Supplemental: Function and constraint in enhancer sequences with multiple evolutionary origins

September 5, 2022

List of Figures

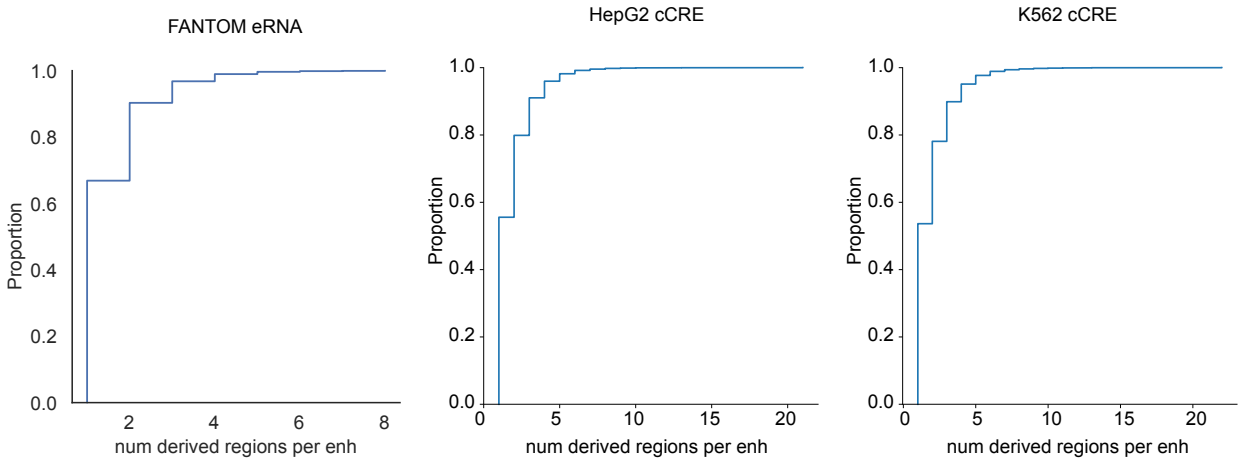

Figure S1: Number of derived regions per complex enhancer

Most complex enhancers have one derived region. Cumulative distribution plots show the number of derived regions as proportion of the total complex enhancer sequences for FANTOM5 eRNA (left, N = 10851), HepG2 cCREs from ENCODE (middle, N = 27289) and K562 cCREs from ENCODE (right, N = 24415). Complex enhancers have a median of one derived region across datasets.

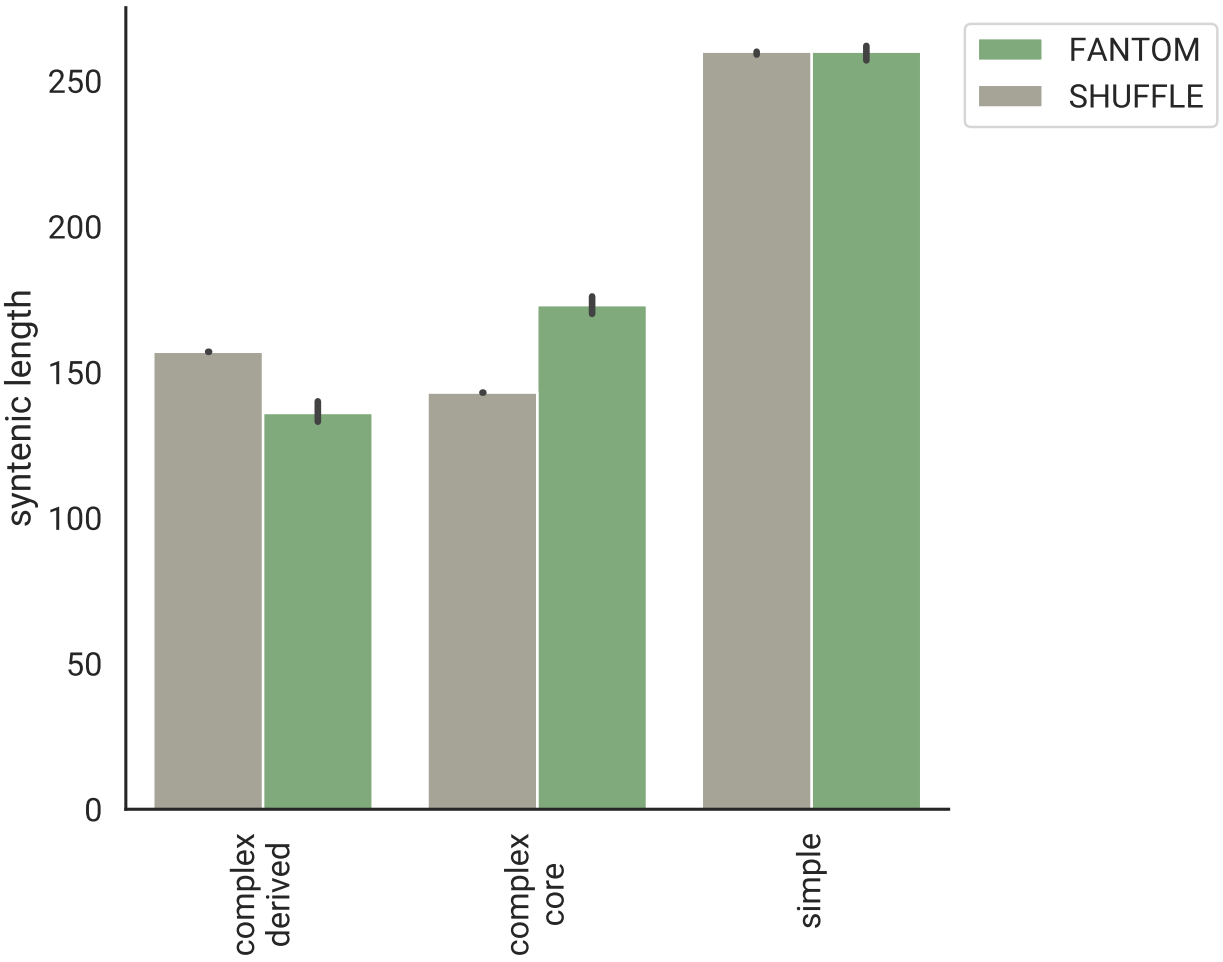

Figure S2: Sequence lengths of derived, core, and simple transcribed enhancers

Derived regions are also shorter than expected from 100 sets of length-, chromosome-, and architecture-matched random non-coding regions (left; median 136 bp derived v. 157 bp shuffled, p = 1.4e-46). Core sequences in complex enhancers are longer than 100x non-coding, chromosome-matched shuffled background cores (right; median 173 bp core v. 143 bp shuffle core, p = 2.4e-75). Sample size is annotated for each bar

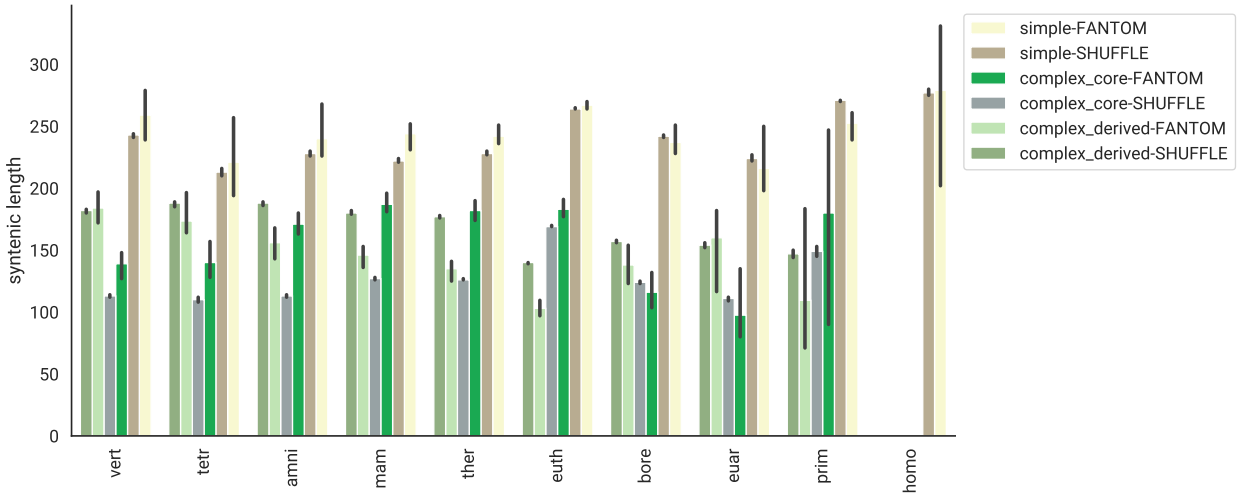

Figure S3: Lengths of derived, core, and simple enhancers versus expectation, stratified by core age

Derived, core and simple sequence lengths stratified by core age (x-axis) and compared with 100x shuffled sequences matched on core sequence age and architecture. Derived sequences are shorter than expected at every age except those with Vertebrate cores. Core sequences from the Eutherian ancestor and older are longer than expected.

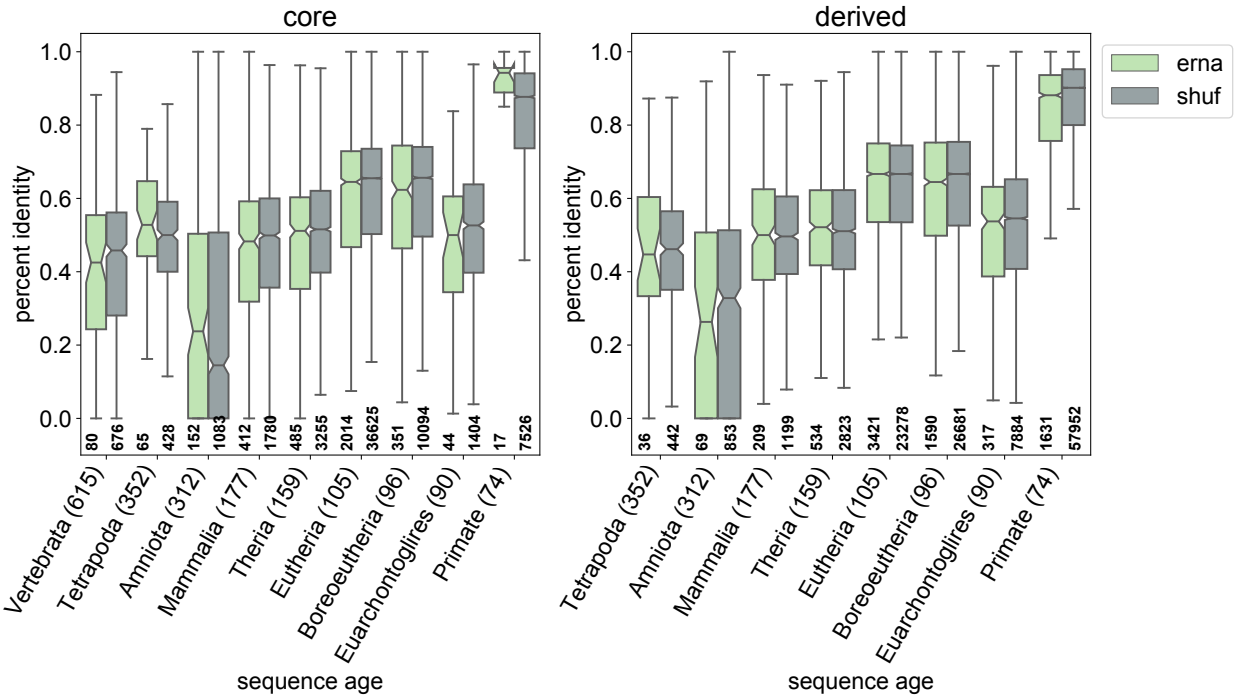

Figure S4: Sequence identity of core and derived regions is consistent with sequence identity of shuffled core and derived sequences

Core and derived sequence identities to their most distant detectable homolog are not significantly different from expected. Core (left) and derived (right) FANTOM sequence identity was quantified as the number of nucleotide mismatches between hg19 and the most distant aligned species (Methods). Stratified by sequence age (x-axis) and compared with their expected sequence identities based on 100x shuffled sequences matched on sequence age and architecture, derived and core sequences do not show significantly different sequence identities (Welch's p-value ¿0.05). Therian and Eutherian cores have slightly lower sequence identity compared with the expectation (median 0.51 Therian core v. 0.52 expected Therian core and 0.64 Eutherian core v. 0.65 expected Eutherian core, Welch's p-value ¡ 0.05). Moreover, the sequence identities are well above the range at which detecting homology becomes challenging for all branches except Amniota, which only contains a very small number of derived regions on it (69) or adjacent branches (36 and 209). These results do not show any evidence of systematic mis-classification of the age of enhancer segments due to varying rates of sequence divergence. The number of elements in each category is annotated below the boxplot. Boxes show the median and interquartile range of sequence identity values. Whiskers reflect 1.5x the interquartile range.

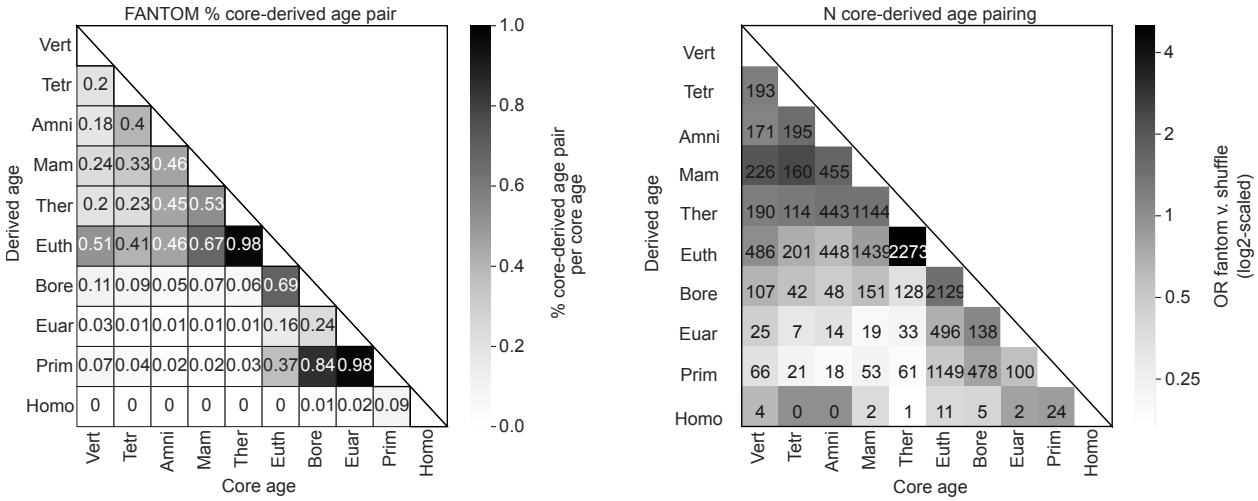

Figure S5: Frequency and count of core-derived sequence age pairs in FANTOM

Frequency (left) and count (right) of core-derived age pairs across complex FANTOM enhancers. Shading in the frequency plot (left) reflects the percentage of age-pairs within a single core age. Cores may have more than one derived sequence of a different age, thus the sum of the columns can be greater than one. Shading in the count plot (right) reflects the enrichment of the core-derived age pair compared with shuffled expectation shown in Figure 3.

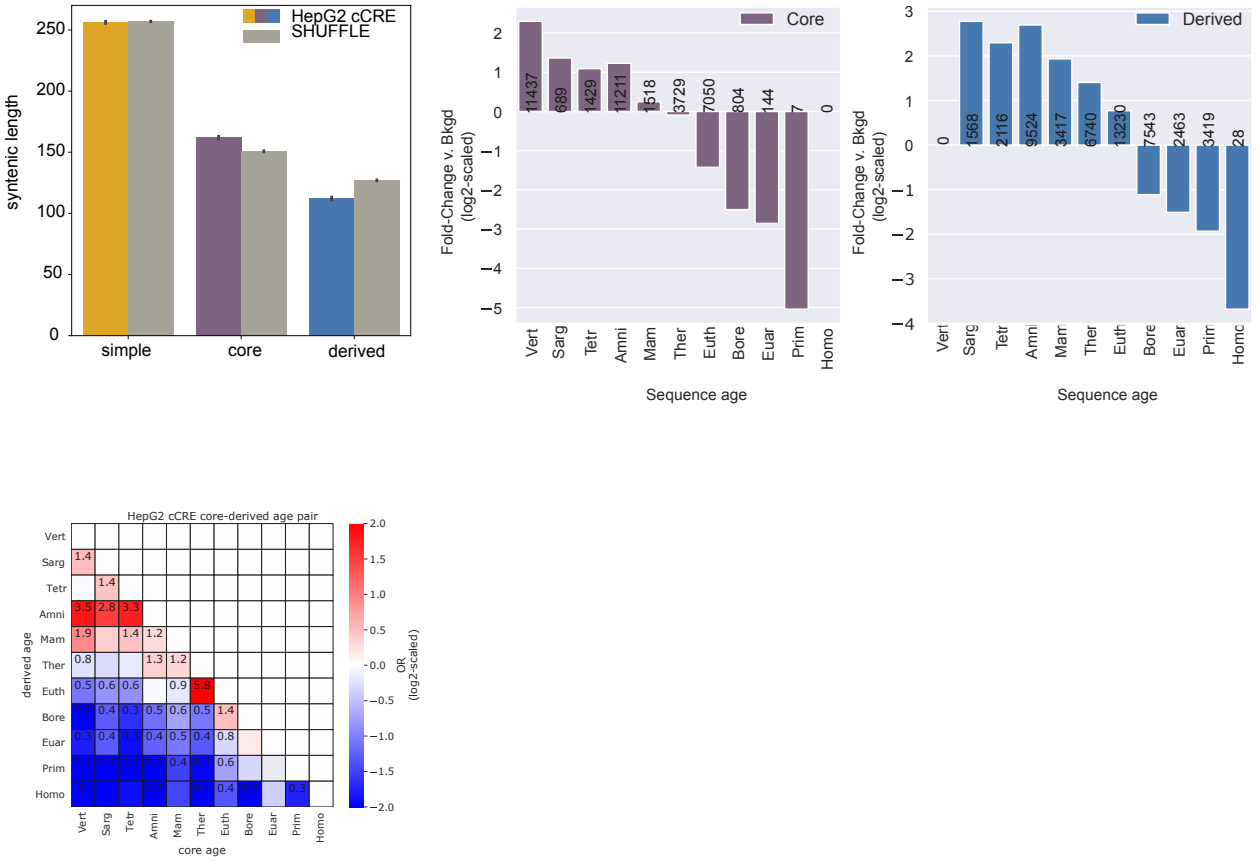

Figure S6: Core and derived evolutionary features in complex HepG2 cCREs recapitulate evolutionary features in FANTOM eRNAs

Derived regions constitute a sizeable portion of complex HepG2 cCREs (N = 27,789 cCREs), are shorter (top left) and older (top right) than expected compared to shuffled complex enhancer architectures (N = 1,047,557). Core sequences from the Mammalian ancestor and older are enriched for derived sequences from the Therian ancestor and older compared with shuffled expectation of core-derived age pairs. These core sequences are also depleted of sequences younger than the Therian ancestor. Core sequences are enriched for the nearest, younger phylogenetic neighbor. Odds ratio of significantly enriched age-pairs (FDR$<$ 0.05) are annotated.

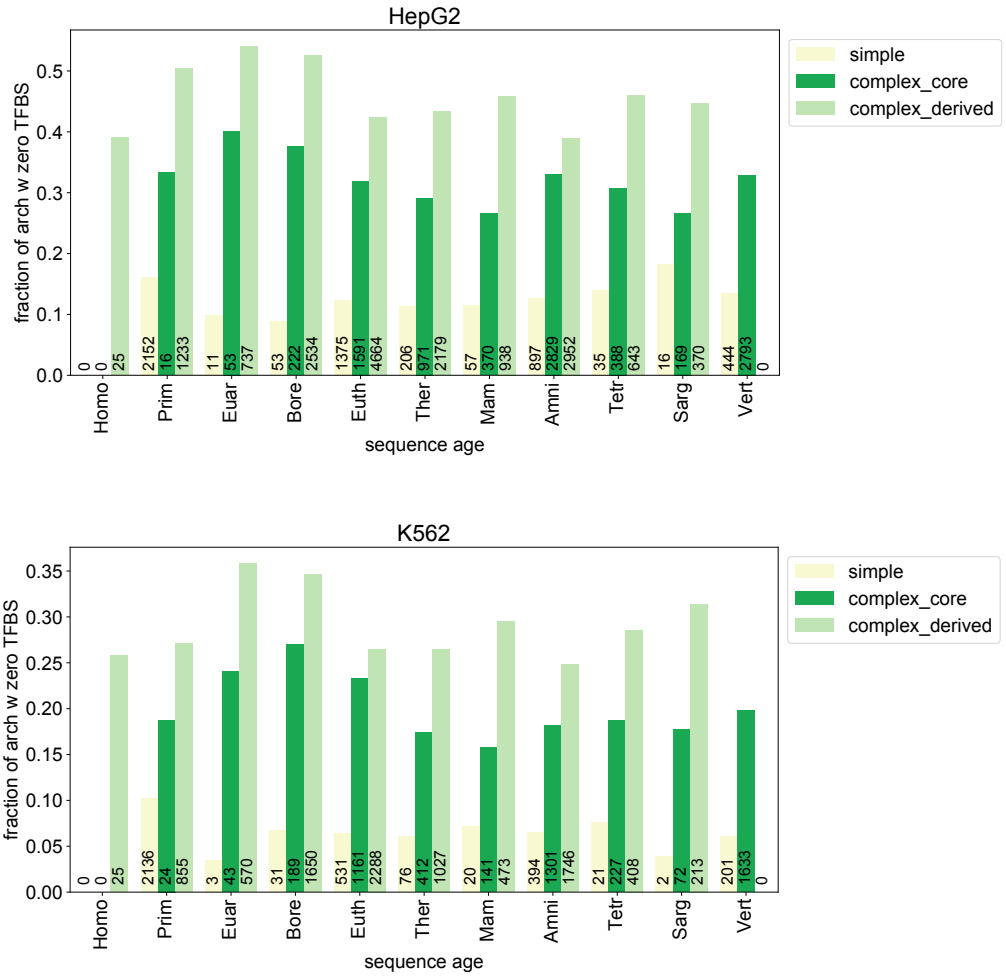

Figure S7: Regions with no TFBS ChIP-seq binding are observed across ages

Similar proportions of derived, core, and simple enhancer sequences have no evidence of TFBS binding within sequence ages in HepG2 and K562 cCREs. K562 cell models generally have fewer elements that do not overlap TFBS, likely because more TFBS ChIP-seq assays have been performed in K562 cells compared with HepG2 cells (249 v. 119 assays, respectively). Enhancer regions are binned according to their syntenic sequence ages. Frequency is calculated as the percent of regions that do not overlap TFBS ChIP-seq peaks within each sequence age and region category. HepG2 is shown above and K562 is shown below. Number of regions with zero TFBS overlap is annotated for each bar.

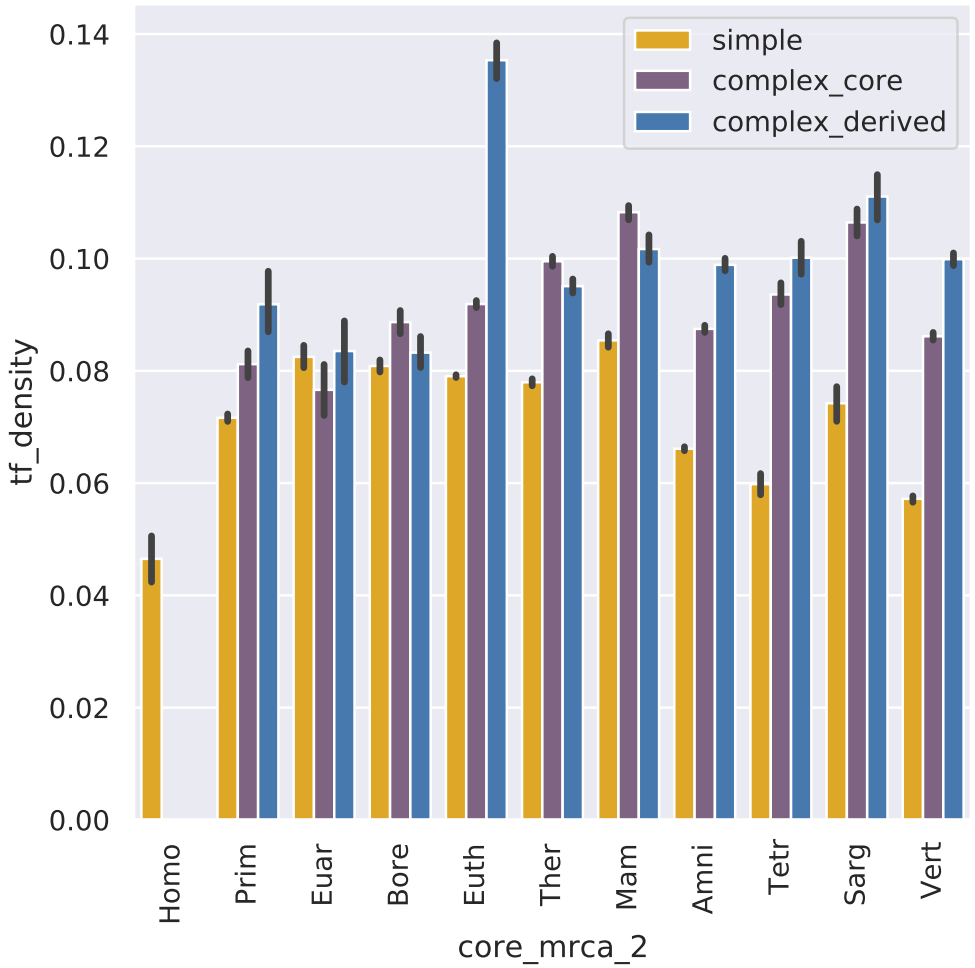

Figure S8: Transcription factor binding site density is similar across ages in HepG2 cCREs

Simple, core, and derived sequences are stratified by core age on the x-axis. TFBS density per architecture and age was measured and plotted on the y-axis.

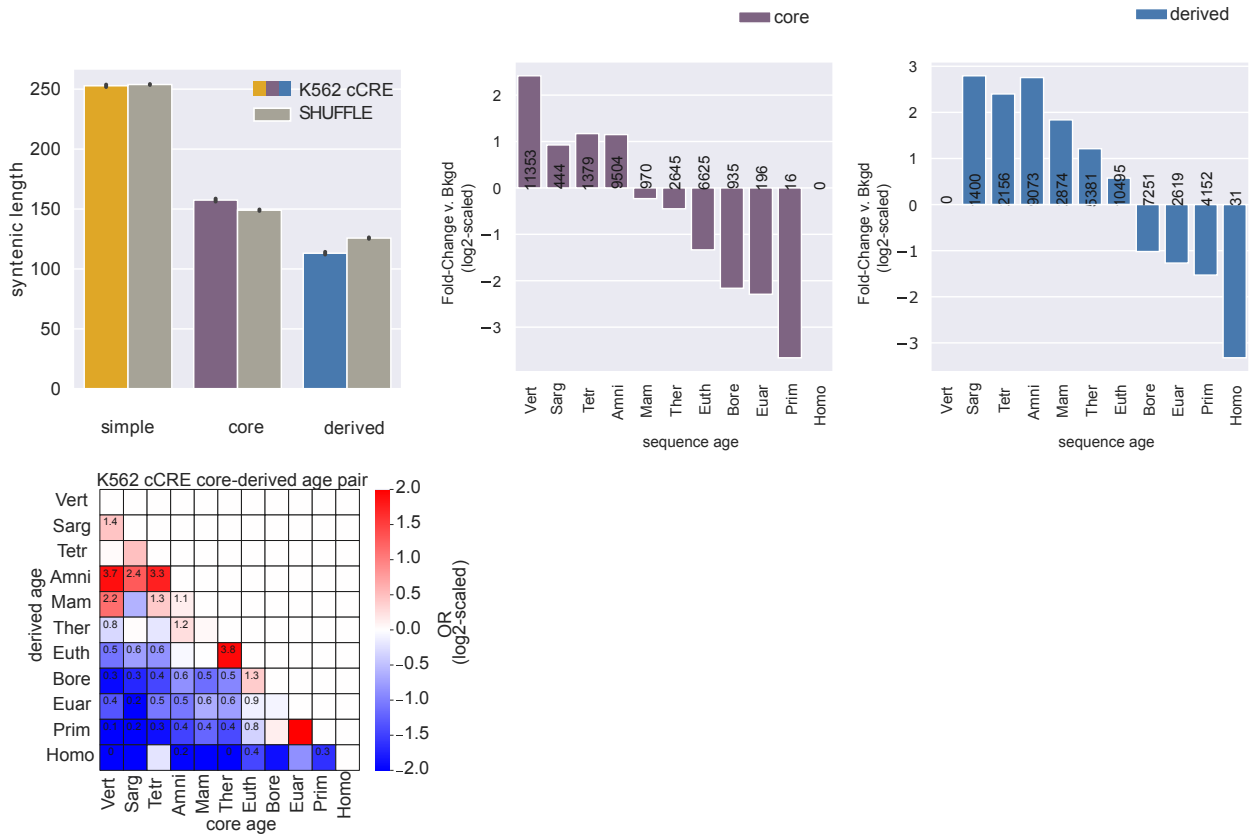

Figure S9: Core and derived evolutionary features in complex K562 cCREs recapitulate evolutionary features in FANTOM eRNAs

Derived regions constitute a sizeable portion of complex K562 cCREs (N = 24,415 cCREs), are shorter (top left) and older (top right) than expected compared to shuffled complex enhancer architectures (N = 473,387 cCREs). Core sequences from the Amniota ancestor and older are enriched for derived sequences from the Mammalian ancestor and older compared with shuffled expectation of core-derived age pairs. These core sequences are also depleted of sequences younger than the Mammalian ancestor. Core sequences are enriched for the nearest, younger phylogenetic neighbor. Odds ratio of significantly enriched age-pairs (FDR$<$ 0.05) are annotated.

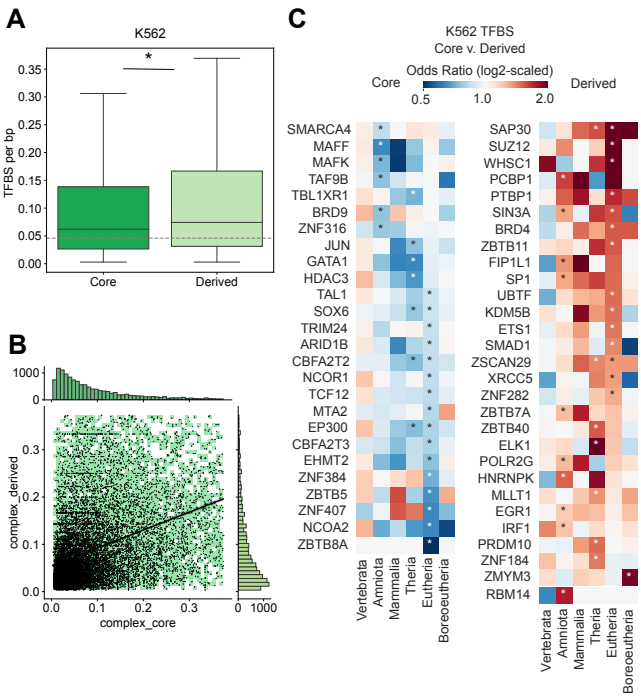

Figure S10: Derived regions have high transcription factor binding site densities and bind different transcription factors compared to core regions in K562 cells

(A) Derived regions (N = 23868) have higher TFBS densities than core regions (N = 20997) (0.074 derived v. 0.062 core TFBS per base pair, Mann Whitney-U p = 3.5e-52). Simple enhancer TFBS density is lower than core and derived regions (0.05 TFBS per base pair) **(B)** TFBS density is positively correlated between core-derived sequence pairs within complex enhancers with evidence of TF binding in both core and derived regions (N = 14142). Color intensity represents the density of core-derived pairs, and the black line is a linear regression fit (slope=0.39, intercept=0.056, r=0.39, p $<$2.2e-238, stderr=0.008; outliers ($>$95th percentile) are not plotted for ease of visualization. **(C)** Derived and core regions of the same age are enriched for binding of different TFs and enrichment patterns are generally consistent across ages. TFBS enrichment for each age was tested using Fisher's exact test; only TFs with at least one significant enrichment (FDR $<$ 0.1) are shown. Vertebrate, Sarcopterygii, and Tetrapod enhancer ancestors were grouped into "Vertebrata". Boreotherian, Euarchontoglires, and Primate enhancer ancestors were grouped into "Boreoeutheria".

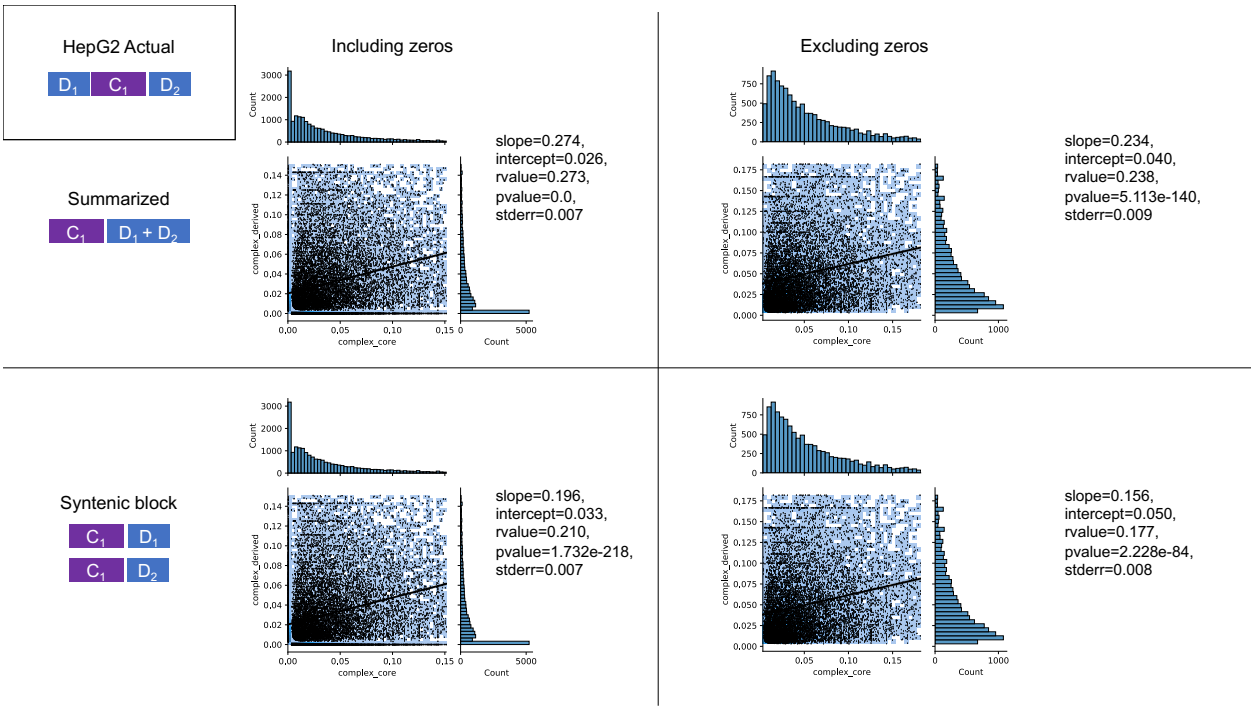

Figure S11: High TFBS density in core regions correlates with high TFBS density in derived regions within the same HepG2 enhancer sequence

We evaluated TFBS density correlations between core and derived sequences of the same enhancer in HepG2 cCREs. TFBS density of core and matched-derived regions per enhancer are plotted on the X- and Y-axis, respectively. Actual enhancers can have more than one core or derived region, so we evaluated our data using two different approaches. In the first (upper) we summarized TFBS density across multiple core and derived regions by summing TFBS density and syntenic length into core and derived groups and quantifying TFBS density in summarized core and derived regions per enhancer sequence. In the second (lower), we quantified TFBS density for every core and derived syntenic region and compared all possible pairs of core and derived syntenic TFBS densities per enhancer. We applied two different thresholds for evaluating TFBS density in core versus matched derived regions; one threshold allowed for regions with no evidence of TFBS in core or derived sequence, but not both (left, “zeros included”), while the other threshold required that TFBS binding was detected in both core and derived sequences within an enhancer (right, “zeros excluded”). Linear regression models were fit for each dataset and model features are annotated for each analysis. Histograms (right and above) display distributions of core and derived TFBS density per analysis.

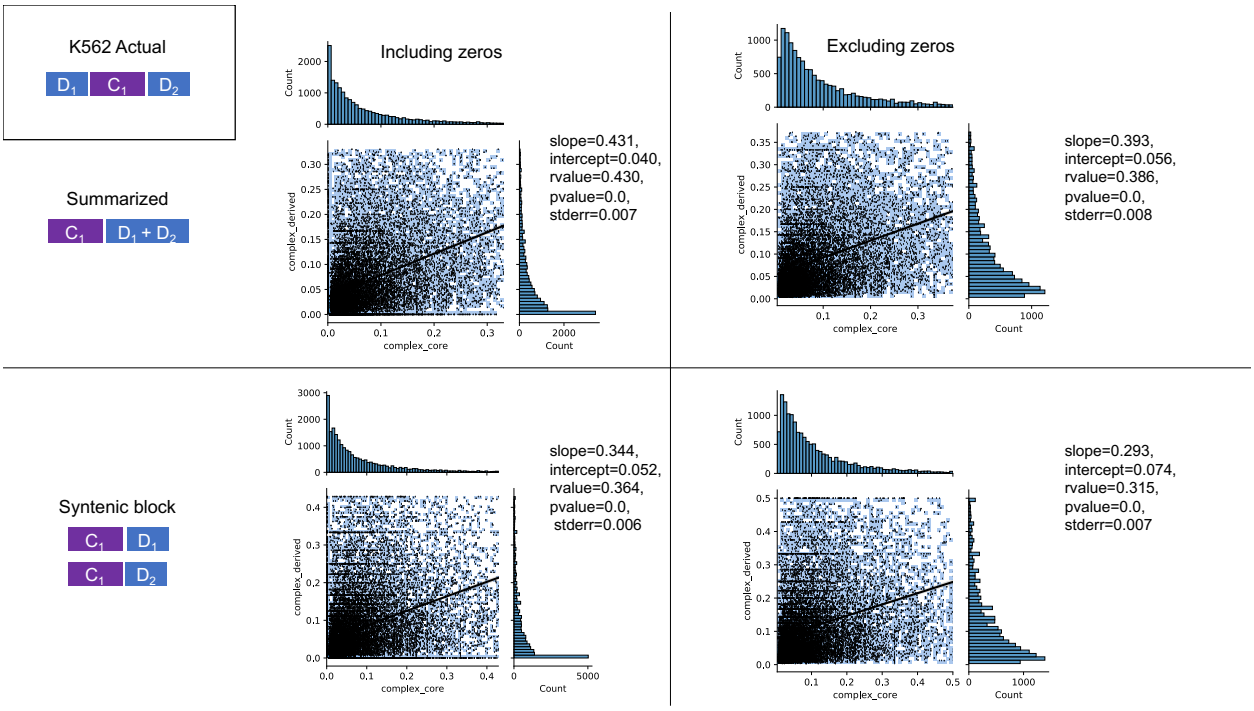

Figure S12: High TFBS density in core regions correlates with high TFBS density in derived regions within the same K562 enhancer sequence

We evaluated TFBS density correlations between core and derived sequences of the same enhancer in K562 cCREs. TFBS density of core and matched-derived regions per enhancer are plotted on the X- and Y-axis, respectively. Actual enhancers can have more than one core or derived region, so we evaluated our data using two different approaches. In the first (upper) we summarized TFBS density across multiple core and derived regions by summing TFBS density and syntenic length into core and derived groups and quantifying TFBS density in summarized core and derived regions per enhancer sequence. In the second (lower), we quantified TFBS density for every core and derived syntenic region and compared all possible pairs of core and derived syntenic TFBS densities per enhancer. We applied two different thresholds for evaluating TFBS density in core versus matched derived regions; one threshold allowed for regions with no evidence of TFBS in core or derived sequence, but not both (left, “zeros included”), while the other threshold required that TFBS binding was detected in both core and derived sequences within an enhancer (right, “zeros excluded”). Linear regression models were fit for each dataset and model features are annotated for each analysis. Histograms (right and above) display distributions of core and derived TFBS density per analysis.

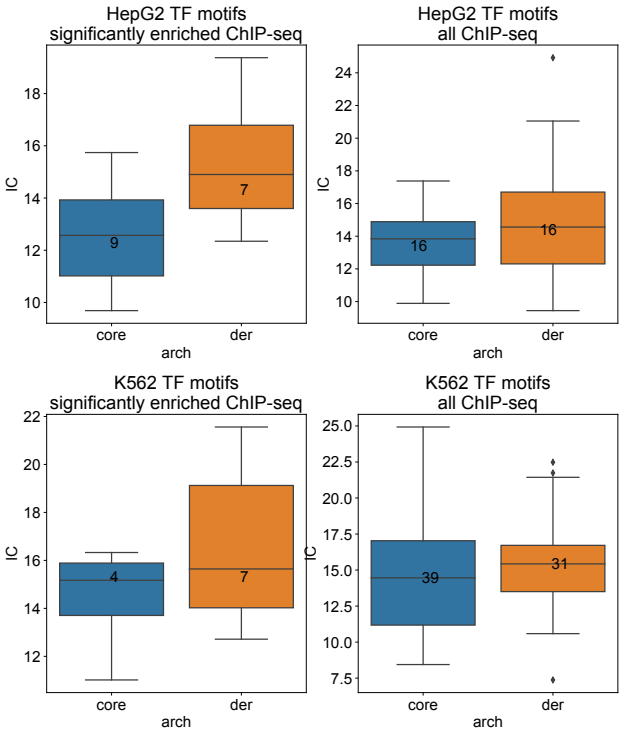

Figure S13: Information content of TFBS motifs in derived sequences is comparable with motifs in core sequences in HepG2 and K562 enhancers

We quantified the information content (IC) of JASPAR core vertebrate non-redundant TF motifs corresponding to significantly enriched HepG2 ChIP-seq signal in core or derived regions. We observed higher IC in derived motifs than in core motifs (upper left panel; median 14.9 derived v. 12.6 core IC, Welch's test p-value = 0.03). We performed a similar analysis in K562 ChIP-seq datasets and found derived TF motifs have higher information content than cores, but this was not significant (lower left panel; median 15.6 derived v. 15.2 core IC, Welch's test p-value = 0.26). Relaxing our criteria, we also evaluated IC for all TF motifs with any enrichment for ChIP-seq binding in core/derived sequences. IC was similar for TF motifs in HepG2 elements (upper right; median 13.8 core and 14.6 derived information content, Welch's test p-value = 0.18) and K562 elements (lower right; median 14.6 core and 15.4 derived information content, Welch's test p-value = 0.35). Together, these data support that derived TF motifs are just as robust to mutations as core motifs.

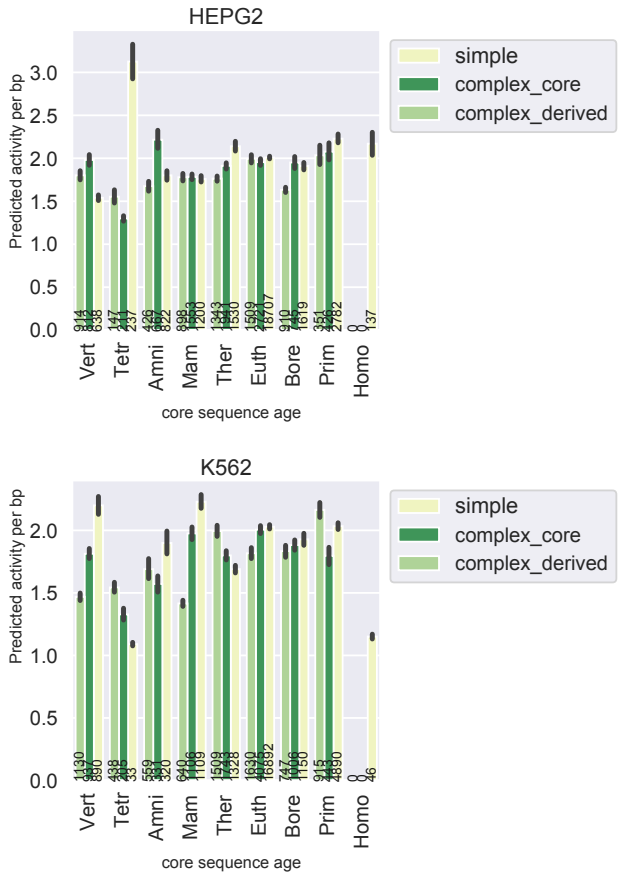

Figure S14: MPRA activity is similar across sequence ages and simple, core, or derived contexts

MPRA predicted activity per bp from Ernst 2016 is similar across ages in K562 and HepG2 cells. Here, predicted activity per bp scores are stratified by core sequence age and simple, core, or derived category. Cell line models and N bp are annotated per bar.

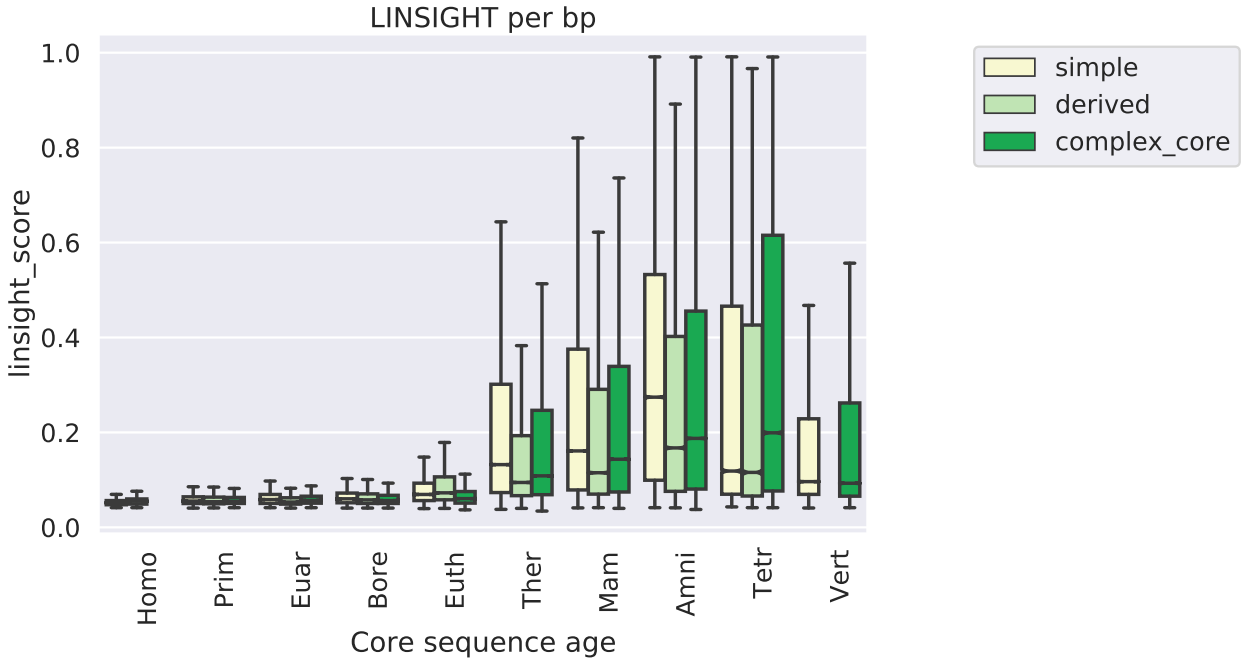

Figure S15: Derived regions experienced weaker purifying selection than cores and simple enhancers of the same age.

Stratified by sequence age, derived regions of complex FANTOM enhancers have lower LINSIGHT scores than core regions of the same age for all sequences older than the Eutherian branch. Number of measurements is annotated per bar.

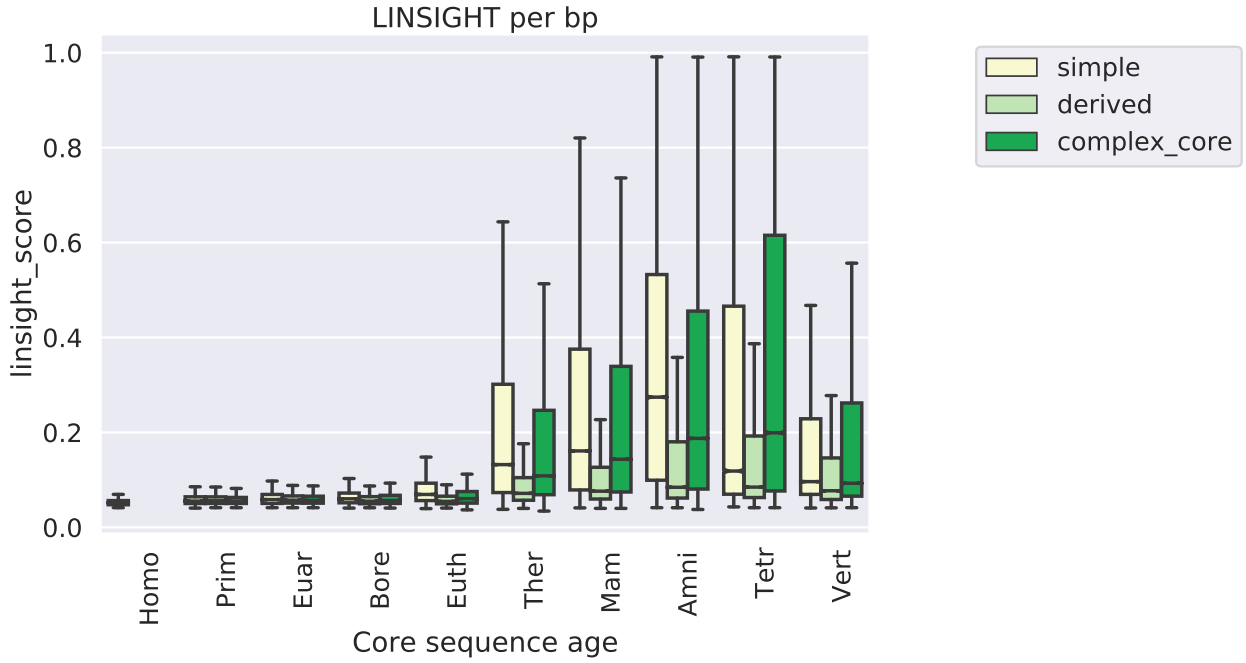

Figure S16: Derived regions experienced weaker purifying selection than cores and simple enhancers with the same age as their corresponding core region.

Stratified by core sequence age, derived regions of FANTOM enhancers have lower LINSIGHT scores than adjacent core regions for all core regions older than Boreotherian. Number of measurements is annotated per bar.

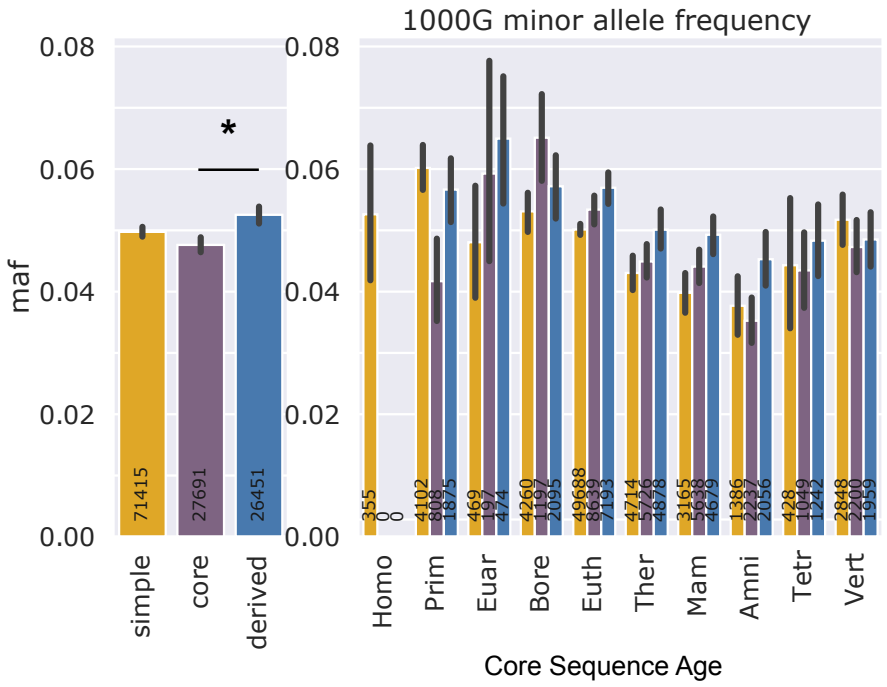

Figure S17: Derived regions have higher minor allele frequencies than core regions across human populations.

Global minor allele frequencies from 1000 Genomes were intersected with FANTOM enhancer components. Singletons were removed. Derived region minor allele frequencies are slightly higher than core region minor allele frequencies (right, mean 0.053 derived v. 0.048 core, derived v. core p = 4.9e-12). Minor allele frequencies stratified by core age and architecture show that derived regions have consistently higher minor allele frequencies compared to core regions at every ancestral origin except Boreotherian. Number of SNPs is annotated per bar.

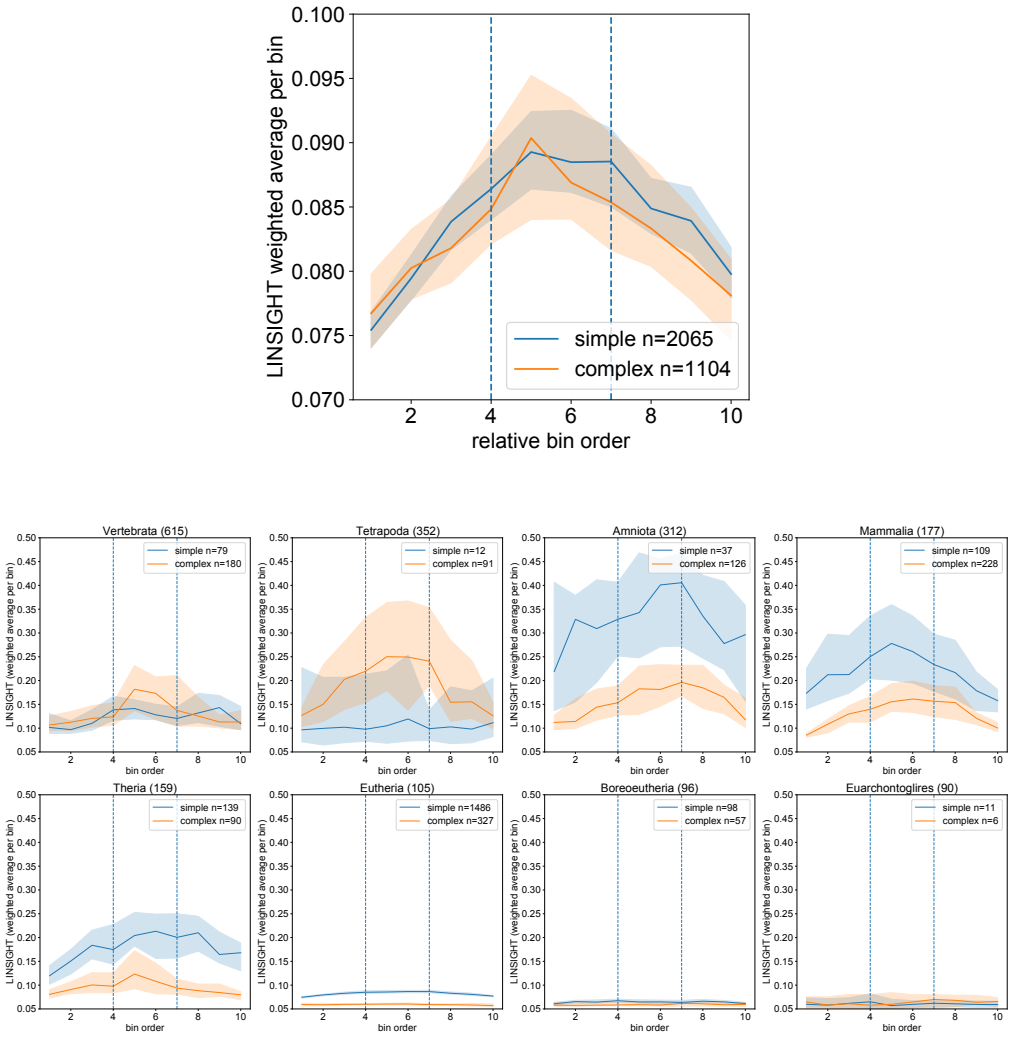

Figure S18: Both complex enhancers with three or more sequence ages and simple enhancers have less purifying selection pressures at sequence edges across ages.

Purifying selection pressures are highest in center of simple, complex enhancer sequences with three or more sequence age regions. LINSIGHT scores in both center four bins of simple (Upper Figure; area in between dashed lines; median weighted average 0.081 outer bins v. 0.89 inner bins, Welch's p-value = 3.4e-24) and complex enhancers (median weighted average 0.80 outer v. 0.86 inner bins, Welch's p-value = 3.4e-24) are significantly higher than outer flank bins (three per side). Selection pressures are higher in the centers of simple enhancers versus complex enhancers (Upper Figure; median 0.089 simple v. 0.086 complex, Welch's p-value = 2.7e-26). Briefly, simple (n = 2065) and complex (n=1104) FANTOM enhancer sequences were matched on sequence length and binned into 10 equally sized bins (median 37 bp per bin). The weighted average LINSIGHT score across bases per bin was calculated and plotted on the y-axis, ordered by bin across the enhancer sequence on the x-axis. Higher selection pressure at the center of sequences is consistent across ages (Lower Figure), multi-age enhancers with three or more age segments and simple enhancers were divided into 10 equally sized bins and the average weighted LINSIGHT score per bin was computed. The centers of both multi-aged and simple enhancer sequences (inner four bins between dashed lines) are more conserved than the flanking edges (outer six bins). Shaded area reflects the 95% confidence interval estimated from 1000 bootstraps.

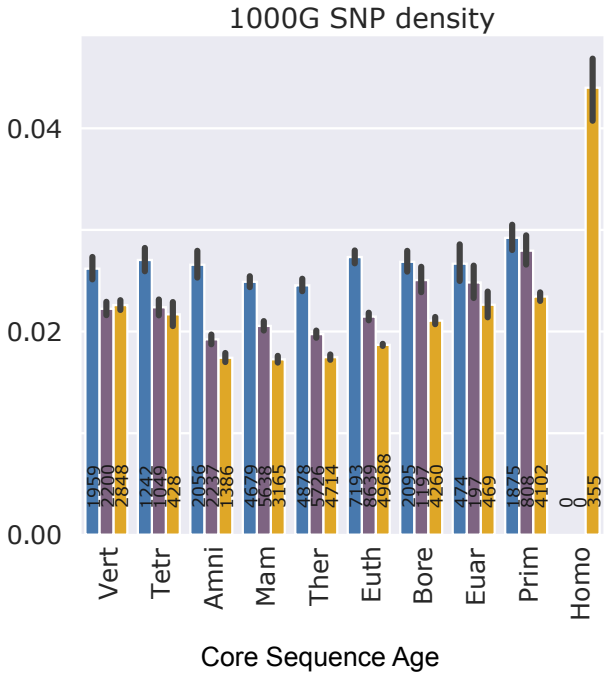

Figure S19: Derived regions have higher SNP densities than adjacent core regions

SNP densities from 1000G were calculated as the number of SNPs in a region divided by the syntenic length. Densities were then stratified by architecture and core age. Number of SNP is annotated per bar.

Cell line	arch	Count zero TF overlap	Total counts	freq zero TF overlap	freq TF overlap
HepG2	derived	23955	44222	0.54	0.46
HepG2	core	9859	29832	0.33	0.67
HepG2	simple	3390	26624	0.13	0.87
K562	derived	16572	40110	0.41	0.59
K562	core	5731	26728	0.21	0.79
K562	simple	1587	22768	0.07	0.93

Figure S20: ChIP-seq TFBS binding frequency in core and derived regions of HepG2 and K562 complex enhancers from ENCODE

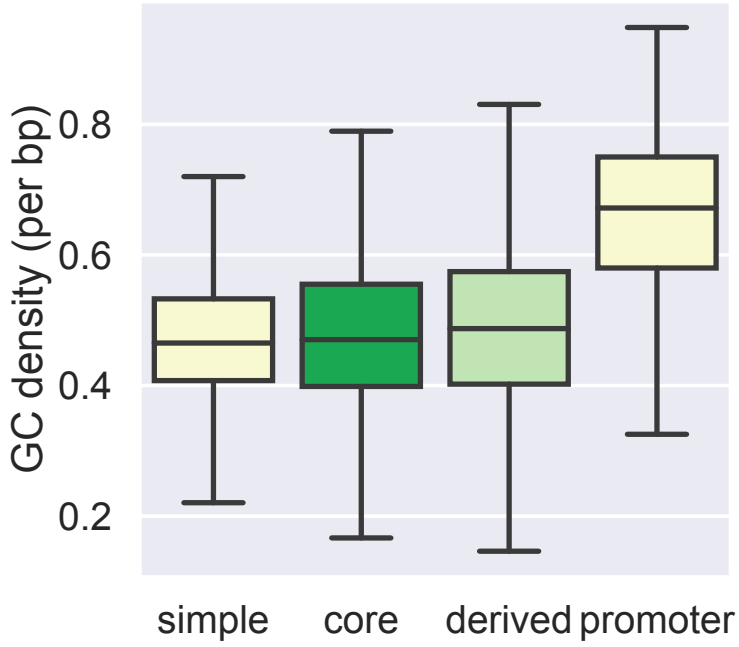

Figure S21: GC density in FANTOM enhancer and promoter regions

GC density was calculated across FANTOM enhancers and promoters as the number of G or C bases divided by the length. Non-exonic enhancers have lower GC density than promoters (N =13781). Derived regions (N = 15357) have slightly higher GC density than core regions (N = 11489) (median 0.49 derived v. 0.47 core GC density; MWU p = 1.7e-12). Simple enhancers (N= 20087) have similar GC density to enhancer cores (median 0.47 GC density)