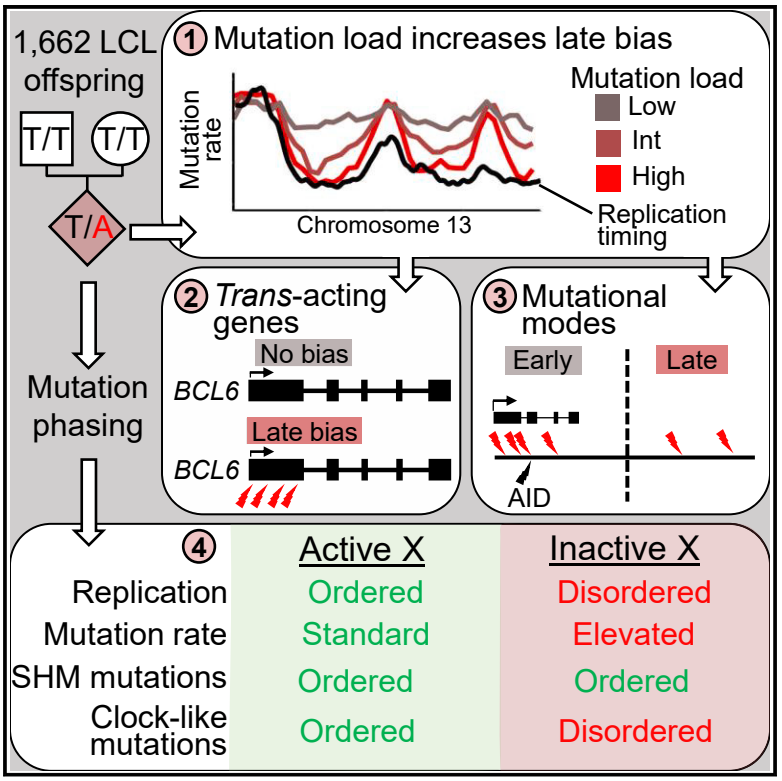
Cell Genomics

Article

The landscape of somatic mutations in lymphoblastoid cell lines

Graphical abstract



Highlights

- Analysis of 885,655 mutations from 1,662 lymphoblastoid cell lines (LCLs)
- Inter-individual variation in the rate and genomic distribution of mutations
- *BCL6* is a candidate modulator of the mutational landscape in LCLs
- Hypermutation of the inactive X chromosome is attributed to DNA polymerase η

Authors

Madison Caballero, Amnon Koren

Correspondence

koren@cornell.edu

In brief

An analysis of the mutational landscape in 1,662 individuals reveals genome-wide variation in mutational loads, genomic distribution, and signatures, all of which appear to be modulated by somatic mutations in *trans*. The inactive X chromosome is unusual in bearing an excess of replication-timing-uncoupled DNA polymerase η-mediated mutations.

Caballero & Koren, 2023, Cell Genomics *3*, 100305
June 14, 2023
https://doi.org/10.1016/j.xgen.2023.100305

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Article

The landscape of somatic mutations in lymphoblastoid cell lines

Madison Caballero[1] and Amnon Koren[1,2,*]
[1]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA
[2]Lead contact
*Correspondence: koren@cornell.edu
https://doi.org/10.1016/j.xgen.2023.100305

SUMMARY

Somatic mutations have important biological ramifications while exerting substantial rate, type, and genomic location heterogeneity. Yet, their sporadic occurrence makes them difficult to study at scale and across individuals. Lymphoblastoid cell lines (LCLs), a model system for human population and functional genomics, harbor large numbers of somatic mutations and have been extensively genotyped. By comparing 1,662 LCLs, we report that the mutational landscape of the genome varies across individuals in terms of the number of mutations, their genomic locations, and their spectra; this variation may itself be modulated by somatic *trans*-acting mutations. Mutations attributed to the translesion DNA polymerase η follow two different modes of formation, with one mode accounting for the hypermutability of the inactive X chromosome. Nonetheless, the distribution of mutations along the inactive X chromosome appears to follow an epigenetic memory of the active form.

INTRODUCTION

Lymphoblastoid cell lines (LCLs) are Epstein-Barr virus (EBV)-transformed peripheral B cells[1,2] that have served as a model for studies in cell biology. Unlike immortalized cell lines derived from tumors, LCLs are generated from untransformed somatic tissue but maintain active proliferation in culture. Straightforward to generate and maintain, LCLs have been made widely available for thousands of individuals and used for major population genetics projects like the HapMap[3] and 1000 Genomes (1kGP)[4] and for studying gene expression, chromatin structure, cytotoxicity, and DNA replication and repair.[2] More specifically, the proliferative nature and karyotypic stability of LCLs make them ideal for studying DNA replication timing—the spatiotemporal pattern of genome replication along S phase.[5–7] In turn, DNA replication timing strongly and specifically correlates with the rates of single-nucleotide mutations, which are more abundant in late-replicating regions.[6,8–12] This relationship is apparent for mutations in LCLs,[6,13] which are comprised of between 40 and 100 germline mutations derived from the donor's parents[14] and a variable number of somatic mutations accumulated in the B cell sample lineage in relation to donor age[15,16] or acquired *in vitro* after transformation in relation to cell line passage number.[17–19] The availability of LCL lines from families provides a particularly effective means of genotyping mutations by comparing the genome sequences of parents and offspring.[6,13]

A mutational pathway active specifically in B cells is somatic hypermutation (SHM), which targets the immunoglobulin (IG) genes to increase antibody diversity.[20,21] SHM initiates with the deamination of cytosine to deoxyuracil via activation-induced cytidine deaminase (AID) at sequence motifs such as WRC(Y)/(R)GYW.[22,23] Repair of deoxyuracils then often involves DNA synthesis by the low fidelity translesion DNA polymerase η, leading to mutations.[24] Polymerase η synthesis can extend to nearby nucleotides to produce proximal A>G/C substitutions with a context preference of 3′ A/T.[25,26] Importantly, SHM is also associated with an elevated mutation rate outside of targeted loci. Such off-target mutations can be initiated by AID itself, which has been shown to target up to 275 highly transcribed hotspot genes.[22,27] However, a recent study[15] suggested that replicative and/or oxidative stress in highly proliferating B cells, together with high expression of polymerase η, leads to off-target mutations in gene-poor, late-replicating genomic regions independent of AID.

Here, we used LCLs to address several important aspects of somatic mutations that have been difficult to address using primary cells. Specifically, the large number of mutations per sample and the availability of LCLs for many families allow a detailed analysis of the somatic mutational landscape and its variation among individuals. These, in turn, can be readily compared with DNA replication timing profiles inferred at high resolution from the very same samples. Last, the availability of parental genome sequences enables phasing of mutations to parental alleles; we previously used LCL family trios in this capacity to demonstrate that the inactive X chromosome replicates its DNA very late and without any discernable spatial pattern.[28] We thus generated a catalog of LCL mutations in 1,662 individuals. Using these data, we identified variation in global mutation load, mostly involving mutations attributed to polymerase η, as an important factor associated with the extent of mutation bias to late-replicating genomic regions; we map this variation to several *trans*-acting genes, in particular *BCL6*, a lymphoid

Table 1. Mutation data sources

Mutation source		Number of offspring or samples	Platform	Approximate coverage	Original genome version	Mutation calling method
LCL	iHART; Ruzzo et al.[32]	1,028	HiSeq X (2 × 150)	35×	hg19	parent-offspring
	1kGP; Byrska-Bishop et al.[33]	602	NovaSeq 6000 (2 × 150)	30×	hg38	parent-offspring
	repeat expansion; Dolzhenko et al.[34]	9	HiSeq X (2 × 150)	30×	hg19	parent-offspring
	Illumina Platinum; Eberle et al.[35]	13	HiSeq 2000 (2 × 100)	50×	hg19	parent-offspring
	Caballero et al. 2022[7]	12	HiSeq X (2 × 150)	15×	hg38	parent-offspring
	Polaris 1000 Genomes Project Consortium et al.[4]	49	HiSeq X (2 × 150)	30×	hg19	parent-offspring
CLL	CLLE-ES, ICGC	151	HiSeq[a]	N/A	hg19	tumor-normal

[a]Further sequencing platform details could not be ascertained.

cancer driver and previously suggested hotspot of AID off-target mutagenesis.[29] A subset of mutations occurred in clusters, which had a different genomic distribution than non-clustered mutations. The inactive X chromosome was subject to hypermutation, likely attributed to polymerase η; unexpectedly, these mutations did not conform to the expected pattern suggested by the replication dynamics or chromatin structure of the inactive X, suggesting the existence of epigenetic memory that influences the mutation landscape.

RESULTS

A catalog of somatic mutations in LCLs

We called LCL mutations by identifying Mendelian errors in parent-offspring allelic inheritance in the genome sequences of 1,662 individuals and their parents using data from six sequencing cohorts (Table 1; Table S1). While cytogenetically normal, 846 samples were generated from donors with autism spectrum disorder, six from unaffected carriers of fragile X syndrome, one from an affected fragile X syndrome patient, one from an affected Friedreich ataxia patient, and two from affected ataxia-telangiectasia patients. None of these samples had global replication timing alterations compared with healthy individuals.[7] We called 885,655 autosomal single-nucleotide variant (SNV) mutations, ranging from 66–8,737 per offspring (median, 408; 0.169 mutations/Mb), consistent with other quantifications of somatic mutations in B cells[15,16] (Figure 1A; Figure S1A). We observed two prominent modes and a long tail of mutation count across offspring; this is consistent with previous LCL mutation calling[19] and likely reflects donor age at blood donation and cell line passage number (neither of which are known for the majority of our samples). Only 0.73% of mutations were functional as predicted by a SNPeff[30] (4.3t) high or moderate variant impact score. Using monozygotic twins, we estimated the fraction of misidentified parental variants as less than 9.66% (STAR Methods; Figures S1B–S1E). Additionally, we used replicate sequencing of 51 samples to estimate the rate of genotyping errors. A median of 93.1% of mutation calls were reproduced in samples resequenced once, while 99.8% of mutations were repeated at least once in a sample resequenced five separate times (Figure S1F; Table S1). To compare LCL mutations with DNA replication timing, we used the same whole-genome sequencing of the 1,662 offspring to infer replication timing profiles from read depth fluctuations along chromosomes.[5,31] In this approach, replication timing is inferred from DNA copy number because early-replicating regions have greater read depth in a population of proliferating cells than late-replicating regions. We merged the data for all cell lines to create a single median "consensus" LCL replication profile used for downstream analyses. This consensus was similar to individual LCL replication timing profiles (Figure 1B; Figure S1H), was highly correlated with LCL replication timing profiles generated by S/G1 sequencing[6] (Pearson's $r = 0.94$) and controls for differences between sequencing cohorts and the more subtle differences between individuals.[5] Low variation among samples in raw DNA copy number fluctuations (Figure S1I) suggested that different samples had similar fractions of actively replicating cells (and differences among samples did not correlate with individual mutation load; $p = 0.82$).

To complement the analysis of LCLs, we incorporated mutations derived from 151 chronic lymphocytic leukemia (CLL) patients. CLL is a malignancy of B cells, neutral to EBV infection,[36,37] and has been studied in depth at the genomic level.[38] CLL comprises two subtypes that differ by the mutational status of the IG heavy chain (*IGHV*) gene; tumors with a mutated *IGHV* (CLL-M) have undergone SHM, while others have an unmutated *IGHV* (CLL-U).[39] Its corollaries to LCL biology, high mutation rate, and the availability of data from many individuals makes CLL a useful comparator with mutations in LCL. Tumor-normal mutation calling and filtering identified 377,605 autosomal mutations with a median of 2,368 mutations per patient (0.98 mutations/Mb; range, 221–5,629; Figure 1A). Generating a CLL replication timing profile either through S/G1 or whole-genome sequencing methods was not feasible because circulating malignant cells are not proliferative[40] and contain many copy number alterations[41,42]; instead, and because replication timing is conserved between closely related cell types,[43,44] we used LCL replication timing to compare with CLL mutations. Indeed, mutation rates in LCL and CLL were highly correlated with LCL replication timing and strongly enriched in late-replicating regions (Figure 1C). In LCL, we confirmed this relationship independently in the two largest population cohorts (Figures S1J and

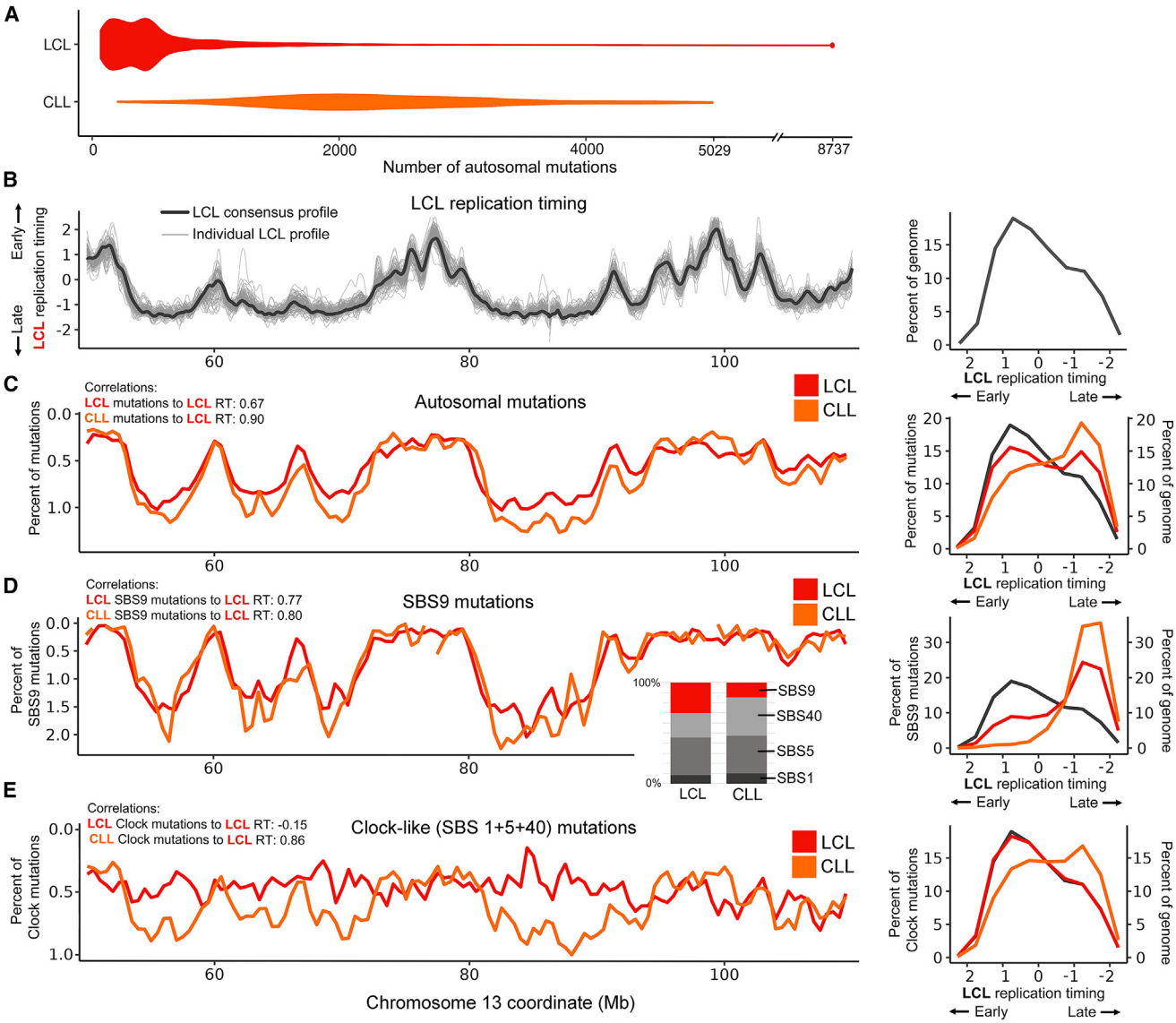


Figure 1. Mutations in LCL and CLL are biased to late replication

(A) The number of autosomal mutations in 1,662 LCLs and 151 CLL tumors.

(B) Left: consensus LCL replication timing profile and 100 individual profiles randomly sampled from 1kGP. Right: distribution of autosomal replication timing values for the LCL consensus profile in 10 bins.

(C) Left: rates of LCL and CLL mutations in 1-Mb sliding windows with a 0.5-Mb step. Correlation values in all panels are Pearson's correlation coefficients for chromosome 13. Right: distribution of autosomal mutations in 10 replication timing bins.

(D and E) As in (C) for LCL and CLL mutations attributed to SBS9 (D) or the clock-like pathway (E). Bar plots: proportions of individual SBS signatures that explain LCL and CLL mutations.

S1K) and in individual samples (Figure S1L). Notably, CLL demonstrated higher correlations with the LCL replication timing profile and greater late replication bias (Figure 1C), suggesting a different mutational landscape in CLL compared with LCL.

To determine which mutational processes were active in LCLs and CLL, we used trinucleotide mutational signature analysis,[45] specifically COSMIC v.3.2 SBS (single base substitution) signatures. To prevent overfitting, we selected a subset of signatures based on biologically expected mutational pathways. In CLL, SBS1, SBS5, SBS9, and SBS40 are established as the predominant mutational signatures.[16,45–47] SBS1, SBS5, and SBS40, are clock-like signatures—highly ubiquitous signatures of unknown etiology that increase in abundance with age.[45,48] The proposed etiology of SBS9 is mutations induced by polymerase η as part of SHM in lymphoid cells.[15,16,45,47] While, to our knowledge, mutational signature analysis has not been performed in LCLs before, we found that the same signatures (SBS1, SBS5, SBS9, and SBS40) best explained LCL mutations with a cosine similarity of 0.96 (compared with 0.97 for

CLL). It is established that SHM is ongoing after EBV transformation in LCLs.[36,49] We also tested for the involvement of SBS8 and SBS18[15] in LCL mutagenesis but only called 1.8% of mutations attributed to SBS18 in CLL and none in LCL, and 6% of mutations as SBS8 in both cell types. Principal component analysis of substitution type frequencies showed an association between mutation spectrum and mutation load but not donor ethnicity (Figure S1M). Notably in this respect, Ng et al.[19] investigated the association between ethnicity and mutation load but found cell culture age to confound the interpretation of this association.

SBS9 was relatively more prevalent in LCL (30.0% ± 0.12% of mutations) than in CLL (14.8% ± 0.15%) (Figure 1D). In both cell types, SBS9 mutations were more abundant in late-replicating regions and closely followed the replication timing profile (after controlling for sequence composition in replication timing bins; Figure 1D). CLL showed greater bias to late replication than LCL despite SBS9 comprising a smaller proportion of mutations. For clock-like mutations, late replication bias and correlation to replication timing were only apparent in CLL (Figure 1E). LCL clock-like mutations were more uniformly distributed and not significantly correlated with replication timing ($p = 0.20$). Taken together, LCLs and CLL share similar mutational pathways but with apparently different proportions and distributions with respect to replication timing.

The mutational landscape varies across individuals in association with global mutation load and SBS9 abundance

Having demonstrated variability in mutation rates, types, and relation to replication timing, we sought to identify additional factors that differ between and within LCL and CLL that could account for such heterogeneity. A major difference between these two cell types is the elevated mutation load (or mutation burden) of CLL as defined by the total number of autosomal mutations per sample. We thus asked whether mutation load itself relates to the replication timing distribution and the spectrum of mutations. To test this, we began by dividing LCLs into three similarly sized groups with ~295,500 mutations each (Figure S2A). A "low-mutation-load" group contained 489 mutations or less per offspring (1,066 offspring); a "high-mutation-load" group had 1,104 or more mutations per offspring (174 offspring); and an "intermediate-mutation-load" group contained the remaining 422 offspring. The relationship of mutation rate to replication timing was substantially more pronounced in the high-mutation-load group, with 4.17-fold more mutations in the latest replicating fraction than the earliest and a greater correlation with LCL replication timing (Figure 2A). In comparison, the intermediate-mutation-load group showed a weaker increase with 1.85-fold more mutations in the latest fraction, while the low-mutation-load group did not show any enrichment for mutations in late-replicating parts of the genome (0.98-fold difference). This was not attributed to statistical power because all groups had a similar number of mutations analyzed. This pattern was also evident for individual offspring, where a greater mutation load corresponded to consistently later replication timing bias, including when offspring were down-sampled to 80 mutations to control for possible power differences among samples (Figures 3A and 3B; Figure S2C).

To test whether these differences between mutation load groups were related to particular mutational signatures, we fit SBS9 and clock-like mutational signatures to the stratified LCL mutation load groups. The proportion of mutations attributed to SBS9 decreased from 43.46% ± 0.22% of mutations in the high-mutation-load group to 25.74% ± 0.19% and down to 21.01% ± 0.18% in the intermediate- and low-mutation-load groups, respectively. This trend was also observable in individual samples as the proportion of SBS9 correlated with mutation load (Pearson's $r = 0.34$, $p < 1 \times 10^{-16}$). Therefore, a high global mutation count in LCLs corresponded to increased SBS9 abundance. With respect to replication timing, the high-mutation-load group showed the greatest enrichment in late-replicating regions for SBS9 and the clock-like category, with 15.1-fold and 1.57-fold more mutations in the latest-replicating fraction compared with the earliest, respectively (Figures 2B and 2C; Figure S2D). This relationship was less pronounced in the intermediate-mutation-load group, with a 4.69-fold increase in SBS9 abundance and a 1.22-fold increase in clock-like abundance. The low-mutation-load group showed enrichment for neither SBS9 nor clock-like mutations in late-replicating regions (Figures 2B and 2C). Together, these findings indicate that the distribution of mutations, most prominently of SBS9 origin, varies in LCLs in accordance with mutation load.

CLL provided an opportunity to further investigate how mutation load and signatures shape the mutational landscape. Because the *IGHV* mutation status of individuals was unreported, we devised a way to use mutational signature analysis as an alternative means of inferring SHM activity and, thus, CLL subtype. Accordingly, we fit CLL mutational signatures to autosomal mutations in individual samples. We assigned 80 samples with a consistent greater than 2% SBS9 contribution (based on the range of 1,000 bootstrap samples) as CLL-M and another 68 samples with 0% SBS9 contribution as CLL-U (Figure S2F). Three remaining samples were ambiguous and not analyzed further. The CLL-M group contained a median of 2,620 mutations per sample (216,451 total mutations; Figure S2G), while the CLL-U group contained a median of 1,986 mutations per sample (138,113 total). This was a significant difference in mutation burden between the two subtypes (two-tailed t test: $p = 1.63 \times 10^{-5}$). In CLL-M samples, a median of 25.4% ± 0.04% of all mutations (591 mutations per sample) were contributed by SBS9, which can fully account for their increased mutation count.

Mutations in CLL-M and CLL-U samples showed exponential-like increases with replication timing (Figure 2D). This effect was slightly stronger in CLL-M (5.54-fold more mutations in the latest replicating fraction than the earliest) than in CLL-U (4.05-fold). More specifically, in CLL-M, as in LCLs, SBS9 contribution was greatly enriched in late-replicating regions, with 18.9-fold more mutations in the latest replicating fraction than the earliest (Figure 2B; Figure S2E). For clock-like mutations, CLL-M and CLL-U showed similar replication timing relationships with 3.32- and 3.69-fold more mutations, respectively, in the latest-replicating fraction than the earliest (Figure 2C).

Having CLL subdivided by *IGHV* mutation status, we then compared high and low mutation load. We divided CLL-M and CLL-U into two mutation load groups each. CLL-M samples with higher mutation loads (28 samples with ≥3,011 mutations) showed greater enrichment for all mutations in late-replicating

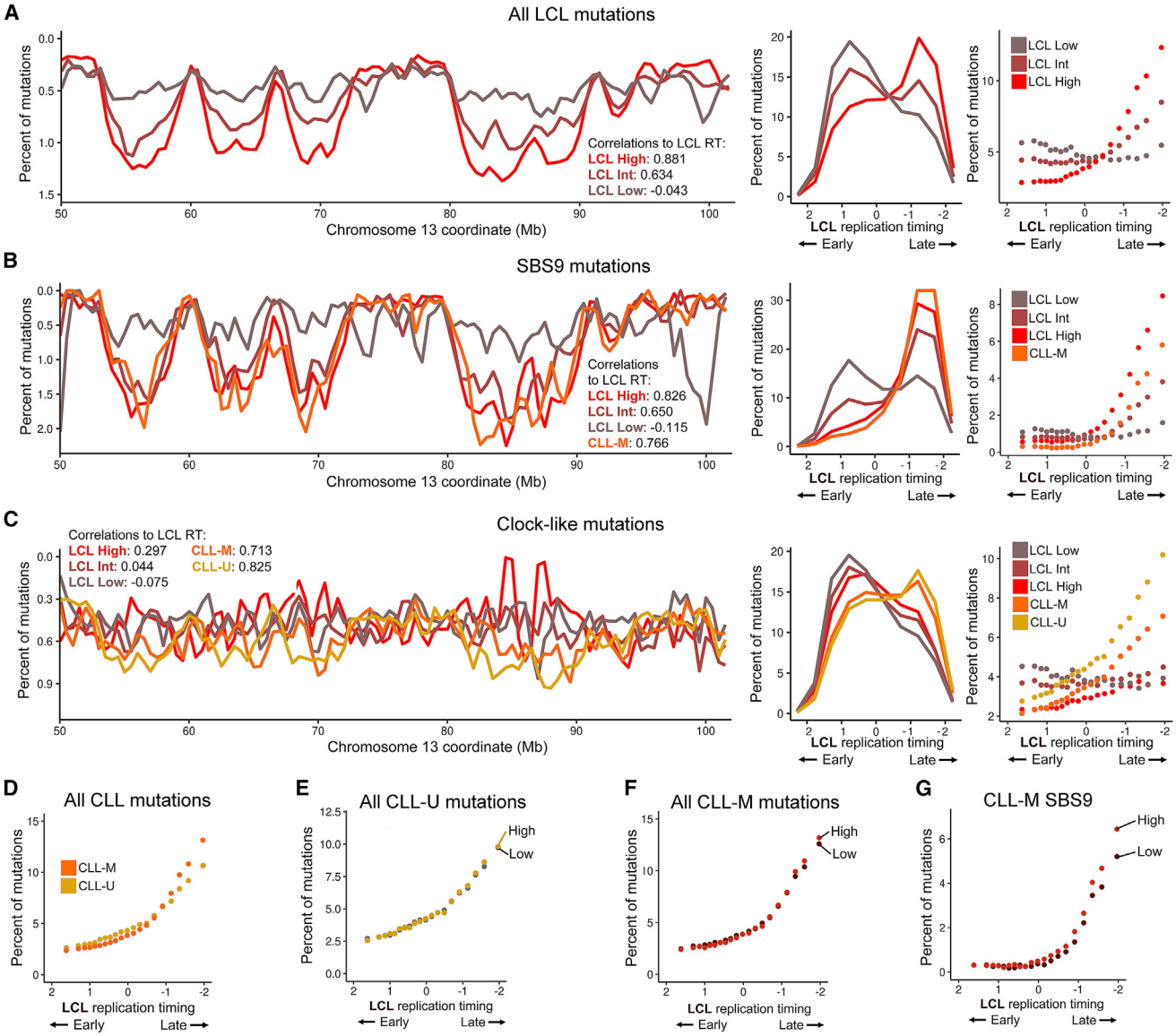


Figure 2. The association of mutation rates and signatures with DNA replication timing varies by mutation load

(A) Left: as in Figure 1C; rates of mutations in 1-Mb sliding windows with a 0.5-Mb step. Correlation values in all panels are Pearson's correlation coefficients. Center: as in Figure 1C; distribution of autosomal mutations compared with replication timing in the high, intermediate, and low LCL mutation load groups. Right: the relationship of autosomal mutation counts to replication timing in the high, intermediate, and low LCL mutation load groups in 20 bins of uniform genome content.

(B and C) As in (A) for mutations attributed to SBS9 (B) or the clock-like pathway (C).

(D) The relationship of autosomal mutation count to replication timing in CLL samples stratified by *IGHV* mutation status.

(E) The distribution of total autosomal mutations in CLL-U samples in high- and low-mutation-load groups.

(F and G) The distribution of total autosomal mutations (F) and mutations attributed to SBS9 (G) as a function of replication timing in the CLL-M high- and low-mutation-load groups.

regions (Figure 2F). Among CLL-M samples, higher mutation load corresponded to greater SBS9 contribution (20.6% ± 0.30% versus 25.24% ± 0.32%) and greater SBS9 enrichment in later-replicating regions (Figure 2G). CLL-U did not show a pronounced change in mutation enrichment in late-replicating regions based on mutation load (Figure 2E), likely because of the diminished mutation load variability. Thus, the distribution of SBS9 again varies with mutation load.

Taken together, we identified global mutation load as a cell-line-specific factor associated with the distribution and spectrum of mutations along the genome. In LCL and CLL-M, elevated mutation load was the product of SBS9 and clock-like mutations, although SBS9 was more prominent. Because we are underpowered to call mutational signature composition in individual samples, we cannot resolve whether mutation load is more directly linked to replication timing bias or

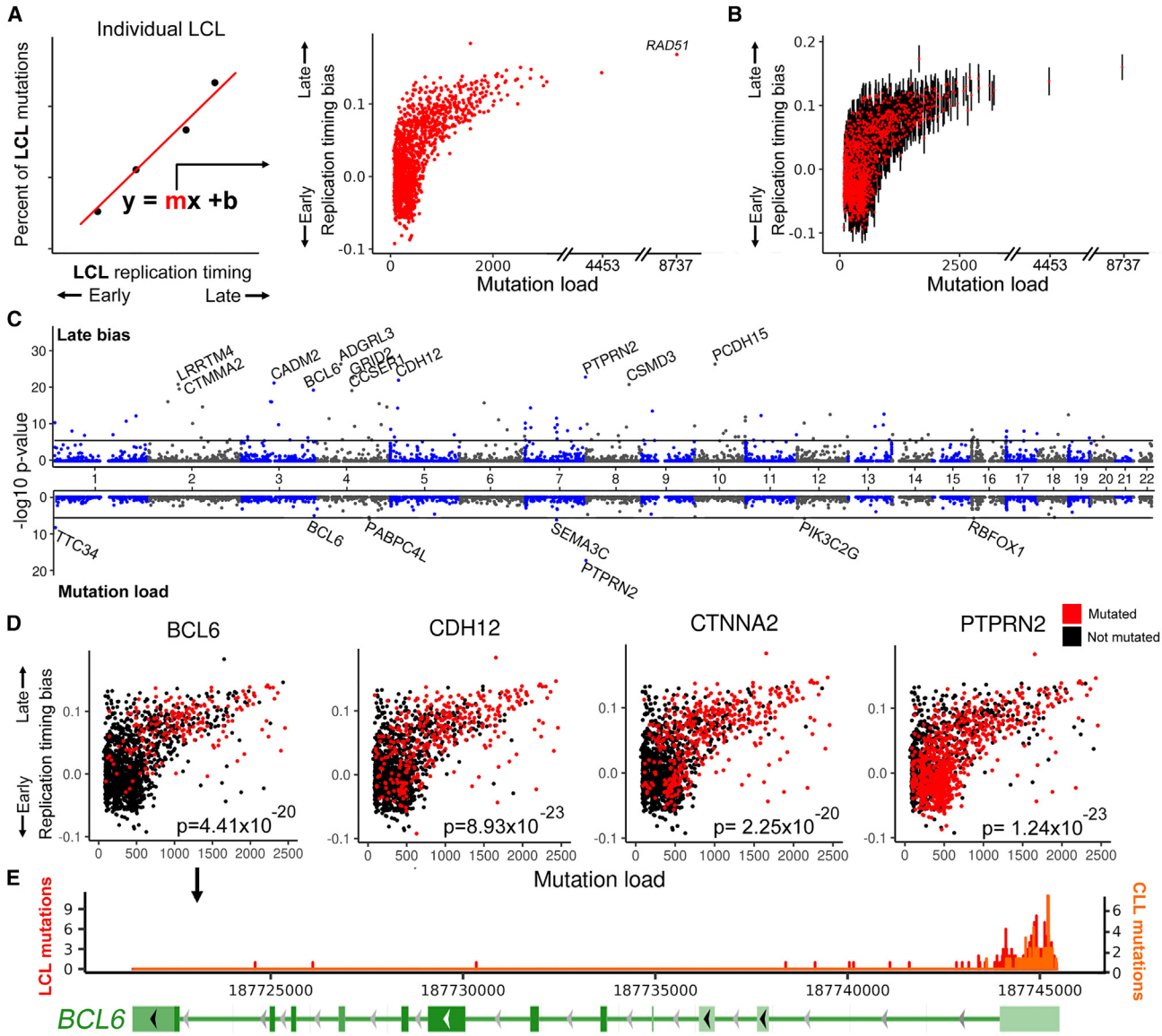


Figure 3. Genes associated with replication timing bias

(A) Left: replication timing bias in individual LCLs is calculated as the linear slope of mutation percentages in four replication timing bins. Right: relationship between replication timing bias and mutation load in individual LCLs.

(B) Down-sampling of individual LCL samples to 80 genome-wide mutations. Red dots: mean slope of 1,000 iterations of sampling for each mutation load. Error bars: standard deviation of sampling.

(C) Top: association of mutated gene frequency to late replication bias of individual samples corrected for mutation load. Black line: Bonferroni-corrected $p < 0.05$ divided by the number of tested genes. The 11 most significant genes are highlighted. Bottom: association of gene mutation frequency to mutation load (corrected for the effect of mutation load on mutation probability). The seven significant genes and *BCL6* are highlighted.

(D) Selected genes from (C), showing mutation status in individual LCLs.

(E) Location of CLL and LCL mutations in *BCL6* (NCBI: NM_001706.5).

to mutational spectra or how these two properties affect each other.

Variation in the mutational landscape is associated with *trans*-acting gene mutations

We next sought possible explanation for how the late replication bias of mutations, particularly SBS9, increases alongside mutation load. A previous study reported a weak correlation between EBV copy number and mutation load in the 1kGP cohort[19] (r = 0.17; $p = 2.32 \times 10^{-5}$); however, in our larger dataset, this correlation seemed weaker and mostly dependent on outlier samples (Figure S2B). Another hypothesis that can explain this association is the presence of (a) *trans*-acting factor(s) that modulates replication timing bias and is itself somatically mutated in some

samples, permitting higher mutation load and SBS9 late replication bias. Mutations affecting DNA repair genes were reported in the 1kGP cohort,[19] and we identified two additional samples with coding mutations in *MSH3*. These samples, however, did not appear to be outliers in mutation load nor replication timing bias, with the possible exception of one sample (HG02683; Figure 3A) with a mutation in *RAD51* that may explain its high mutation load (although this sample was not an outlier for late replication bias).

To more systematically search for potential *trans*-acting effectors of the mutational landscape, we associated LCL mutations at the level of genes with individuals' mutational late replication bias, normalizing for mutation load (STAR Methods; Figure 3A). We identified 97 candidates significantly associated with late replication bias, including several cancer risk genes, such as *CSMD3* and *CTNNA2* (Figures 3C and 3D; Table S1). Of particular interest was *BCL6* (B cell lymphoma 6), a transcription factor that promotes proliferation of B cells after the onset of SHM by repressing genes that would otherwise arrest the cell cycle as a result of elevated DNA damage.[50] *BCL6* is also the prototypical off-target SHM hotspot[29,51] (and the only hotspot gene among the candidates we identified).

We further associated gene mutation frequency with mutation load itself. Seven genes were significantly associated with mutation load, the strongest of which was *PTPRN2* (Figure 3C), a protein-tyrosine phosphatase associated with proliferation and cancer.[52] *PTPRN2* was also significantly associated with late replication bias (Figure 3C) and was mutated in 906 LCLs (Figure 3D), far more than its size or replication timing would predict. *BCL6* was also associated with mutation load ($p = 8.9 \times 10^{-6}$), but just below the corrected significance threshold.

Focusing on *BCL6*, we identified 345 mutations in the gene sequence among 192 LCLs. In the high-mutation-load group, *BCL6* mutations were found in 52.3% of samples compared with only 17.8% and 2.1% in the low- and intermediate-mutation-load group, respectively. This could not be explained by differences in sample mutation load per se because high-mutation-load samples had, on average, 6.1-fold more mutations than low-mutation-load samples, whereas *BCL6* mutations were 24.9-fold more common. *BCL6* mutations were similarly associated with replication timing bias in individual samples, present in 20.7% of the 906 samples with a late replication bias (score > 0 in Figure 3A) compared with only 5.7% among samples with early or no replication bias. This again could not be explained by the higher mutation numbers in samples with a late replication bias because the latter had, on average, 1.58-fold more mutations, whereas *BCL6* mutations were 3.63-fold more common. Mutations in *BCL6*, which is a driver of CLL,[53] were also found in 26.5% of CLL samples and were far more common in CLL-M (48.8% of samples) than in CLL-U (1.5%); the latter is consistent with *BCL6* mostly affecting SBS9 mutations.

Mutations that alter *BCL6*'s amino acid sequence were rare; only two were discovered in LCL and one in CLL. However, in LCL and CLL, *BCL6* mutations were highly enriched in the first exon, itself part of the gene's 5′ UTR (Figure 3E). This region of *BCL6* is the binding location for negative autoregulation,[54] suggesting that mutations in this region can block the downregulation of *BCL6*. An attractive possibility is that *BCL6* mutations arise in LCL culture and promote a higher mutation load as well as an altered mutational landscape manifesting in late replication bias and, in LCL and CLL-M, an enrichment of SBS9 mutations. Moreover, such mutations may be selected for during LCL culture, making *BCL6* an equivalent to *BCOR* (*BCL6* corepressor) mutations that are selected for in iPS (induced pluripotent stem) cell culture[55]; indeed, BCOR functions together with BCL6 to repress cell-cycle arrest in cells with active SHM.

Clustered SBS9 mutations are associated with AID activity in early-replicating regions

Consistent with a recent report,[15] the majority of SBS9 mutations in our datasets appeared to be unrelated to AID-induced C>N mutations and showed a different genomic distribution (late compared to early replication) than AID-dependent off-target SHM. However, other studies have described clusters of mutations in B lymphocyte cancers that are enriched in early-replicating regions, near promoters and enhancers of actively transcribed genes, and within 100 bp of C>N mutations.[56,57] These clusters were proposed to result from polymerase η activity near DNA lesions induced by AID or APOBEC (apolipoprotein B mRNA-editing enzyme, catalytic polypeptide) cytidine deaminases. Indeed, off-target AID-mediated mutations have been shown to localize preferentially to highly expressed genes through a proposed interaction with RNA polymerase.[20,24] We thus asked whether SBS9 mutation clusters distribute differently than non-clustered mutations across chromosomes and with regard to gene activity. We considered two or more SBS9-context mutations (A>G/C substitutions with a context preference of 3′ A/T) within 500 bp of each other as a cluster. We identified 26,759 such clusters in LCLs and 1,736 in CLL-M, encompassing 37.01% and 7.13% of total SBS9-context mutations, respectively. We simulated SBS9 mutation clustering in different genomic regions by considering sequence composition and the replication timing bias of SBS9 mutation (STAR Methods). Compared with the simulated expectation, mutation clustering in LCL and CLL-M was significantly elevated across all replication timing bins but, importantly, was relatively more abundant in early-replicating regions (Figures 4A and 4B). In contrast, non-clustered mutations were relatively more abundant in late-replicating regions (Figure 4B). Furthermore, clustered mutations in early-replicating regions were associated more than expected with nearby C>N mutations in the AID motif context (Figure 4C). In contrast, non-clustered mutations were far less enriched near AID-context mutations. Similarly, a greater proportion of clustered mutations in CLL-M were relatively closer to gene transcription start sites (TSSs) than non-clustered mutations (Figure 4D). Clustered mutations were enriched downstream of the TSS of highly expressed genes in CLL-M, although no such relationship was apparent in LCL nor for non-clustered mutations (Figures 4D, 4E, and 4G). Of previously described AID hotspot genes,[58] TSS proximity was observed for a subset. *BCL6* was the most mutated hotspot gene, comprising 15.4% of the LCL and 46.2% of the CLL-M clustered mutations. TSS association was, however, also observed independent of *BCL6* (Figure 4F).

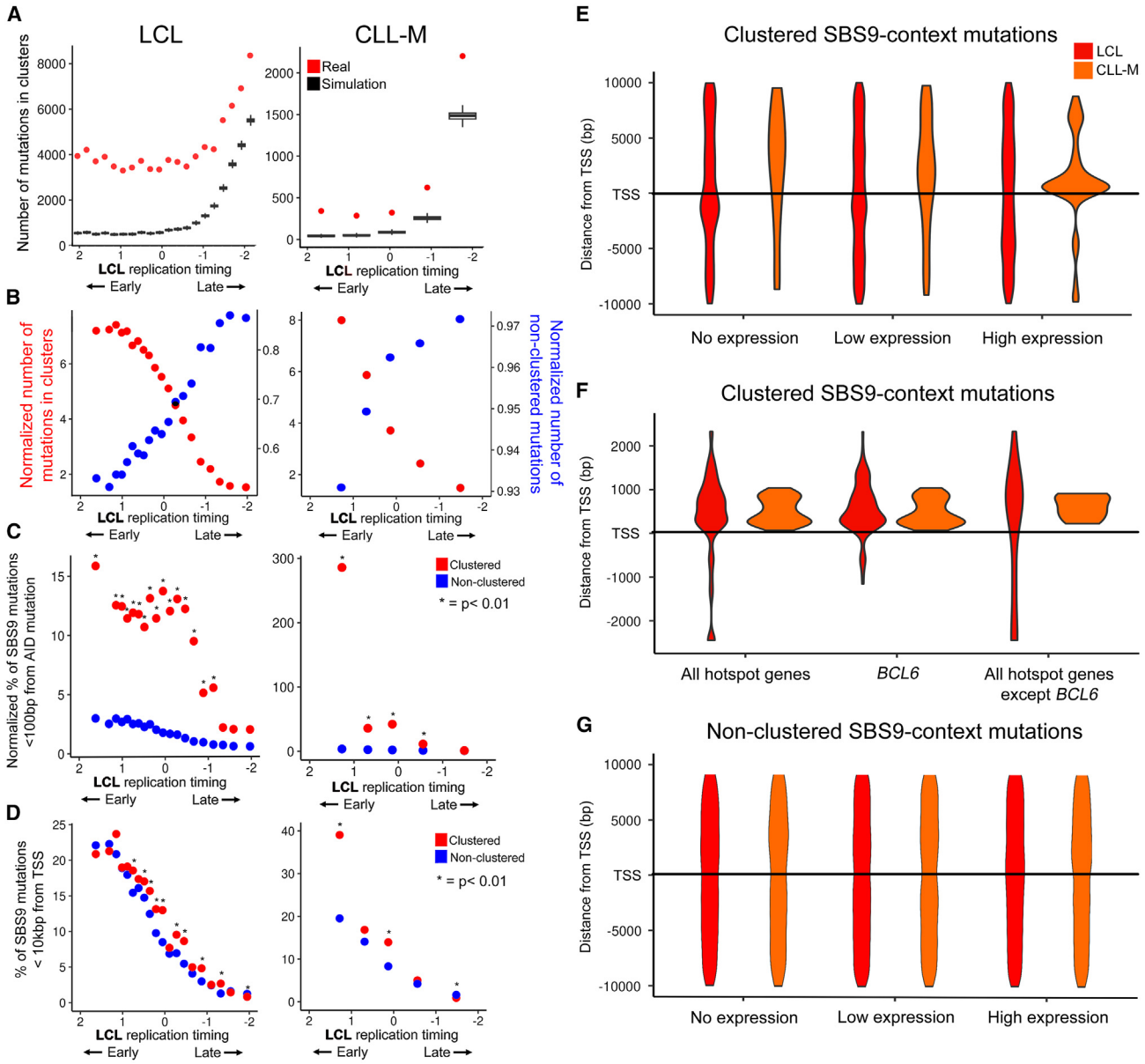


Figure 4. Two modes of SBS9-context mutations

(A) The number of clustered SBS9-context mutations in LCL and CLL-M compared with 1,000 iterations of sampled and clustered SBS9-context motifs in the genome equal to the number of SBS9-context mutations in 20 (for LCL) or five (for CLL-M) replication timing bins.

(B) The number of LCL and CLL-M clustered/non-clustered SBS9-context mutations normalized by the simulated mean number of clustered/non-clustering mutations within replication timing bins as calculated in (A).

(C) The percent of LCL and CLL-M SBS9-context mutations within 100 bp of an AID-context mutation in replication timing bins; values are normalized by a simulated mean percentage based on SBS9 and AID-context mutations rate and motif availability in replication timing bins. p values represent clustered versus non-clustered mutations from a Fisher's exact test of normalized values and the number of clustered/non-clustered mutations per bin.

(D) The percent of clustered/non-clustered LCL and CLL-M SBS9-context mutations within 10 kb of a TSS. p values are calculated with Fisher's exact test.

(E–G) Frequency of 1,044 LCL and 91 CLL clustered (E and F) and non-clustered (G) SBS9-context mutations around all protein-coding gene TSSs (E and G) or the subset (91 genes in LCL and 18 in CLL) of 275 off-target SHM hotspots (F).

Taken together, there appear to be two mutational modes of SBS9: early-replicating clustered mutations that are associated with AID-context C>N mutations and, at least partially, with active gene transcription; and late-replicating non-clustered mutations that are likely AID independent, as previously suggested.[15] Of note, the majority of SBS9 mutations in LCL and CLL-M conforming to the latter mode suggests that AID-independent mutagenesis is more prevalent for off-target mutations in B cells. Last, differing activities of these two mutational modes in different cell types may explain why

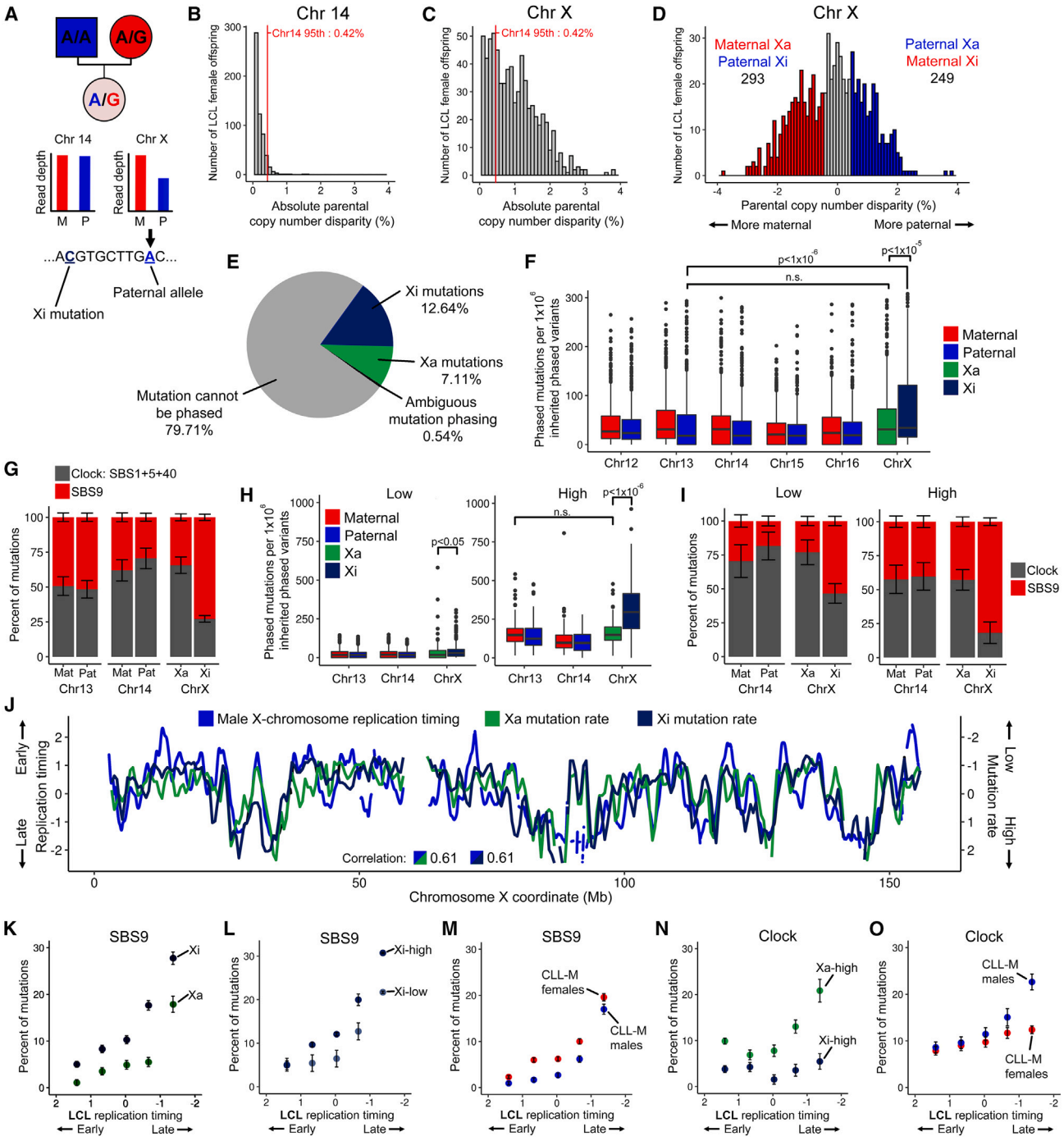


Figure 5. Unique mutational processes on the Xi chromosome

(A) Identification of Xi parental identity and mutation phasing.

(B) The absolute parental read depth disparity in LCL female offspring on chromosome 14. Disparity was calculated as the absolute difference of paternal and maternal median read depth of inherited phaseable variants divided by their combined median depth.

(C) The elevated absolute parental read depth disparity on the X chromosome in female LCL offspring. Xi was identified in females with a disparity greater than the 95th percentile value from chromosome 14.

(D) Xi parental identity classification among females with an identifiable Xi as described in (C). Xi is the parental homolog with the lower read depth.

(E) The number of phased X chromosome mutations in females with an identifiable Xi.

(legend continued on next page)

SBS9 is more prevalent in late-replicating regions in CLL than in LCL.

Hypermutation of the inactive X chromosome is the result of SBS9

We described above multiple interacting factors that shape the mutational landscape, including DNA replication timing, mutational processes, mutation clustering, and cell line-specific mutation load. As a case in point, we examined these factors from the perspective of chromosome inactivation. The female inactive X chromosome (Xi) replicates late in S phase with no discernable replication timing pattern,[28] which is distinct from the active X chromosome (Xa), the male X chromosome, and autosomes. This, and the tight link between replication dynamics and the mutational landscape, led us to predict that Xi would also have unusual mutational properties. Consistently, Xi has been inferred to be hyper-mutated in cancers.[59,60] In our female LCL and CLL samples, we also found that the X chromosome demonstrated significantly higher mutation rate than autosomes (Figures S3A and S3B). Interestingly, the female X chromosome also showed a greater abundance of SBS9 compared with autosomes (Figures S3C and S3D).

The family-based configuration of the LCLs enabled phasing mutations and separate investigation of the mutational landscapes of Xa and Xi, in contrast to previous studies that investigated Xi mutations by male-female comparisons or with limited expression-phased mutations.[59,60] Xi has been shown to be clonally propagated[61–63] and, thus, expected to be detectable in at least a subset of the 746 female LCL offspring. While phasing inherited variants enables discriminating parental chromosome pairs, functional data are required to identify the Xi chromosome. To this end, we devised an approach using the replication timing data itself, as inferred from sequencing read depth; because of its later replication, Xi is expected to demonstrate a lower median copy number compared with Xa (Figure 5A). Indeed, female X chromosomes showed greater parental copy number disparity than autosomes, which we used as a benchmark for assigning X chromosome identity (specifically, for samples with greater than the 95th percentile disparity on chromosome 14, the autosome with the closest number of phaseable inherited variants to the X chromosome; Figures 5B and 5C). This approach yielded reproducible Xi assignments in all 17 replicate sequenced offspring for which assignments could be made. In addition, paternal Xi identity for NA12878 was consistent with RNA expression analyses[64,65] and with our previous classification.[28] Thus, the Xi chromosome can be identified, and mutations it harbors can be called, from the genome sequence data itself. Accordingly, we identified the Xi in 542 of 746 female offspring (72.65%), of which 293 were paternally X inactivated, and 249 were maternally X inactivated (Figure 5D; Table S1).

Being able to phase the X chromosomes across cell lines, we systematically quantified how mutation rate and processes differed between Xa and Xi. We phased mutations by identifying mutant alleles on the same sequencing read or mate pair as a phaseable inherited variant (Figure 5A). Among 542 females with an identifiable Xi, we phased 6,005 (19.75%) X chromosome mutations, of which 3,844 (64.01%) were assigned to Xi (Figure 5E). We confirmed that the mutation rate of Xi was 1.78-fold higher ($p < 1 \times 10^{-5}$) than that of Xa and significantly higher than any autosome ($p < 1 \times 10^{-6}$) (Figure 5F; Figure S3E); the mutation rate of Xa was not significantly different than autosomes (Figure 5F). With regard to mutational processes, the proportions of mutations explained by SBS9 (34.36% ± 2.49%) and the clock-like mutational category (65.64% ± 5.94%) were similar between Xa and autosomes (Figure 5G; Figure S3F). On Xi, however, only 27.16% ± 2.38% of mutations were attributable to the clock-like category, while 72.84% ± 2.27% were attributable to SBS9 (Figure 5G). The elevated mutation rate on Xi can thus be predominantly attributed to SBS9.

Given our observation that mutation load relates to SBS9 enrichment in late-replicating regions, we hypothesized that increased mutation load in a cell line would correspond to a disproportionately greater Xi mutation rate and SBS9 abundance. We split the 542 LCL offspring with an identifiable Xi into a low-mutation-load group with fewer than 832 autosomal mutations (433 offspring) and a high-mutation-load group (remaining 109 offspring). Each group contained ~157,000 autosomal mutations. As predicted, X chromosome mutations were proportionally more abundant in the high-mutation-load group, comprising 11.10% of mutations compared with 8.25% in the low-mutation-load group. Using phased mutations, we further found that 67.33% of X chromosome mutations in the high-mutation-load group were located on Xi compared with 58.14% in the low group (Figure 5H). As a control, Xa showed the same mutation rate as autosomes in both groups (Figure 5H). This confirms that Xi has an elevated mutation load. As further hypothesized, we found that SBS9 on Xi was strongly elevated in the high-mutation-load group, at 81.72% ± 2.71% of Xi mutations compared with 53.37% ± 3.44% in the low-mutation-rate group (Figure 5I). In addition, SBS9 on Xi was higher than on Xa, comprising 38.92% more mutations in the high load group compared with 30.33% in the low group. Taken together, X

(F) Xa and Xi mutation rate compared with maternal and paternal homologous autosomes with the most similar number of inherited phaseable variants to chromosome X. Mutation rate was calculated as the number of phased mutations normalized by the number of inherited phaseable variants on each chromosome homolog pair. p values were calculated from a two-tailed t test.
(G) Proportions of mutational pathways on maternal and paternal homologous autosomes and Xa/Xi.
(H) As in (F), the mutation rate of phased mutations in high and low autosomal mutation load groups.
(I) As in (G), the proportions of mutational pathways in high and low autosomal mutation load groups.
(J) Pearson correlations of Xa and Xi regional mutation rate (calculated as in Figure 1K and further normalized by the number of inherited phaseable sites in each window) to male X chromosome replication timing.
(K–O) Abundance of mutational pathways on the X chromosome in five replication timing bins: SBS9 abundance for Xa/Xi mutations (K), Xi mutations in the high and low autosomal mutation load groups (L), CLL-M male and female patients (M) and clock-like mutation abundance for Xa/Xi mutations in the high autosomal mutation load groups (N) and CLL-M male and female patients (O).
In all panels, error bars represent the standard error of signature fit.

chromosome inactivation is associated with an elevated mutation load driven by SBS9, creating a distinct mutational landscape on Xi. This disparity of mutation load and SBS9 composition relative to Xa is particularly pronounced in cell lines with a greater global mutational load.

Association of mutational pathways with X chromosome-specific replication programs

We showed above that the elevated mutation load and SBS9 abundance on Xi were consistent with its late replication. We next investigated how mutations relate to the random replication pattern of Xi. If replication timing modulates mutation rate, the random replication of Xi would predict a random, uniform distribution of mutations. Using the 542 LCL offspring with an identifiable Xi, we assessed regional mutation rates of phased mutations in 1-Mb sliding windows with a 0.5-Mb step. As expected, for Xa, regional mutation rate correlated with male X chromosome replication timing (r = 0.61) at similar levels as phased autosomal mutations to autosomal replication timing (Figure S3G). Unexpectedly, regional Xi mutation rate demonstrated an equally high correlation to male X chromosome replication timing (r = 0.61; Figure 5J; Figure S3G). This suggests that Xi mutation distribution follows the ordered replication timing pattern of Xa rather than the random pattern of Xi.

Given the unanticipated result of ordered Xi mutations in LCL, we sought to validate these findings in CLL. Although we were unable to phase CLL mutations, we compared X chromosome mutations across male and female patients to estimate the mutational landscape of Xi. For autosomes, regional mutation rates in males and females near-equally correlated with replication timing (Figure S3H). However, in contrast to LCLs, this correlation was reduced for X chromosome mutations in female CLL patients (r = 0.67 among females, 0.76 among males; Figure S3H). By analyzing CLL-M and CLL-U separately, we found that the correlation for X chromosome regional mutation rate in CLL-U female patients (r = 0.46) was diminished compared with males (r = 0.70) and autosomes (Figure S3I). Such reduced correlation was not observed in CLL-M females (Figure S3J). Because CLL-U samples lack SHM and SBS9 mutations, we infer that clock-like mutations are randomly distributed on Xi, while SBS9 mutations more closely follow the Xa replication pattern.

To study the distribution of SBS9 mutations on Xi, we split phased mutations into five bins based on the male X chromosome replication timing. In LCLs, Xa and Xi mutations showed similarly high enrichment for SBS9 in late-replicating regions of the male X chromosome (Figure 5K). Late-replicating enrichment was stronger for Xi mutations in the high (6.21-fold more) versus low (4.28-fold) autosomal mutation load groups (Figure 5L). Thus, the disordered replication timing of Xi does not directly relate to SBS9 mutation rate in LCLs. To validate this in CLL-M, we expected equal enrichment for SBS9 in late-replicating regions in males and females. Indeed, female CLL-M X chromosome mutations were similarly enriched in late-replicating regions (10.41-fold) as males (12.29-fold; Figure 5M). Thus, in LCL and CLL, Xi SBS9 mutation distribution follows the ordered pattern of Xa replication timing.

Last, we examined clock-like mutations on Xi, focusing specifically on the LCL offspring with high autosomal mutation loads (because we only observed late-replication enrichment of clock-like mutations in those; Figure 2C). Xa clock-like mutations in the high-load group were enriched in late-replicating regions of the male X chromosome (2.11-fold; Figure 5N). However, in contrast to SBS9, Xi clock-like mutations were more uniformly distributed with respect to male X chromosome replication timing (0.99-fold; Figure 5N). This supported the hypothesis that clock-like mutations are randomly distributed on Xi. We again validated these results in CLL-M; CLL-M females demonstrated a striking reduction of clock-like mutations in late-replicating regions of the male X chromosome (1.57-fold) compared with CLL-M males (2.63-fold; Figure 4O). Taken together, LCL and CLL suggest that the replication pattern of Xi may directly relate to clock-like, but not necessarily polymerase η, mutations.

DISCUSSION

Taking advantage of the extensive genotyping of LCLs from families in several population-scale cohorts, together with the relatively large number of mutations in LCLs and the availability of matched replication timing profiles, we reveal several novel patterns related to the locations, types, and contexts of somatic mutations. We find that B cell mutation load and mutation clustering, particularly driven by DNA polymerase η, each associate with the replication timing biases of mutation locations. Greater mutation load corresponded to greater late replication bias, whereas clustered mutations were relatively enriched in early-replicating regions. The hypermutability of the Xi chromosome is predominantly attributed to SBS9, but the distribution of mutations on Xi was unexpectedly divorced from its DNA replication dynamics. These results, together with the description of multi-level mutational heterogeneity between LCL, CLL, and other cell types,[66] reveals that mutational processes are highly complex in terms of their interactions with genomic and epigenomic properties, in particular DNA replication dynamics.

Our study design enabled analyzing inter-individual variation and performing genetic association studies of the mutational landscape. We found that the wide variation in mutational load among samples does not merely represent different timescales of activity of mutational processes; instead, mutational load correlated with the extent of mutational late replication bias as well as with the proportion of mutation belonging to the SBS9 signature. We confirmed this observation among individual LCLs and in CLL, where mutations were identified using a different methodology. The similar mutation numbers analyzed across mutation load groups and the down-sampling of mutations in individual LCLs indicate that the association between mutation load and the mutation landscape is not the result of low statistical power. Instead, mutation load and the mutational landscape appear to be correlated attributes that are inherent to individual samples. We consider several possible mechanisms to explain this variability. First, past mutations may inherently increase the probability and skew the distribution of future mutations in a type of mutational feedback loop; for instance, because of local recruitment of mutagenic DNA repair pathways. However, the observation that SBS9 mutational clustering decreases with higher mutation load implies that mutation rate increases in late-replicating regions are not driven by proximal changes, arguing against this mechanism.

Instead, we favor a model by which the mutation of (a) *trans*-acting factor(s) increases the global mutation rate and also underlies the shift of mutations toward later-replicating genomic regions and/or particular mutational signatures. As this mutation increases in clonal frequency, possibly because of compounding effects of the mutated gene(s) on cell proliferation, we would observe greater late replication bias for newly acquired somatic mutations. We were able to test this directly by performing a genome-wide association study for mutation load and for mutational late replication bias, identifying *BCL6*, a transcription factor that prevents cell-cycle arrest under the tremendous DNA damage of SHM.[50] Further investigation of *BCL6* will clarify its role in mutagenesis, while studies in other cell types could test whether comparable effects take place, potentially mediated by other genes. It is important to note that we are unable to determine whether mutation load, replication timing bias, and mutational signatures are causally related to each other or, rather, all independently stem from the same underlying genetic perturbation. Because we were underpowered to call mutational signatures in individual samples, we could not perform an independent association analysis for signature composition.

Previous studies have reported *BCL6* mutations in up to ~30% of human B cells because of off-target SHM.[50] This is consistent with our observation of high *BCL6* mutation rate in LCLs as well as with most (or all) of these mutations occurring in clusters linked to AID motifs. Thus, off-target SHM may cause the initial mutations in *BCL6*, which then proceed to facilitate non-clustered SBS9 mutations during subsequent cell divisions. While previous studies have interpreted hotspot gene mutations and preference for expressed genes as suggestive of off-target biases, an alternative possibility is that mutations in certain genes drive B cell proliferation and are therefore observed more often. Last, the co-variation of mutation load and the mutational landscape has an important methodological implication because it may confound the interpretation of the genomic distributions of mutational signatures. For example, a collection of high-mutation-load LCLs would produce different conclusions about SBS9 or clock-like mutation abundance than a collection of low-mutation-load LCLs. It is therefore vital to control for mutation load when evaluating genomic properties of mutations.

While most SBS9 mutations occurred sporadically in late-replicating regions, a second mode of SBS9 mutations was clustered in early-replicating genomic regions and at least partially associated with gene activity. These findings are consistent with previous reports[15,56,57] and with DNA polymerase η being recruited to DNA either through AID-mediated lesions or to sites impacted by replication stress or other forms of genotoxic stress. Our results extend these findings and show that both modes can co-occur. SBS9 may therefore be regarded as being due to a specific mutational mechanism; however, this mechanism may be derived from more than one etiology, necessitating its more specifically stratified analysis in future studies.

SBS9 also accounted for the elevated mutation rate on the Xi chromosome. Unexpectedly, SBS9 mutations on Xi followed the early/late replication pattern of the Xa chromosome rather than the random Xi replication pattern. This suggests that replication timing may not directly modulate where SBS9 mutations occur. Instead, some yet unidentified correlated factor that is otherwise unaltered on Xi and serves as an epigenetic "memory" of its pre-inactivation state may explain the landscape of SBS9. Because gene expression, chromatin structure, and chromosome conformation are all effectively lost on Xi alongside replication timing programming,[67,68] it is difficult for us to speculate on the nature of such a factor.

A major and still not fully answered question in the mutagenesis field pertains to the mechanism(s) that lead(s) to preferential mutation accumulation in late-replicating regions. From a biochemical point of view, this could be related to the activity of DNA repair pathways,[11] *trans*-lesion DNA polymerases, dNTP levels,[69] or other factors. Our results support the idea that there is no singular mechanism that can explain this association. Rather, mutational landscapes are shaped by composites of pathways with varied associations to the replication program. Overall, the combination of DNA replication timing, mutational pathways, mutational load, rate of clustering, and other factors shape the complex landscape of genomic mutations. Given that replication timing itself is a polymorphic trait in humans,[5,70] we would further predict that different people would have different mutational patterns in different genomic regions.

Limitations of the study

While much can be gleaned from associating various genetic and epigenetic properties across many individuals, our study had limited power for analyzing certain individual properties, such as mutational signatures (especially in samples with low mutation load). We also lack information for gene expression (in particular, of genes such as *BCL6* or of EBV) and for donor age and cell culture history and passage for most of the samples we analyzed. Generating and incorporating this information in future cohorts or analyses will shed further light on the contributions of such factors (and potentially others) to the mutational associations we report. This study also focused on single-nucleotide mutations, while other mutation types, such as insertions or deletions (indels) or copy number alterations, would also be of interest for future studies. Last, we focused here on two B cell systems (and in an accompanying paper,[66] colon cancer cell lines), but a more comprehensive analysis of additional B cell and non-B cell cancer types (EBV positive and EBV negative) and of non-cancerous somatic tissue or cell lines promises to draw a more complete picture of mutational patterns and their molecular causes.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Genomic data sources and mutation calling
  - LCL replication timing profiles
  - Mutation signatures
  - EBV copy number

- Identifying genes associated with late replication timing bias and mutation load
- Clustering mutations
- Determining Xi parental identity and phasing mutations

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100305.

ACKNOWLEDGMENTS

This work was funded by the National Institutes of Health (award DP2-GM123495 to A.K.), the National Science Foundation (award MCB-1921341 to A.K.), and the United States-Israel Binational Science Foundation (award 202108 to A.K. and I. Simon). We thank Matthew Edwards for assistance with data analysis.

AUTHOR CONTRIBUTIONS

M.C. analyzed data, and A.K. provided supervision. M.C. and A.K. wrote the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 27, 2022
Revised: February 3, 2023
Accepted: March 28, 2023
Published: May 2, 2023; corrected online: August 30, 2023

REFERENCES

1. Neitzel, H. (1986). A routine method for the establishment of permanent growing lymphoblastoid cell lines. Hum. Genet. *73*, 320–326. https://doi.org/10.1007/BF00279094.
2. Hussain, T., and Mulherkar, R. (2012). Lymphoblastoid cell lines: a continuous in vitro source of cells to study carcinogen sensitivity and DNA repair. Int. J. Mol. Cell. Med. *1*, 75–87.
3. International HapMap Consortium; Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., et al. (2003). The international HapMap project. Nature *426*, 789–796. https://doi.org/10.1038/nature02168.
4. 1000 Genomes Project Consortium; Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.
5. Koren, A., Handsaker, R.E., Kamitaki, N., Karlić, R., Ghosh, S., Polak, P., Eggan, K., and McCarroll, S.A. (2014). Genetic variation in human DNA replication timing. Cell *159*, 1015–1026. https://doi.org/10.1016/j.cell.2014.10.025.
6. Koren, A., Polak, P., Nemesh, J., Michaelson, J.J., Sebat, J., Sunyaev, S.R., and McCarroll, S.A. (2012). Differential relationship of DNA replication timing to different forms of human mutation and variation. Am. J. Hum. Genet. *91*, 1033–1040. https://doi.org/10.1016/j.ajhg.2012.10.018.
7. Caballero, M., Ge, T., Rebelo, A.R., Seo, S., Kim, S., Brooks, K., Zuccaro, M., Kanagaraj, R., Vershkov, D., Kim, D., et al. (2022). Comprehensive analysis of DNA replication timing across 184 cell lines suggests a role for MCM10 in replication timing regulation. Hum. Mol. Genet. *31*, 2899–2917, ddac082. https://doi.org/10.1093/hmg/ddac082.
8. Agarwal, I., and 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. USA *116*, 17916–17924. https://doi.org/10.1073/pnas.1900714116.
9. Chen, C., Qi, H., Shen, Y., Pickrell, J., and Przeworski, M. (2017). Contrasting determinants of mutation rates in germline and soma. Genetics *207*, 255–267. https://doi.org/10.1534/genetics.117.1114.
10. Tomkova, M., Tomek, J., Kriaucionis, S., and Schuster-Böckler, B. (2018). Mutational signature distribution varies with DNA replication timing and strand asymmetry. Genome Biol. *19*, 129. https://doi.org/10.1186/s13059-018-1509-y.
11. Supek, F., and Lehner, B. (2015). Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature *521*, 81–84. https://doi.org/10.1038/nature14173.
12. Woo, Y.H., and Li, W.-H. (2012). DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. Nat. Commun. *3*, 1004–1008. https://doi.org/10.1038/ncomms1982.
13. Michaelson, J.J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., Jian, M., Liu, G., Greer, D., Bhandari, A., et al. (2012). Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell *151*, 1431–1442. https://doi.org/10.1016/j.cell.2012.11.019.
14. Sasani, T.A., Pedersen, B.S., Gao, Z., Baird, L., Przeworski, M., Jorde, L.B., and Quinlan, A.R. (2019). Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. Elife *8*, e46922. https://doi.org/10.7554/eLife.46922.
15. Machado, H.E., Mitchell, E., Øbro, N.F., Kübler, K., Davies, M., Leongamornlert, D., Cull, A., Maura, F., Sanders, M.A., Cagan, A.T.J., et al. (2022). Diverse mutational landscapes in human lymphocytes. Nature, 1–9. https://doi.org/10.1038/s41586-022-05072-7.
16. Zhang, L., Dong, X., Lee, M., Maslov, A.Y., Wang, T., and Vijg, J. (2019). Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. Proc. Natl. Acad. Sci. USA *116*, 9014–9019. https://doi.org/10.1073/pnas.1902510116.
17. Tan, Q., Ku, W., Zhang, C., Heyilimu, P., Tian, Y., Ke, Y., and Lu, Z. (2018). Mutation analysis of the EBV-lymphoblastoid cell line cautions their use as antigen-presenting cells. Immunol. Cell Biol. *96*, 204–211. https://doi.org/10.1111/imcb.1030.
18. Conrad, D.F., Keebler, J.E.M., DePristo, M.A., Lindsay, S.J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C.L., Torroja, C., Garimella, K.V., et al. (2011). Variation in genome-wide mutation rates within and between human families. Nat. Genet. *43*, 712–714. https://doi.org/10.1038/ng.862.
19. Ng, J., Vats, P., Fritz-Waters, E., Padhi, E.M., Payne, Z.L., Leonard, S., Sarkar, S., West, M., Prince, C., Trani, L., et al. (2021). De novo variant calling identifies cancer mutation profiles in the 1000 Genomes Project. Hum. Mutat. *43*, 1979–1993. https://doi.org/10.1002/humu.24455.
20. Kenter, A.L., Kumar, S., Wuerffel, R., and Grigera, F. (2016). AID hits the jackpot when missing the target. Curr. Opin. Immunol. *39*, 96–102. https://doi.org/10.1016/j.coi.2016.01.008.
21. Papavasiliou, F.N., and Schatz, D.G. (2002). Somatic hypermutation of immunoglobulin genes: merging mechanisms for genetic diversity. Cell *109*, S35. https://doi.org/10.1016/S0092-8674(02)00706-7.
22. Álvarez-Prado, Á.F., Pérez-Durán, P., Pérez-García, A., Benguria, A., Torroja, C., de Yébenes, V.G., and Ramiro, A.R. (2018). A broad atlas of somatic hypermutation allows prediction of activation-induced deaminase targets. J. Exp. Med. *215*, 761–771. https://doi.org/10.1084/jem.20171738.
23. Tang, C., Krantsevich, A., and MacCarthy, T. (2022). Deep learning model of somatic hypermutation reveals importance of sequence context beyond hotspot targeting. iScience *25*, 103668. https://doi.org/10.1016/j.isci.2021.103668.
24. Maul, R.W., and Gearhart, P.J. (2010). Aid and somatic hypermutation. Adv. Immunol. *105*, 159–191. https://doi.org/10.1016/S0065-2776(10)05006-6.
25. Matsuda, T., Bebenek, K., Masutani, C., Hanaoka, F., and Kunkel, T.A. (2000). Low fidelity DNA synthesis by human DNA polymerase-eta. Nature *404*, 1011–1013. https://doi.org/10.1038/35010014.
26. Mayorov, V.I., Rogozin, I.B., Adkison, L.R., and Gearhart, P.J. (2005). DNA polymerase η contributes to strand bias of mutations of A versus T in

immunoglobulin Genes1. J. Immunol. *174*, 7781–7786. https://doi.org/10.4049/jimmunol.174.12.7781.

27. Wang, Q., Oliveira, T., Jankovic, M., Silva, I.T., Hakim, O., Yao, K., Gazumyan, A., Mayer, C.T., Pavri, R., Casellas, R., et al. (2014). Epigenetic targeting of activation-induced cytidine deaminase. Proc. Natl. Acad. Sci. USA *111*, 18667–18672. https://doi.org/10.1073/pnas.1420575111.

28. Koren, A., and McCarroll, S.A. (2014). Random replication of the inactive X chromosome. Genome Res. *24*, 64–69. https://doi.org/10.1101/gr.161828.113.

29. Shen, H.M., Peters, A., Baron, B., Zhu, X., and Storb, U. (1998). Mutation of BCL-6 gene in normal B cells by the process of somatic hypermutation of ig genes. Science *280*, 1750–1752. https://doi.org/10.1126/science.280.5370.1750.

32. Ruzzo, E.K., Pérez-Cano, L., Jung, J.-Y., Wang, L.K., Kashef-Haghighi, D., Hartl, C., Singh, C., Xu, J., Hoekstra, J.N., Leventhal, O., et al. (2019). Inherited and de novo genetic risk for autism impacts shared networks. Cell *178*, 850–866.e26. https://doi.org/10.1016/j.cell.2019.07.015.

33. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell *185*, 3426–3440.e19. https://doi.org/10.1016/j.cell.2022.08.004.

34. Dolzhenko, E., van Vugt, J.J.F.A., Shaw, R.J., Bekritsky, M.A., van Blitterswijk, M., Narzisi, G., Ajay, S.S., Rajan, V., Lajoie, B.R., Johnson, N.H., et al. (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res. *27*, 1895–1903. https://doi.org/10.1101/gr.225672.117.

35. Eberle, M.A., Fritzilas, E., Krusche, P., Källberg, M., Moore, B.L., Bekritsky, M.A., Iqbal, Z., Chuang, H.-Y., Humphray, S.J., Halpern, A.L., et al. (2017). A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. Genome Res. *27*, 157–164. https://doi.org/10.1101/gr.210500.116.

30. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly *6*, 80–92. https://doi.org/10.4161/fly.19695.

31. Koren, A., Massey, D.J., and Bracci, A.N. (2021). TIGER: inferring DNA replication timing from whole-genome sequence data. Bioinformatics *37*, 4001–4005. https://doi.org/10.1093/bioinformatics/btab166.

36. Dolcetti, R., and Carbone, A. (2010). Epstein-Barr virus infection and chronic lymphocytic leukemia: a possible progression factor? Infect. Agent. Cancer *5*, 22. https://doi.org/10.1186/1750-9378-5-22.

37. Hallek, M., Cheson, B.D., Catovsky, D., Caligaris-Cappio, F., Dighiero, G., Döhner, H., Hillmen, P., Keating, M., Montserrat, E., Chiorazzi, N., et al. (2018). iwCLL guidelines for diagnosis, indications for treatment, response assessment, and supportive management of CLL. Blood *131*, 2745–2760. https://doi.org/10.1182/blood-2017-09-806398.

38. Kipps, T.J., Stevenson, F.K., Wu, C.J., Croce, C.M., Packham, G., Wierda, W.G., O'Brien, S., Gribben, J., and Rai, K. (2017). Chronic lymphocytic leukaemia. Nat. Rev. Dis. Primers *3*, 16096. https://doi.org/10.1038/nrdp.2016.96.

39. Crombie, J., and Davids, M.S. (2017). IGHV mutational status testing in chronic lymphocytic leukemia. Am. J. Hematol. *92*, 1393–1397. https://doi.org/10.1002/ajh.24808.

40. Decker, T., Schneller, F., Hipp, S., Miething, C., Jahn, T., Duyster, J., and Peschel, C. (2002). Cell cycle progression of chronic lymphocytic leukemia cells is controlled by cyclin D2, cyclin D3, cyclin-dependent kinase (cdk) 4 and the cdk inhibitor p27. Leukemia *16*, 327–334. https://doi.org/10.1038/sj.leu.2402389.

41. Mosquera Orgueira, A., Antelo Rodríguez, B., Díaz Arias, J.Á., González Pérez, M.S., and Bello López, J.L. (2019). New recurrent structural aberrations in the genome of chronic lymphocytic leukemia based on exome-sequencing data. Front. Genet. *10*, 854. https://doi.org/10.3389/fgene.2019.00854.

42. Edelmann, J., Holzmann, K., Tausch, E., Saunderson, E.A., Jebaraj, B.M.C., Steinbrecher, D., Dolnik, A., Blätte, T.J., Landau, D.A., Saub, J., et al. (2020). Genomic alterations in high-risk chronic lymphocytic leukemia frequently affect cell cycle key regulators and NOTCH1-regulated transcription. Haematologica *105*, 1379–1390. https://doi.org/10.3324/haematol.2019.217307.

43. Rivera-Mulia, J.C., Buckley, Q., Sasaki, T., Zimmerman, J., Didier, R.A., Nazor, K., Loring, J.F., Lian, Z., Weissman, S., Robins, A.J., et al. (2015). Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells. Genome Res. *25*, 1091–1103. https://doi.org/10.1101/gr.187989.114.

44. Yaffe, E., Farkash-Amar, S., Polten, A., Yakhini, Z., Tanay, A., and Simon, I. (2010). Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture. PLoS Genet. *6*, e1001011. https://doi.org/10.1371/journal.pgen.1001011.

45. Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N., et al. (2020). The repertoire of mutational signatures in human cancer. Nature *578*, 94–101. https://doi.org/10.1038/s41586-020-1943-3.

46. Yaacov, A., Vardi, O., Blumenfeld, B., Greenberg, A., Massey, D.J., Koren, A., Adar, S., Simon, I., and Rosenberg, S. (2021). Cancer mutational processes vary in their association with replication timing and chromatin accessibility. Cancer Res. *81*, 6106–6116. https://doi.org/10.1158/0008-5472.CAN-21-2039.

47. Maura, F., Degasperi, A., Nadeu, F., Leongamornlert, D., Davies, H., Moore, L., Royo, R., Ziccheddu, B., Puente, X.S., Avet-Loiseau, H., et al. (2019). A practical guide for mutational signature analysis in hematological malignancies. Nat. Commun. *10*, 2969. https://doi.org/10.1038/s41467-019-11037-8.

48. Alexandrov, L.B., Jones, P.H., Wedge, D.C., Sale, J.E., Campbell, P.J., Nik-Zainal, S., and Stratton, M.R. (2015). Clock-like mutational processes in human somatic cells. Nat. Genet. *47*, 1402–1407. https://doi.org/10.1038/ng.3441.

49. Laskov, R., Yahud, V., Hamo, R., and Steinitz, M. (2011). Preferential targeting of somatic hypermutation to hotspot motifs and hypermutable sites and generation of mutational clusters in the IgVH alleles of a rheumatoid factor producing lymphoblastoid cell line. Mol. Immunol. *48*, 733–745. https://doi.org/10.1016/j.molimm.2010.10.009.

50. Yang, H., and Green, M.R. (2019). Epigenetic programing of B-cell lymphoma by BCL6 and its genetic deregulation. Front. Cell Dev. Biol. *7*, 272. https://doi.org/10.3389/fcell.2019.00272.

51. Jantus Lewintre, E., Reinoso Martín, C., García Ballesteros, C., Pendas, J., Benet Campos, C., Mayans Ferrer, J.R., and García-Conde, J. (2009). BCL6: somatic mutations and expression in early-stage chronic lymphocytic leukemia. Leuk. Lymphoma *50*, 773–780. https://doi.org/10.1080/10428190902842626.

52. Sorokin, A.V., Nair, B.C., Wei, Y., Aziz, K.E., Evdokimova, V., Hung, M.-C., and Chen, J. (2015). Aberrant expression of proPTPRN2 in cancer cells confers resistance to apoptosis. Cancer Res. *75*, 1846–1858. https://doi.org/10.1158/0008-5472.CAN-14-2718.

53. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. *47*, D941–D947. https://doi.org/10.1093/nar/gky1015.

54. Pasqualucci, L., Migliazza, A., Basso, K., Houldsworth, J., Chaganti, R.S.K., and Dalla-Favera, R. (2003). Mutations of the BCL6 proto-oncogene disrupt its negative autoregulation in diffuse large B-cell lymphoma. Blood *101*, 2914–2923. https://doi.org/10.1182/blood-2002-11-3387.

55. Rouhani, F.J., Zou, X., Danecek, P., Badja, C., Amarante, T.D., Koh, G., Wu, Q., Memari, Y., Durbin, R., Martincorena, I., et al. (2022). Substantial somatic genomic variation and selection for BCOR mutations in human

induced pluripotent stem cells. Nat. Genet. *54*, 1406–1416. https://doi.org/10.1038/s41588-022-01147-3.

56. Supek, F., and Lehner, B. (2017). Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. Cell *170*, 534–547.e23. https://doi.org/10.1016/j.cell.2017.07.003.
57. Mas-Ponte, D., and Supek, F. (2020). DNA mismatch repair promotes APOBEC3-mediated diffuse hypermutation in human cancers. Nat. Genet. *52*, 958–968. https://doi.org/10.1038/s41588-020-0674-6.
58. Jiang, Y., Soong, T.D., Wang, L., Melnick, A.M., and Elemento, O. (2012). Genome-wide detection of genes targeted by non-Ig somatic hypermutation in lymphoma. PLoS One *7*, e40332. https://doi.org/10.1371/journal.pone.0040332.
59. Akdemir, K.C., Le, V.T., Kim, J.M., Killcoyne, S., King, D.A., Lin, Y.-P., Tian, Y., Inoue, A., Amin, S.B., Robinson, F.S., et al. (2020). Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. Nat. Genet. *52*, 1178–1188. https://doi.org/10.1038/s41588-020-0708-0.
60. Jäger, N., Schlesner, M., Jones, D.T.W., Raffel, S., Mallm, J.-P., Junge, K.M., Weichenhan, D., Bauer, T., Ishaque, N., Kool, M., et al. (2013). Hypermutation of the inactive X chromosome is a frequent event in cancer. Cell *155*, 567–581. https://doi.org/10.1016/j.cell.2013.09.042.
61. Tukiainen, T., Villani, A.-C., Yen, A., Rivas, M.A., Marshall, J.L., Satija, R., Aguirre, M., Gauthier, L., Fleharty, M., Kirby, A., et al. (2017). Landscape of X chromosome inactivation across human tissues. Nature *550*, 244–248. https://doi.org/10.1038/nature24265.
62. Kucera, K.S., Reddy, T.E., Pauli, F., Gertz, J., Logan, J.E., Myers, R.M., and Willard, H.F. (2011). Allele-specific distribution of RNA polymerase II on female X chromosomes. Hum. Mol. Genet. *20*, 3964–3973. https://doi.org/10.1093/hmg/ddr315.
63. McDaniell, R., Lee, B.-K., Song, L., Liu, Z., Boyle, A.P., Erdos, M.R., Scott, L.J., Morken, M.A., Kucera, K.S., Battenhouse, A., et al. (2010). Heritable individual-specific and allele-specific chromatin signatures in humans. Science *328*, 235–239. https://doi.org/10.1126/science.1184655.
64. Marinov, G.K., Williams, B.A., McCue, K., Schroth, G.P., Gertz, J., Myers, R.M., and Wold, B.J. (2014). From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res. *24*, 496–510. https://doi.org/10.1101/gr.161034.113.
65. Wainer Katsir, K., and Linial, M. (2019). Human genes escaping X-inactivation revealed by single cell expression data. BMC Genom. *20*, 201. https://doi.org/10.1186/s12864-019-5507-6.
66. Caballero, M., Boos, D., and Koren, A. (2023). Cell type specificity of the human mutation landscape with respect to DNA replication dynamics. Cell Genomics *3*. https://doi.org/10.1016/j.xgen.2023.100315.
67. Splinter, E., de Wit, E., Nora, E.P., Klous, P., van de Werken, H.J.G., Zhu, Y., Kaaij, L.J.T., van Ijcken, W., Gribnau, J., Heard, E., and de Laat, W. (2011). The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. Genes Dev. *25*, 1371–1383. https://doi.org/10.1101/gad.633311.
68. Lee, J.T. (2011). Gracefully ageing at 50, X-chromosome inactivation becomes a paradigm for RNA and chromatin control. Nat. Rev. Mol. Cell Biol. *12*, 815–826. https://doi.org/10.1038/nrm3231.
69. Kenigsberg, E., Yehuda, Y., Marjavaara, L., Keszthelyi, A., Chabes, A., Tanay, A., and Simon, I. (2016). The mutation spectrum in genomic late replication domains shapes mammalian GC content. Nucleic Acids Res. *44*, 4222–4232. https://doi.org/10.1093/nar/gkw268.
70. Ding, Q., Edwards, M.M., Wang, N., Zhu, X., Bracci, A.N., Hulke, M.L., Hu, Y., Tong, Y., Hsiao, J., Charvet, C.J., et al. (2021). The genetic architecture of DNA replication timing in human pluripotent stem cells. Nat. Commun. *12*, 6746. https://doi.org/10.1038/s41467-021-27115-9.
71. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443. https://doi.org/10.1038/s41586-020-2308-7.
72. GTEx Consortium; Laboratory Data Analysis &Coordinating Center LDACC—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx eGTEx groups; NIH Common Fund; Jo, B., Mohammadi, P., Park, Y., Parsana, P., et al.; Biospecimen Collection Source Site—NDRI (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213. https://doi.org/10.1038/nature24277.
73. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at arXiv. https://doi.org/10.48550/arXiv.1303.3997.
74. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.
75. Bergstrom, E.N., Huang, M.N., Mahto, U., Barnes, M., Stratton, M.R., Rozen, S.G., and Alexandrov, L.B. (2019). SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. BMC Genom. *20*, 685. https://doi.org/10.1186/s12864-019-6041-2.
76. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993. https://doi.org/10.1093/bioinformatics/btr509.
77. Manders, F., Brandsma, A.M., de Kanter, J., Verheul, M., Oka, R., van Roosmalen, M.J., van der Roest, B., van Hoeck, A., Cuppen, E., and van Boxtel, R. (2022). MutationalPatterns: the one stop shop for the analysis of mutational processes. BMC Genom. *23*, 134. https://doi.org/10.1186/s12864-022-08357-3.
78. Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., et al. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at bioRxiv. https://doi.org/10.1101/201178.
79. Yuen, R.K.C., Merico, D., Cao, H., Pellecchia, G., Alipanahi, B., Thiruvahindrapuram, B., Tong, X., Sun, Y., Cao, D., Zhang, T., et al. (2016). Genome-wide characteristics of de novo mutations in autism. NPJ Genom. Med. *1*, 160271–1602710. https://doi.org/10.1038/npjgenmed.2016.27.
80. Meyer, D., C Aguiar, V.R., Bitarello, B.D., C Brandt, D.Y., and Nunes, K. (2018). A genomic perspective on HLA evolution. Immunogenetics *70*, 5–27. https://doi.org/10.1007/s00251-017-1017-3.
81. denovo-db. denovo-db.gs.washington.edu.
82. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium; Getz, G., Korbel, J.O., Stuart, J.M., Jennings, J.L., Stein, L.D., Perry, M.D., Nahal-Bose, H.K., Ouellette, B.F.F., Li, C.H., et al. (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93. https://doi.org/10.1038/s41586-020-1969-6.
83. Massey, D.J., Kim, D., Brooks, K.E., Smolka, M.B., and Koren, A. (2019). Next-generation sequencing enables spatiotemporal resolution of human centromere replication timing. Genes *10*, E269. https://doi.org/10.3390/genes10040269.
84. Blokzijl, F., Janssen, R., van Boxtel, R., and Cuppen, E. (2018). MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med. *10*, 33. https://doi.org/10.1186/s13073-018-0539-0.
85. Mandage, R., Telford, M., Rodríguez, J.A., Farré, X., Layouni, H., Marigorta, U.M., Cundiff, C., Heredia-Genestar, J.M., Navarro, A., and Santpere, G. (2017). Genetic factors affecting EBV copy number in lymphoblastoid cell lines derived from the 1000 Genome Project samples. PLoS One *12*, e0179446. https://doi.org/10.1371/journal.pone.0179446.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
iHART cohort LCLs	Ruzzo et al. 2019[32]	https://doi.org/10.1016/j.cell.2019.07.015
1kGP cohort LCLs	Byrska-Bishop et al. 2022[33]	https://doi.org/10.1016/j.cell.2022.08.004
Repeat expansion cohort LCLs	Dolzhenko et al. 2017[34]	https://doi.org/10.1101/2Fgr.225672.117
Illumina platinum cohort LCLs	Eberle et al. 2017[35]	https://doi.org/10.1101/gr.210500.116
Additional LCLs	Caballero et al. 2022[7]	https://doi.org/10.1093/hmg/ddac082
Polaris kids cohort LCLs	Auton et al. 2015[4]	https://doi.org/10.1038/nature15393
CLL mutations	ICGC/PCAWG	https://dcc.icgc.org/projects/CLLE-ES
LCL consensus replication timing profile	This manuscript	supplemental information
LCL S/G1 replication timing profile	Koren et al. 2012[6]	https://doi.org/10.1016/j.ajhg.2012.10.018
gnomAD V3	Karczewski et al. 2020[71]	https://gnomad.broadinstitute.org/
COSMIC v3.2 SBSsignatures	Alexandrov et al. 2020[45]	https://cancer.sanger.ac.uk/signatures/
Protein coding genes (hg38)	NCBI	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.39_GRCh38.p13/
Median LCL and whole-blood gene expression	GTEx[72]	https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz
Software and algorithms		
Picard Tools (v1.138)		http://broadinstitute.github.io/picard/
BWA-mem (v0.7.17)	Li H. 2013[73]	http://arxiv.org/abs/1303.3997
GATK (v4.1.4.0)	McKenna et al. 2010[74]	https://gatk.broadinstitute.org/hc/en-us
vcf-liftover		https://github.com/hmgu-itg/VCF-liftover
SigProfilerMatrixGenerator (v1.2)	Bergstrom et al. 2019[75]	https://github.com/AlexandrovLab/SigProfilerMatrixGenerator
Samtools (v1.6)	Li 2010[76]	https://doi.org/10.1093/bioinformatics/btr509
TIGER	Koren et al. 2021[31]	https://github.com/TheKorenLab/TIGER
MutationalPatterns (v3.8.0)	Manders et al. 2022[77]	https://bioconductor.org/packages/release/bioc/html/MutationalPatterns.html
ClusteredMutations		https://cran.r-project.org/web/packages/ClusteredMutations/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources, additional code, and data should be directed to and will be fulfilled by the lead contact, Amnon Koren (koren@cornell.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

All mutation calls in hg38 coordinates (along with sample IDs, mutational context, replication timing values, and allele frequencies) for LCL and CLL are available in Data S1. Consensus replication timing profiles are available in Data S2. Code used for analyses, replication timing generation, and mutation calling is available on Mendeley (Mendeley Data: https://doi.org/10.17632/2hwhv32gs2.4). Additional tables are available in Table S1: Details of LCL offspring (parents, cohort, sex, and autosomal/X chromosome mutation load), details on repeat LCL sequencing including NA12878, genes associated with late bias or mutation load in LCL, Xi parental identity and score in female LCLs, details on replication timing filtering in individual LCLs, and predicted CLL *IGHV* mutation status.

METHOD DETAILS

Genomic data sources and mutation calling

LCL genomic data sources

Mutations in the 1662 LCL offspring were sourced from six cohorts (Table 1). These offspring were matched to 989 pairs of fully genotyped parents, as 377 families contained two or more offspring. Eight families covered three generations. The largest cohort was iHART[32] and included 1028 offspring with or without a diagnosis of autism. While iHART samples included both LCL and whole blood samples, only LCL offspring were included in this study, although for parental data we also considered whole blood samples (1.2% of parents). The second-largest LCL mutation cohort was sourced from the 1000 Genomes Project (1kGP) and contained 602 trios.[33] We used 49 offspring from the Polaris project Kids cohort[4] as replicate samples as all overlapped the 1kGP cohort. An additional nine offspring were sourced from the Repeat Expansion (RE) cohort[34] and 13 offspring from the Illumina Platinum[35] family; of those, two (NA12878 and NA12877) overlapped with 1kGP samples and were used for primary analyses instead of the latter due to their higher read depth (~50x compared to ~30x). Finally, we used 12 LCL that we previously described.[7] Briefly, these samples were obtained from the Coriell Institute, sequenced on an Illumina HiSeq X (2 × 150bp) to a depth of approximately 15X, and aligned to hg38 with BWA-mem[73] (v0.7.17) similar to the other LCL cohorts.

LCL genotyping

In order to ultimately identify mutations, we first genotyped LCL offspring and parents. Genotypes for iHART samples were obtained from Ruzzo et al. 2019.[32] All other LCL cohorts were genotyped by us using the GATK (v4.1.4.0) best practices for germline short variant discovery.[74,78] Briefly, BAM files were recalibrated and aligned around common insertions and deletions with 'BaseRecalibrator' and 'IndelRealigner'. Next, gVCF files were generated from all recalibrated BAM files using 'HaplotypeCaller'. gVCFs were then merged into families with 'CombineGVCFs' and joint genotyped with 'GenotypeGVCFs'. Finally, SNVs were recalibrated with 'VariantRecalibrator'. We note that genotype calling for the iHART cohort differed from the above in that all samples were jointly genotyped, and variants were removed if they had a depth of <10X, a genotype quality of <25, or an alternative allele frequency of <0.2; we subsequently applied equal or stricter filtering metrics to all samples when identifying mutations, hence ruling out an effect of these differences in iHART genotyping on our analyses.

For samples originally aligned and genotyped in hg19 (approximately half of all samples), genotypes were lifted-over to hg38 coordinates using vcf-liftover (https://github.com/hmgu-itg/VCF-liftover, only liftover within the same chromosome were allowed). We removed genotypes in samples originally aligned to hg38 at coordinates without an hg19 equiv to compensate for the reduction of genotypes following liftover. This eliminated approximately 1.9% of all sites.

LCL mutation calling

Candidate mutations were identified as single nucleotide Mendelian errors between parent and offspring alleles. The following steps were based on previously established family-based mutation calling methods from Yuen et al. 2016.[79] Mutations on the autosomes and X chromosome in female offspring were identified as heterozygous genotypes (for the reference allele and an alternate allele) in offspring where parents were homozygous for the reference allele. For the X chromosome in male offspring, mutations were identified as sites with only an alternate allele where the mother is homozygous for the reference allele. Next, we filtered mutations with a Fisher's exact test Phred-scaled p value (FS) < 60.0, RMS mapping quality (MQ) < 40.0, Wilcoxon rank-sum test *Z* score of mapping qualities (MQRankSum)<-12.5 or read position (RPRS)<-8.0, symmetric odds ratio (SOR) > 3, and a Phred-scaled quality score (QUAL) < 30. We excluded sites that did not pass variant quality score recalibration. To remove sub-clonal mutations and potential technical errors, we eliminated candidate mutations for which the mutant (alternate) allele frequency was <0.2. We removed likely inherited variants where either parent contained reads matching the mutant allele. Finally, to eliminate possible false-positive mutation calls caused by somatic deletions in the offspring (and hence reduced genotyping accuracy), we eliminated candidate mutations in cases where the offspring read depth was <10% of the combined parental read depth (again, adjusted for the X chromosome in male offspring) at the mutation site. After this initial hard filtering, 4.4 million candidate mutations were called across all 1662 offspring.

Next, we removed candidate mutations based on genomic location. We first removed 61,479 candidate mutations around the HLA locus (chr6:28477797–33548354 in hg38) due to the high propensity for genotyping errors stemming from high local polymorphism density.[80] Similarly, we removed 63,547 mutations around the immunoglobulin heavy locus (*IGHV*, chr14:105580000-106880000 in hg38), which is hyper-mutated in LCLs. Next, we removed 587,511 mutations within gaps >25Kb in the LCL replication timing profile (see section LCL replication timing profiles). Regions of the genome removed for HLA and *IGHV* were also removed from the LCL reference RT profile.

To further eliminate inherited variants, we implemented a last filtering step to remove mutations based on population allele frequency. Specifically, we removed mutations with a gnomAD[71] V3 frequency of >0.001. We did not use a frequency of zero as many of our samples (including all 1kGP individuals), and their somatic mutations, are represented in gnomAD. We also filtered mutations occurring in more than 30 of the 1662 offspring. In total, 2,826,985 candidate mutations were eliminated through this allele frequency filtering. After all filtering steps, 885,655 autosomal and 42,061 X chromosome mutations remained in the 1662 non-replicate LCL offspring. The mean variant allele frequency of autosomal mutations was 0.40.

For each mutation, trinucleotide context was generated with SigProfilerMatrixGenerator,[75] and replication timing values at mutations sites were calculated with the R function 'approx' using the linear method. Mutation allele frequency did not vary by substitution

type nor by trinucleotide context. The mean variant allele frequency among mutation in trinucleotide contexts ranged from 0.40 to 0.46.

LCL mutation validation

Parent-offspring mutation calling carries a risk of falsely identifying an inherited variant as a *de novo* mutation. This could stem, for instance, from failing to identify the inherited alleles in a parent due to a somatic deletion or false-negative genotyping. To quantify the proportion of false mutations that are inherited variants, we analyzed mutation calls in 73 monozygotic (MZ) twin pairs. MZ twins share all inherited alleles and germline mutations but have unique somatic mutations (Figure S1B). Although parent-offspring mutation calling cannot distinguish somatic from germline mutations, having an estimate for one of those enables to estimate the other. Specifically, based on all samples from denovo-db,[81] the average human contains 65.5 autosomal germline mutations, is similar to other estimations of approximately 70 autosomal mutations per generation.[14] In contrast, in this study, MZ pairs shared between 81 and 245 autosomal mutations (median:113; Figure S1C, D). Thus, the excess number (above 65.5) of MZ twin shared mutations provides a rough estimate of the number of falsely called mutations that are likely inherited variants (Figure S1E). We thus predicted that between 1.85% and 27.2% of autosomal mutations in MZ twins are inherited variants (median: 9.66%; Figure S1E). This is likely an overestimate, as the paternal age among MZ twins was relatively high (median: 32.26 years, range: 20.43–78.51), thus increasing the expected number of germline mutations.

We also estimated false mutation calls derived from technical errors by analyzing genotype calls in 51 offspring that were resequenced by different groups on different platforms (Table S1). We compared mutant alleles of samples in the main dataset to the GVCF of the replicate. A mutation was considered validated if the mutant allele was found in the replicate sample at any frequency. A median of 93.1% of autosomal mutations were supported by their replicate sample (range: 65.1–98.7%; Figure S1F). The mutations that could not be validated did not show a strong enrichment toward late replication timing and, therefore, should not have influenced our results (Figure S1G). We further validated mutation calls in the offspring sample NA12878. The Illumina Platinum cohort sample of NA12878 was used as part of the main dataset (of 1662 offspring), and the 1kGP NA12878 sample was used for validation (and counted as part of the 51 replicate sample analysis mentioned above). We sourced four other replicate sequencings of NA12878 (Table S1) and found that 98.8% of mutations were supported by at least one alternate source.

CLL mutation data

Mutations in CLL patients were obtained from the ICGC/PCAWG cohorts CLLE-ES. Alignment and mutation calling for tumor samples (peripheral blood-derived) and normal samples was performed by PCAWG using their pipeline[82] in hg19. We only included mutations called from 151 patients with whole genome sequencing. This provided 371,252 autosomal mutations and 23,130 X chromosome mutations.

Before filtering, all mutations were lifted-over to hg38 coordinates using vcf-liftover (https://github.com/hmgu-itg/VCF-liftover, only liftover within the same chromosome were allowed). We then removed mutations around the HLA and IGHV loci and in gaps of the LCL replication timing profile. Hence, we used two LCL replication timing profiles in our analyses: one in which regions filtered from the LCL offspring dataset were removed, and another in which regions filtered from the CLL dataset were removed. We interpolated replication timing values for the final 355,474 autosomal and 22,131 X chromosome mutations with the CLL-filtered LCL reference replication timing profile and determined trinucleotide contexts in an identical manner to LCLs.

LCL replication timing profiles

The LCL consensus replication profile was generated using TIGER[31] from median read count data from all 1662 offspring. First, uniquely mapping reads were extracted from aligned BAM files of each sample. For samples aligned to hg19, BAM coordinates were lifted to hg38 in an identical manner to mutations. We compensated for lift-over by modifying TIGER to exclude hg38 coordinates with no hg19 equiv when creating 2.5Kb windows of uniquely alignable sequence. We tested the effect of this method by comparing the replication timing profiles of 22 samples originally aligned to hg38 with those aligned to hg19 and lifted-over to hg38. The lifted replication timing profile in all samples on all autosomes was nearly identical (Pearson's $r > 0.99$) to the one aligned to hg38.

Using default TIGER parameters, the liftover-corrected 2.5Kb windows were GC-corrected and normalized to an autosomal genome copy number of two. We eliminated sub-clonal aneuploidies in individual offspring by filtering out whole chromosomes with an average autosomal copy number of >2.2 or <1.8, an X chromosome copy number of >2 or <1.6 for female offspring, and an X chromosome copy number of >1.2 or <0.8 for male offspring. This removed 34 chromosomes from 23 samples. We removed suspected small copy number alterations by filtering out 2.5Kb windows with an exceptionally high or low median copy number across all offspring and within individual offspring. We first removed autosomal and female X chromosome windows across all offspring with a median copy number ±0.6 than that chromosome's median copy number (as calculated from all offspring). The cutoff was ±0.4 for the X chromosome in male offspring. We then filtered out windows in individual offspring with a copy number ±0.6 than that chromosome's median copy number (as calculated in the individual offspring). The cutoff was ±0.3 for the X chromosome in male offspring. We next calculated autocorrelation for all offspring using the MATLAB command "autocorr" and removed whole chromosomes for samples with abnormally high autocorrelation. This removed 51 chromosomes in 26 samples. Finally, we discarded the two offspring, HG02523 and NA12344, as they had more than six individual chromosomes removed.

After filtering, we took the median GC-corrected data in 2.5Kb each window across all offspring. For the X chromosome, we calculated separate medians using only male or female offspring. Replication timing values were generated by smoothing the median

GC-corrected data with a cubic smoothing spline (MATLAB command 'csaps', smoothing parameter: 1×10^{-17}). Only regions of >20 continuous 2500bp windows were included. Smoothing was not performed over data gaps >100Kb or reference genome gaps >50Kb. The smoothed profiles were then normalized to an autosomal mean of zero and a standard deviation of one. For analyses on the X chromosome, we generated an X chromosome replication timing profile considering only male LCL offspring.

We compared our median LCL replication timing profile to a replication profile of NA12878 generated by sequencing S and G1 phase DNA.[83] The S/G1 coordinates were interpolated to TIGER window coordinates with the MATLAB function 'interp1'.

Mutation signatures

We fit COSMIC v3.2 SBSsignatures[45] 1, 5, 9, and 40 (and SBS8 and 18 in a separate analysis) to all autosomal mutations using the MutationalPatterns[84] command 'fit_to_signatures'. Following best-practices,[47] we corrected COSMIC SBSsignatures by adjusting the 96 trinucleotide frequencies by the relative abundance of trinucleotide frequencies between the filtered and unfiltered genome. We used cosine similarity (MutationalPatterns command 'cos_sim') to assess the confidence of signature fit which compares the original trinucleotide frequencies of mutations to reconstructed frequencies based on predicted signature contributions. A value of one indicates an identical reconstruction. We additionally performed 1000 bootstrap sampling when fitting signatures using the MutationalPatterns command 'fit_to_signatures_bootstrapped'. We used the standard deviation of 1000 bootstrap samples as the standard error for signature contribution. Standard errors for clock-like mutations (SBS1, 5, and 40) were calculated using standard error in the difference of the means (the square-root of the sum of variances).

EBV copy number

We assessed if mutation load and mutation landscape were associated with EBV copy number using established methods.[85] For the 602 LCL offspring of the 1kGP cohort, we calculated the mean depth of uniquely mapping reads to the decoy hg38 EBV genome and normalized by the mean copy number of autosomes as calculated for replication timing analyses. Normalized EBV copy numbers ranged from 3.88 to 448.92 with a mean of 33.12.

Identifying genes associated with late replication timing bias and mutation load

We identified individual LCL mutational replication timing bias by calculating the proportion of mutations in four replication timing bins. We used the linear slope of proportions as a representation for replication timing bias and calculated PCs using the R command 'prcomp.' Gene associations were calculated using the binary state of whether at least one mutation in a sample fell within the range of a protein coding gene (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.39_GRCh38.p13/) against individual replication timing biases. Mutation functionality was not considered. p-value of association was calculated with the R command 'lm' and individual autosomal mutation load was input as a covariate. 97 genes showed significant association with late replication biases and were mutated in at least 50 samples (list available in Table S1). We also performed association analysis for mutation load versus the binary state of whether at least one mutation in a sample fell within the range of a protein coding gene. To control for the effect of mutation load on gene mutations status, we normalized the mutated/unmutated binary state by the mutation loads of individual samples. Seven genes showed significant association with mutation load and were mutated in at least 50 samples (list available in Table S1).

Clustering mutations

We clustered SBS9-context mutations using the 'ClusteredMutations' (https://cran.r-project.org/web/packages/ClusteredMutations/) command 'showers.' The minimum cluster size was two mutations, and the maximum distance between SBS9-context mutations was 500bp. We simulated autosomal SBS9-context mutations of matched mutation rates in 20 (for LCL) or 5 (for CLL-M, to account for fewer mutations) replication timing bins. Within the replication timing range of each bin, we performed 1000 random samples of SBS9-context motifs from the genome without replacement, matching the number of SBS9-context mutations in the bin. The sampled mutations were then clustered identically as described above for real mutations. Similar simulations were performed incorporating distance to AID-context mutations. In addition to sampling and clustering SBS9-context motifs, we sampled AID-context motifs in replication timing bins equal to the number of AID-context mutations in the bin. In each simulation, we calculated the proportion of clustered/non-clustered SBS9-context motifs within 100bp of an AID-context motif.

We evaluated the distance of SBS9-context mutations to the TSS of 22,337 protein-coding genes adjusted for gene directionality. We interpolated LCL replication timing values at the TSS. Analyses were repeated for the TSS of 275 off-target SHM hotspots.[22] Gene expression of LCLs (median expression across 144 LCLs) and whole-blood for CLL (median expression across 338 samples) were sourced from GTEx.[72]

Determining Xi parental identity and phasing mutations

We phased Mendelian inherited single nucleotide variants in female LCL offspring. For each variant, we required the offspring and parents to have a read depth ≥ 5, MQ > 30, FS < 60.0, MQRankSum>−12.5, RPRS>−8.0, and SOR<3. In the heterozygous offspring genotype, we required the alternate allele frequency to be greater than 0.3. We calculated parental copy number disparity as the absolute difference of mean sequencing read depth for paternal and maternal alleles divided by their combined read depth. To determine a threshold for identifying X-inactivation, we used the 95th percentile of parental copy number disparity on chromosome

14. This chromosome was chosen as it contained the most comparable number of phaseable variants as chromosome X. The parental identity of Xi was assigned to the parental homolog with the lower mean sequencing read depth.

We phased mutations occurring on the same read or mate-pair as a phaseable inherited variant. We first determined the read names containing the maternal and paternal alleles using the Samtools[76] (v1.6) command 'mpileup.' We repeated this process to identify read names containing the mutation alleles. We phased mutations where read names containing mutation alleles exclusively matched those phased to one parent. If mutation alleles matched read names phased to both parents, the mutation was considered ambiguous. We calculated mutational signature contributions on phased chromosomes as described above using the biologically relevant LCL signatures corrected for individual chromosome trinucleotide content.