Emergence of new enhancers at late DNA replicating regions

Paola Cornejo-Páramo[1,2], Veronika Petrova[1,2^], Xuan Zhang[1^], Robert S. Young[3], Emily S. Wong[1,2*]

[1] Victor Chang Cardiac Research Institute, Darlinghurst, NSW, Australia
[2] School of Biotechnology and Biomolecular Sciences, UNSW Sydney, Australia
[3] Usher Institute, University of Edinburgh, Edinburgh, UK

*Correspondence: e.wong@victorchang.edu.au

Abstract

Unlike protein-coding regions, enhancers generally evolve rapidly. By examining enhancer turnover across mammalian species and in multiple tissue-types, we uncovered a relationship between the emergence of novel enhancers and genome organization, reflected by DNA replication time. While enhancers are most abundant in euchromatic regions, new enhancers emerged almost twice as often in late compared to early replicating regions, independent of transposable elements. Using a deep learning model, we demonstrate that new enhancers are enriched for mutations that alter transcription factor (TF) binding. Overall recent enhancers often exhibit neutral evolution, eQTL enrichment, and greater tissue specificity than their evolutionarily conserved counterparts. Accordingly, transcription factors that bind to these enhancers, inferred by their binding sequences, are also more recently evolved and more tissue-specific in gene expression. A similar relationship with DNA replication time is observed in multiple cancer types. Somatic mutations in cancer are consistently elevated in enhancers that have undergone turnover compared to those that remain unchanged. These results demonstrate a relationship between DNA replication time and enhancer evolution across diverse time scales, suggesting that these observations may be time-invariant principles of genome evolution.

Introduction

Enhancers are *cis*-regulatory elements that are important as modulators of spatiotemporal gene expression by acting as integrators of trans-acting signals by recruiting transcription factors (TFs) and other effector molecules. Enhancers are typically rapidly evolving and are frequently species-specific. For example, most human enhancers are not found in the mouse (Blow et al. 2010; Schmidt et al. 2010; Villar et al. 2015; Arnold et al. 2014). The factors responsible for enhancer turnover are not well understood. Although transposable elements (TE) have been linked to enhancer turnover in mammals by expanding cis-regulatory elements across the genome, they do not account for the majority of recently evolved mammalian enhancers, many of which seem to have originated from ancestral sequences without prior biochemical activity in the same tissue (Bourque et al. 2008; Lynch et al. 2011; Lowe et al. 2007; Roller et al. 2021; Young et al. 2015). Local point mutations in non-regulatory sequences can give rise to novel enhancer activity, suggesting a possible mechanism for generating tissue-specific enhancers (White et al. 2013; Duque and Sinha 2015; Smith et al. 2013; Arnold et al. 2013; Zemojtel et al. 2011; Stone and Wray 2001).

Here, we sought to understand the genome principles governing enhancer evolution and the role genome organization plays in shaping mammalian gene regulation. Given the connections between DNA replication time, chromatin structure, and DNA substitutions, we investigate enhancer evolution as a function of changes in genome organization (Schmidt et al. 2010; Costantini and Bernardi 2008; Eyre-Walker 1992). The DNA replication timing program, which refers to the temporal order of DNA replication during the S-phase, is closely linked to the spatial organization of chromatin in the nucleus and transcriptional activity (Rivera-Mulia and Gilbert 2016). One of the most significant predictors of local mutational density is DNA replication time (Hodgkinson and Eyre-Walker 2011; Zhao and Boerwinkle 2002; Hwang and Green 2004). Mutations are elevated during late DNA replication time, corresponding to late S-phase, likely due to errors made in DNA replication and a reduced ability of DNA repair mechanisms to function effectively during late replication (Supek and Lehner 2015). The trend of increased mutational burden during the time course of DNA replication extends throughout eukaryotic evolution (Lang and Murray 2011). In hominids and rodents, mutation rates are 20-30% higher at late compared to early replication domains (Chen et al. 2010).

In mammals and other warm-blooded vertebrates, DNA replication time is also linked to large regional variation in GC base composition (known as ‘isochores’) (Bernardi 2000, 1993; Eyre-Walker and Hurst 2001). Mutational biases, which partly explains the origin of isochores (Eyre-Walker and Hurst 2001; Francino and Ochman 1999; Lercher et al. 2002), are distinct between early and late replicating regions (Schmidt et al. 2010; Costantini and Bernardi 2008; Eyre-Walker 1992). Late replicating sequences harbour biased substitution patterns towards A and T

nucleotides (Chen et al. 2010; Pink and Hurst 2010; Stamatoyannopoulos et al. 2009), where the primary contributor to this bias is the deamination of methyl-cytosine at CpG sites resulting in C > T transitions. C > A mutations at both CpG and non-CpG sites are also known to contribute (Fryxell and Zuckerkandl 2000; Agarwal and Przeworski 2019). These AT-rich and late replicating domains have been linked to facultative heterochromatin and tissue-specific gene expression (Vinogradov 2003, 2005).

We assessed enhancer turnover relative to genomic features that demarcate large-scale chromatin domains, including GC content and DNA replication timing. Using lineage-specific functional genomics data, we captured enhancers at various evolutionary distances, across species evolution and in cancer. We further investigated whether new enhancers are associated with the *de novo* creation of TF binding sites and that this process differs across chromatin domains. We examined whether these changes can be driven by non-adaptive mechanisms.

Results

Enhancer evolution as a function of genome structure

The rate of evolution across the genome is not uniform; for instance, DNA replication time serves as a key determinant of nucleotide changes on a megabase scale within the genome (Hodgkinson and Eyre-Walker 2011; Zhao and Boerwinkle 2002; Hwang and Green 2004). In light of these differences, we sought to understand how genome structure contributes to enhancer turnover. To determine genomic features that explain the evolutionary turnover of enhancers, we constructed a non-linear classifier using a boosted tree approach, XGBoost (Chen and Guestrin 2016). We used this to examine the local relationship between the rate of human liver enhancers turnover with a broad range of genetic and epigenetic features across the human genome at 1Mb wide non-overlapping segments.

Conserved and recent human liver enhancers were obtained from multi-species histone mark ChIP-seq data (Acc: E-MTAB-2633)(Villar et al. 2015). Enhancers were defined as those regions which overlapped a H3K27ac ChIP-seq peak but not a H3K4me3 peak. We compare enhancers across mammals to distinguish their degree of evolutionary conservation. Conserved enhancers (n = 13329) were defined as human enhancers with conserved histone enrichment in at least two of 18 other mammalian species (**Figure S1**). Recently evolved enhancers (n = 10434) were defined as those which either did not align to another mammalian species profiled or did not possess a similar histone mark. Enhancer turnover was calculated per bin as the log ratio of the number of species-specific enhancers to the number of conserved enhancers. Features included the number of CpG sites, DNA methylation, local mutation rates, gene density, GC content, DNA replication time, the rate of transposable elements turnover, chromatin accessibility (ATAC-seq), CTCF binding and histone marks including H3K79me2, H2AFZ, H3K9ac, H3K27me3, H3K9me3, H3K36me3, H4K20me1, H3K4me1, H3K4me2, H3K4me3 (**Methods**). Highly concordant results are obtained using germline replication time estimated from testis and ovary ATAC-seq data (**Methods**).

To assess the importance of each feature to enhancer turnover, we calculated for each feature an average marginal contribution using SHAP (Lundberg and Lee 2017). Hallmarks of euchromatin, including numbers of ATAC-seq peaks and the presence of the histone marks H3K4me1, H3K4me2, H3K79me2 and H4K20me1 were most strongly linked to enhancer conservation (**Fig. 1A**). In contrast, regions involving rapidly evolving enhancers were associated with an increased rate of TE turnover, increased numbers of CpGs, increased methylation of CpGs, late DNA replication time and increased local mutation rates (**Fig. 1B-C; Fig. S2A-B**). Mean percent of methylated reads at CpGs was higher in regions harbouring more lineage-specific compared to conserved enhancers (**Fig. 1D**). Rates of enhancer and promoter turnover were largely concordant across the genome although promoters were more slowly evolving than enhancers (**Fig. S2C**), consistent with earlier reports (Villar et al. 2015). The rate of non-coding background turnover was most strongly linked to TEs, and like *cis*-regulatory elements, are negatively associated with early DNA replication (**Fig. S2A and D**). Fast evolving enhancer regions were linked to increased

turnover in non-coding background suggesting these sequences may share a similar mode of evolution.

While a higher rate of enhancer turnover appeared more commonly associated with less active parts of the genome, there are exceptions. We identified a subset of rapidly evolving regions with high gene density (≥ 40 genes per 1 Mb) (**Fig. S2E**). To investigate this, we performed functional gene term analyses at early and late DNA replicating regions where gene density exceeded 40 genes per 1 Mb. We found in both cases these regions were enriched for gene families. Early replicating gene dense regions were enriched for chemokine and kallikrein genes families, whereas at late replicating domains gene dense regions were overrepresented for G protein-coupled receptors gene families, particularly olfactory receptors (**Fig. S3; Supplemental Tables S1-2**)(Hiratani et al. 2010). Both sets of regions were enriched in recently evolved TEs (**Fig. S3C**)

To investigate these associations more directly, we directly compared the genomic and epigenomic features overlapping candidate enhancer elements. These results were highly consistent results to the above findings (**Fig. S2F-H**). Overall, features associated to euchromatic regions are associated to enhancer conservation, while novel enhancers are mainly associated with heterochromatic features at both 1Mb and enhancer element levels. At early replication time, accessible chromatin and histone modifications were associated with enhancer conservation. Enhancer turnover was generally linked to high local mutation rate, high DNA methylation and high TE turnover at later replicating domains, reflecting structural domains in the genome (**Fig. 1, Fig. S2**).

Germline replication time is associated with the rate of enhancer turnover across the genome

Genetic changes occurring in the germline provide a key source of novel genetic variation that is the substrate for species evolution. Based on this premise, we examined multi-tissue enhancer turnover in mouse using germline DNA replication time. Mouse-specific and conserved enhancers were defined using histone mark ChIP-seq data based on multi-species comparisons (cat, dog, horse, macaque, marmoset, opossum, pig, rabbit, rat) (Roller et al. 2021). Enhancers were termed evolutionarily conserved if they possess the same histone mark in at least two other species in the same tissue (liver, brain, testis, muscle). Active enhancers were elements enriched in H3K27ac and not overlapping regions enriched in H3K4me3, while poised enhancers were defined as regions enriched only in H3K4me1 (Roller et al. 2021; Creyghton et al. 2010) (**Fig. 2A-B**).

To track replication timing dynamics across the mouse genome, we used Repli-Seq data for two germline stages, male and female primordial germ cells (PGC) (n = 2) and spermatogonia stem cells (SSC) (n = 2), as well as 22 other independent mouse cell lines across ten early stages of embryogenesis (Hiratani et al. 2010; Yehuda et al. 2018). We assessed DNA replication time dynamics across cell types by performing *k*-means clustering of cell-type specific replication timing data across the mouse genome partitioned into 200 kb regions (**Methods)**. This resulted in 18 clusters and revealed that approximately half of the mouse genome was consistently early or late replicating across both germline and developmental cell types: 25% of the genome replicated early, whereas 32% were late replicating (early: RT > 0.5, late: RT < -0.5, n = 8966 blocks of 200 kb) (**Fig. 2C**).

Based on germline replication timing, we calculated the probability of observing new enhancers at early and late replication times. We found the fastest rate of enhancer turnover occurred at late replicating domains – hence the emergence of new enhancers was negatively correlated with germline replication time. New enhancers were 1.8 times more common, proportionately, at late compared to early replicating regions, even though the total number of enhancers were highest at early replicating loci (**Fig. 2D-F, Fig. S4**). Newly evolved enhancers were consistently later replicating for both poised and active enhancers across four mouse tissues (brain, liver, muscle, testis) (**Fig. 2G**). Liver and testis enhancers were significantly faster evolving than brain and muscle (t = -6.61 and -4.86; p = 1.88×10^{-07} and 2.96×10^{-05} for poised and active enhancers,

respectively) (**Fig. 2H**). Enhancer turnover and replication time was most highly correlated with germline rather than somatic replication timing data (R^2 = 0.95 compared to R^2 = 0.62) (**Fig. 2E; Fig. S5**), however, similar trends were observed using somatic developmental replication time (**Fig. S6**).

As transposable elements (TE) are widespread across the genome and have been implicated in the turnover of *cis*-regulatory elements, we compared the relationship between TE evolution and replication timing. To do this, we calculate the ratio of the number of new subfamilies of TE and ancestral TE subfamilies across the mouse genome. Like enhancers, our analyses revealed new subfamilies of TEs were more abundant in late replicating regions (R^2 = 0.98). Recent enhancers were also more likely to overlap to lineage specific retrotransposons (**Fig. 2I**). Removing enhancers overlapping TE (~55% of enhancers), the correlation between the rates of enhancer and replication time evolution was unchanged (R^2 = 0.92). This result suggests the rate of enhancer turnover is not wholly dependent on TE, but that TE and enhancer evolution are both strongly correlated to DNA replication time across large chromatin domains (Roller et al. 2021; Yehuda et al. 2018; Jacques et al. 2013).

To more generally assess the proportion of recently evolved enhancers that could have emerged from recent copy number duplication events, we clustered enhancers based on their sequence similarity (**Methods**). While copy number variants associated with homologous recombination events or TEs are expected to share similarity and group together, sequences emerging from mutations of ancestral sequences should not. Using a significance cut-off of $E = 1 \times 10^{-6}$ and a sequence coverage threshold of greater than 20% of the query sequence to detect homology among recently evolved enhancers, we find the proportions of singleton enhancers are 75.92% and 77.04% for human and mouse enhancers, respectively. Fewer singletons overlapped repetitive elements, including TE, compared to non-singleton enhancers (Fisher's exact test, $p = 3.56\times10^{-40}$, odds ratio = 0.13 for human enhancers; Fisher's exact test, $p = 6.8 \times 10^{-171}$, odds ratio = 0.63 for mouse enhancers) (**Fig. S7**). Our results confirm that the vast majority of recently evolved mouse and human enhancers were unlikely to have arisen through tandem duplications.

Taken together, while recently evolved enhancers were disproportionately enriched in late replication regions, most recently evolved enhancers were in early replicating domains. Late germ line DNA replication time is not only associated with an increased number of lineage-specific enhancers but also with new subfamilies of TE. Most lineage-specific enhancers appeared to have emerged by mechanisms other than tandem duplications. This is in contrast to newly evolved genes where existing regulatory sequences are often repurposed following a local or global duplication event (Goode et al. 2010; Rebeiz et al. 2011; Frankel et al. 2012).

Sequence divergence predicts changes to TF binding during mammalian evolution

Given enhancers can emerge from non-regulatory sequences, we sought to understand the impact of evolutionary sequence variants on enhancer turnover. To address this, we trained a deep learning model to infer liver-specific TF binding sites based solely on cis-regulatory sequence input (Cochran et al. 2022). We used a model architecture developed by Cochran *et al*. (Cochran et al. 2022) and retrained the model to classify TF binding sites at enhancers that have experienced evolutionary turnover between human and mouse (**Methods**). The model takes orthologous human and mouse DNA sequences, each 500 bp in length, as inputs and investigates whether evolutionary differences have resulted in genetic changes that affect TF binding. We hypothesize that alterations in TF binding could be potential drivers of enhancer turnover. If mutations have contributed to this turnover, we anticipate a higher predicted frequency of TF binding at new enhancers (i.e., those with recently acquired enhancer histone marks) compared to evolutionarily conserved, but non-regulatory sequences.

The neural network uses a gradient reversal layer (a domain adaptive network architecture) to remove the effect of species-specific genome background between human and mouse to optimize the learning of functional sequences shared by both species (**Methods**) (Cochran et al. 2022). The

model was trained using DNA sequences of regions bound by liver-specific TFs (HNF4A, CEPBA), as determined by ChIP-seq, but excluding regions harbouring any human and mouse-specific enhancers. We used the trained model to predict TF binding at enhancers that aligned between human and mouse where an enhancer is identified, by histone marks, in one species but not its orthologous sequence.

The sequence-based model identified significantly higher number of HNF4A/CEPBA TF binding sites at human enhancers compared to the mouse orthologs without enhancer histone marks, suggesting that genetic variation between the sequences is associated with the gain or loss of functional TF binding sites ($p = 5.78 \times 10^{-30}$; OR = 1.95) (**Fig. 3A**). In contrast, when comparing mouse-specific enhancers to their aligned non-enhancer sequences in humans, the opposite pattern emerged ($p = 1.27 \times 10^{-77}$; OR = 3.82) (**Fig. 3A)**. Overall proportion of species-specific enhancers with predicted HNF4A/CEPBA binding sites increased at late replicating regions (**Fig. 3B).**

Late replicating enhancers are tissue-specific and neutrally evolving

Next, we investigated whether tissue-specific enhancer activity was associated with late replicating regions. Regions associated with dynamically regulated heterochromatin and nucleosome formation potential have been linked to tissue specific gene expression (Vinogradov 2003, 2005; Holmquist 1987; Rhind and Gilbert 2013). By comparing across the four mouse tissue types, we found tissue-specific enhancers were more likely to be late replicating than enhancers active in more than one tissue (Fisher's exact test, $p = 2.2 \times 10^{-16}$, odds ratio = 0.28, **Fig. 4A**). Because tissue-specific control of gene expression is especially critical during development, further we hypothesized that enhancers which drive developmental promoter activity were likely later replicating compared to enhancers driven by a housekeeping promoter. Indeed, enhancers associated to developmental promoter activity via massively parallel reporter assays in *Drosophila* were later replicating than enhancers associated with the housekeeping promoter (**Fig. 4B;** Mann-Whitney *U*-test, $p = 8.5 \times 10^{-6}$; **Methods**). This pattern was consistent across different chromosomes and was independent of the endogenous location of the promoters (not shown). This suggested that enhancers regulating developmental gene expression generally replicate later in S-phase than housekeeping gene-linked enhancers.

Further considering this link between tissue-specificity enhancer activity and late DNA replication, we examined whether enhancers located at late replicating regions were also associated with the binding of more tissue-specific TFs. Using an established index of tissue-specificity of gene expression, tau (Yanai et al. 2005), we examined tissue-specific TF expression for 477 and 360 genes in human and mouse across 27 and 19 tissues, respectively (Kryuchkova-Mostacci and Robinson-Rechavi 2017). TFs were partitioned into five groups based on the relative enrichment their motifs at enhancers from early and late replication time. The TFs with the most enriched motifs at late replicating enhancers were significantly more likely to show tissue-specific expression patterns (**Fig. 4C and D**).

Given that younger enhancers are correlated with late DNA replication time in mammals, we further considered whether the evolutionary age of TFs also correlates with replication time. To do this, we assigned an evolutionary age to each human TF (Domazet-Lošo and Tautz 2010), and partitioned TF motifs into five groups based in their relative enrichment at early and late replicating enhancers. We found TFs binding to motifs at late replicating enhancers were more evolutionary recent than those at early replicating regions. Motifs of TFs of metazoan-origin were enriched at late replicating enhancers, while motifs of more ancient TFs tended to be more prevalent at early replicating enhancers (**Fig. 4E**). A similar trend was observed when the motif enrichment was calculated using replication time predicted by testis ATAC-seq data (**Fig. S8 and S9**).

To address whether new and conserved enhancers make qualitatively different contributions to transcription, we measured the relative enrichment of 226,768 significant liver *cis*-eQTLs from the GTEx consortium (GTEx V7) at human recent and conserved enhancers. Recently evolved enhancers harboured significantly more eQTLs than conserved enhancers (Fisher's exact test, p <

2×10^{-16}, odds ratio = 1.2) – consistent with recent reports of an increased frequency of eQTLs at new promoters in the human genome (Young et al. 2022). To further understand the selective pressures at these enhancers, we calculated a derived allele frequency (DAF) score in 10 bp windows across the human genome using human population variation data from the whole-genome sequencing of the Icelandic population, a relatively isolated population (deCODE)(Gudbjartsson et al. 2015). We plotted DAF for recently evolved and conserved liver enhancers by centring each enhancer based on functional motifs to increase power to detect purifying selection (**Fig. 4F and G**, **Methods**). We analysed the DAF relative to genome-wide background and found weak evidence for purifying selection at motifs of both conserved and human-specific liver enhancers against the genome background. In contrast, a strong signature of purifying selection was detected at human conserved promoters (**Fig. 4H, Fig. S10A and B**). These results revealed that enhancer sequences were largely neutrally evolving in the human population regardless of their evolutionary history, even though certain motifs may be under selective constraint. Although the DAF scores were significantly different from genome background at recent enhancers, this difference may be due to a higher frequency of common variants rather than a depletion of rare variants (**Fig. S10C and D**).

GC content isochores shape the type of TF binding sites at enhancers

As GC isochores are closely linked to replication timing domains in mammals (Costantini and Bernardi 2008; Eyre-Walker 1992; Schmegner et al. 2007), we examined how changes in GC content could influence where different types of TFs bind in the genome. We examined base substitutions using the common ancestor of *Homo* and *Pan* with DNA replication domains from the H9 cell line (**Fig. S11)**. Consistent with earlier report, AT bias at late DNA replication time was independent of the deamination of methylated CpGs despite CpG density being a significant contributor to C>T mutations at late replicating domains (Kenigsberg et al. 2016) (**Fig. 5A**).

We considered whether GC bias impacts the types of TF binding sites found at enhancers located within early versus late replication domains. We characterized the type and frequencies of TF binding sites between early and late replicating human liver enhancers and find replication timing was associated with a profound shift in the frequencies of TF binding motifs (**Fig. 5B-C**). As expected, the most prevalent motifs at late replicating enhancers were more AT-rich, while motifs at early replicating enhancers were more GC-rich (**Fig. 5B-C**). For example, the high GC-content motifs of ZNF460, KLF2 and KLF6 were highly enriched in early replicating enhancers while high AT-content motifs of PRRX1, ISX and DLX6 were enriched at late replicating enhancers (**Fig. 5D**). A similar pattern existed at mouse enhancers and at promoters (**Fig. S12-13**). We observed similar trends among motifs belonging to the same TF class. For example, homeodomain factor motifs, which act as key regulators in development, were predominantly enriched in late replicating enhancers (**Fig. 5E**). Using a non-linear classifier, we showed that frequency of TF binding sites could reliably classify the replication time of enhancers (**Methods**) (**Fig. 5F**). Our balanced model obtained an accuracy of 74% in predicting the replication time of human liver enhancers and this performance was directly attributed to a change in the GC composition of enhancers (**Fig. 5F**). This trend was replicated in mouse (accuracy = 68%) (**Fig. S14**).

Because the nucleotide frequencies of motifs at enhancers are an indirect measure of TF binding, we asked whether the observed trends are reflected by TF binding *in vivo*. Using the binding locations of 71 DNA binding proteins in the human K562 cell line determined by ChIP-seq, we found the density of binding sites against H9 replication times were typically bimodally distributed. We separated the binding sites into two Gaussian components using mixture modelling resulting in two set of binding sites for each of the 71 binding proteins – one at early replicating regions and one at later replicating sites (**Fig. S15**, **Supplemental Table S3**). The binding sites located at earlier replicating regions were less sensitive to the GC content of its binding motif than sites at later replicating regions. GC-rich binding motifs were generally earlier replicating than AT-rich TF binding motifs (**Fig. 5G**). A consistent relationship was observed between the nucleotide content at TF binding sites and simulated germ line DNA replication time (**Fig. S16**).

Enhancer turnover is associated with late DNA replication time and a higher mutational burden in cancer

We asked whether a relationship between DNA replication timing and cis-regulatory element turnover exists in cancer cells. DNA replication time has been reported to be largely stable between normal and cancer cells (Du et al. 2019; Blumenfeld et al. 2017). We compared replication time in healthy and cancer cell states across four different cancer types (**Fig. 6A**). Enhancers were identified in normal breast, prostate, thyroid and preleukemic cells and in their diseased state (Corces et al. 2016; Sanghi et al. 2021; ENCODE Project Consortium 2012; Taberlay et al. 2016, 2014). For breast and prostate cancer, we relied on established cancer cell lines and called candidate enhancers based on enhancer chromatin marks (**Fig. 6A; Fig. S17A; Supplemental Table S4**) (**Methods**). To accurately calculate somatic mutations rates, we used matched tumor and healthy samples from the same individual to annotate enhancers where data is available.

For AML and thyroid cancer, enhancers were defined as distal chromatin accessible regions (± 1kb from TSS). We used ATAC-seq data for matched cancer and normal thyroid samples from three individuals (Acc: GSE162515; C1, C7, C8) (Sanghi et al. 2021), and ATAC-seq files for the matched pre-leukemic and blast cells from three individuals (Acc: GSE74912; SU484, SU501, SU654) (Corces et al. 2016). ATAC-seq processing and peak calling are described in **Supplemental Methods**.

Distal peaks were annotated as: a) new enhancers in cancer – peaks present in all cancer samples but not normal samples; b) enhancers lost in cancer – peaks present in all normal samples but not in cancer; c) enhancers common to both – peaks present in all three normal and three cancer samples. Mutect2 (Benjamin et al. 2019) was used to annotate somatic variants at ATAC-seq peaks. Prostate cancer variants were annotated using the whole genome sequence of the cancer cell line. For further details on variant calling see **Supplemental Methods**.

We used replication time data for LNCaP, PrEC and MCF-7 (Acc: GSE98730, GSM923442) (Du et al. 2019). We predicted thyroid replication time from ATAC-seq data using Replicon as described above (Acc: GSE170176). For blood, the mean ATAC-seq signal of the union of accessible regions across the pre-leukemic samples was used to predict replication time.

We defined a 'gain' of enhancer as those characterized in cancer cell lines but not in the non-diseased state. In turn, those existing in the normal cell state but not in cancer were defined as 'lost'. Enhancers present in both healthy and diseased cell states were termed 'unchanged' (**Fig. 6A and B; Supplemental Table S4**). We compared enhancer gains and losses across different replication timepoints and found that these results were consistent with cross-species evolutionary results; enhancers turnover was highest at late DNA replicating domains relative to background (i.e. the number of unchanged enhancers) (**Fig. 6C-F, Fig. S17B-G**). This trend was unaffected by differences in recombination breakpoints (Baca et al. 2013) (**Fig. S18**).

To infer the role of mutations on enhancer turnover, we compared cancer variants at gained, lost and unchanged enhancers in thyroid, AML, and prostate cancer. In thyroid and acute myeloid leukaemia (AML) where matched patient samples were available, somatic variants were called using ATAC-seq data from healthy and cancer tissues/cells for six individuals on a per individual basis (**Methods**). In prostate cancer, we leveraged somatic variants from the whole-genome sequencing of the prostate cancer genome cell line (Seim et al. 2017).

Mutation numbers were elevated for enhancers that were gained or lost compared to unchanged enhancers across the three cancer types for which we had SNP data (**Fig. 6G-L; Fig. S17H-K and Supplemental Table S4**). The trend was consistent across all individuals for thyroid cancer and AML (**Fig. S19**). We restricted our analysis to single nucleotide polymorphisms (SNPs) where common genetic variants, defined using multiple large-scale population cohorts, were filtered out (**Methods**). In prostate cancer, 48,161 putative somatic mutations were identified across 37,482

enhancers (13.91%) where on average 1.46 mutations were observed per enhancer (0.08% of total sequence length).

In summary, enhancer turnover was observed to be highest in late replicating regions, both during species evolution and consistently across all cancer types. Notably, these regions also displayed the highest rates of somatic mutations, demonstrating a strong correlation between the two phenomena.

Discussion

In this study, we demonstrated the significance of genome organization on enhancer evolution in the context of species evolution and cancer. While most enhancers were identified in early replicating regions, comparative analyses showed that young enhancers were almost twice as likely to replicate later than conserved enhancers. During late replication time, local mutation rates, DNA methylation sites, and TE evolution positively correlated with enhancer turnover. Intriguingly, similar patterns in cis-regulatory evolution are evident in both species evolution, spanning millions of years, and in cancer cells, occurring over months or a few years perhaps suggesting that the interplay between genome structure and enhancer evolution is a general feature of regulatory turnover although the mechanisms by which they occur may differ.

Genetic changes during evolution can create or abolish TF binding sites which are associated with the emergence of cis-regulatory elements or the decommissioning of existing ones. Theoretically, the short length of TF binding motifs and their degeneracy allows for the rapid emergence and fixation of TF binding motifs (White et al. 2013; Duque and Sinha 2015; Smith et al. 2013; Arnold et al. 2013; Zemojtel et al. 2011; Stone and Wray 2001). We assessed whether sequence differences predict changes to TF binding at enhancers that have experienced turnover between human and mouse. Although we do not expect all species-specific differences to be explained by variants in the binding sites for two TFs, species-specific gain or loss of enhancers were significantly linked to base changes that impact TF binding. These changes are more prevalent at late DNA replication times, suggesting that cis-regulatory elements created at late replicating domains may evolve more rapidly due to elevated mutation rates in these regions. Except at TF binding sites, even conserved enhancers in our study displayed limited evidence for purifying selection. This appears to align with the known robustness properties of enhancers.

Furthermore, tissue-specific enhancers were predominantly located in late replicating regions, where they tended to be recognized by tissue-specific transcription factors (TFs). This is consistent with the known properties of GC isochores whose position mirror replication domains (Costantini and Bernardi 2008; Schmegner et al. 2007) and are linked to tissue-specific gene expression differences in mammals (De and Babu 2010; Du et al. 2019). We showed that highly organized isochore patterns in mammalian genomes influenced the genome location and frequencies of TF binding sites, with specific types of TFs more likely to be recruited at certain replication timing domains due to GC isochores. For instance, tissue-specific TFs are frequently linked to AT-rich, late replicating binding sites, which may explain tissue-specific gene expression in GC-depleted regions (Vinogradov 2003, 2005).

Finally, results suggest that evolutionary innovation in gene regulatory modules may be more likely to emerge from regulatory elements at later replication domains in a tissue-restricted manner. Late replicating regions are associated with developmental enhancer activity and are enriched for developmentally relevant TF binding sites related to body patterning. Our evidence implies that transcriptional changes occurring at late replicating domains may have played a pivotal role in the evolution of the bilaterian body plan and embryonic development of multicellular organisms. Notably, replication timing is dynamic between cell types and varies between germline and somatic cell types (Ryba et al. 2010; Hiratani et al. 2010; Siefert et al. 2017). Approximately 30% of the human genome switches between replication timing domains across 26 human cell lines (Rivera-Mulia et al. 2015), allowing enhancers emerging in germ cells at late replication time to shift to earlier replicating domains in differentiated cell types, where they can influence gene activity.

Methods

All analyses were performed on the human and mouse genome assemblies hg19 and mm10 unless specified otherwise. R v4.0.0 (R Core Team 2020) was used. Species-specific ChIP-seq datasets used are E-MTAB-2633 and E-MTAB-7127 (Villar et al. 2015; Roller et al. 2021).

Non-linear boosted tree-based classification

Datasets used in the XGBoost prediction of enhancer turnover include liver DNA methylation, liver chromatin accessibility, CTCF binding and histone marks including H3K79me2, H2AFZ, H3K9ac, H3K27me3, H3K9me3, H3K36me3, H4K20me1, H3K4me1, H3K4me2, H3K4me3 and H3K9me2 (**Supplemental Table S5**). Datasets were downloaded in hg19 coordinates, if available, or mapped using USCS liftOver from hg38 to hg19. Human promoters were regions enriched in H3K4me3 and H3K27ac or only enriched in H3K4me3 (Villar et al. 2015). Conserved human liver promoters were defined as regions enriched in H3K4me3 and H3K27ac present in human, macaque and marmoset (Villar et al. 2015).

We generated non-overlapping 1 Mb genomic windows using bedtools makewindows across the main chromosomes in hg19 reference assembly (no scaffolds) (bedtools v2.28.0)(Quinlan and Hall 2010). GC content of every window was calculated using bedtools nuc (Quinlan and Hall 2010). Numbers of CpG sites and methylated CpG sites were counted for every window. We calculated the number of histone mark peaks, ATAC-Seq peaks, genes, CTCF binding sites, enhancers, promoters, random regions, and TEs in every 1 Mb window. Rates of enhancers, promoters, TE and random regions turnover were calculated as the log ratio of the number of recent vs conserved elements. For continuous features we calculated a mean value for every genomic window (DNA replication time, local mutation rate and DNA methylation). Local mutation rates were estimated using rare population variants using 1000 Genomes Phase 3 data with bcftools (v1.16) excluding indels filtering for MAF < 0.05 at a resolution of 100kb and averaged for every window. We used H9 replication time (Acc: GSE137764) (Zhao et al. 2020).

An XGBoost model (Chen and Guestrin 2016) was trained on the features described above to predict enhancer turnover (xgboost v1.5.0, Python v3.9.12). Test data split was 0.2 of total dataset. Model was trained using the following parameters: eta = 0.01, nround = 5000, early_stopping_rounds = 20, subsample = 0.5, objective = reg:squarederror, subsample= 0.5, base_score = np.mean (y_train), eval_metric = logloss. Feature importance was calculated using SHAP (Lundberg and Lee 2017) (shap v0.39.0, Python v3.9.12).

To analyse similar features at the enhancer level, we calculated density and overlap with ATAC-seq peaks, histone marks, CTCF binding sites, human specific TEs, and genes for every enhancer. Additionally, we calculated CpG and methylated CpG density at enhancers, along with mean local mutation rate, methylation at CpG sites, and replication time per enhancer. A matrix was then constructed with all these features, and enhancers were classified as recent (coded as 1) or conserved (coded as 0) (n = 9984 and n = 12956 recent and conserved enhancers, respectively).

An XGB model was trained on 80% of enhancers to predict enhancer class with parameters eta = 0.01, n_estimators = 100, objective = "binary:logistic", subsample= 0.5, base_score = 0.5, eval_metric = "auc". The model was tested on the 20% of enhancers not used for training (AUC = 0.68). Feature importance was calculated with SHAP (xgboost v1.5.2, SHAP v0.41.0 and python v3.6.8).

Replication time analysis

Z-score transformed replication timing data was obtained from human ESC H9 (Zhao et al. 2020), mouse primordial germ cells (PGC) and spermatogonial stem cells (SSC) (Acc: GSE109804)(Yehuda et al. 2018), and 22 mouse cell lines differentiated from ES cells (Hiratani et al. 2010). Early and late replication elements were denoted as those with mean times > 0.5 and < -

0.5, respectively. The replication time of every enhancer was calculated by averaging the replication time of the regions they overlap. Difference in the probability of recent enhancers between the early and late replication time was calculated as:

$$\frac{P(Recent\ enhancer \mid Late\ RT)}{P(Recent\ enhancer \mid Early\ RT)}$$

Clustering of mouse replication time data

Replication time data from 22 early embryonic mouse cell lines differentiated from mouse embryonic stem cells were transferred to mm10 coordinates (USCS liftOver tool) and overlapped with the replication time regions from PGC (n = 2 cell lines) and SSP cell lines (n = 2 cell lines) (Hiratani et al. 2010; Yehuda et al. 2018). Mean replication time were calculated for every 200 kb region across the mouse genome (n = 8966 replication time bins across all cell types). Mean replication time values were centred and scaled using the function “scale” (R Core Team 2020). The function “Mclust” from the R package mclust was used to estimate the best number of *k*-means clusters of 200 kb replication time bins (G = 1:k.max, modelNames = mclust.options("emModelNames"), where k.max is 20) (mclust version 5.4.6) (Scrucca et al. 2016). The best number of clusters was selected based on its Bayesian Information Criterion (BIC) (k = 18, BIC = -37291.04). Replication time bins clusters were obtained with the function kmeans (18, iter.max = 20). Cell types were clustered using hierarchical clustering with k = 9 (function hclust from base R, method = “complete”, distance = “euclidean”).

Tissue specific linear model

To test for differences in tissue-type specific evolutionary rates we used a linear model. Using the formula logFC_enh ~ mean_RT + tissue_pair + tissue_pair:mean_RT, where logFC_enh represent the log (number of recent enhancers / number of conserved enhancers) values, tissue_pair is the tissue pair codified as binary (liver and testis = 1, brain and muscle = 0) and mean_RT is the mean germline replication time. The interaction effect was tested using the t-test statistic.

Transposable elements

Transposable elements from RepBase (v27.04 for mouse and human) (Bao et al. 2015) were annotated using RepeatMasker (v4.0.6 for mouse, v4.1.0 for human) using the sensitive search setting for mouse (‘-s’) (Smit et al. 2013). Species-specific elements, as classified by RepBase, were termed ‘recent’ and the remaining TE termed ‘ancestral’.

Developmental enhancer analysis

Summit coordinates of *Drosophila* enhancers determined using STARR-seq on S2 cell line were downloaded (Acc: GSE57876)(Zabidi et al. 2015). Housekeeping and developmental promoter were of *RpS12* and *even-skipped* TF, respectively. We defined housekeeping and developmental enhancers as the most highly ranked 200 enhancers for each promoter based on their STARR-seq score. Fly DNA replication time profiles for S-phase in S2-DRSC cells were used (Acc: GSE41350)(Lubelsky et al. 2014). Enhancers summits were extended by 250 bp upstream and downstream and a mean replication time calculated for each enhancer.

Neural network model

A domain adaptive neural network architecture was used to remove background sequence biases between human and mouse genomes at TF binding sites (Cochran et al. 2022). Input data was generated by splitting in the mouse (mm10) and human (hg38) genomes into 500 bp windows, with 50 bp offset. After excluding all regions containing human- and mouse-specific enhancers and their orthologous region in the other species, we than trained the model as described in Cochan et al. (Cochran et al. 2022). Liver human and mouse HNF4a and CEPBA ChIP-seq peak data taken from from (Schmidt et al. 2010) were remapped to hg38 or mm10, respectively. For every TF we trained two sets of models: one using human as the source species and one using mouse as the source species. For each species, peaks were converted to binary labels for each window in the genome: “bind” (1) if any peak centre fell within the window, “unbind” (0) otherwise. We constructed balanced datasets for training by using all bound regions and an equal number of randomly unbound sampled (without replacement). Sequence data was one hot encoded. Human

and mouse genome sequences excluding chr 1 and chr 2 were used for model training. Genome windows from chr 2 were used for testing. Genome windows from chr 1 were used for validation. For each TF and source species, 15 models were trained to monitor reproducibility. Final models were selected based on maximal auPRCs. Test data set comprised of species-specific enhancers centered on the 500bp of each enhancer. For predictions using the model, we used a probability >= 0.9. For genome assembly remapping, USCS 'liftOver' with minMatch = 0.6 was used.

Natural selection analyses

Human genome variation data was retrieved from the deCODE whole-genome sequencing study of the Icelandic population (Gudbjartsson et al. 2015). Derived Allele Frequency (DAF scores) of every segregating SNP was calculated and alleles were defined as either rare (< 1.5% population frequency) or common (> 5% frequency) as previously described (Young et al. 2022). The number of rare and common alleles in 10 bp windows were centred with respect to the locations of functional liver-specific TF binding motifs from the database funMotifs, 75 unique funMotifs were counted (Umer et al. 2019). These counts were normalised for the average rates with 2–4 kb upstream and downstream flanking regions and confidence intervals obtained by performing 100 bootstrap replicates of sampling the motif locations with replacement. Odds ratios of rare against common alleles between enhancers (and promoters) and size-matched background genomic regions selected at random were calculated in 10 bp windows as before. Odds ratio confidence intervals and p-values were obtained using Fisher's exact test. Only autosomes were considered.

Copy number analyses

Homology was assessed using blastn with the option -max_target_seqs N (blast+/2.11.0)(Camacho et al. 2009), this option was used to retrieve the maximum number of hits for every enhancer; N represents the number of enhancers in every dataset, 10434 and 80904 for human and mouse, respectively. Hits were filtered by E-value < 1×10^{-6} and query coverage > 20. We defined singleton enhancers as the enhancers without significant similarity to other enhancers.

Cancer data analyses

ChIP-seq data from the prostate cancer cell line, LNCaP, was used to annotate enhancers (Acc: GSE73783) (Taberlay et al. 2016). For normal prostate epithelial cells (PrEC), enhancers were defined in an unsupervised manner using chromHMM (Acc: GSE57498) (Taberlay et al. 2014). We used ChIP-seq data of histone marks in breast cancer cell line, MCF-7, to annotate enhancers (Acc: GSE96352, GSE86714), and normal epithelial breast cells (patients epithelium samples) (Acc: GSE139697, GSE139733). LNCaP, MCF-7 and breast epithelium enhancers were defined as regions enriched in H3K27ac or H3K4me1 excluding proximal regions (± 1kb from TSS). ChIP-seq data from the prostate cancer cell line, LNCaP, was used to annotate enhancers (Acc: GSE73783) (Taberlay et al. 2016). For normal prostate epithelial cells (PrEC), enhancers were defined in an unsupervised manner using a collection of histone marks using chromHMM (Acc: GSE57498) (Taberlay et al. 2014). ChIP-seq data of histone marks in breast cancer cell line, MCF-7, was used to annotate enhancers (Acc: GSE96352, GSE86714), and normal epithelial breast cells (patients epithelium samples) (Acc: GSE139697, GSE139733). LNCaP, MCF-7 and breast epithelium enhancers were defined as regions enriched in H3K27ac or H3K4me1 excluding proximal regions (± 1kb from TSS).

Estimation of DNA replication time using ATAC-seq

ATAC-seq from human ovary, human and mouse testes were used for replication time prediction using Replicon v0.9 (Gindin et al. 2014) (Acc: ENCSR872MAO, GSE172988, GSE187731). ATAC-seq signal was normalised to a mean of 0 and unit variance. Replicon was run with default options on every chromosome (excluding scaffolds). The predicted replication time values were multiplied by -1 to match the direction of Repli-Seq RT values.

Code availability

https://github.com/ewonglab/enhancer_turnover

Acknowledgements

We thank M.Roller and P.Flicek for assistance with metadata access; J.Wong, C.Du, S.Clark S.Mahony, and L.Louis for helpful discussions, and Q.Wang for assistance with data processing. PCP is supported by a UNSW International Postgraduate Scholarship. VP is supported by an Australian Government Research Training Stipend Scholarship. ESW is supported by a NHMRC Investigator Grant (GNT2009309) and ARC Discovery Project (DP200100250).

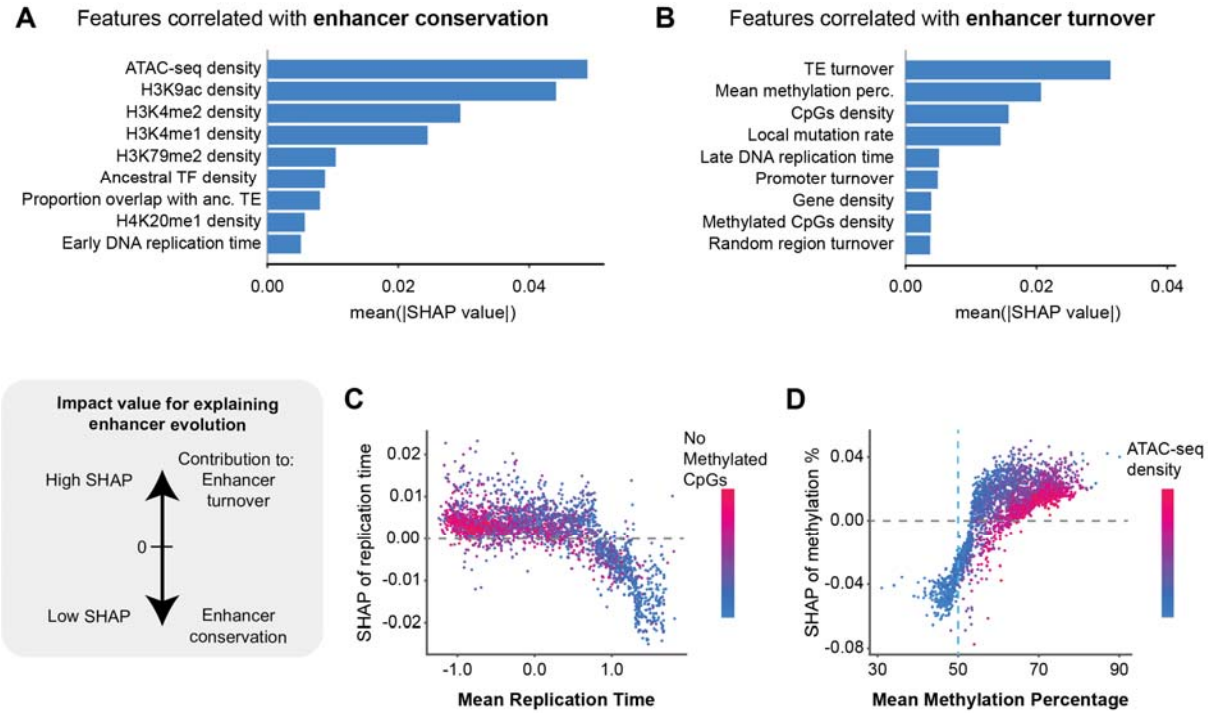


Figure 1. Genomic and epigenomic features associated with enhancer turnover. (A) The average marginal contribution of features (SHAP) that explain enhancer conservation in human liver using XGBoost**. (B)** The average marginal contribution of features that explain enhancer turnover. **(C)** SHAP score of replication time in explaining enhancer turnover where every dot is a 1 Mb window. The colour scale represents the number of methylated CpGs **(D)** SHAP scores of mean methylation percentage at CpGs per 1 Mb window in explaining enhancer turnover. The colour scale represents the number of ATAC-seq peaks.

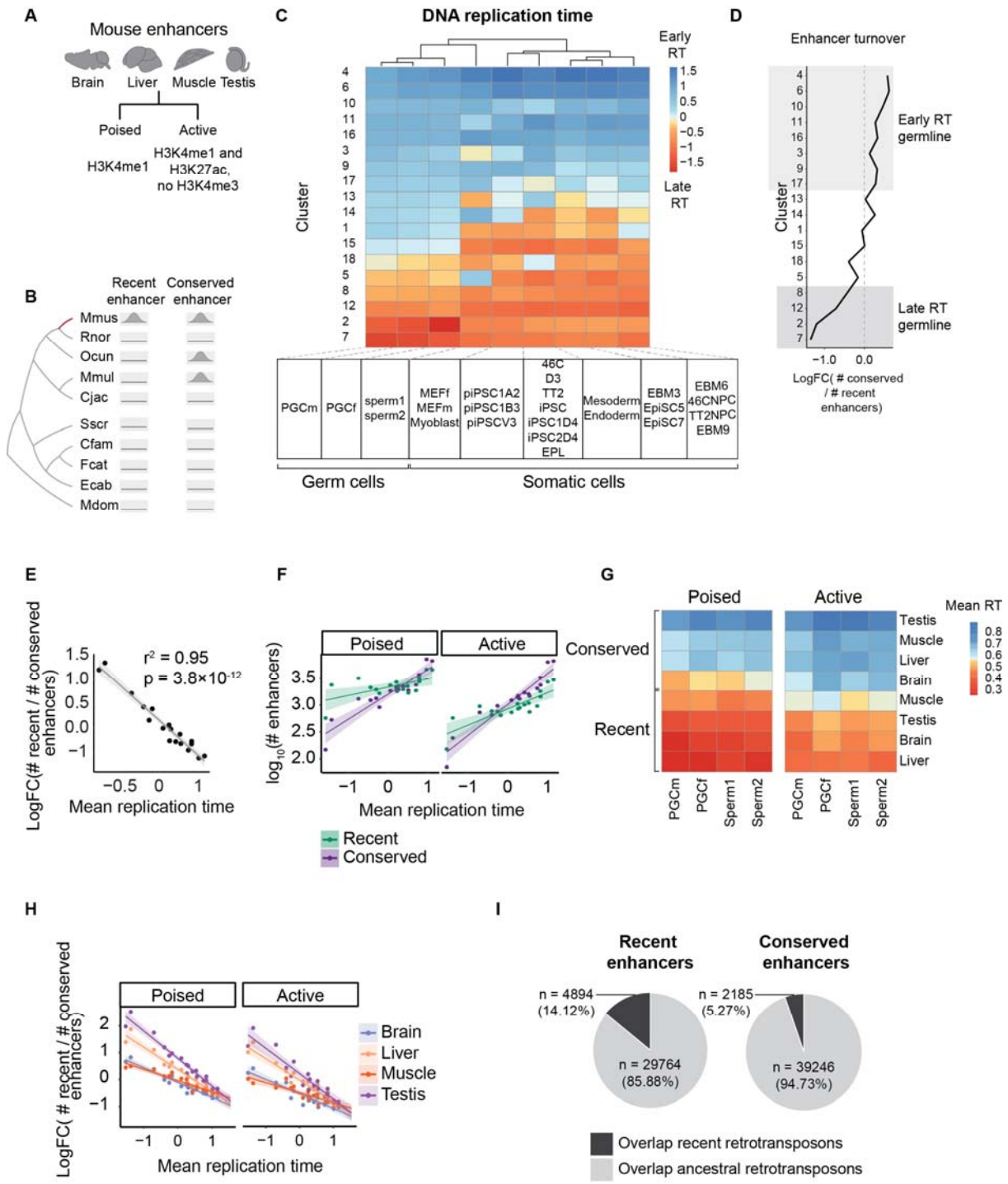


Figure 2. Turnover of enhancers increase in late replicating genomic regions. (A). Mouse enhancers are defined based on combinations of histone marks. **(B)** Definition of mouse recent and conserved enhancers (Roller et al. 2021). Recent enhancers were defined as regions that do not align to nine other mammalian genomes in the same tissue (Roller et al. 2021). Conserved enhancers are aligned to enhancers in the same tissue in at least two other species (Roller et al. 2021). Mmus = *Mus musculus*, Rnor = *Rattus norvegicus*, Ocun = *Oryctolagus cuniculus*, Mmul = *Macaca mulatta*, Cjac = *Callithrix jacchus*, Sscr = *Sus scrofa*, Cfam = *Canis familiaris*, Fcat = *Felis catus*, Ecab = *Equus caballus* and Mdom = *Monodelphis domestica*. **(C)** Replication time across 200 kb blocks of the mouse genome (n = 8966 blocks) in PGC (n = 2 cell lines), SSC cells (n = 2

cell lines) and early somatic cell types (n = 22 cell lines). Row clustering (blocks) was carried out with *k*-means clustering, column are cell type clusters generated with hierarchical clustering. **(D)** Enhancer turnover is the log fold change of conserved enhancers vs. recent enhancers for the 200 kb clusters in **(C)**. Shaded areas in **(D)** indicate the clusters that are consistently early (top) or late (bottom) replicating in PGC and SSC lines. All mouse tissues were included (brain, liver, muscle, and testis). **(E)** Scatterplot of mean germline replication time and across the 18 clusters shown in **(C)**. R^2 and p-value are indicated. **(F)** Scatterplot of germline mean replication time and $\log_{10}$-transformed numbers of recent and conserved enhancers. Shaded region represents the 95% confidence interval of the line of best fit. **(G)** Heatmaps of mean PGC and SSC replication time of poised (left) and active (right) mouse enhancers separated by tissue and type (recent or conserved). **(H)** Scatterplots of germline mean replication time and enhancer turnover by tissue and enhancer type (poised or active). Shaded region represents the 95% confidence interval of the line of best fit. **(I)** Pie charts of the number of recent/conserved enhancers overlapping recent/ancestral retrotransposons (Fisher's exact test, $p < 2.2 \times 10^{-16}$, odds ratio = 2.95).

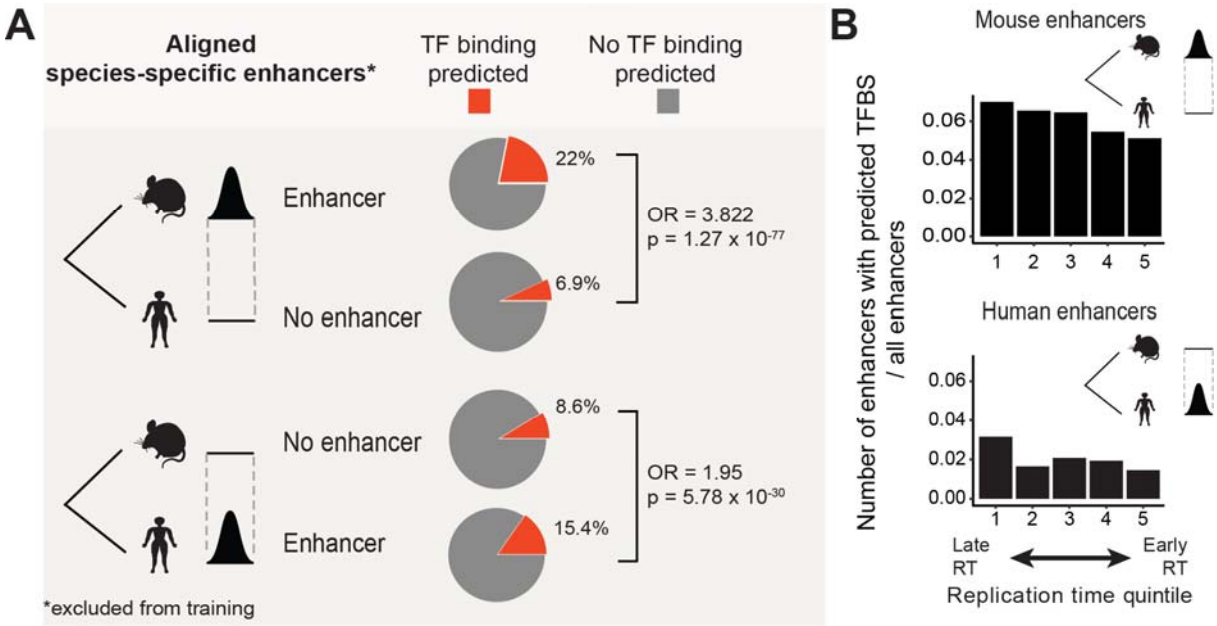


Figure 3. Deep leaning model links changes in TF binding sites with enhancer turnover. (A) Modified deep learning domain adaptive model trained with HNF4A and CEBPA binding sites in mouse and human genomes (Cochran et al. 2022). Prediction on species-specific enhancers and their aligned non-enhancer sequences in the other species (not included in training data) (Model details in **Supplemental Text**). The pie charts show the percentage of enhancers and matched non-enhancer regions with predicted HNF4A and CEBPA TFBSs with a probability threshold ≥ 0.9. Fisher's exact test odds ratio and p-value shown for each enhancer vs non-enhancer comparison. **(C)** Proportion of mouse- and human-specific enhancers with predicted TFBSs divided by the total number of enhancers across replication time quintiles.

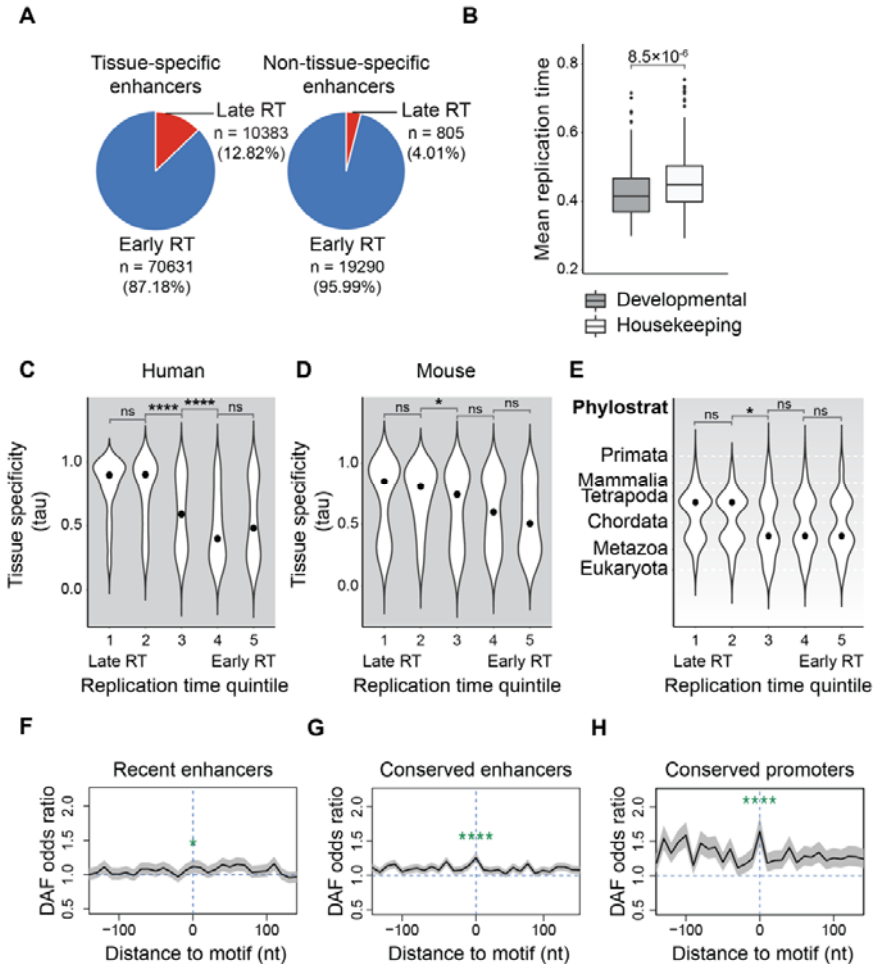
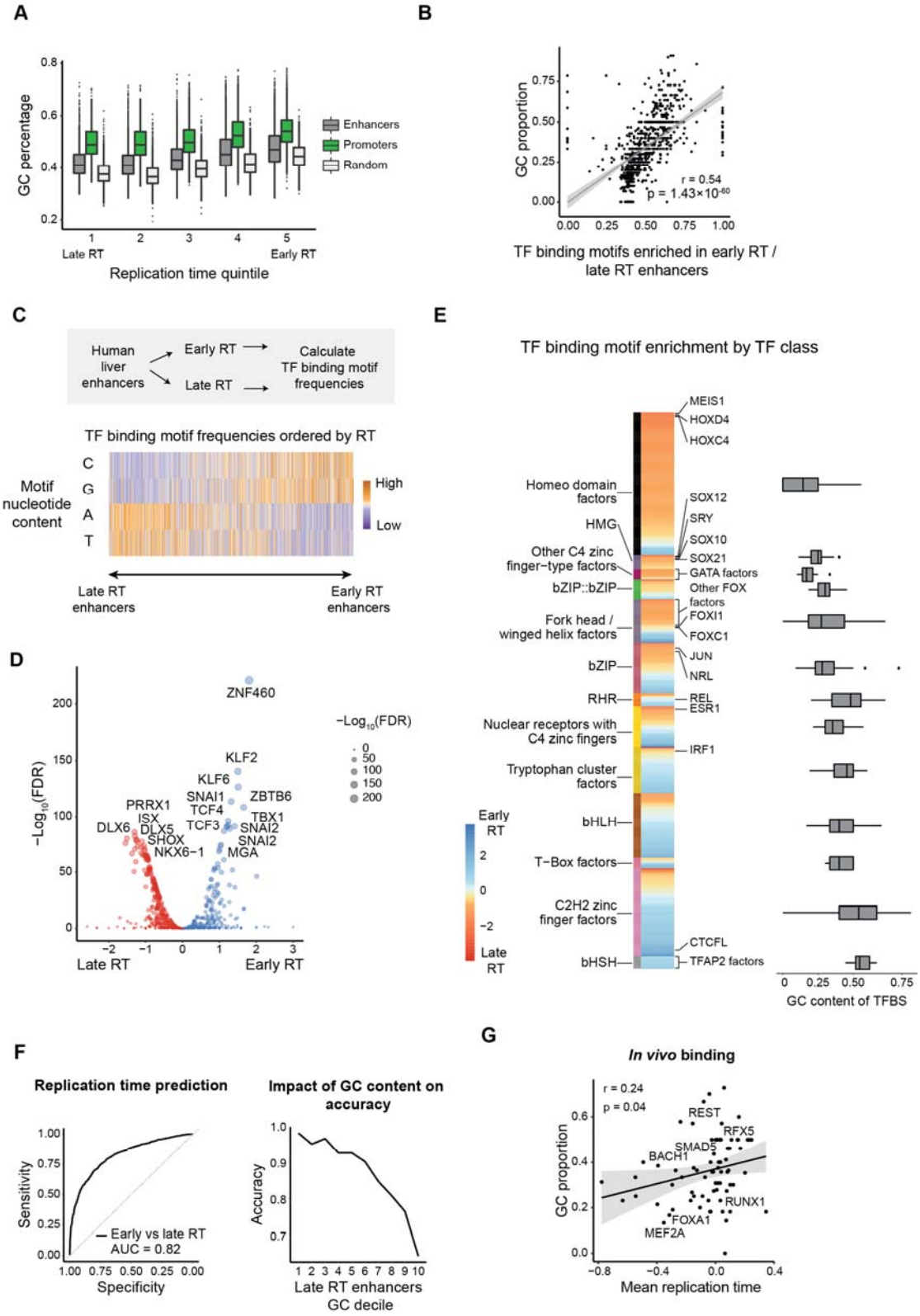


Figure 4. Enhancer tissue-specificity and selection pressures. (A) Proportion of early and late replicating enhancers for tissue-specific and non-tissue-specific mouse enhancers (all tissues) (Fisher's exact test, $p = 2.2 \times 10^{-16}$, odds ratio = 0.28). **(B)** Mean replication time of developmental and housekeeping fruit fly enhancers (Mann-Whitney *U*-test, alternative = "greater", n = 200 enhancers each class). **(C, D)** Violin plots of tissue-specific expression scores (tau values) of human and mouse TFs separated into five quintiles depending on their respective motif enrichments at early versus late replicating enhancers (Mann-Whitney *U*-test, pairwise comparison of later vs. earlier replicating quintile, alternative = "greater", significance code 'ns' P > 0.05 and '*' P ≤ 0.05)**. (E)** Violin plot of the evolutionary age of TFs separated into five quintiles depending on their motif enrichment at early versus late replicating enhancers (Mann-Whitney *U*-test, pairwise comparison of later vs. earlier replicating quintile, alternative = "greater", significance code 'ns' P > 0.05 and '*' P ≤ 0.05). **(F-H)** Derived Allele Frequency (DAF) odds ratio for recently evolved **(F)** and conserved human liver enhancers **(G)** and conserved promoters **(H)** compared to background genomic regions as a measure of selection pressure. Promoters and enhancers were centred based on the location of liver-specific functional motifs (Fisher's exact test, significance code '*' P ≤ 0.05 and '****' P ≤ 0.0001).



5 **Figure 5. Genome wide and enhancers nucleotide composition and replication time. (A)** GC percentage of human liver enhancers and promoters and random genomic regions across replication time quintiles (random regions were sampled from non-genic regions of the genome excluding also liver promoters and enhancers) n=10945 enhancers n=5724 promoters. **(B)**

Scatterplot of the proportion of GC for TF binding motifs (each dot) based on relative enrichment at early versus late replicating human liver enhancers. Motifs enriched within late replicating enhancers are to the left and those enriched at early replicating enhancers are on the right. Pearson correlation r and p-value are shown. **(C)** Heatmap of the proportion of GC/AT nucleotide content of TF binding motifs ordered based on their relative enrichment at early versus late replicating human liver enhancers (n = 5538 each replication time). Each column shows a human TF binding motif from the JASPAR database. **(D)** Enriched JASPAR motifs between early and late replicating liver enhancers. X-axis shows the relative enrichment for each motif at early versus late replicating enhancers; y-axis represent $-\log_{10}$ (Fisher's exact test, FDR). **(E)** Relative enrichment of TF binding motifs at early versus late replicating liver enhancers grouped by TF Class. The GC content of the motifs belonging to every TF Class is shown on the right. Only TF Classes with more than ten TFs are shown. **(F)** Receiver Operating Characteristic (ROC) curve for the prediction of early versus late replication time in liver enhancers based on their TF binding motif composition using XGBoost (left). Motifs used for prediction are the same as the ones shown in **(C).** Change in machine learning model prediction accuracy across late enhancers GC content deciles (right). Late replicating enhancers are divided into deciles based on their GC content. Smallest decile represents the lowest GC content. **(G)** Scatterplot of DNA replication time versus GC content of ChIP-seq binding sites in the cell line K562 for 71 TF ChIP-seq datasets. Mean replication time of binding sites in the later replicating cluster for each TF is shown (**Methods**, **Fig. S15**). Pearson correlation coefficient (r) and p-value are shown. Shaded region represents the 95% confidence interval of the line of best fit.

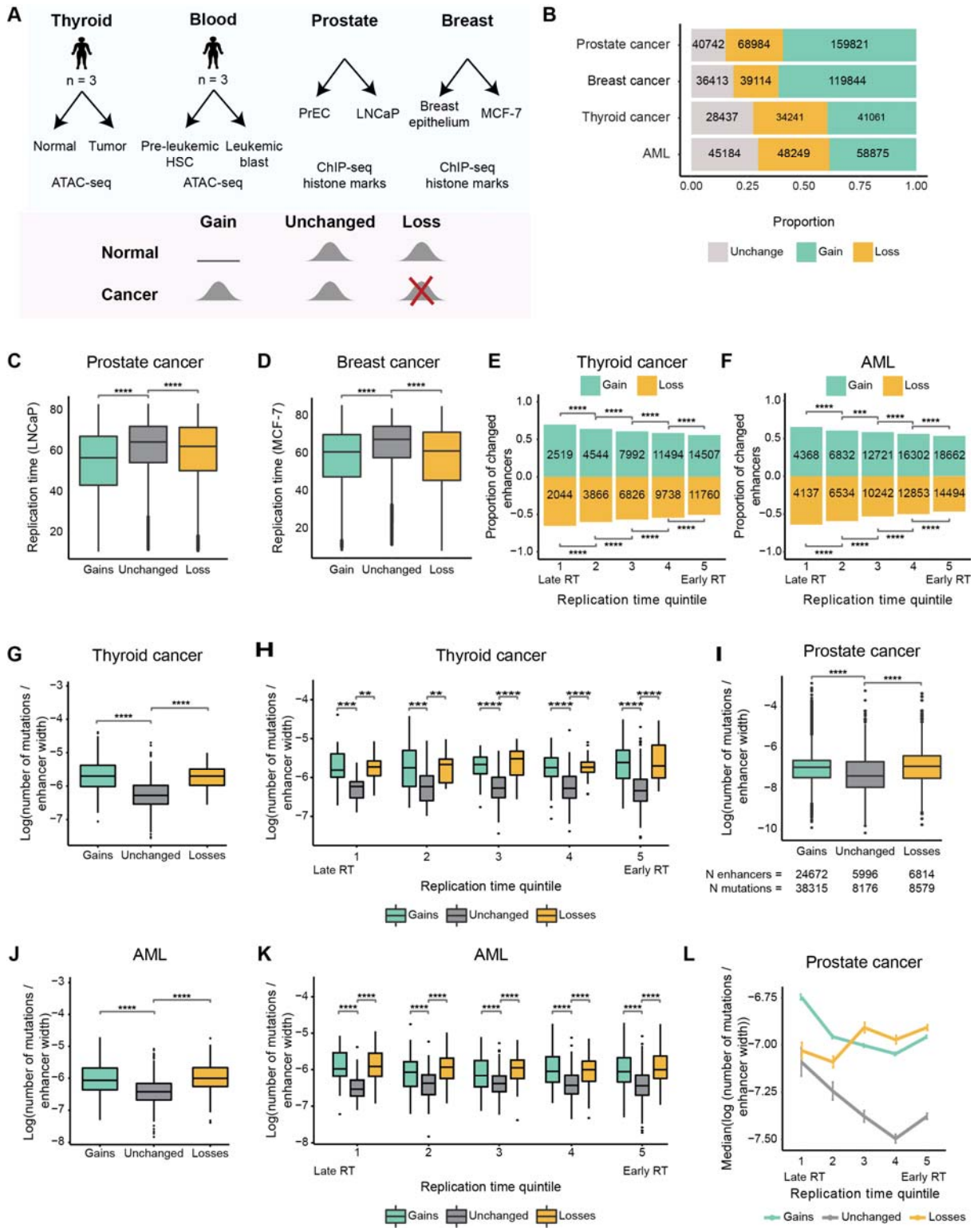


Figure 6. Enhancer turnover is associated to replication time in cancer. (A) Overview of the enhancer data sets in cancer and matched normal tissues and cell lines (top). Gained, unchanged and lost enhancers were defined in each cancer type (bottom). **(B)** Proportion of unchanged, gained and lost enhancers in each cancer type. **(C-D)** Replication time of gains, unchanged enhancers and loses in prostate **(C)** and breast **(D)** cancer (Mann-Whitney *U*-test). A similar trend is present in thyroid cancer and AML (**Fig. S17 B and C**). **(E-F)** proportion of enhancer gains and losses in thyroid cancer **(E)** and AML **(F)** relative to the number of unchanged enhancers across replication time quintiles (Fisher's exact test, alternative = "greater", compared to unchanged

enhancers). **(G)** Log transformed number of mutations normalised by enhancer width in thyroid cancer (Mann-Whitney *U*-test). **(H)** Log transformed number of mutations normalised by enhancer width in thyroid cancer across replication time quintiles **(**Mann-Whitney *U*-test**). (I)** Plot as **(G)** for prostate cancer. **(J)** Plot as **(G)** for AML. **(K)** Plot as **(H)** for AML**. (L)** Median log transformed number of mutations normalised by enhancer width at prostate cancer gains, unchanged enhancers, and losses across replication time quintiles (error bars represent standard error). Significance notation: 'ns' $P > 0.05$; '***' $1 \times 10^{-4} < P \leq 1 \times 10^{-3}$; '****' $P \leq 1 \times 10^{-4}$. Number of enhancers and mutations in each cancer type can be found in **Supplemental Table S3**.

References:

Agarwal I, Przeworski M. 2019. Signatures of replication timing, recombination, and sex in the spectrum of rare variants on the human X chromosome and autosomes. *Proc Natl Acad Sci U S A* **116**: 17916–17924.

Arnold CD, Gerlach D, Spies D, Matts JA, Sytnikova YA, Pagani M, Lau NC, Stark A. 2014. Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nat Genet* **46**: 685–692.

Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. 2013. Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* **339**: 1074–1077.

Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, Park K, Kitabayashi N, MacDonald TY, Ghandi M, et al. 2013. Punctuated evolution of prostate cancer genomes. *Cell* **153**: 666–677.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**: 11.

Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. 2019. *Calling Somatic SNVs and Indels with Mutect2*. Bioinformatics http://biorxiv.org/lookup/doi/10.1101/861054 (Accessed October 14, 2022).

Bernardi G. 2000. Isochores and the evolutionary genomics of vertebrates. *Gene* **241**: 3–17.

Bernardi G. 1993. The vertebrate genome: isochores and evolution. *Mol Biol Evol* **10**: 186–204.

Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2010. ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet* **42**: 806–810.

Blumenfeld B, Ben-Zimra M, Simon I. 2017. Perturbations in the Replication Program Contribute to Genomic Instability in Cancer. *Int J Mol Sci* **18**: E1138.

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew J-L, Ruan Y, Wei C-L, Ng HH, et al. 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* **18**: 1752–1762.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

Chen C-L, Rappailles A, Duquenne L, Huvet M, Guilbaud G, Farinelli L, Audit B, d'Aubenton-Carafa Y, Arneodo A, Hyrien O, et al. 2010. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res* **20**: 447–457.

Chen T, Guestrin C. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, New York, USA https://dl.acm.org/doi/10.1145/2939672.2939785 (Accessed September 2, 2021).

Cochran K, Srivastava D, Shrikumar A, Balsubramani A, Hardison RC, Kundaje A, Mahony S. 2022. Domain-adaptive neural networks improve cross-species prediction of transcription factor binding. *Genome Res* **32**: 512–523.

Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, Snyder MP, Pritchard JK, Kundaje A, Greenleaf WJ, et al. 2016. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet* **48**: 1193–1203.

Costantini M, Bernardi G. 2008. Replication timing, chromosomal bands, and isochores. *Proc Natl Acad Sci U S A* **105**: 3433–3437.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**: 21931–21936.

De S, Babu MM. 2010. A time-invariant principle of genome evolution. *Proc Natl Acad Sci U S A* **107**: 13004–13009.

Domazet-Lošo T, Tautz D. 2010. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468**: 815–818.

Du Q, Bert SA, Armstrong NJ, Caldon CE, Song JZ, Nair SS, Gould CM, Luu P-L, Peters T, Khoury A, et al. 2019. Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. *Nat Commun* **10**: 416.

Duque T, Sinha S. 2015. What does it take to evolve an enhancer? A simulation-based study of factors influencing the emergence of combinatorial regulation. *Genome Biol Evol* **7**: 1415–1431.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Eyre-Walker A. 1992. Evidence that both G + C rich and G + C poor isochores are replicated early and late in the cell cycle. *Nucleic Acids Res* **20**: 1497–1501.

Eyre-Walker A, Hurst LD. 2001. The evolution of isochores. *Nat Rev Genet* **2**: 549–555.

Francino MP, Ochman H. 1999. Isochores result from mutation not selection. *Nature* **400**: 30–31.

Frankel N, Wang S, Stern DL. 2012. Conserved regulatory architecture underlies parallel genetic changes and convergent phenotypic evolution. *Proc Natl Acad Sci U S A* **109**: 20975–20979.

Fryxell KJ, Zuckerkandl E. 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol Biol Evol* **17**: 1371–1383.

Gindin Y, Meltzer PS, Bilke S. 2014. Replicon: a software to accurately predict DNA replication timing in metazoan cells. *Front Genet* **5**: 378.

Goode DL, Cooper GM, Schmutz J, Dickson M, Gonzales E, Tsai M, Karra K, Davydov E, Batzoglou S, Myers RM, et al. 2010. Evolutionary constraint facilitates interpretation of genetic variation in resequenced human genomes. *Genome Res* **20**: 301–310.

Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, Besenbacher S, Magnusson G, Halldorsson BV, Hjartarson E, et al. 2015. Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* **47**: 435–444.

Hiratani I, Ryba T, Itoh M, Rathjen J, Kulik M, Papp B, Fussner E, Bazett-Jones DP, Plath K, Dalton S, et al. 2010. Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. *Genome Res* **20**: 155–169.

Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* **12**: 756–766.

Holmquist GP. 1987. Role of replication time in the control of tissue-specific gene expression. *Am J Hum Genet* **40**: 151–173.

Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci* **101**: 13994–14001.

Jacques P-É, Jeyakani J, Bourque G. 2013. The Majority of Primate-Specific Regulatory Sequences Are Derived from Transposable Elements. *PLoS Genet* **9**: e1003504.

Kenigsberg E, Yehuda Y, Marjavaara L, Keszthelyi A, Chabes A, Tanay A, Simon I. 2016. The mutation spectrum in genomic late replication domains shapes mammalian GC content. *Nucleic Acids Res* **44**: 4222–4232.

Kryuchkova-Mostacci N, Robinson-Rechavi M. 2017. A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform* **18**: 205–214.

Lang GI, Murray AW. 2011. Mutation Rates across Budding Yeast Chromosome VI Are Correlated with Replication Timing. *Genome Biol Evol* **3**: 799–811.

Lercher MJ, Smith NGC, Eyre-Walker A, Hurst LD. 2002. The Evolution of Isochores: Evidence From SNP Frequency Distributions. *Genetics* **162**: 1805–1810.

Lowe CB, Bejerano G, Haussler D. 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc Natl Acad Sci U S A* **104**: 8005–8010.

Lubelsky Y, Prinz JA, DeNapoli L, Li Y, Belsky JA, MacAlpine DM. 2014. DNA replication and transcription programs respond to the same chromatin cues. *Genome Res* **24**: 1102–1114.

Lundberg S, Lee S-I. 2017. A Unified Approach to Interpreting Model Predictions. http://arxiv.org/abs/1705.07874.

Lynch VJ, Leclerc RD, May G, Wagner GP. 2011. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat Genet* **43**: 1154–1159.

Pink CJ, Hurst LD. 2010. Timing of Replication Is a Determinant of Neutral Substitution Rates but Does Not Explain Slow Y Chromosome Evolution in Rodents. *Mol Biol Evol* **27**: 1077–1086.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma Oxf Engl* **26**: 841–842.

R Core Team. 2020. R: A language and environment for statistical computing.

Rebeiz M, Jikomes N, Kassner VA, Carroll SB. 2011. Evolutionary origin of a novel gene expression pattern through co-option of the latent activities of existing regulatory sequences. *Proc Natl Acad Sci U S A* **108**: 10036–10043.

Rhind N, Gilbert DM. 2013. DNA Replication Timing. *Cold Spring Harb Perspect Biol* **5**: a010132.

Rivera-Mulia JC, Buckley Q, Sasaki T, Zimmerman J, Didier RA, Nazor K, Loring JF, Lian Z, Weissman S, Robins AJ, et al. 2015. Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells. *Genome Res* **25**: 1091–1103.

Rivera-Mulia JC, Gilbert DM. 2016. Replication timing and transcriptional control: beyond cause and effect—part III. *Curr Opin Cell Biol* **40**: 168–178.

Roller M, Stamper E, Villar D, Izuogu O, Martin F, Redmond AM, Ramachanderan R, Harewood L, Odom DT, Flicek P. 2021. LINE retrotransposons characterize mammalian tissue-specific and evolutionarily dynamic regulatory regions. *Genome Biol* **22**: 62.

Ryba T, Hiratani I, Lu J, Itoh M, Kulik M, Zhang J, Schulz TC, Robins AJ, Dalton S, Gilbert DM. 2010. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res* **20**: 761–770.

Sanghi A, Gruber JJ, Metwally A, Jiang L, Reynolds W, Sunwoo J, Orloff L, Chang HY, Kasowski M, Snyder MP. 2021. Chromatin accessibility associates with protein-RNA correlation in human cancer. *Nat Commun* **12**: 5732.

Schmegner C, Hameister H, Vogel W, Assum G. 2007. Isochores and replication time zones: a perfect match. *Cytogenet Genome Res* **116**: 167–172.

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**: 1036–1040.

Scrucca L, Fop M, Murphy TB, Raftery AE. 2016. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J* **8**: 289–317.

Seim I, Jeffery PL, Thomas PB, Nelson CC, Chopin LK. 2017. Whole-Genome Sequence of the Metastatic PC3 and LNCaP Human Prostate Cancer Cell Lines. *G3 Bethesda Md* **7**: 1731–1741.

Siefert JC, Georgescu C, Wren JD, Koren A, Sansam CL. 2017. DNA replication timing during development anticipates transcriptional programs and parallels enhancer activation. *Genome Res* **27**: 1406–1416.

Smit AFA, Hubley R, Green P. 2013. RepeatMasker. http://www.repeatmasker.org.

Smith RP, Riesenfeld SJ, Holloway AK, Li Q, Murphy KK, Feliciano NM, Orecchia L, Oksenberg N, Pollard KS, Ahituv N. 2013. A compact, in vivo screen of all 6-mers reveals drivers of tissue-specific expression and guides synthetic regulatory element design. *Genome Biol* **14**: R72.

Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009. Human mutation rate associated with DNA replication timing. *Nat Genet* **41**: 393–395.

Stone JR, Wray GA. 2001. Rapid Evolution of cis-Regulatory Sequences via Local Point Mutations. *Mol Biol Evol* **18**: 1764–1770.

Supek F, Lehner B. 2015. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521**: 81–84.

Taberlay PC, Achinger-Kawecka J, Lun ATL, Buske FA, Sabir K, Gould CM, Zotenko E, Bert SA, Giles KA, Bauer DC, et al. 2016. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome Res* **26**: 719–731.

Taberlay PC, Statham AL, Kelly TK, Clark SJ, Jones PA. 2014. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. *Genome Res* **24**: 1421–1432.

Umer HM, Smolinska-Garbulowska K, Marzouka N, Khaliq Z, Wadelius C, Komorowski J. 2019. funMotifs: Tissue-specific transcription factor motifs. 683722. https://www.biorxiv.org/content/10.1101/683722v1 (Accessed April 27, 2022).

Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ, et al. 2015. Enhancer Evolution across 20 Mammalian Species. *Cell* **160**: 554–566.

Vinogradov AE. 2003. Isochores and tissue specificity. *Nucleic Acids Res* **31**: 5212–5220.

Vinogradov AE. 2005. Noncoding DNA, isochores and gene expression: nucleosome formation potential. *Nucleic Acids Res* **33**: 559–563.

White MA, Myers CA, Corbo JC, Cohen BA. 2013. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc Natl Acad Sci* **110**: 11952–7.

Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**: 650–659.

Yehuda Y, Blumenfeld B, Mayorek N, Makedonski K, Vardi O, Cohen-Daniel L, Mansour Y, Baror-Sebban S, Masika H, Farago M, et al. 2018. Germline DNA replication timing shapes mammalian genome composition. *Nucleic Acids Res* **46**: 8299–8310.

Young RS, Hayashizaki Y, Andersson R, Sandelin A, Kawaji H, Itoh M, Lassmann T, Carninci P, Consortium F, Bickmore WA, et al. 2015. The frequent evolutionary birth and death of functional promoters in mouse and human. *Genome Res* gr.190546.115.

Young RS, Talmane L, Marion de Procé S, Taylor MS. 2022. The contribution of evolutionarily volatile promoters to molecular phenotypes and human trait variation. *Genome Biol* **23**: 89.

Zabidi MA, Arnold CD, Schernhuber K, Pagani M, Rath M, Frank O, Stark A. 2015. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**: 556–559.

Zemojtel T, Kielbasa SM, Arndt PF, Behrens S, Bourque G, Vingron M. 2011. CpG deamination creates transcription factor-binding sites with high efficiency. *Genome Biol Evol* **3**: 1304–1311.

Zhao PA, Sasaki T, Gilbert DM. 2020. High-resolution Repli-Seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. *Genome Biol* **21**: 76.

Zhao Z, Boerwinkle E. 2002. Neighboring-Nucleotide Effects on Single Nucleotide Polymorphisms: A Study of 2.6 Million Polymorphisms Across the Human Genome. *Genome Res* **12**: 1679–1686.