechanisms and functional implications. *Nature Reviews Genetics* (2014).
1152 doi:10.1038/nrg3481
1153 76. True, J. R. & Carroll, S. B. Gene co-option in physiological and morphological evolution.
1154 *Annual Review of Cell and Developmental Biology* (2002).
1155 doi:10.1146/annurev.cellbio.18.020402.140619
1156 77. Telis, N., Aguilar, R. & Harris, K. Selection against archaic hominin genetic variation in
1157 regulatory regions. *Nat Ecol Evol* (2020). doi:10.1038/s41559-020-01284-0
1158 78. Prescott, S. L. *et al.* Enhancer Divergence and cis-Regulatory Evolution in the Human and
1159 Chimp Neural Crest. *Cell* **163**, 68–84 (2015).
1160 79. Lieberman, P. The Evolution of Human Speech: Its Anatomical and Neural Bases. *Curr.*
1161 *Anthropol.* **48**, 39–66 (2007).
1162 80. Mariën, P. *et al.* Consensus paper: Language and the cerebellum: An ongoing enigma.
1163 *Cerebellum* (2014). doi:10.1007/s12311-013-0540-5
1164 81. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-3.0. *RepeatMasker Open-3.0*

1165 (1996).
1166 82. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*
1167 (2012). doi:10.1038/nmeth.1923
1168 83. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE
1169 project. *Genome Res.* (2012). doi:10.1101/gr.135350.111
1170 84. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes.
1171 *Nucleic Acids Res.* (2019). doi:10.1093/nar/gky955
1172


a
ESCs
Osteoblasts
NPCs
Number of oligos
2000
0
450
Number of barcodes per oligo
b
rep1 vs. rep2
rep1 vs. rep3
rep2 vs. rep3
R = 0.77
R = 0.76
R = 0.78
R = 0.89
R = 0.88
R = 0.92
R = 0.94
R = 0.95
R = 0.96
P < 10 -100
log 2(RNA/DNA)
Relative density
c
All sequences
RNA/DNA > 3
Normalized Pearson's R
Percent of barcodes sampled
ESCs
NPCs
Osteoblasts
d
ESCs
P = 3x10 -8
Osteoblasts
P = 0.25 (putative pos.)
P = 1x10 -3 (ESC + NPC)
NPCs
P = 2x10 -6
RNA/DNA
Scrambled
Positive
Putative positive
ESC + NPC positive
Type of control


ESCs
Osteoblasts
NPCs
0%
100%
a
Whole genome
b
Fixed derived variants
c
Active
d
Differentially active
Active TSS
Flanking Active TSS
Bivalent/Poised TSS
Flanking Biv. TSS/Enh
Enhancers
Genic enhancers
Bivalent enhancers
Strong transcription
Weak transcription
Transcr. at gene 5' and 3'
Repressed PolyComb
Weak Rep. PolyComb
Quiescent/Low
Heterochromatin


a
ESCs
Osteo.
NPCs
R = 0.95
P = 2.4x10 -5
0.8
-0.8
0.8
-0.8
Fold-change (2nd sequence, log 2)
Fold-change (1st sequence, log 2)
b
ESCs
Osteo.
NPCs
R = 0.67
P = 3.7x10 -4
1.8
-2.4
1.5
-1.5
Fold-change (Luciferase assay, log 2)
Fold-change (lentiMP RA, log 2)


a
ESCs
P = 3.3x10 -3
Osteoblasts
P = 0.03
NPCs
P = 0.28
Absolute predicted TF binding difference (modern vs archaic, log2)
10
0
Acitve
Differentially active
Active
Differentially active
Active
Differentially active
b
ESCs
R = 0.43
P = 1.2x10 -3
(4)
(16)
(22)
(10)
Predicted TF binding difference (modern vs archaic, log2)
Expression fold-change (modern vs archaic, log2)
c
Osteoblasts
R = 0.23
P = 0.02
(15)
(25)
(40)
(20)
d
NPCs
R = 0.35
P = 2.4x10 -3
(10)
(22)
(26)
(14)
e
Pearson's R
-log10(FDR)
KLF6
EGR2
SP3
ZNF341
ZNF467
VEZF1
MAZ
SP1
KLF15
ZBTB17
ZNF263
ZNF467
VEZF1
MAZ
SP1
KLF15
ZBTB17
ZNF263
ZNF880
SP4
PRDM9
ZNF281
KLF3
WT1
f
SP3
NPCs
R = 0.67
P = 5x10 -4
Expression fold-change (modern vs archaic, log2)
Predicted binding difference (modern vs archaic, log2)
g
ZNF281
NPCs
R = 0.61
P = 0.02