RESEARCH ARTICLE

Figures and figure supplements

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

Carly V Weiss *et al*

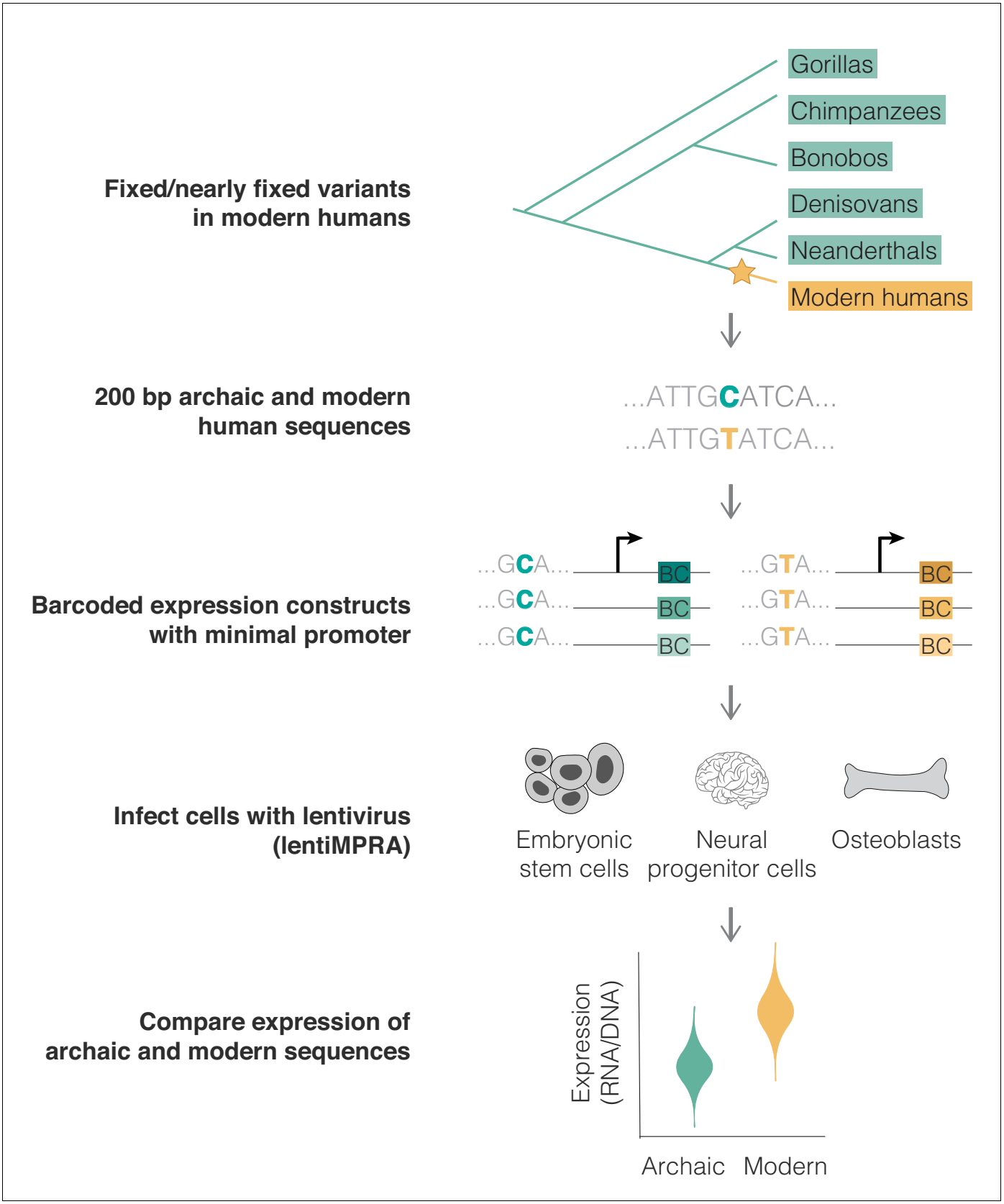

Figure 1. Using lentivirus-based MPRA (lentiMPRA) to identify variants driving differential expression in modern humans.

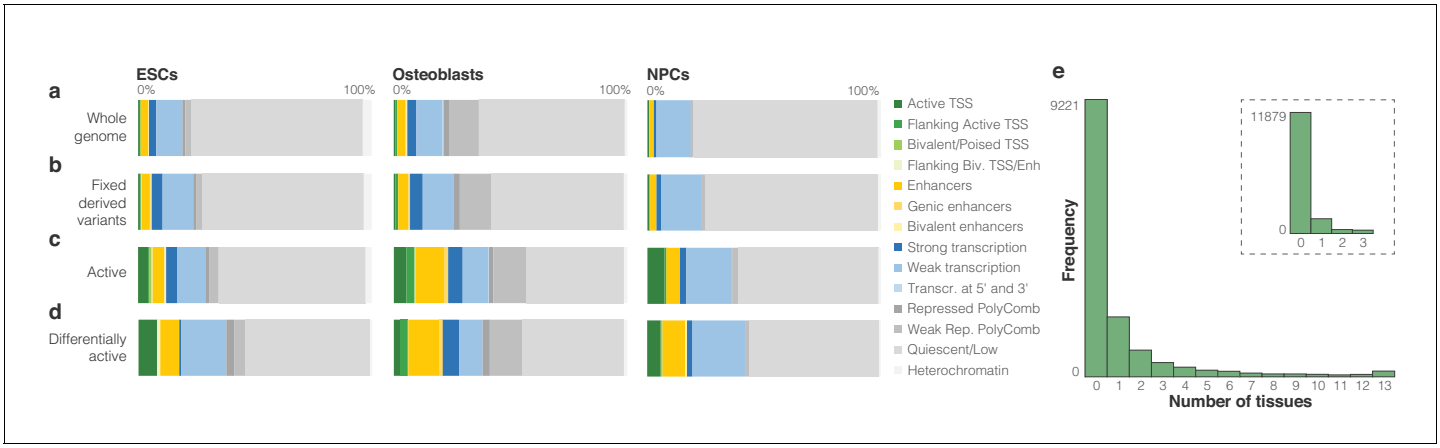

Figure 1—figure supplement 1. Classification of chromHMM annotations for different groups of variants. Relative percentage of bases in each chromHMM (***Ernst and Kellis, 2012***; ***Kundaje et al., 2015***) category throughout the entire genome (**a**), in fixed or nearly fixed modern human-derived variants (**b**), in active sequences (**c**), and in differentially active sequences (**d**), per cell type. See Discussion for cell-type specificity and enhancer enrichment. (**e**) Histogram of the number of tissues and number of sequences with transcription start site- (TSS) or enhancer-related chromHMM marks for all 14,042 sequences. Tissues and cell types investigated include embryonic stem cells (ESCs), osteoblasts, neural progenitor cells (NPCs), mesenchymal stem cells, monocytes, skin fibroblasts, brain hippocampus, skeletal muscle, heart left ventricle, sigmoid colon, ovary, fetal lung, and liver. Inset shows data for ESC, osteoblast, and NPC only.

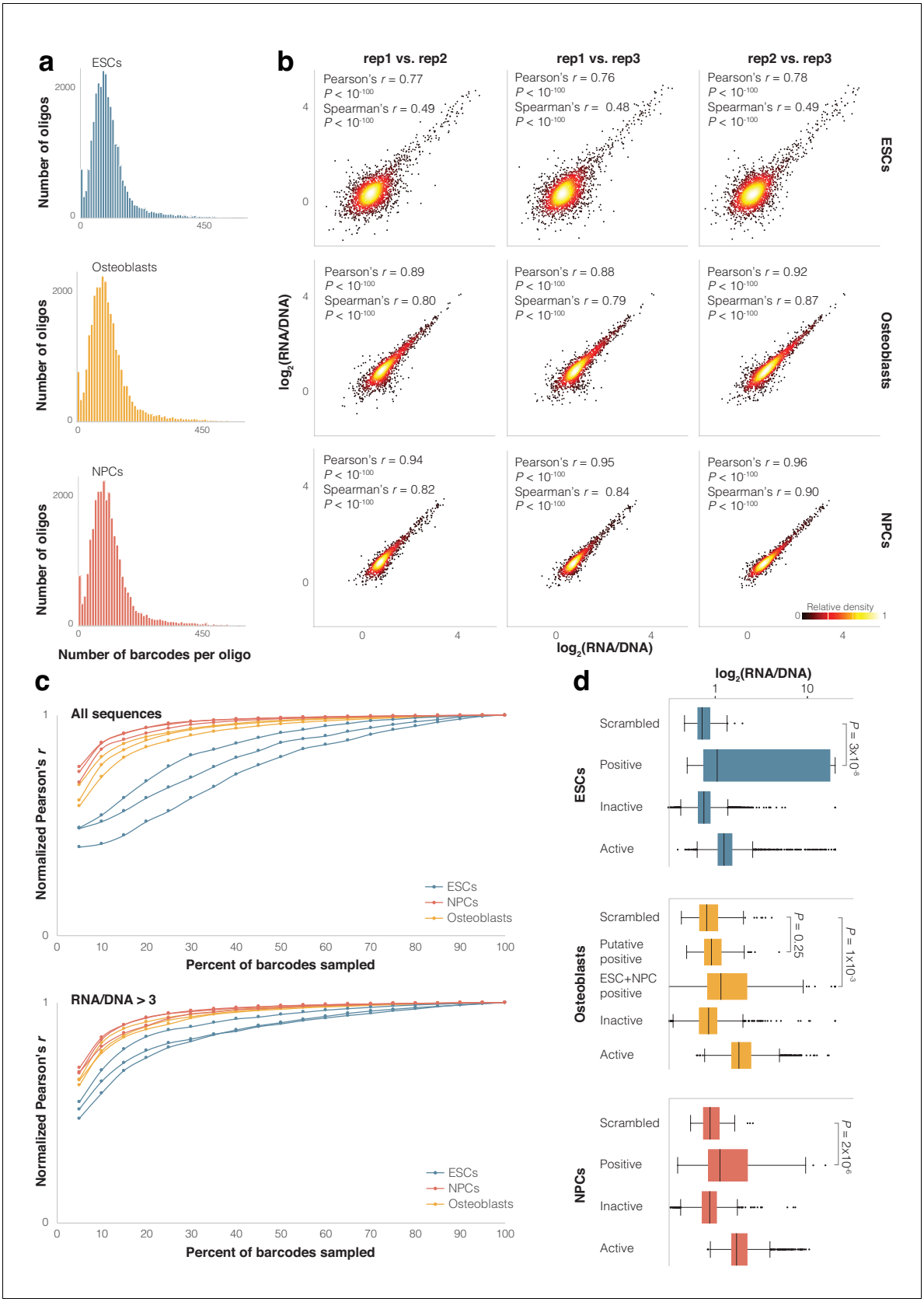

Figure 1—figure supplement 2. Reproducibility of lentivirus-based MPRA (lentiMPRA) data. (a) Distribution of number of barcodes per each sequence. (b) Replicate-by-replicate correlation of expression (RNA/DNA). Each point represents an active sequence. (c) Simulations of barcode downsampling

Figure 1—figure supplement 2 continued on next page

Figure 1—figure supplement 2 continued

showing Pearson's correlation of expression (RNA/DNA) between replicates. Upper panel shows all sequences and lower panel shows sequences with higher expression (RNA/DNA >3). Pearson's *r* values are normalized to maximum Pearson's *r* observed for each pair of replicates. (**d**) Box plots of scrambled, positive control, inactive and active sequences. One-sided *t*-test p-values are shown. Boxes show interquartile range (IQR), black line within box shows median, whiskers show 1.5× IQR from box borders, points show outliers.

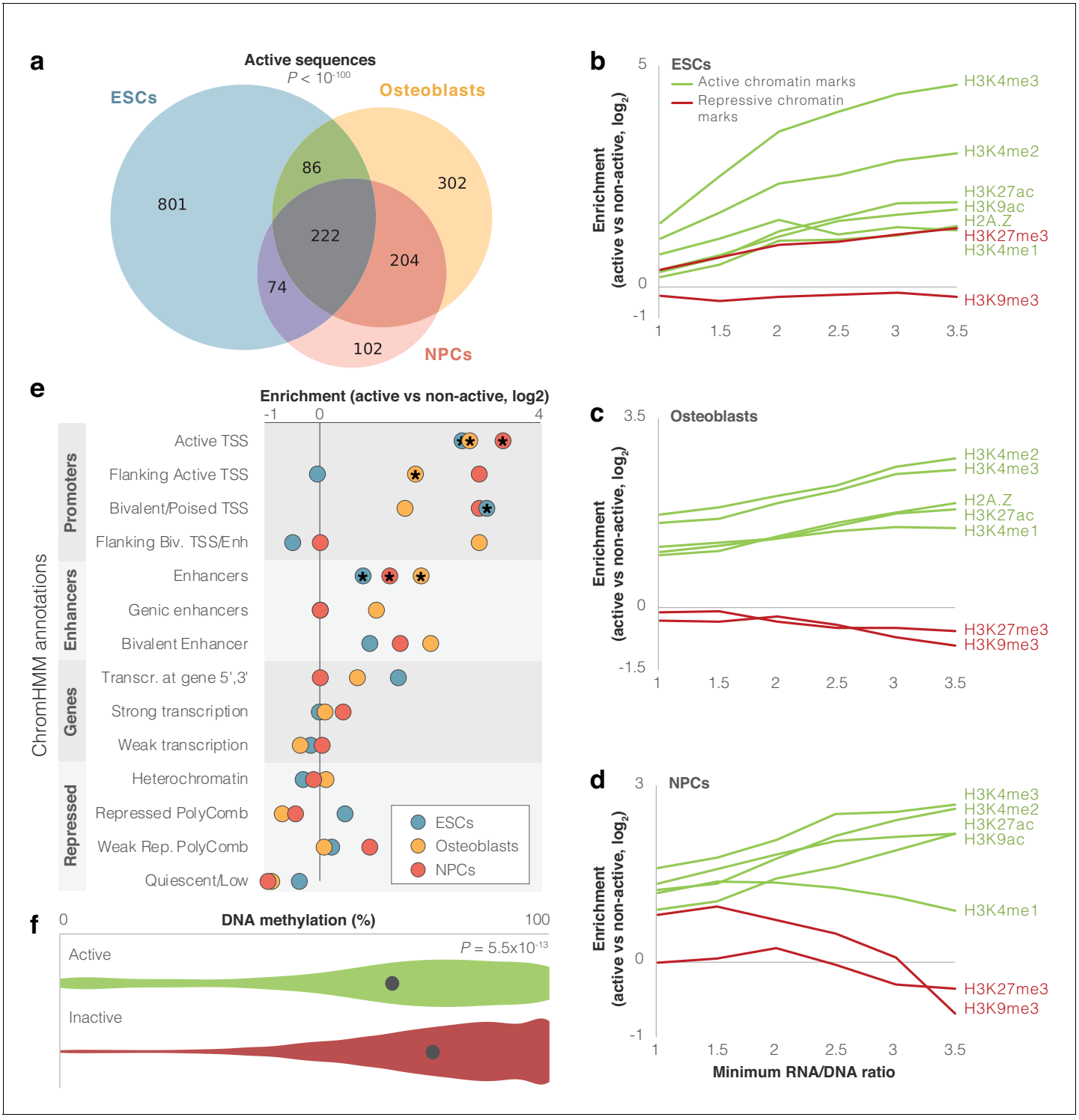

Figure 2. Identification of modern human sequences promoting expression in lentivirus-based MPRA (lentiMPRA). (**a**) Overlap between cell types of active sequences. Super Exact test p-value is shown for the overlap of the three groups. (**b-d**) Enrichment levels of active and repressive histone modification marks within active sequences. Enrichment is computed compared to inactive sequences. The enrichment of H3K27me3 in embryonic stem cells (ESCs) possibly reflects the presence of this mark in bivalent genes, which become active in later stages of development (***Blanco et al., 2020***). For confidence intervals, see ***Supplementary file 2***. (**e**) Enrichment of differentially active sequences in various chromatin-based genomic annotations. Missing circles reflect no differentially active sequences in that category. Stars mark significant enrichments (false discovery rate [FDR] <0.05). (**f**) Violin plots of DNA methylation levels for active (green) vs. inactive (red) sequences in osteoblasts. Methylation levels per sequence were computed as the mean methylation across all modern and archaic human bone methylation samples. The circle marks mean methylation across all sequences in each group. *t*-test p-value is shown.

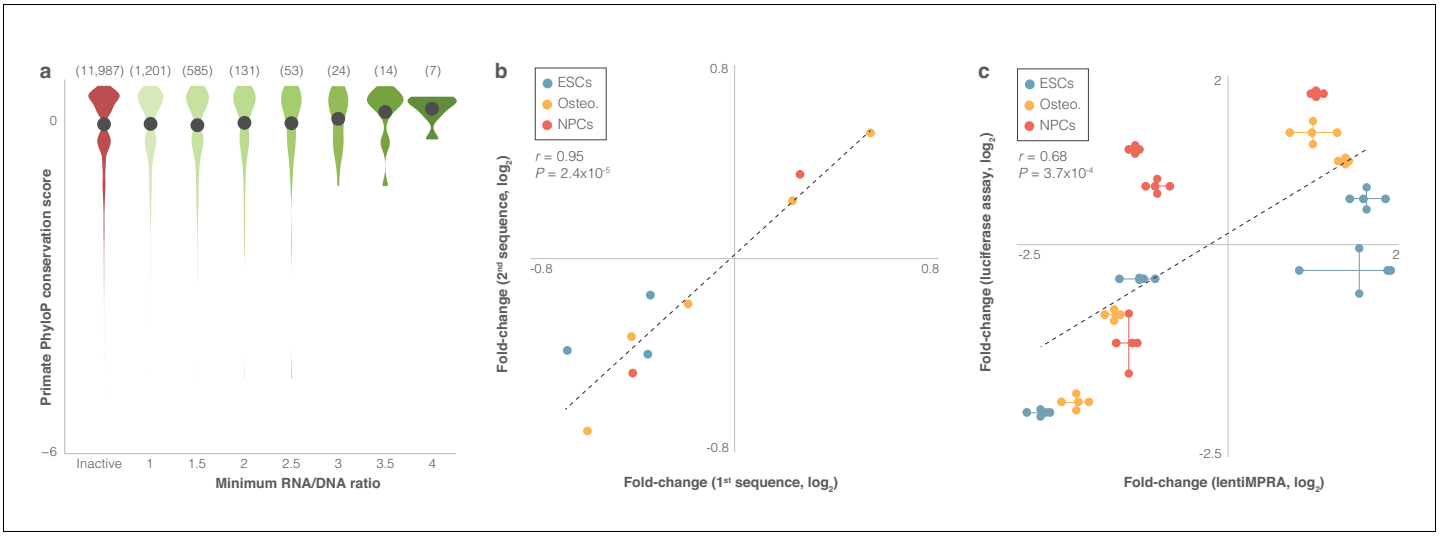

Figure 2—figure supplement 1. Differential expression is replicated across overlapping sequences and in a reporter assay validation. (**a**) Primate PhyloP conservation scores in inactive sequences and active sequences with increasingly higher RNA/DNA ratios (maximum RNA/DNA across the three cell types). Dots signify mean conservation per bin. Numbers in parentheses show number of sequences per bin. (**b**) Expression fold-change of overlapping pairs of sequences. Pearson's *r* and p-value are presented. (**c**) Expression fold-change of lentivirus-based MPRA (lentiMPRA) vs. luciferase assay. Each pair of points connected by a vertical line represents two replicates in the luciferase assay. Each triplet of points connected by a horizontal line represents three lentiMPRA replicates. Pearson's *r* and p-value are presented.

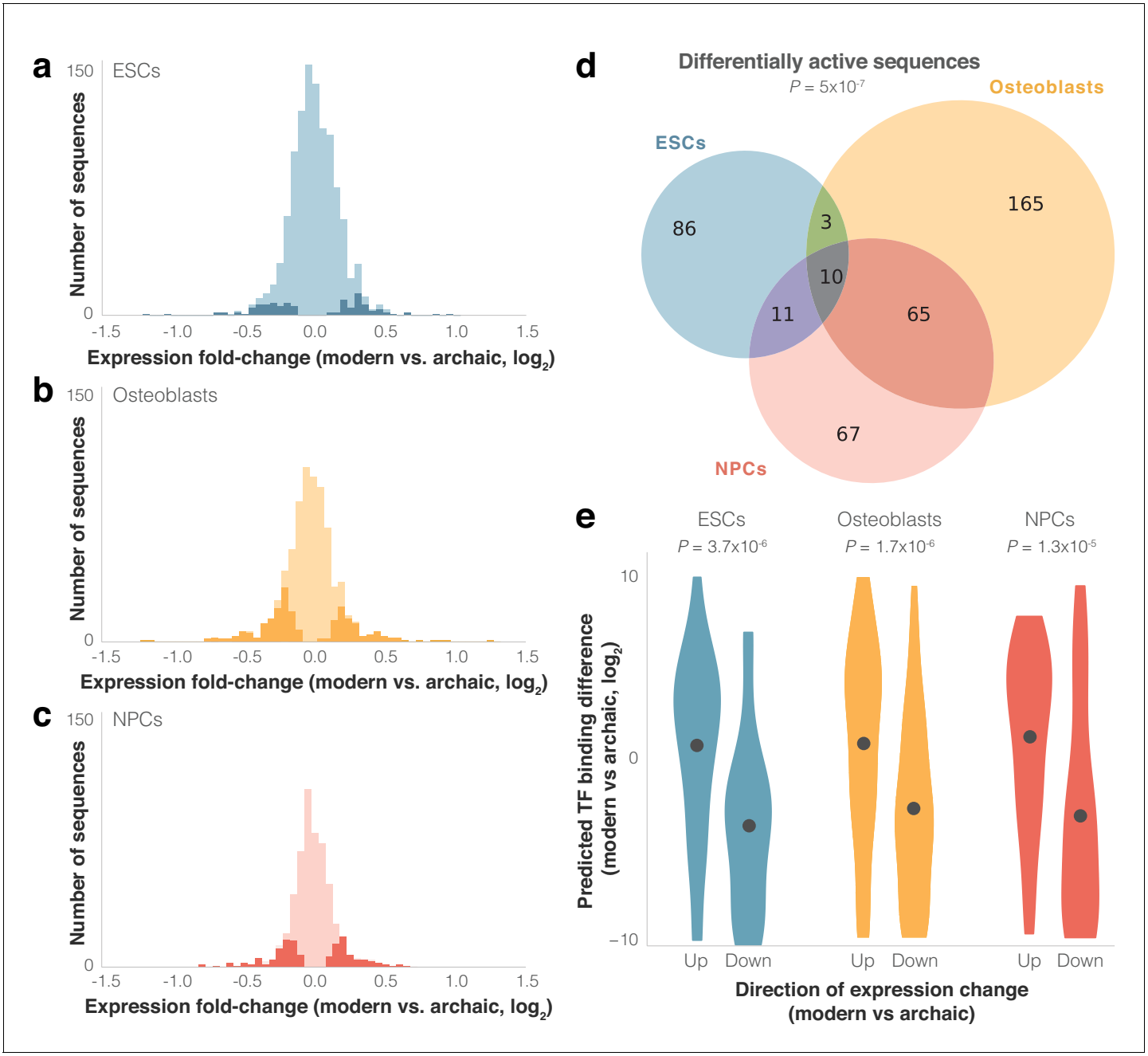

Figure 3. Differential activity of derived modern human sequences. (**a–c**) Distributions of expression fold-changes (RNA/DNA) of active (light) and differentially active (dark) sequences in each cell type. (**d**) Overlap of differentially active sequences between cell types. Super Exact test p-value is presented for the overlap of the three groups compared to active sequences. In the 10 sequences that were differentially active across all three cells types, the direction of fold-change was identical across all cell types ($p=1.9\times10^{-3}$, binomial test). (**e**) Violin plots of predicted transcription factor (TF) binding score difference between modern and archaic sequences. Positive scores represent increased binding in the modern sequence. Points show mean.

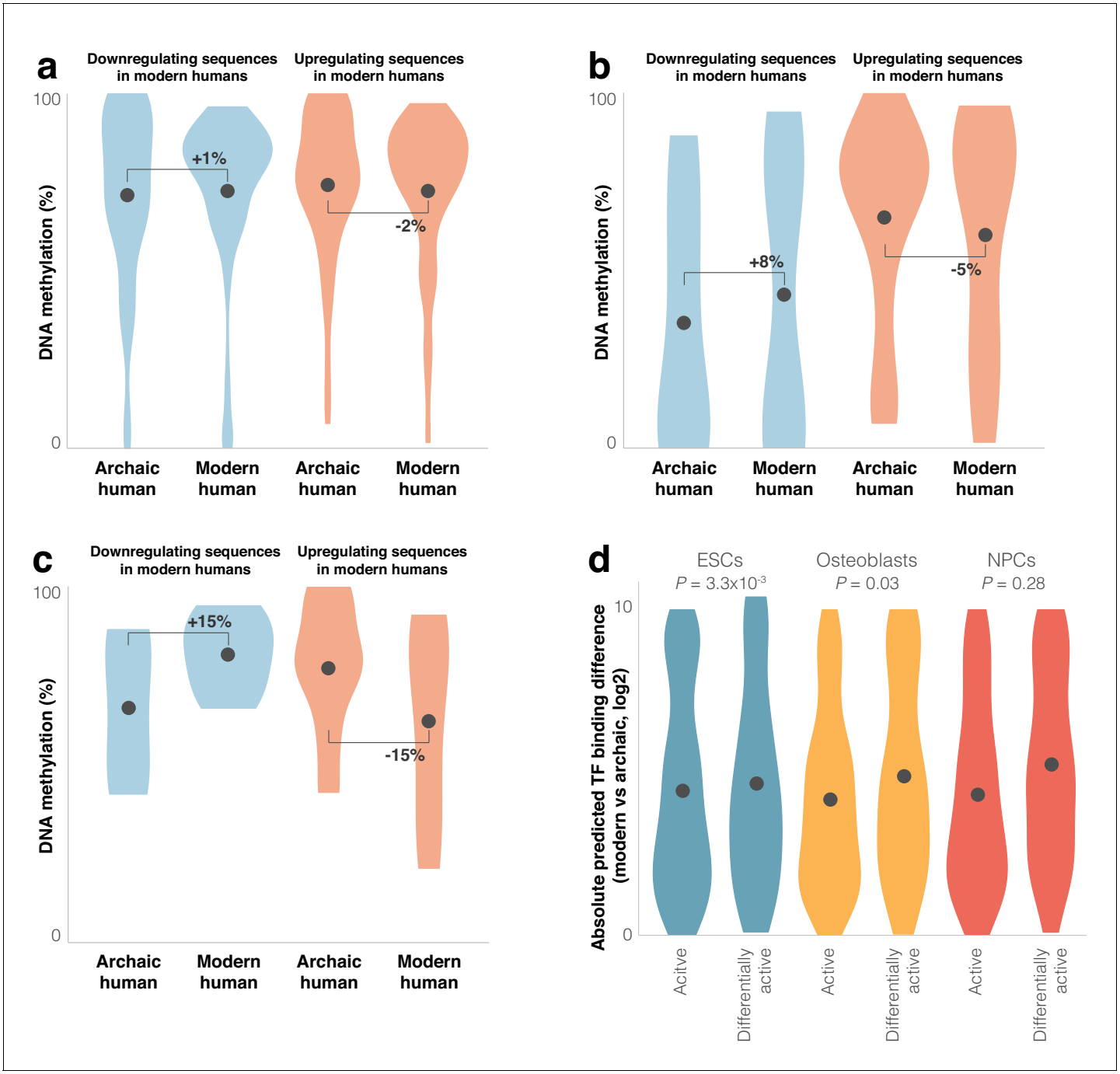

Figure 3—figure supplement 1. Differential activity is associated with differential DNA methylation and transcription factor (TF) binding. (**a–c**) Violin plots of DNA methylation levels in modern and archaic human bone methylation samples, for differentially active (**a**), promoter differentially active (**b**), and CpG-poor promoter differentially active (**c**) sequences in osteoblasts. Promoter sequences are sequences between 5 kb upstream and 1 kb downstream of a transcription start site (TSS). CpG-poor promoter sequences were defined as the bottom 50% promoter sequences. (**d**) Violin plots of absolute predicted TF binding score difference between modern and archaic sequences. Points show mean.

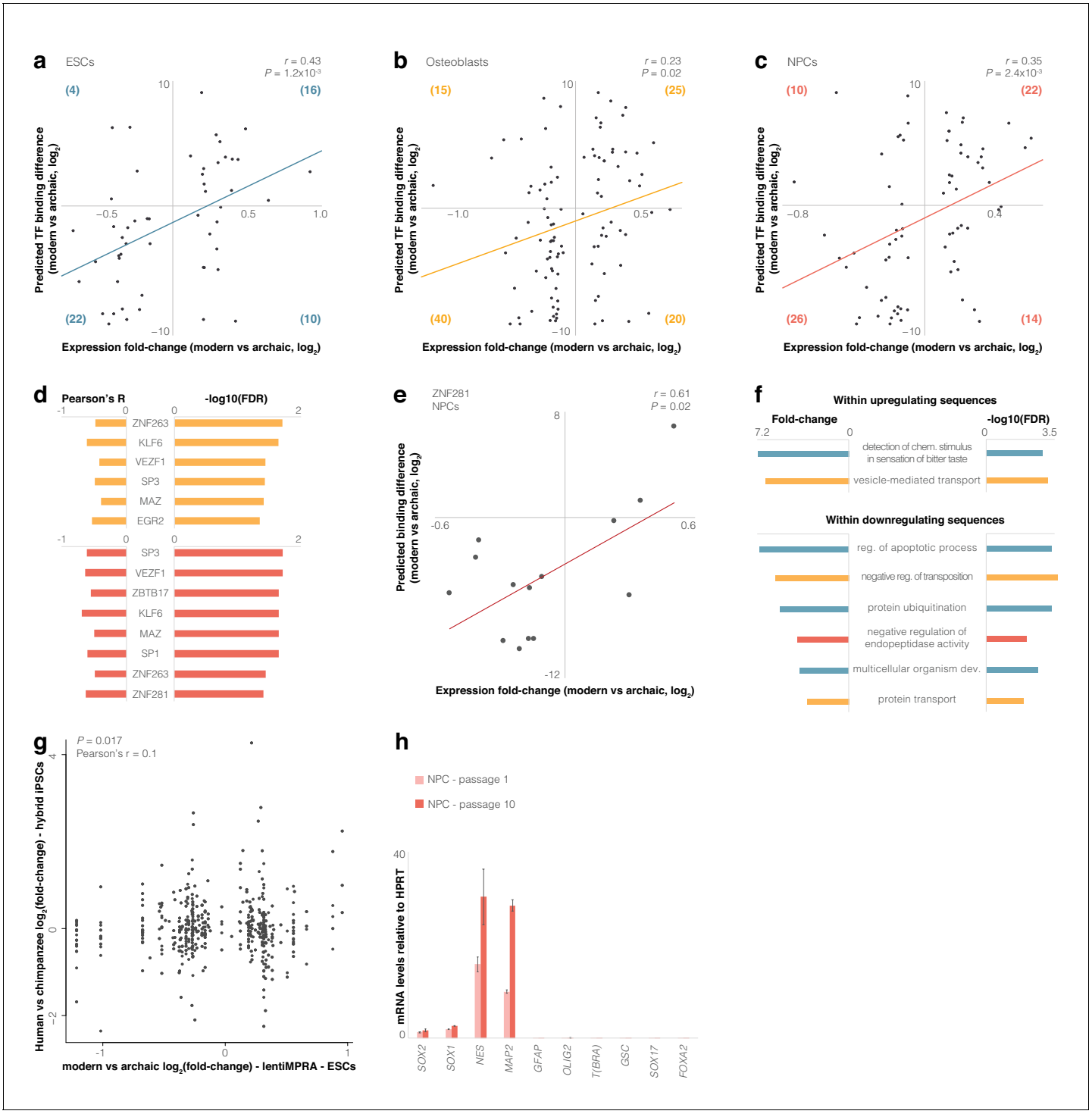

Figure 3—figure supplement 2. Predicted transcription factor (TF) binding is correlated with differential activity. (a–c) Expression fold-change vs. predicted TF binding fold-change for each sequence. Positive scores represent increased binding in the modern sequence. Parentheses show number of points in each quadrant with a score difference >0. (**d**) Pearson's correlation between differential expression and predicted differential binding affinity. Only significant TFs (false discovery rate [FDR] ≤0.05, ***Supplementary file 3***) are shown for osteoblasts (yellow) and neural progenitor cells (NPCs) (red). (**e**) Expression fold-change vs. predicted TF binding fold-change for ZNF281 in NPCs. Pearson's *r* and p-value are shown. (**f**) Enriched Gene Ontology terms for embryonic stem cells (ESCs) (blue), osteoblasts (yellow), and NPCs (red). (**g**) Expression fold-change of differentially active sequences compared to the *cis*-regulatory expression fold-change between human and chimpanzee of genes associated with these sequences. *cis*-regulatory expression changes were taken from hybrid human-chimpanzee induced pluripotent stem cells (iPSCs) (***Gokhman et al., 2021***). (**h**) RT-qPCR validation of NPCs at passage 1 (pink) and passage 10 (red). Expression levels are normalized to *HPRT* expression.

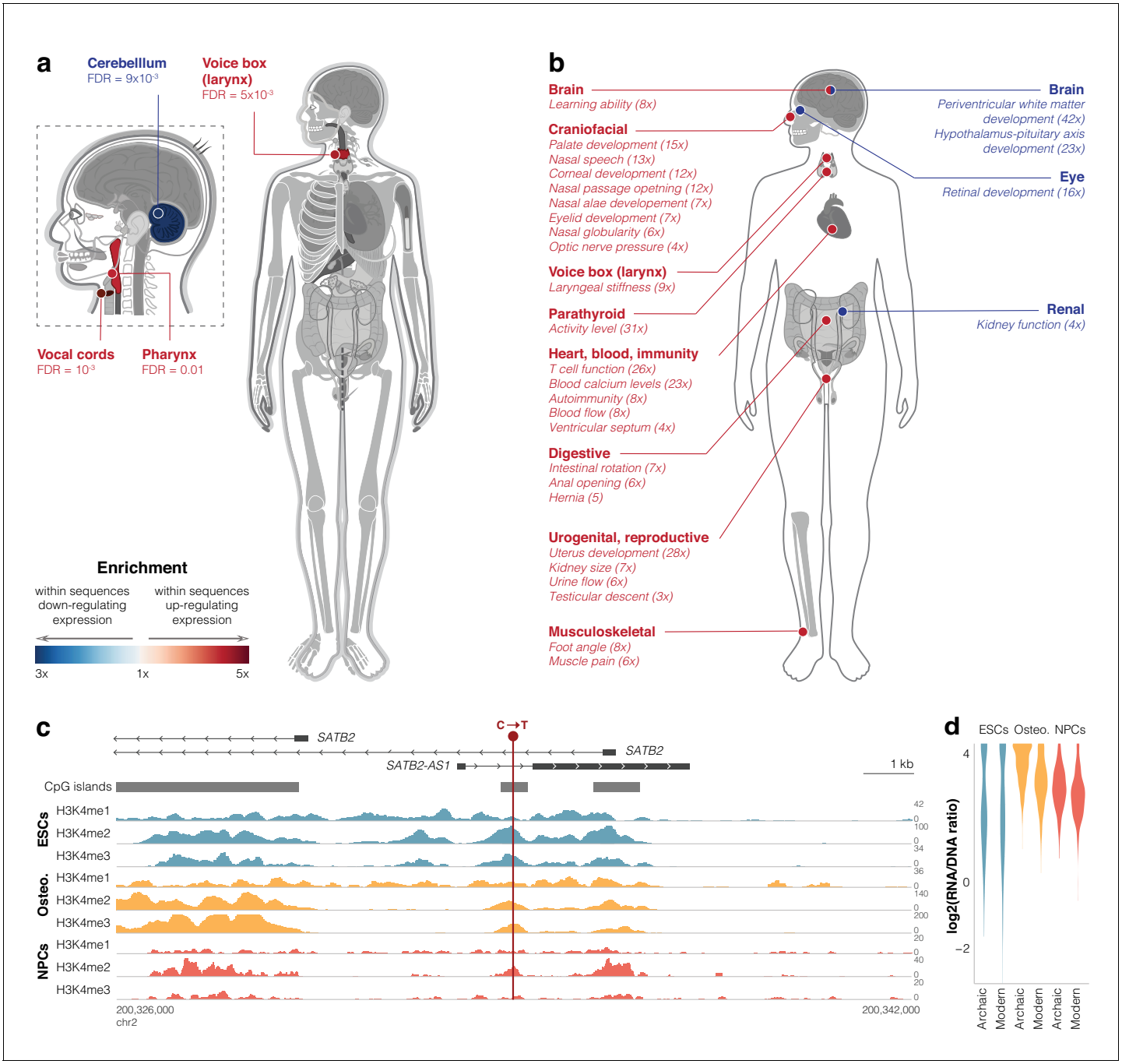

Figure 4. Differentially active sequences are linked to genes affecting the vocal tract and brain. (**a**) Gene ORGANizer enrichment map showing body parts that are significantly over-represented within genes linked to differentially active sequences (false discovery rate [FDR] <0.05). Organs are colored according to the enrichment scale. See ***Supplementary file 4*** for cell types. (**b**) Human Phenotype Ontology (HPO) phenotypes significantly enriched (FDR <0.05) within differentially active sequences. Fold enrichment is shown in parentheses. See ***Supplementary file 4*** for cell types. (**c**) CpG islands and read density of active histone modification marks (***Kundaje et al., 2015***) around the differentially active sequence in *SATB2* (GRCh37 genome version). (**d**) Violin plots of archaic vs. modern activity of the differentially active sequence in *SATB2*.