Enhancer turnover in cancer and species evolution are associated with DNA replication timing

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
^ equal contributions

Abstract

Enhancers change rapidly during evolution, but the mechanisms by which new enhancers originate in the genome are mostly unknown. Not all regions of the genome evolve at the same rate and mutations are elevated at late DNA replication time. To understand the role played by mutational processes in enhancer evolution, we leveraged changes in mutation rates across the genome. By examining enhancer turnover in matched healthy and tumor samples in human individuals, we find while enhancers are most common in early replicating regions, new enhancers emerged more often at late replicating regions. Somatic mutations in cancer are consistently elevated in enhancers that have experienced turnover compared to those that are maintained. A similar relationship with DNA replication time is observed in enhancers across mammalian species and in multiple tissue-types. New enhancers appeared almost twice as often in late compared to early replicating regions, independent of transposable elements. We trained a deep learning model to show that new enhancers are enriched for mutations that modify transcription factor (TF) binding. New enhancers are also typically neutrally evolving, enriched in eQTLs, and are more tissue-specific than evolutionarily conserved enhancers. 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. These results demonstrate a relationship between mutation rate, DNA replication time and enhancer evolution across multiple time scales, suggesting these observations are time-invariant principles of genome evolution.

Introduction

Enhancers are *cis*-regulatory elements that are important as modulators of spatiotemporal gene expression. They act as as integrators of trans-acting signals by recruiting transcription factors (TFs) and other effector molecules. Unlike promoters, enhancers are typically rapidly evolving and are frequently species-specific. For example, most human enhancers are not found in the mouse [1–4] and the mechanisms responsible for enhancer turnover are not well understood. Expansions of *cis*-regulatory elements across the genome via transposable elements (TE) have been associated with enhancer turnover in mammals, however they do not explain the majority of recently evolved mammalian enhancers, many of which appear to have emerged from ancestral sequences without biochemical activity in the same tissue [5–9]. New enhancer activity can emerge via mutations of non-regulatory sequences, suggesting that local point mutations can create novel tissue-specific enhancers [10–15].

One of the most significant predictors of local mutational density is DNA replication time, which refers to the temporal order of DNA replication during the S-phase [16–18]. Replication time is associated with changes to CpG and non-CpG substitutions across the genome [19–22]. Mutations are elevated during late DNA replication time, corresponding to late S-phase, thought to be due to errors made in DNA replication and a reduced ability of DNA repair mechanisms to function effectively during late replication [23]. The trend of an increased mutational burden during the time course of DNA replication is deeply conserved and extends throughout eukaryotic evolution [24]. In hominids and rodents, mutation rates are 20-30% greater at late compared to early replication domains [19].

Mutational biases, not only mutational frequencies, are distinct between early and late DNA replicating domains in mammalian genomes [2,25,26]. Compared to early replicating regions, late replicating sequences are biased towards A and T base substitutions [19,20,27]. This bias appears to be due largely to the deamination of methyl-cytosine at CpG sites resulting in increased C to T transitions, combined with C to A mutations at CpG and non-CpG sites [28,29]. Large regional variations in GC base composition are found across mammalian genomes, and are known as ‘isochores’ [30–32]. Isochores position mirror replication domains [25,33]. Variation in GC content is correlated with genomic features including gene expression, recombination rates and 3D chromatin topology. Specifically, AT-rich and late replicating domains are linked to dynamic heterochromatin and tissue-specific gene expression [34,35].

Due to the deep links between replication time and DNA substitutions, we aim to understand these relationships with enhancer evolution. If mutations make a sizable contribution to enhancer evolution, we expect new enhancers to be more common at late replication time. To determine whether mutational processes associated with DNA replication time play a role in enhancer turnover in cancer, we discriminated between lost and gained enhancers across four cancer types and compared changes in mutation rates. We resolved the rate of mammalian enhancers evolutionary turnover using lineage-specific functional genomics data to test for similar mechanisms across vast timescales. By comparing enhancers across species, we capture enhancers at varying degrees of evolutionary divergence to measure their turnover with respect to DNA replication time.

Results

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

Not all regions of the genome evolve at the same rate. DNA replication time is a major predictor of local mutation rates along the genome at megabase scale with genetic changes providing novel variation that is the substrate for evolution. DNA replication time is reported to be largely stable between normal and cancer cells [36,37]. We compared replication time in healthy and cancer cell states across four different cancer types (**Fig. 1A**). Enhancers, based on distal chromatin accessible regions, were identified in matched patient primary cancer and healthy samples in six individuals (n = 3 for each cancer type) for thyroid cancer and acute myeloid leukemia (AML) where preleukemic cells were considered as the healthy counterpart in AML. We also identified enhancers in both normal breast and prostate and in their diseased states based on the presence/absence of histone modifications denoting active enhancers [38–42] (**Fig. 1A, Fig. S1A, Table S1**) (**Methods**).

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. 1A-B**). We compared enhancer gains and losses across different replication timepoints and show 1) unchanged enhancers were, on average, more early replicating than enhancers that have experienced turnover ($p < 2.2 \times 10^{-16}$ in prostate and breast cancer, $p \leq 1 \times 10^{-4}$ in thyroid cancer and AML) 2) the proportion of enhancers, relative to unchanged, that have experienced turnover was tend to be greatest at the latest replicating domains (**Fig. 1C-F, Fig. S1B-G**). This trend was unaffected by differences in recombination breakpoints [43] (**Fig. S2**).

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 patient-matched 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 [44].

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 (Mann-Whitney *U*-test for gains or losses versus unchanged, $p < 6.6 \times 10^{-147}$ in prostate; $p \leq 2 \times 10^{-16}$ in thyroid; $p \leq 2 \times 10^{-16}$ in AML; effect size r = 0.2, 0.4-0.5, 0.4, respectively) (**Fig. 1G-L, Fig. S1H-K, Table S1**). The trend was consistent across all individuals for thyroid cancer and AML (**Fig. S3**). In all cancer types, we restricted our analysis to single nucleotide polymorphisms (SNPs) and common genetic variants were filtered out. 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).

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

We hypothesized that the same DNA replication timing-linked mechanisms which promote regulatory element turnover between healthy and tumor cells play a similar role in species evolution. Based on this premise, we examined multi-tissue enhancer turnover in mouse using germline DNA replication time. Mouse-specific (termed 'recent') 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) [8]. 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 [8,45] (**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 [46,47]. 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 consistent with cancer evolution. 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. 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). Thus, 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. Recent enhancers were 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 [8,47,48].

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 = 1×10^{-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 = 3.56×10^{-40}, odds ratio = 0.13 for human enhancers; Fisher's exact test, p = 6.8×10^{-171}, odds ratio = 0.63 for mouse enhancers) (**Fig. S7**). Our results confirm that the vast majority of recently evolved mouse and human enhancers are unlikely to have arisen through tandem duplications.

Sequence divergence predicts functional changes in species-specific enhancers

Given many enhancers arise in non-regulatory sequences, we sought to understand the impact of evolutionary sequence variants on enhancer turnover. To address this question, we trained a deep learning model to infer liver-specific TF binding sites based solely on *cis*-regulatory sequence input [49]. We then tested the model's ability to classify TF binding sites at enhancers that have experienced evolutionary turnover between human and mouse (**Methods**). The model uses a domain adaptive network architecture which involves a gradient reversal layer to remove the effect of species-specific genome background between human and mouse (**Supplemental Text**) [49]. 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 in one species but not its ortholog. If mutations causally impact enhancer turnover through altering HNF4A and CEPBA binding motifs we would expect to observe an enrichment for their binding sites at enhancers compared to the non-enhancer sequence ortholog.

The sequence-based model identified significantly more HNF4A and CEPBA TF binding sites at human enhancers compared to the mouse aligned non-enhancer sequence ($p = 5.78 \times 10^{-30}$; OR = 1.95) (**Fig. 3A**). The opposite was true while comparing mouse-specific enhancers to their aligned non-enhancer sequence in human ($p = 1.27 \times 10^{-77}$; OR = 3.82) (**Fig. 2A)**. Furthermore, the overall proportion of species-specific enhancers with predicted HNF4A or CEPBA binding sites was higher at late replicating regions, suggesting that mutations are linked to the loss and gain of functional TF binding sites (**Fig. 3B)**. We note that we do not expect that all such species-specific enhancers can be explained by changes to HNF4A and CEPBA binding due to the large number of trans-acting factors that are active in liver. Moreover, variants that impact factors beyond direct binding (e.g. DNA shape) can also change enhancer activity.

Late replicating enhancers are tissue-specific and neutrally evolving

GC poor, late replicating regions are associated with dynamically regulated heterochromatin and nucleosome formation potential and linked to tissue specific gene expression [34,35,50,51] . Thus, we asked whether tissue-specific enhancer activity was similarly associated. 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. 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 [52], we examined tissue-specific TF expression for 477 and 360 genes in human and mouse across 27 and 19 tissues, respectively [53]. 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-D**). Consistent with this, certain TF classes are more likely to be recruited to enhancers at different replication domains due to GC isochores

(**Supplementary Text**). For example, overrepresented motifs at late replicating enhancers are AT-rich and these motifs are associated with tissue-specific TFs and homeodomain TFs play an important role in body patterning.

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 [54], 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. S8-9**).

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, $p < 2 \times 10^{-16}$, odds ratio = 1.2) – consistent with recent reports of an increased frequency of eQTLs at new promoters in the human genome [55]. 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)[56]. 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-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-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-D**).

Enhancer evolution is associated with CpG methylation at late replication time

Although DNA replication time is a major predictor of local mutation rates along the genome, the connections between genomic features are complex. We sought to examine the relationships between multiple genetic and epigenetic features and the turnover of human enhancers.

We defined recently evolved and evolutionary conserved human liver enhancers by comparing enhancers across mammals to distinguish their degree of evolutionary conservation [3]. Enhancers were characterized based their overlap with H3K27ac but not H3K4me3 histone marks. Conserved enhancers (n = 13329) were defined as human enhancers with conserved activity in at least two of 18 other mammalian species (**Fig. S11**). Recently evolved enhancers (n = 10434) were defined as those which either did not align to another mammalian species profiled or did not show a similar histone mark. Human promoters were defined as regions enriched in H3K4me3 and H3K27ac or only enriched in H3K4me3 [3]. 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 (ESC H9), the rate of promoter and transposable elements turnover, chromatin accessibility (ATAC-seq), CTCF binding and histone marks including H3K79me2, H2AFZ,

H3K9ac, H3K27me3, H3K9me3, H3K36me3, H4K20me1, H3K4me1, H3K4me2, H3K4me3 and H3K9me2 (**Methods**).

To assess the average consequence of each feature to enhancer turnover, we performed an ordinary least squares linear multiple regression at 1Mb non-overlapping segments across the human genome and calculated an average marginal contribution using SHAP for each feature (**Table S2**) [57]. Hallmarks of euchromatin, including numbers of ATAC-seq peaks and the presence of the histone marks H3K4me2, H3K9ac and H3K4me1 were most strongly linked to enhancer conservation (**Fig. S12A-G**). In contrast, rapidly evolving enhancers were associated with increased numbers of CpGs, increased methylation of CpGs, late DNA replication time, increased local mutation rates and rates of TE turnover (**Fig. 5A, Fig. S12A-B and H-J**). These results are consistent with an association of enhancer turnover with the deamination of methylated CpG sites leading to C to T transitions due to increased DNA methylation levels at late replication time. However, all substitution types increase with DNA replication timing not only at CpG sites [19]. Rates of enhancer and promoter turnover were largely concordant across the genome although promoters were more slowly evolving than enhancers (**Fig. 5B**) consistent with earlier reports [3]. Promoter turnover showed a similar relationship with replication time to enhancers (**Fig. S13; Table S3**). Unlike promoters, the rate of turnover in randomly sampled non-coding regions did not clearly correlate with enhancer turnover (**Fig. S12K**).

While a higher rate of enhancer turnover was generally associated with later replication time at less transcriptionally active parts of the genome, there are exceptions. Certain highly gene dense regions displayed rapid rates of enhancer turnover (*t*-test, $t = 2.183$ and $p = 2.9 \times 10^{-02}$) (**Fig. 5C**). 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 comprised of expanded gene families. Early replicating gene dense regions were enriched for kallikrein and zinc finger gene families, whereas at late replicating domains gene dense regions were overrepresented for G protein-coupled receptors gene families, particularly olfactory receptors (**Fig. S5, Tables S13-14**)[46]. Both sets of regions were enriched in recently evolved TEs (**Fig. S13C**).

In summary, our analyses reveal candidate causal mechanisms for enhancer turnovers include higher local mutation rates, DNA methylation of CpG sites as well as TE turnover. The relationship with replication time is not exclusive to enhancers and impacts other cis-regulatory elements including promoters.

Discussion

The results of this study reveal the significance of genome organization on the evolution of enhancers both at the level of species evolution and in the context of cancer. Although most enhancers were found at early replicating regions, comparative analyses revealed young enhancers were almost twice as likely to replicate later than conserved enhancers. Genomic features linked to enhancer conservation across the genome include open chromatin and the density of histone marks. At late replication time, local mutation rates, DNA methylated sites and TE evolution are positively associated with enhancer turnover. The relationship with replication timing is not exclusive to enhancers; it also affects other cis-regulatory elements such as promoters.

Enhancers in cancer follow a similar evolutionary pattern. Gained and lost enhancers in cancer were more often at late DNA replication time and harboured significantly more mutations than unchanged enhancers suggesting mutations have causal roles. The finding was consistent across four different types of cancers independent of the presence of recombination breakpoints which

have been associated with increased inversions, deletions and translocation in cancer genomes [36,58].

One means by which an enhancer can be created or decommissioned is through changes to TF binding. Theoretically, the short length of TF binding motifs and their degeneracy of binding allows TF binding motifs to emerge quickly in evolution and become fixed in the population [10–15]. Indeed, increased methylation levels at late replication time linked to high rates of C to T transitions at CpG dinucleotides due to the spontaneous deamination of methylated cytosines have been linked to the creation of new highly specific TF binding sites [14]. Using a deep learning model, we tested whether sequence differences predict changes to TF binding at enhancers that have experienced turnover events between human and mouse. We traced species-specific gain or loss of enhancers to changes in nucleotides that impact TF binding. The result is significant because it demonstrates that mutational changes across evolution results in *cis*-regulatory sequence changes at liver-specific TFs that can alter TF binding. Moreover**,** these changes were more common at late DNA replication time suggesting that once cis-regulatory elements are created at late replicating domains, they should also evolve more rapidly due to the elevated mutation rates at these regions.

Our results indicate tissue-specific enhancers are more common at late replicating regions, and these are recognised by tissue-specific TFs. Replication timing domains are closely matched to GC isochores and are linked to tissue-specific differences in gene expression in mammals [36,58]. The highly organized isochore patterns in mammalian genomes influences the positions of known TF binding sites across the genome. Specific types of TFs are more likely to be recruited at certain replication timing domains due to GC isochores. For example, tissue-specific TFs appeared frequently linked to AT-rich, late replicating binding sites. These findings help explain the tissue-specificity of gene expression at GC-depleted regions [34,35]. However, *Drosophila*, which do not possess isochores, also appeared to regulate developmental enhancer activity at later replication time compared to house-keeping enhancer activity.

Across the whole genome, enhancers appear to be largely neutrally evolving except at TF binding sites. Perhaps, most surprisingly, even conserved enhancers in our study did not show strong evidence for purifying selection except at TF binding sites. In contrast, promoters appear subjected to a much stronger level of purifying selection [9]. Our findings are consistent with the known properties of robustness within enhancers and suggest that both promoters and enhancers are subjected to increased mutations at late replication time contributing to increased rates of turnover relative to the total number of cis-regulatory elements within the replication domain.

Following these lines of reasoning, our findings would suggest that evolutionary innovation in gene regulatory modules may be more likely to emerge from regulatory elements at later replication domains, and that this is more likely to occur in a tissue-restricted manner. Late replicating regions were associated with development enhancer activity and enriched for developmentally relevant TF binding sites associated with body patterning. Based on the enrichment of developmental TFs, transcriptional changes occurring at late replicating domains may have been key to the evolution of the bilaterian body plan and the embryonic development of multicellular organism. It is also important to note that replication timing is dynamic between cell types and varies between germline and somatic cell types [46,59,60]. Around 30% of the human genome switches between replication timing domains across 26 human cell lines [61]. These dynamics can shift enhancers emerging in germ cells at late replication time to earlier replicating domains in differentiated cell types where there is a greater chance that it can influence gene activity.

In summary, the relationship of DNA replication timing with genome structure and mutational processes is correlated with the birth and death of enhancers, the distribution and evolution TF binding motifs, tissue specific gene expression and binding proteins across the genome. Similar

genome-wide mechanisms contributing to *cis*-regulatory evolution play out in species evolution, in the scale of millions of years, and in cancer cells, across months or several years, suggesting these findings are general features of genome evolution and regulatory turnover.

Methods

All analyses were performed on the mouse and human genome assemblies mm10 and hg19 unless specified otherwise. R v4.0.0 [62] was used.

Cancer data analyses

ChIP-seq data from the prostate cancer cell line, LNCaP, was used to annotate enhancers (Acc: GSE73783) [41]. For normal prostate epithelial cells (PrEC), enhancers were defined in an unsupervised manner using chromHMM (Acc: GSE57498) [42]. 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) [41]. For normal prostate epithelial cells (PrEC), enhancers were defined in an unsupervised manner using a collection of histone marks using chromHMM (Acc: GSE57498) [42]. 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).

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) [39], and ATAC-seq files for the matched pre-leukemic and blast cells from three individuals (Acc: GSE74912; SU484, SU501, SU654) [38]. 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 [63] 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) [36]. 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.

Cross-species replication time analysis

Species-specific ChIP-seq datasets used are E-MTAB-2633 and E-MTAB-7127 [3,8]. Z-score transformed replication timing data was obtained from human ESC H9 [64], mouse primordial germ cells (PGC) and spermatogonial stem cells (SSC) (Acc: GSE109804)[47], and 22 mouse cell lines differentiated from ES cells [46]. 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) [46,47]. 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" [62]. 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) [65]. 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) [66] were annotated using RepeatMasker (v4.0.6 for mouse, v4.1.0 for human) using the sensitive search setting for mouse ('-s') [67]. 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)[68]. 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)[69]. 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 [49]. 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. [49]. Liver human and mouse HNF4a and CEPBA ChIP-seq peak data taken from from [2] 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 [56]. 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 [55]. 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 [70]. 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)[71], 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 $< 1 \times 10^{-6}$ and query coverage > 20. We defined singleton enhancers as the enhancers without significant similarity to other enhancers.

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 [72] (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.

Multiple regression

A multiple linear regression was used to access the relationship between enhancer turnover and other features including liver DNA methylation, liver chromatin accessibility, CTCF binding and histone marks including H3K79me2, H2AFZ, H3K9ac, H3K27me3, H3K9me3, H3K36me3, H4K20me1, H3K4me1, H3K4me2, H3K4me3 and H3K9me2 (**Table S5**). Datasets were downloaded in hg19 coordinates, if available, or mapped using USCS liftOver from hg38 to hg19.

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)[73]. GC content of every window was calculated using bedtools nuc [73]. 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) [64].

A multiple linear regression model was fitted using enhancer turnover as the dependent variable and all other features described above as independent variables. The model was fitted using the function OSL from statsmodels version 0.12.2 in python. Feature importance was calculated using SHAP [57] (shap v0.39.0, Python v3.9.12). Multiple regression was also performed using promoter turnover as the dependent variable.

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), ARC Discovery Project (DP200100250) and a Snow Fellowship.

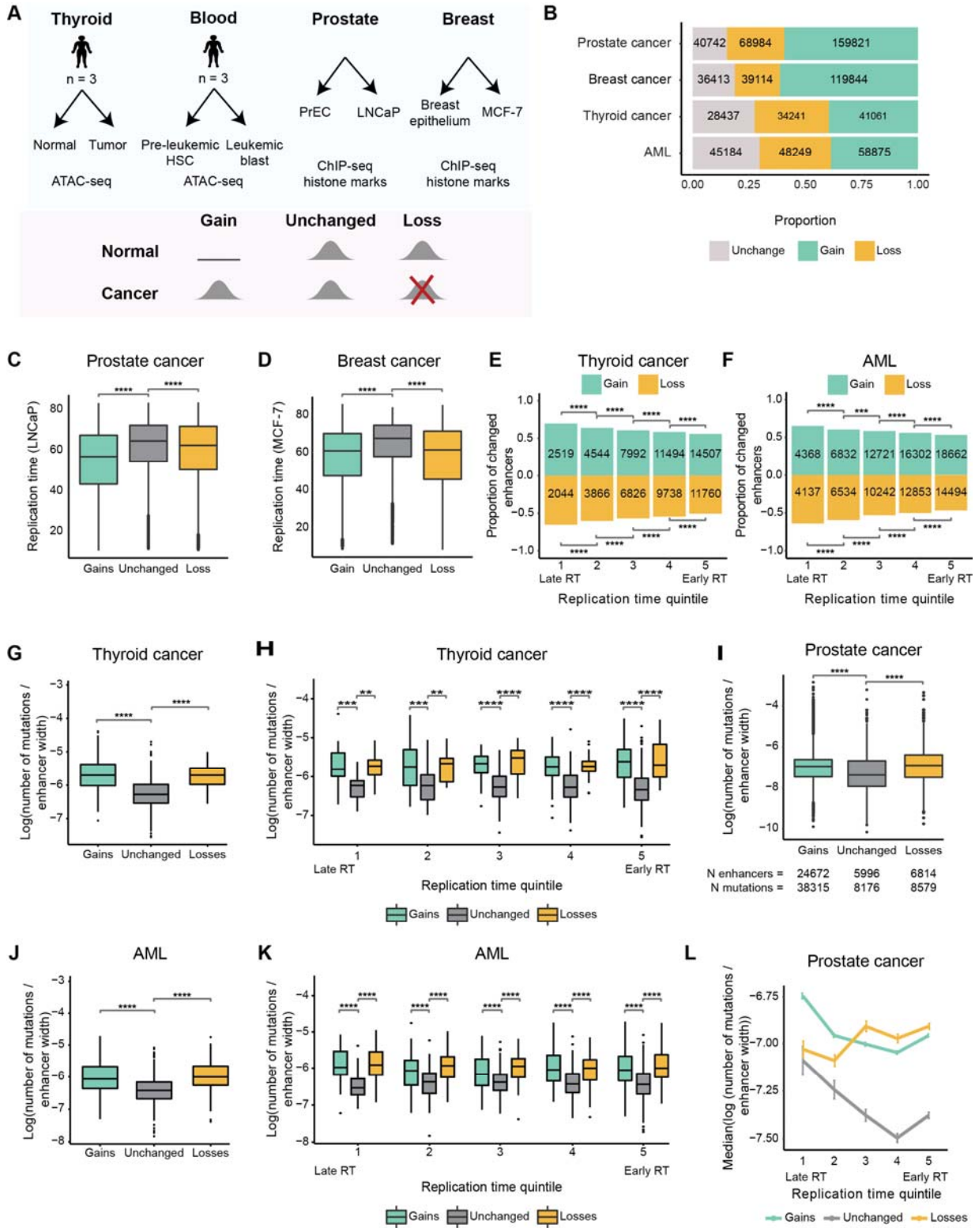

Figure 1. Enhancer turnover is associated to replication time in cancer (A) Overview of the enhancer data sets in cancer and matched normal tissues and cell lines (top). Gained, unchanged and lost enhancers were defined in each cancer type (bottom). **(B)** Proportion of unchanged, gained and lost enhancers 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. S11 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 **Table S1**.

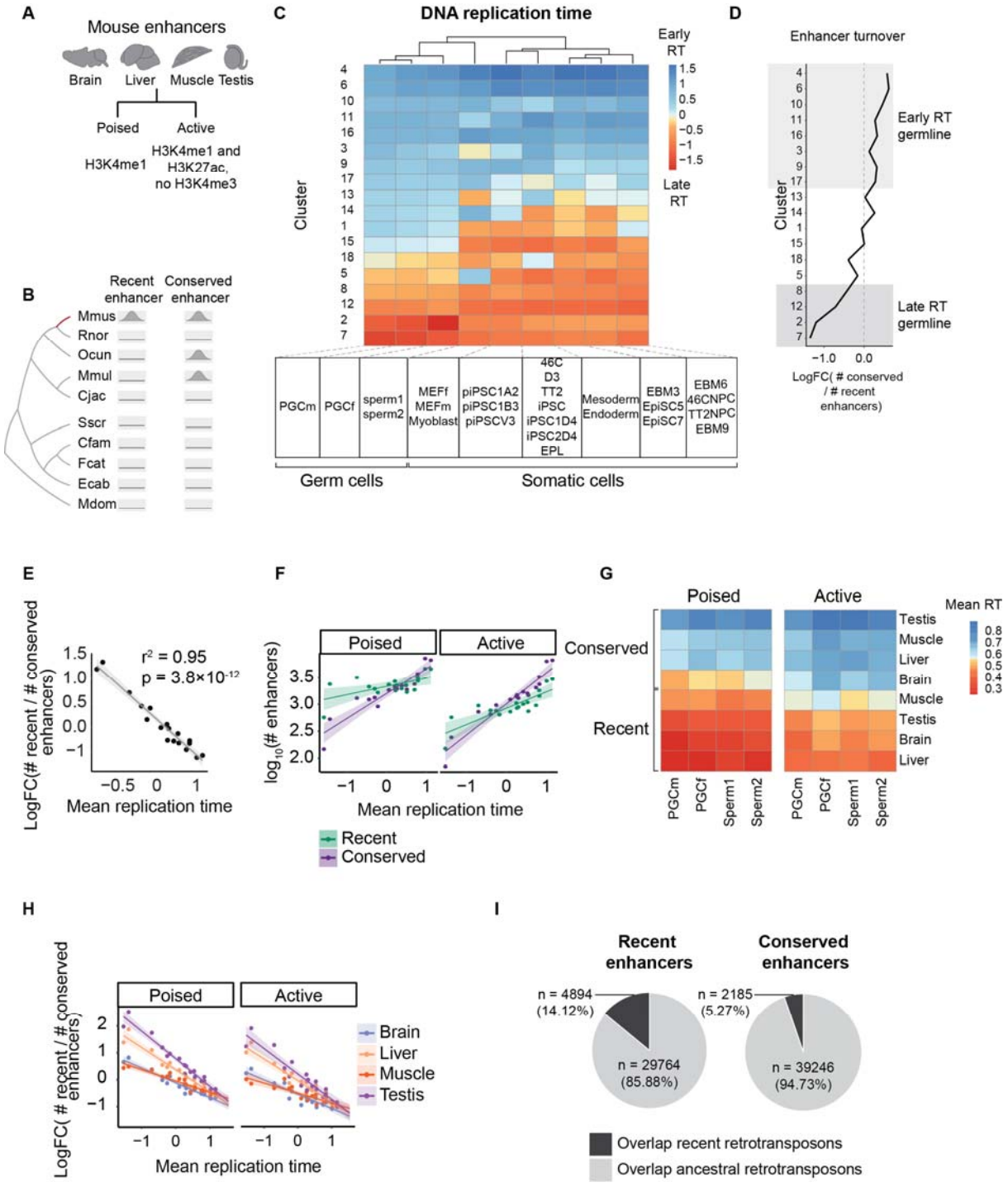

Figure 2. Increased turnover of enhancers increase in late replicating genomic regions (A) Mouse enhancers are defined based on combinations of histone marks. **(B)** Definition of mouse recent and conserved enhancers [8]. Recent enhancers were defined as regions that do not align to nine other mammalian genomes in the same tissue [8]. Conserved enhancers are aligned to enhancers in the same tissue in at least two other species [8]. 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).

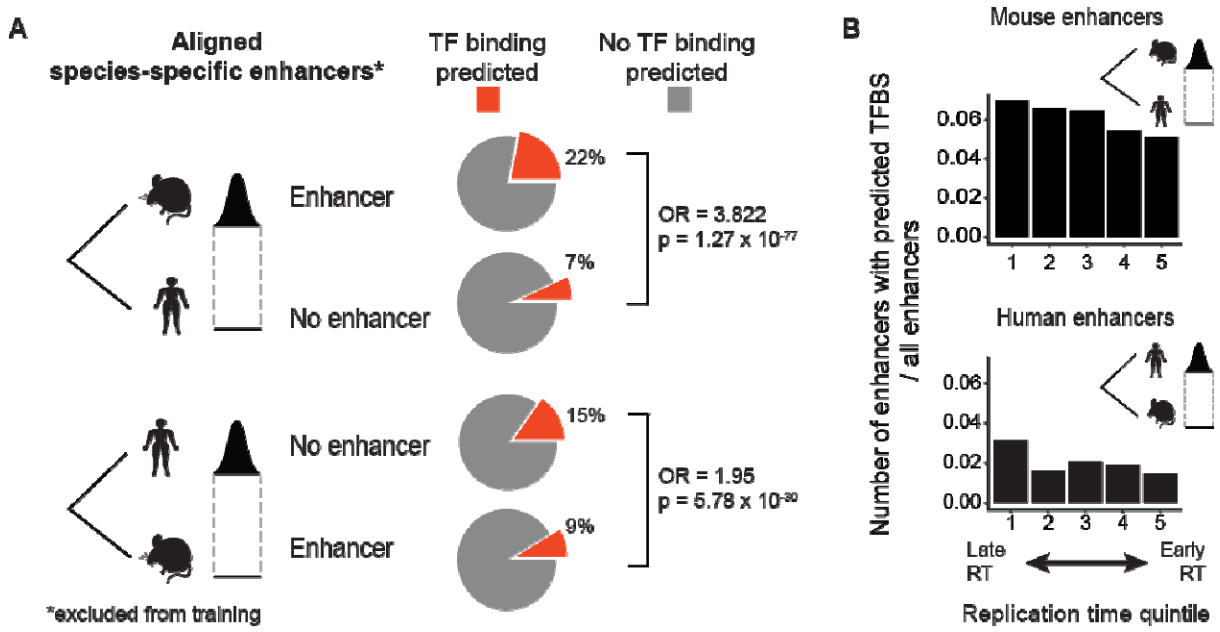

Figure 3. Deep learning model links changes in TF binding sites with enhancer turnover (A) Modified deep learning domain adaptive model trained with HNF4A and CEBPA binding sites in mouse and human genomes [49]. Prediction on species-specific enhancers and their aligned non-enhancer sequences in the 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.

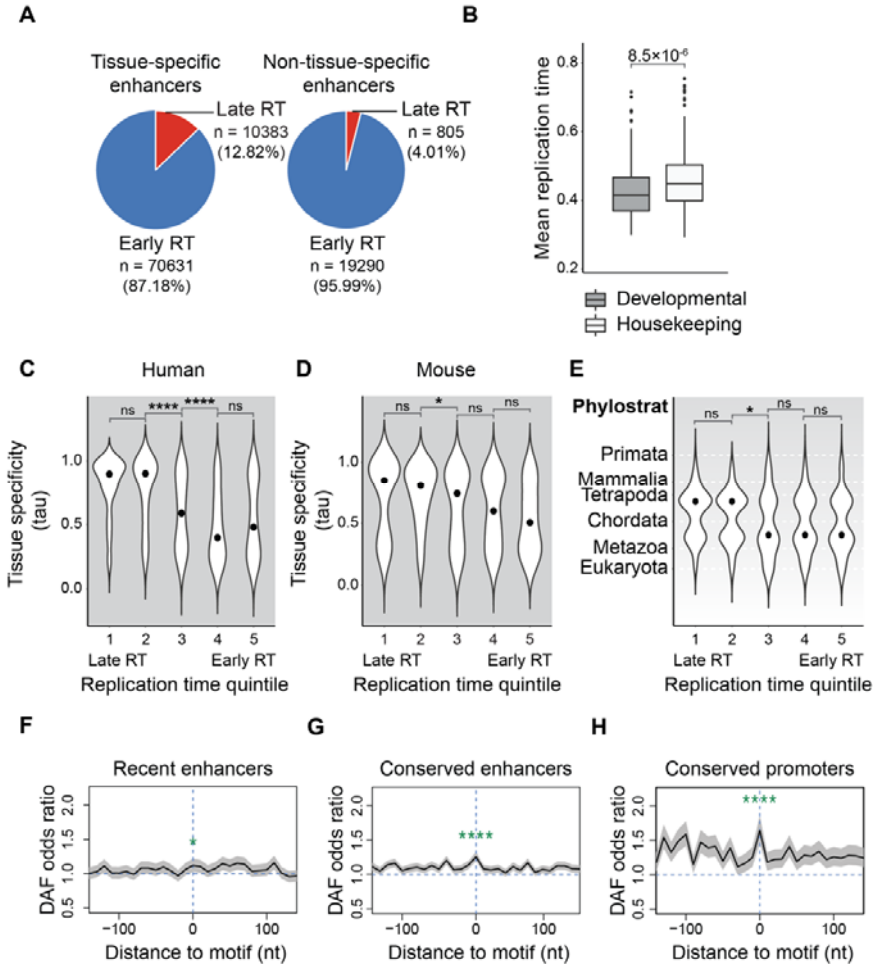

Figure 4. Enhancer tissue-specificity and selection pressures **(A)** Proportion of early and late replicating enhancers for tissue-specific and non-tissue-specific mouse enhancers (all tissues) (Fisher's exact test, $p = 2.2 \times 10^{-16}$, odds ratio = 0.28). **(B)** Mean replication time of developmental 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 '*' $P \leq 0.05$). **(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 '*' $P \leq 0.05$). **(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 '*' $P \leq 0.05$ and '****' $P \leq 0.0001$).

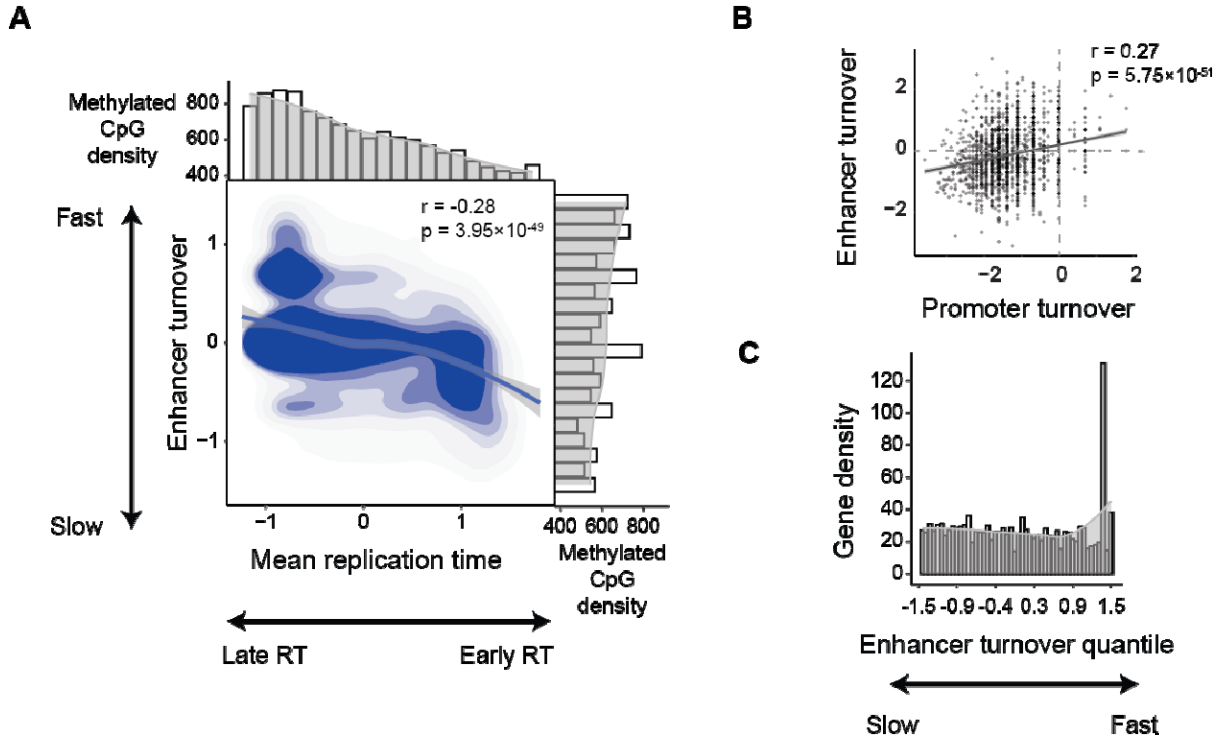

Figure 5. Genomic and epigenomic features associated with enhancer turnover (A) Density plot between enhancer turnover rate and mean DNA replication time across 1Mb genomic bins (Pearson's r shown). The colour scale represents the probability distribution of the 1Mb bins. The marginal distributions show of the number of methylated CpGs across 20 bins of enhancer turnover and mean replication time values. Only bins with enhancer turnover in the range from -1.5 to 1.5 are shown (94% bins). **(B)** Scatterplot of enhancers turnover and promoter turnover across 1Mb bins. Pearson's r shown. **(C)** Gene density values across enhancer turnover. The mean gene density was calculated across 50 bins of enhancer turnover.

References:

1. Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, et al. ChIP-Seq identification of weakly conserved heart enhancers. Nat Genet. 2010;42:806–10.

2. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science. 2010;328:1036–40.

3. Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, et al. Enhancer Evolution across 20 Mammalian Species. Cell. 2015;160:554–66.

4. Arnold CD, Gerlach D, Spies D, Matts JA, Sytnikova YA, Pagani M, et al. Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. Nat Genet. 2014;46:685–92.

5. Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res. 2008;18:1752–62.

6. Lynch VJ, Leclerc RD, May G, Wagner GP. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat Genet. 2011;43:1154–9.

7. Lowe CB, Bejerano G, Haussler D. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proc Natl Acad Sci U S A. 2007;104:8005–10.

8. Roller M, Stamper E, Villar D, Izuogu O, Martin F, Redmond AM, et al. LINE retrotransposons characterize mammalian tissue-specific and evolutionarily dynamic regulatory regions. Genome Biol. 2021;22:62.

9. Young RS, Hayashizaki Y, Andersson R, Sandelin A, Kawaji H, Itoh M, et al. The frequent evolutionary birth and death of functional promoters in mouse and human. Genome Res. 2015;gr.190546.115.

10. White MA, Myers CA, Corbo JC, Cohen BA. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. Proc Natl Acad Sci [Internet]. 2013 [cited 2013 Jul 3]; Available from: http://www.pnas.org/content/early/2013/06/28/1307449110

11. Duque T, Sinha S. What does it take to evolve an enhancer? A simulation-based study of factors influencing the emergence of combinatorial regulation. Genome Biol Evol. 2015;7:1415–31.

12. Smith RP, Riesenfeld SJ, Holloway AK, Li Q, Murphy KK, Feliciano NM, et al. A compact, in vivo screen of all 6-mers reveals drivers of tissue-specific expression and guides synthetic regulatory element design. Genome Biol. 2013;14:R72.

13. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. Science. 2013;339:1074–7.

14. Zemojtel T, Kielbasa SM, Arndt PF, Behrens S, Bourque G, Vingron M. CpG deamination creates transcription factor-binding sites with high efficiency. Genome Biol Evol. 2011;3:1304–11.

15. Stone JR, Wray GA. Rapid Evolution of cis-Regulatory Sequences via Local Point Mutations. Mol Biol Evol. 2001;18:1764–70.

16. Hodgkinson A, Eyre-Walker A. Variation in the mutation rate across mammalian genomes. Nat Rev Genet. 2011;12:756–66.

17. Zhao Z, Boerwinkle E. Neighboring-Nucleotide Effects on Single Nucleotide Polymorphisms: A Study of 2.6 Million Polymorphisms Across the Human Genome. Genome Res. 2002;12:1679–86.

18. Hwang DG, Green P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci. Proceedings of the National Academy of Sciences; 2004;101:13994–4001.

19. Chen C-L, Rappailles A, Duquenne L, Huvet M, Guilbaud G, Farinelli L, et al. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. Genome Res. 2010;20:447–57.

20. Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. Human mutation rate associated with DNA replication timing. Nat Genet. 2009;41:393–5.

21. Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, Sunyaev SR, et al. Differential Relationship of DNA Replication Timing to Different Forms of Human Mutation and Variation. Am J Hum Genet. 2012;91:1033–40.

22. Herrick J. Genetic Variation and Dna Replication Timing, or Why Is There Late Replicating Dna? Evolution. 2011;65:3031–47.

23. Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature. 2015;521:81–4.

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

25. Costantini M, Bernardi G. Replication timing, chromosomal bands, and isochores. Proc Natl Acad Sci U S A. 2008;105:3433–7.

26. Eyre-Walker A. Evidence that both G + C rich and G + C poor isochores are replicated early and late in the cell cycle. Nucleic Acids Res. 1992;20:1497–501.

27. Pink CJ, Hurst LD. Timing of Replication Is a Determinant of Neutral Substitution Rates but Does Not Explain Slow Y Chromosome Evolution in Rodents. Mol Biol Evol. 2010;27:1077–86.

28. Fryxell KJ, Zuckerkandl E. Cytosine deamination plays a primary role in the evolution of mammalian isochores. Mol Biol Evol. 2000;17:1371–83.

29. Agarwal I, Przeworski M. 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. 2019;116:17916–24.

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

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

32. Eyre-Walker A, Hurst LD. The evolution of isochores. Nat Rev Genet. 2001;2:549–55.

33. Schmegner C, Hameister H, Vogel W, Assum G. Isochores and replication time zones: a perfect match. Cytogenet Genome Res. 2007;116:167–72.

34. Vinogradov AE. Isochores and tissue specificity. Nucleic Acids Res. 2003;31:5212–20.

35. Vinogradov AE. Noncoding DNA, isochores and gene expression: nucleosome formation potential. Nucleic Acids Res. 2005;33:559–63.

36. Du Q, Bert SA, Armstrong NJ, Caldon CE, Song JZ, Nair SS, et al. Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. Nat Commun. Nature Publishing Group; 2019;10:416.

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

38. Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet. Nature Publishing Group; 2016;48:1193–203.

39. Sanghi A, Gruber JJ, Metwally A, Jiang L, Reynolds W, Sunwoo J, et al. Chromatin accessibility associates with protein-RNA correlation in human cancer. Nat Commun. Nature Publishing Group; 2021;12:5732.

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

41. Taberlay PC, Achinger-Kawecka J, Lun ATL, Buske FA, Sabir K, Gould CM, et al. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. Genome Res. 2016;26:719–31.

42. Taberlay PC, Statham AL, Kelly TK, Clark SJ, Jones PA. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. Genome Res. 2014;24:1421–32.

43. Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, et al. Punctuated evolution of prostate cancer genomes. Cell. 2013;153:666–77.

44. Seim I, Jeffery PL, Thomas PB, Nelson CC, Chopin LK. Whole-Genome Sequence of the Metastatic PC3 and LNCaP Human Prostate Cancer Cell Lines. G3 Bethesda Md. 2017;7:1731–41.

45. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A. 2010;107:21931–6.

46. Hiratani I, Ryba T, Itoh M, Rathjen J, Kulik M, Papp B, et al. Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. Genome Res. 2010;20:155–69.

47. Yehuda Y, Blumenfeld B, Mayorek N, Makedonski K, Vardi O, Cohen-Daniel L, et al. Germline DNA replication timing shapes mammalian genome composition. Nucleic Acids Res. 2018;46:8299–310.

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

49. Cochran K, Srivastava D, Shrikumar A, Balsubramani A, Hardison RC, Kundaje A, et al. Domain-adaptive neural networks improve cross-species prediction of transcription factor binding. Genome Res. 2022;32:512–23.

50. Holmquist GP. Role of replication time in the control of tissue-specific gene expression. Am J Hum Genet. 1987;40:151–73.

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

52. Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics. 2005;21:650–9.

53. Kryuchkova-Mostacci N, Robinson-Rechavi M. A benchmark of gene expression tissue-specificity metrics. Brief Bioinform. 2017;18:205–14.

54. Domazet-Lošo T, Tautz D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature. Nature Publishing Group; 2010;468:815–8.

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

56. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet. 2015;47:435–44.

57. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. ArXiv170507874 Cs Stat [Internet]. 2017 [cited 2021 Sep 1]; Available from: http://arxiv.org/abs/1705.07874

58. De S, Babu MM. A time-invariant principle of genome evolution. Proc Natl Acad Sci U S A. 2010;107:13004–9.

59. Ryba T, Hiratani I, Lu J, Itoh M, Kulik M, Zhang J, et al. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Res. 2010;20:761–70.

60. Siefert JC, Georgescu C, Wren JD, Koren A, Sansam CL. DNA replication timing during development anticipates transcriptional programs and parallels enhancer activation. Genome Res. 2017;27:1406–16.

61. Rivera-Mulia JC, Buckley Q, Sasaki T, Zimmerman J, Didier RA, Nazor K, et al. Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells. Genome Res. 2015;25:1091–103.

62. R Core Team. R: A language and environment for statistical computing. Vienna, Austria.: R Foundation for Statistical Computing; 2020.

63. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling Somatic SNVs and Indels with Mutect2 [Internet]. Bioinformatics; 2019 Dec. Available from: http://biorxiv.org/lookup/doi/10.1101/861054

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

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

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

67. Smit AFA, Hubley R, Green P. RepeatMasker [Internet]. 2013. Available from: http://www.repeatmasker.org

68. Zabidi MA, Arnold CD, Schernhuber K, Pagani M, Rath M, Frank O, et al. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. Nature. 2015;518:556–9.

69. Lubelsky Y, Prinz JA, DeNapoli L, Li Y, Belsky JA, MacAlpine DM. DNA replication and transcription programs respond to the same chromatin cues. Genome Res. 2014;24:1102–14.

70. Umer HM, Smolinska-Garbulowska K, Marzouka N, Khaliq Z, Wadelius C, Komorowski J. funMotifs: Tissue-specific transcription factor motifs [Internet]. bioRxiv; 2019 [cited 2022 Apr 27]. p. 683722. Available from: https://www.biorxiv.org/content/10.1101/683722v1

71. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.

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

73. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinforma Oxf Engl. 2010;26:841–2.