Genome-wide mutational signatures of immunological diversification in normal lymphocytes

Authors: Heather E Machado[1], Emily Mitchell[1,2]†, Nina F Øbro[2,3,4]†, Kirsten Kübler[5-7]†, Megan Davies[2,3,8], Francesco Maura[9], Daniel Leongamornlert[1], Mathijs A. Sanders[1,10], Alex Cagan[1], Craig McDonald[2,3,11], Miriam Belmonte[2,3,11], Mairi S. Shepherd[2,3], Robert J. Osborne[1,12], Krishnaa Mahbubani[3,13,14], Iñigo Martincorena[1], Elisa Laurenti[2,3], Anthony R Green[2,3], Gad Getz[5-7,15], Paz Polak[16], Kourosh Saeb-Parsy[13,14], Daniel J Hodson[2,3], David Kent[2,3,11]*, Peter J Campbell[1,2]*

Affiliations:
[1] Wellcome Sanger Institute, Hinxton, United Kingdom
[2] Wellcome MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge, United Kingdom
[3] Department of Hematology, University of Cambridge, Cambridge, United Kingdom
[4] Department of Clinical Immunology, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark
[5] Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
[6] Center for Cancer Research, Massachusetts General Hospital, Charlestown, Massachusetts, USA
[7] Harvard Medical School, Boston, Massachusetts, USA
[8] Cambridge Molecular Diagnostics, Milton Road, Cambridge, United Kingdom
[9] Sylvester Comprehensive Cancer Center, Miami, Florida, USA
[10] Department of Hematology, Erasmus MC Cancer Institute, Rotterdam, The Netherlands
[11] York Biomedical Research Institute, University of York, Wentworth Way, York, United Kingdom
[12] Biofidelity, 330 Cambridge Science Park, Milton Road, Cambridge, United Kingdom
[13] Department of Surgery, University of Cambridge, Cambridge, United Kingdom
[14] NIHR Cambridge Biomedical Research Centre, Cambridge Biomedical Campus, Cambridge, United Kingdom
[15] Department of Pathology, Massachusetts General Hospital, Boston, Massachusetts, USA
[16] Oncological Sciences, Icahn School of Medicine at Mount Sinai, New York, USA

†Authors contributed equally

*Correspondence to: Peter J Campbell (pc8@sanger.ac.uk) and David Kent (david.kent@york.ac.uk)

Abstract:

A lymphocyte suffers many threats to its genome, including programmed mutation during differentiation, antigen-driven proliferation and residency in diverse microenvironments. After developing protocols for single-cell lymphocyte expansions, we sequenced whole genomes from 717 normal naive and memory B and T lymphocytes and hematopoietic stem cells. Lymphocytes carried more point mutations and structural variation than stem cells, accruing at higher rates in T than B cells, attributable to both exogenous and endogenous mutational processes. Ultraviolet light exposure and other sporadic mutational processes generated hundreds to thousands of mutations in some memory lymphocytes. Memory B cells acquired, on average, 18 off-target mutations genome-wide for every one on-target *IGV* mutation during the germinal center reaction. Structural variation was 16-fold higher in lymphocytes than stem cells, with ~15% of deletions being attributable to off-target RAG activity.

One Sentence Summary:

The mutational landscape of normal lymphocytes chronicles the off-target effects of programmed genome engineering during immunological diversification and the consequences of differentiation, proliferation and residency in diverse microenvironments.

Main Text:

The adaptive immune system depends upon programmed somatic mutation to generate antigen receptor diversity. T lymphocytes use RAG-mediated deletion to generate functional T-cell receptors (TCRs); B lymphocytes also use RAG-mediated deletion to rearrange immunoglobulin (Ig) heavy and light chains, followed by AID-mediated somatic hypermutation and class-switch recombination to further increase diversity (*1*–*3*). The machinery undertaking this physiological genome editing is tightly regulated, switched on at specific stages of lymphocyte maturation and targeted to specific regions of the genome through chromatin interaction and characteristic sequence motifs.

Off-target activity of proteins generating genomic diversity in lymphocytes can produce mutations in cancer genes with unintended consequences. Whole genome sequencing of malignant lymphoid tumors has revealed high numbers of off-target mutations with signatures resembling those seen at antigen receptors. For example, RAG-mediated deletions at genes regulating B-cell maturation are common in acute lymphoblastic leukemia (*4*, *5*); off-target AID-mediated somatic hypermutation is common in diffuse large B-cell lymphoma and chronic lymphocytic leukemia (*6*–*9*); translocations of oncogenes to immunoglobulin loci occurring during class-switch recombination are common in multiple myeloma (*10*).

It remains unclear whether the burden and signatures of somatic mutations seen in lymphoid cancers would be equivalent in normal lymphocytes, not least because driver mutations in cancers can stall lymphocyte differentiation at stages where physiological genome editing is active (*11*, *12*). While single-cell sequencing of 54 B lymphocytes has demonstrated increasing mutation burden with age and evidence of off-target somatic hypermutation (*13*), our novel protocols for producing single-cell derived colonies from naive and memory B and T cells enables the generation of genomes free of whole-genome amplification artifacts, allowing accurate quantification of mutation burdens, signatures and distributions across cells of several lymphocyte subsets.

Whole genome sequencing of B and T lymphocyte colonies

We and others have shown that growing single hematopoietic stem cells into colonies *in vitro* can enable accurate identification of all classes of somatic mutation using genome sequencing, avoiding the artifacts introduced by whole genome amplification (*14*–*16*). In order to achieve this for lymphocytes, we developed protocols for expanding flow-sorted single naive and memory B and T lymphocytes *in vitro* to colonies of 30-2000+ cells (**Fig. 1A**; **Methods**).

We obtained peripheral blood, spleen and bone marrow samples from four individuals aged 27-81 years, as well as tonsillar tissue from two four-year old children and cord blood from one neonate (**Table S1**). All individuals studied were hematopoietically normal and healthy; one subject had a history of inflammatory bowel disease (Crohn's disease) treated with azathioprine and the two tonsil donors had a history of tonsillitis. We focused on four main classes of lymphocytes: naive B lymphocytes, memory B lymphocytes, CD4+ and CD8+ naive T lymphocytes, and CD4+ and CD8+ memory T lymphocytes (**Fig. S1**). In one subject we also expanded T-regulatory cells. From five of the subjects, we isolated and expanded hematopoietic stem and

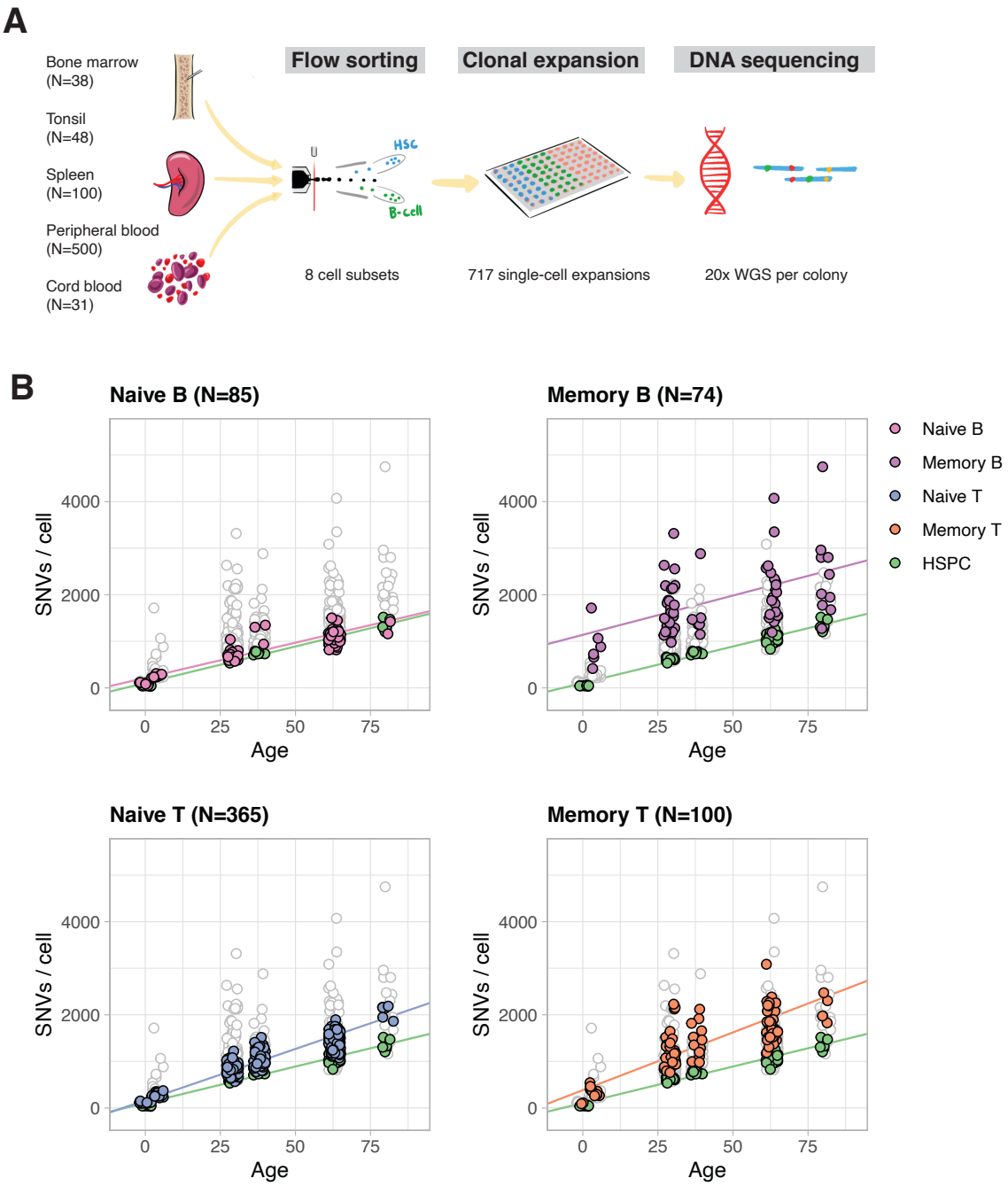

Fig. 1. Experimental design and lymphocyte mutation burden with age. (A) Schematic of the experimental design. (B) SNV mutation burden per genome for the four main lymphocyte subsets (colored points), compared with HSPCs (green points). Each panel has all genomes plotted underneath in white with grey outline.

progenitor cells (HSPC) to provide a baseline for comparison of mutation burden and signatures in the lymphocytes. We performed whole genome sequencing to an average depth of ~20x, and called somatic mutations using standard, benchmarked bioinformatic algorithms. Average telomere lengths were also estimated from the sequencing data. The final dataset analyzed here comprises 717 whole genomes (**Table S2**).

High mutation burden of memory lymphocytes and an increased mutation rate in T cells

The overall burden of both single nucleotide variants (SNVs) and insertion/deletions (indels) per cell varied extensively across the dataset, influenced predominantly by age and cell type (**Fig. 1B**), which we quantified with linear mixed effects models. The burden of base substitutions increased linearly with age across all cell types, but the rate of mutation accumulation differed across cell types ($p=1x10^{-4}$ for age-cell type interaction; linear mixed effects model). HSPCs accumulated mutations at ~16 mutations/cell/year ($CI_{95\%}$=13-19), similar to previous estimates (*15*, *16*). Naive and memory B cells showed broadly similar rates of mutation accumulation (naive B: 15 mutations/cell/year, $CI_{95\%}$=12-18; memory B cells: 17 mutations/cell/year, $CI_{95\%}$=6-28). T cells, though, had higher mutation rates (naive T: 22 mutations/cell/year, $CI_{95\%}$=19-25; memory T cells: 25 mutations/cell/year, $CI_{95\%}$=17-32). Overall, this suggests that there are clock-like mutational processes adding mutations at steady rates, with different rates in each lymphocyte subset.

Additionally, there was a significant increase in the burden of base substitutions in lymphocytes that could not be explained by age, especially for memory lymphocytes. Compared to HSPCs, naive B and T lymphocytes had an average of 110 ($CI_{95\%}$=5-216) and 59 ($CI_{95\%}$=-35-153) extra SNVs per cell, respectively, beyond the effects of age. Memory B and T lymphocytes had an even more pronounced excess of mutations, carrying an average of 1034 ($CI_{95\%}$=604-1465) and 277 ($CI_{95\%}$=5-549) more mutations than HSPCs respectively. Regulatory T cells were similar to memory T cells in mutation burden. This extra burden of base substitutions presumably represents variants acquired during differentiation: approximately one hundred from HSPC to naive lymphocyte and hundreds to thousands from naive to memory lymphocyte.

While these estimates show that the population average for mutation burden increases with lymphocyte differentiation, we found that the variance in mutation burden across cells also massively increased with differentiation. Thus, compared to a standard deviation of 70 SNVs/cell for HSPCs within a given individual, the values estimated for memory B and T lymphocytes were 820 SNVs/cell and 592 SNVs/cell respectively ($p<10^{-16}$ for heterogeneity of variance across cell types).

Indels showed similar patterns, accumulating at an average of 0.7/cell/year in HSPCs ($CI_{95\%}$=0.5-0.9; p=0.02 for age-cell type interaction) with lymphocytes, especially memory B and T cells, carrying an excess burden compared to hematopoietic stem cells (**Fig. S2**).

Driver mutations, defined as those under positive selection, are frequently present in ageing normal tissues (*17–20*), including blood (*21–25*). We did not observe any significant enrichment of non-synonymous variants in driver genes associated with age-related clonal hematopoiesis or lymphoma in our normal lymphocytes, however.

Mutational signatures in B and T lymphocytes are distinct from HSPCs

In order to determine whether the excess mutations observed in lymphocyte subsets were due to a specific mutational process, we extracted mutational signatures across lymphocyte compartments (**Fig. 2**). Like HSPCs, the vast majority of mutations in naive B and T cells were derived from two mutational signatures. One of these, SBS1, is caused by spontaneous

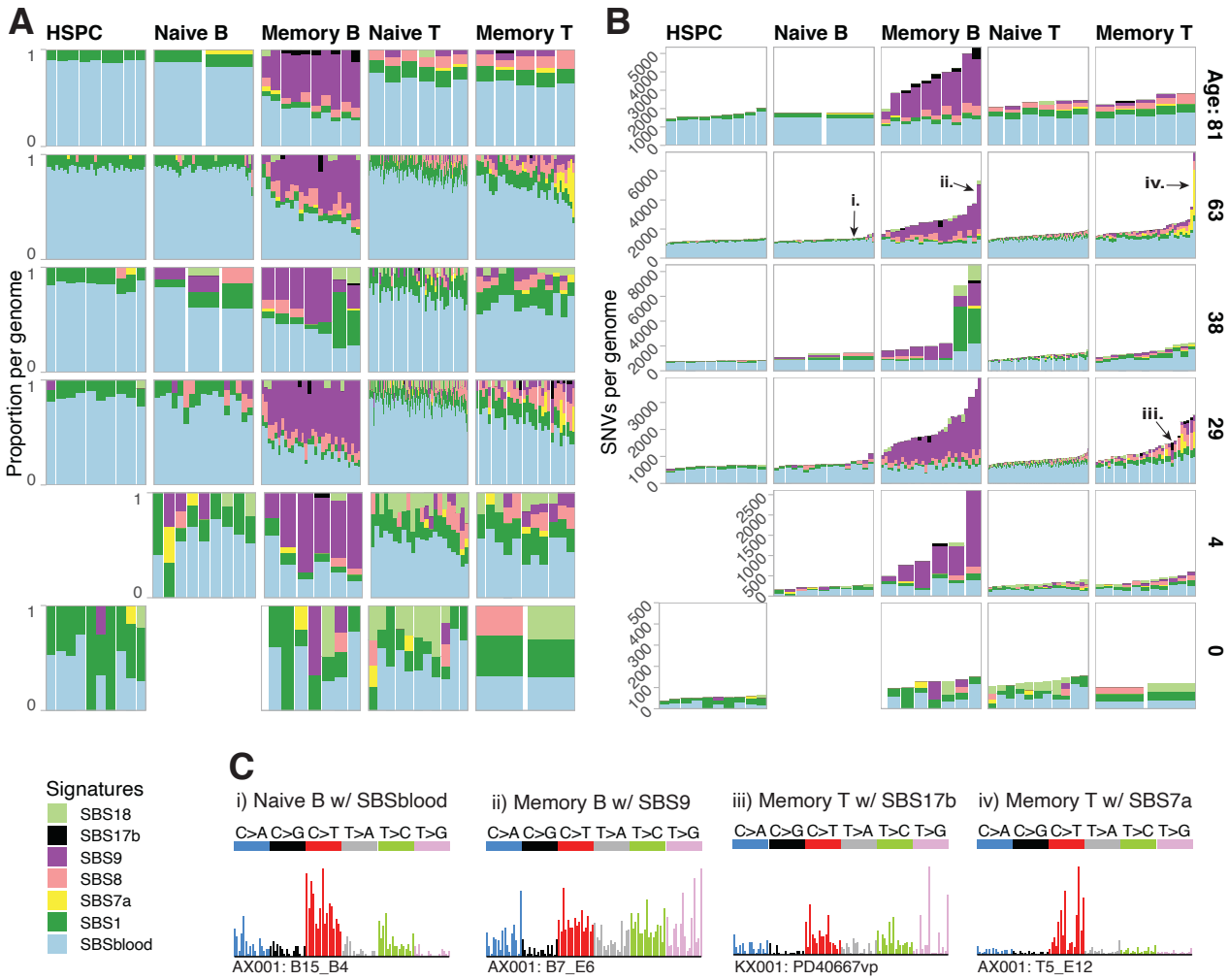

Fig. 2. Mutational processes in lymphocytes. (A) The proportion of SNVs and (B) SNV burden per mutational signature. Each column represents one genome. Signatures are identified by the programs *hdp* and *sigprofiler* and attribution per genome is performed by *sigfit*. Per genome, signatures with a 90% CI lower bound of less than 1% are excluded from plotting. (C) Mutational spectra of single genomes enriched in the specified mutational signature. The specific genome plotted is identified with the corresponding roman numeral in panel (B). The trinucleotide contexts of the mutations, which compose the x-axis, are ordered as in Fig. S3.

deamination of methylated cytosines (*26*, *27*), and accounted for 14% of mutations in HSPCs and naive B and T lymphocytes. Nearly all remaining somatic mutations in these cellular compartments had the typical signature of endogenous mutations in HSPCs (*15*, *16*), which we term 'SBSblood' (**Fig. S3**). The burden of both signatures correlated linearly with age (**Fig. S4**), suggesting that they represent clock-like endogenous mutational processes.

For memory B and T lymphocytes, the absolute numbers of mutations attributed to these two endogenous signatures were broadly similar to those seen in naive B and T lymphocytes (**Fig. 2**). The hundreds to thousands of extra mutations seen in memory B and T lymphocytes derived from additional mutational signatures: SBS7a, SBS8, SBS9, and SBS17b. While signatures SBS8 and SBS9 show correlations with age, SBS7 and SBS17a do not, consistent with them being

sporadic (**Fig. S4**). SBS7a and SBS17b likely represent exogenous mutational processes, discussed in the next section, while SBS9 is differentiation-associated, discussed thereafter.

Mutational signatures as clues to historical tissue residency of circulating lymphocytes

SBS7 is the canonical signature of ultraviolet light damage, the predominant mutational process in melanoma (*28*), normal skin (*18*, *29*) and mycosis fungoides (*30*, *31*), a T-cell lymphoma derived from skin-resident memory T cells (*32*). The signature we extracted in memory lymphocytes matches the features of SBS7, with a predominance of C>T substitutions in a dipyrimidine context, transcriptional strand bias and a high rate of CC>TT dinucleotide substitutions (**Fig. S5**). We found a substantial contribution of SBS7 (>10% of mutations; mean=757/cell, range 205-2783) and CC>TT dinucleotide substitutions in 9/100 memory T cells. Interestingly, memory lymphocytes with high SBS7 had significantly shorter telomeres than other memory T cells (p=0.01, Fisher's method; **Fig. S6**), indicative of increased proliferation. UVB radiation, the most mutagenic, only penetrates human skin to a depth of 10-50µm (*33*), which suggests that these memory T lymphocytes were skin-resident at some stage during their life. That such a high fraction (9/100) of memory T lymphocytes from peripheral blood exhibit a skin-specific mutational process suggests that skin-resident T lymphocytes represent a large and dynamic population, frequently recirculating via the blood system (*34*, *35*).

A second unexpected signature in memory lymphocytes was SBS17. This signature has been observed in cancers of the stomach and esophagus (*36*) and occasionally in B (*36*) and T cell lymphomas (*37*). This signature, characterized by T>G mutations in a TpT context, accounted for >10% of mutations (4SD above mean) in 3/74 memory B and 1/100 memory T lymphocytes. SBS17 has been linked to 5-flurouracil chemotherapy in metastatic cancers (*38*, *39*), but its occurrence in esophageal and gastric cancers (as well as our samples here) is independent of treatment. If its incidence in upper gastrointestinal tract cancers is caused by some unknown local mutagen, then the presence of SBS17 in memory lymphocytes may again represent evidence of a specific microenvironmental exposure associated with tissue residency in the gastric and/or esophageal mucosa.

Signatures of the germinal center reaction

The extensive variation in memory B cell mutation burden is primarily explained by mutational signature SBS9, which accounts for 42% of mutations (mean, 780 mutations/cell), at times tripling the baseline mutation burden. This signature is characterized by mutations at A:T base-pairs, especially T>G in a TpW context, and has been found in both healthy (*13*) and malignant post-germinal center B cells (*6*–*9*). As reported for lymphoid malignancies (*8*, *40*, *41*), we found that SBS9 has a different spectrum to that of somatic hypermutation at immunoglobulin loci (**Fig. 3A**) and that the two mutational signatures have very different distributions across the genome (**Fig. S7**).

In our normal memory B lymphocytes, the number of SBS9 mutations genome-wide showed a strong linear correlation with the number of mutations in the productive V(D)J rearrangement of

the *IGHV* gene, despite their different spectra (**Fig. 3B**). Strikingly, 57% of the variation in the number of SBS9 mutations in memory B cells genome-wide could be explained by the number of mutations in *IGHV* (R^2=0.57, p=4x10^{-9}, linear regression). The density of mutations was 270,000-fold greater at the *IGHV* locus than for SBS9 mutations genome-wide, confirming the precise targeting of somatic hypermutation to antibody regions. Nonetheless, the genome is large, and even this high degree of mutational targeting means that every 1 on-target IGV mutation is accompanied by an average of 18 SBS9 mutations elsewhere in the genome.

Another feature of the germinal center reaction is increased telomerase activity in B cells (*42*, *43*). We therefore estimated telomere lengths from the genome sequencing data for our dataset. Whereas HSPCs and other lymphocyte subsets showed tightly clustered telomere lengths and the expected decrease with aging, telomere lengths in memory B cells were longer, more variable and actually increased with age (excluding tonsil samples; R^2=0.13, p=3x10^{-3}, linear regression; **Fig. S8**). Telomere lengths also correlated linearly with the number of SBS9 mutations genome-wide (excluding tonsil samples; R^2=0.37, p=3x10^{-8}, linear regression; **Fig. 3C**). These data confirm that telomeres do lengthen in the germinal center reaction, and provide further evidence that off-target SBS9 mutations are generated in the germinal center.

A replicative-stress model of SBS9 mutation

The cytosine deaminase AID initiates on-target somatic hypermutation at immunoglobulin loci, which generates damage (and consequent mutation) at C:G base-pairs. On-target mutations at A:T base-pairs during SHM arise through errors introduced during translesion bypass of AID-deaminated cytosines by polymerase-eta (*44–48*), which has an error spectrum weighted towards a TpW context (*49*, *50*). As has been noted in lymphoid malignancies (*8*, *36*, *44*, *51*), off-target SBS9 has a different spectrum from on-target, AID-mediated somatic hypermutation, something we also observe in normal lymphocytes. In particular, SBS9 has a paucity of mutations at C:G base-pairs and an enrichment of T mutations in TpW context (**Fig. 3A**), which makes the role of AID unclear. The genome-wide distribution of off-target AID-induced deamination has been measured directly (*52*), and shows a predilection for highly transcribed regions with active chromatin marks, which tend to be early-replicating.

To explore whether genomic regions with high SBS9 burden show the same distribution, we used general additive models to predict SBS9 burden from 36 genomic features, including gene density, chromatin marks and replication timing across 10kb genome bins. After model selection, 18 features were included in the regression (R^2=0.20; **Fig. 3D**, **Table S3**). Replication timing is by far the strongest predictor, with increased mutation density in late-replicating regions, individually accounting for 17% of the variation in the genomic distribution of SBS9 (**Fig. 3E**). In contrast, replication timing accounted for only 0.6% of variation in density of SBSblood/SBS1 mutations in memory B cells and 0.1% in HSPCs. The next 4 strongest predictors of SBS9 distribution were all broadly related to inactive versus active regions of the genome (distance from CpG islands, gene density, GC content, and LAD density: individual R^2 0.09, 0.07, 0.05, and 0.02, respectively). For each variable, mutation density increased in the direction of less active genomic regions, in contrast to AID-induced deamination, which occurs in actively transcribed regions (*52*).

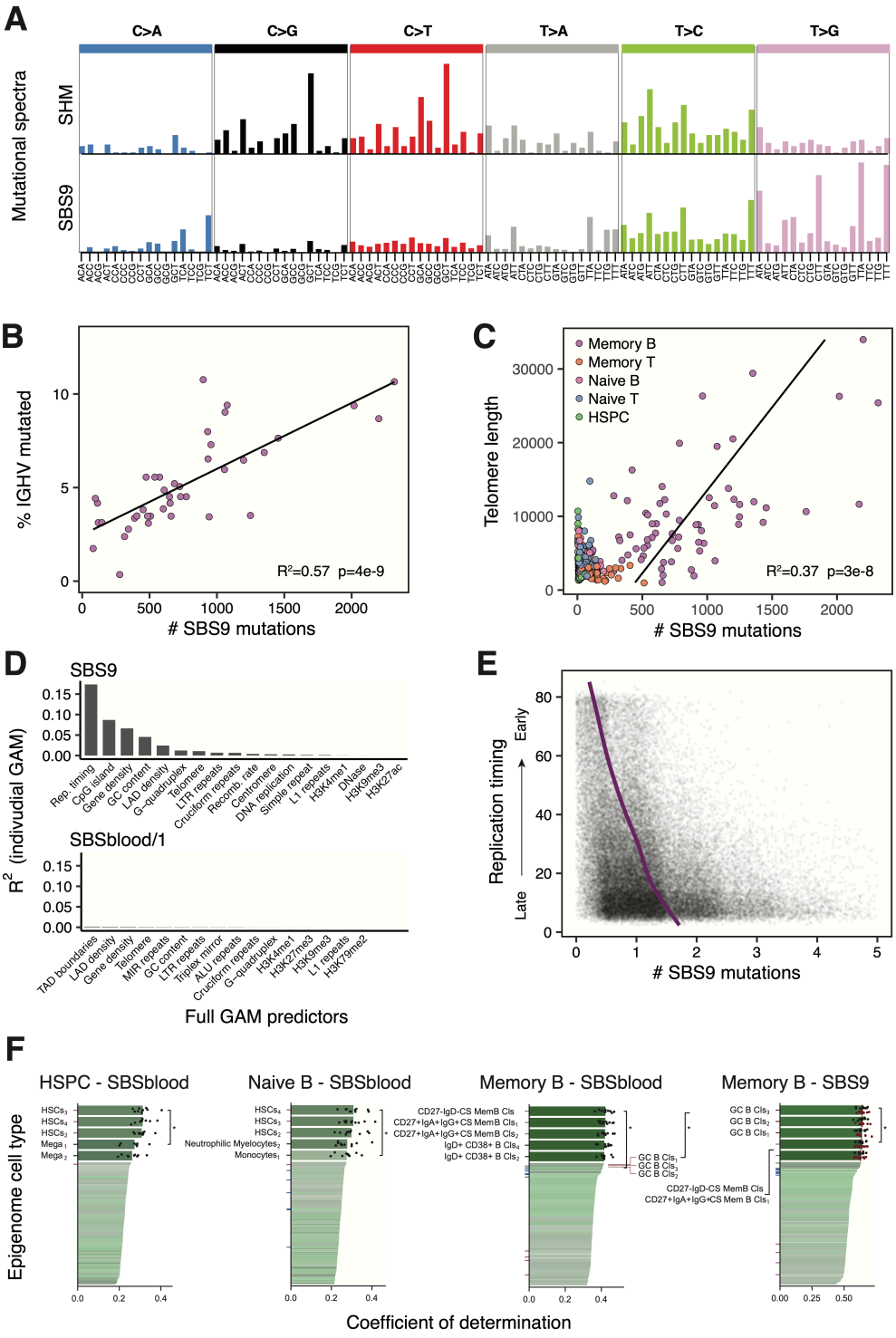

Fig. 3. Correlation of SBS9 with genomic attributes and timing of mutational processes. (A) Mutational spectra of the SBS9 signature and the SHM signature, the latter identified independently by de-novo extraction from 1MB bins of memory B cell mutations. (B) Correlation of SBS9 and the extent of SHM as measured by the proportion of the IGHV mutated in the productive rearrangement of memory B cells. (C) Correlation of SBS9 and telomere length per genome. The regression line is for memory B cells. (D) Explanatory power of each genomic feature found significant in the full GAM model (expressed as the R^2 of the individual GAM model) for predicting number of SBS9 mutations

(top) or number of SBSblood/1 mutations per 10Kb window. (E) Replication timing and the number of SBS9 mutations per 10Kb window. The purple line is the GAM regression prediction. The x-axis is truncated at 5, excluding 0.3% of the data. Points have random noise (-0.5 to 0.5) on the SBS9 mutation measurement to facilitate visualization. (F) Performance of prediction of genome-wide mutational profiles (number of mutations indicated) attributable to particular mutational signatures from histone marks of 149 epigenomes representing distinct blood cell types and different phases of development (subscripts indicate replicates); ticks are colored according to the epigenetic cell type (purple, HSC; blue, naive B cell; grey, memory B cell; maroon, GC B cell); black points depict values from ten-fold cross validation; p-values were obtained for the comparison of the 10-fold cross validation values using the two-sided Wilcoxon test (Cls, cells; CS, class switched; GC, germinal center; HSC, hematopoietic stem cell; Mem, memory). To compare signatures in memory B cells, we trained models on 33,950 mutations with the highest SBS9 probability (maroon dots) and found that the models using germinal center B cells epigenomes were significantly better than for an equally sized set of SBSblood mutations ($p=1.1x10^{-5}$). Mega: megakaryocyte.

Taken together, our data demonstrate that SBS9 accumulates during the germinal center reaction, evidenced by its tight correlation with both on-target SHM and telomere lengthening. However, the relative sparsity of mutations at C:G base-pairs and the distribution of SBS9 to late-replicating, repressed regions of the genome make it difficult to argue that AID is involved. Instead, we hypothesize that SBS9 arises from polymerase-eta bypass of other background DNA lesions induced by the high levels of replicative and oxidative stress experienced by germinal center B cells. Normally, mismatch repair and other pathways would accurately correct such lesions, but the high expression of polymerase-eta in germinal center cells (*53*) provides the opportunity for error-prone translesion bypass to compete. The enrichment of SBS9 in late-replicating, gene-poor, repressed regions of the genome, regions where mismatch repair is typically less active (*54*, *55*), would be consistent with this as a model of SBS9 mutation. The shift from T>C transitions seen in the SHM signature (*36*, *44*, *56*) and polymerase eta in vitro spectrum (*49*, *50*), to T>G transversions in SBS9 (**Fig. 3A**) may be related to the strong transversion bias observed in late replication (*57*).

Association with cell-type-specific epigenetic marks reveals timing of mutational processes

Among human cell types, lymphocytes are unusual for passing through functionally distinct, long-lived differentiation stages with on-going proliferative potential. Since variation in mutation density across the genome is shaped by chromatin state, a cell's specific distribution of somatic mutation provides a record of the past epigenetic landscape of its ancestors back to the fertilized egg (*58–60*). We thus hypothesized that the distribution of clock-like signatures will inform on the cell types present in a given cell's ancestral line-of-descent. In contrast, the distribution of sporadic or episodic signatures can inform on the differentiation stage exposed to that particular mutational process.

We used Random Forest regression to model the distribution of somatic mutations across the genome compared to 149 epigenomes representing 48 distinct blood cell types and differentiation stages (*61–63*). We found that mutations resulting from the clock-like signature SBSblood in HSPCs correlated best with histone marks from hematopoietic stem cells (p=0.002, Wilcoxon test; **Fig. 3F**), consistent with mutation accumulation in undifferentiated cells. Notably, SBSblood mutational profiles in naive B cells also correlated better with the epigenomes of

hematopoietic stem cells than naive B cells (p=0.004; **Fig. 3F**). This implies that the majority of SBSblood mutations in naive B cells were acquired pre-differentiation, consistent with on-going production of these cells from the HSPC compartment throughout life and a relatively short-lived naive B differentiation state. In contrast, SBSblood mutations in naive T cells mapped best to the epigenomes of $CCR7^{+}/CD45RO^{-}/CD25^{-}/CD235^{-}$ naive T cells (p=0.049; **Fig. S9**), consistent with a long-lived, numerically predominant pool of naive T cells generated in the thymus during early life. For memory B cells, SBSblood most closely correlated with histone marks from that cell type and not earlier differentiation stages (p=0.02; **Fig. 3F**), suggesting that the majority of their lineage has been spent as a memory B cell.

For the sporadic mutational processes, SBS9 mutations most closely correlated with germinal center B cell epigenomes (p=0.049; **Fig. 3F**). This is consistent with our finding of a correlation between SBS9 and other germinal center-associated processes (SHM and telomere lengthening), providing further evidence that SBS9 arises as a by-product of the germinal center reaction. For SBS7, the signature of ultraviolet light exposure seen in memory T cells, the genomic distribution more tightly associated with epigenomes of differentiated T cells than naive T cells (**Fig. S9**), supporting the hypothesis that SBS7 mutations accumulate in differentiated T cells.

V(D)J recombination (but not CSR) machinery is associated with off-target structural variants

Both V(D)J recombination and class-switch recombination (CSR) generate large deletions in the Ig/TCR gene regions during lymphocyte maturation. This programmed genome engineering is associated with off-target structural variation (SV) in human lymphoid malignancies (*4*, *5*, *64*) and murine cell models (*65*), and genomic analysis of associated binding motifs suggests substantial mutagenic potential (*66*); however, rates and patterns of SVs have not been studied in normal human lymphocytes.

We found 1037 SVs across 635 lymphocytes, of which 85% occurred in Ig/TCR regions, consistent with 1-2 V(D)J recombination events per lymphocyte and 0-2 class-switch recombination events per memory B cell. Excluding Ig/TCR gene regions, B and T lymphocytes carried more SVs than HSPCs, with 103/609 (17%) of lymphocytes having at least one off-target SV (compared to a single SV in 82 HSPCs; $p=9x10^{-5}$, Fisher exact test). Memory B and T lymphocytes had higher non-Ig/TCR SV burdens than their respective naive subsets (27% memory B versus 5% naive B cells; 25% memory T versus 15% naive T cells; $p=1x10^{-5}$). Although we saw occasional instances of more complex abnormalities, including chromoplexy (*67*) (**Fig. 4A**) and cycles of templated insertions (*64*), most non-Ig/TCR SVs were deletions (49%), several of which affected genes mutated in lymphoid malignancies (**Fig. 4B**).

V(D)J recombination is mediated by RAG1 and RAG2 cutting at an 'RSS' DNA motif comprising a heptamer and nonamer with intervening spacer. 24% of non-Ig/TCR and 96% of Ig/TCR SVs had a full RSS motif or the heptamer within 50bp of a breakpoint (**Fig. 4C-D**). Taking into account the baseline occurrence of these motifs using genomic controls, we estimate that 12% of non-Ig/TCR and 84% of Ig/TCR SVs were RAG-mediated, especially deletions (~15% of non-Ig/TCR deletions). As expected, the RSS motif was typically internal to the breakpoint (62% and 91% for non-Ig/TCR and Ig/TCR SVs). We observed a rapid decay in the enrichment of RAG motifs with distance from

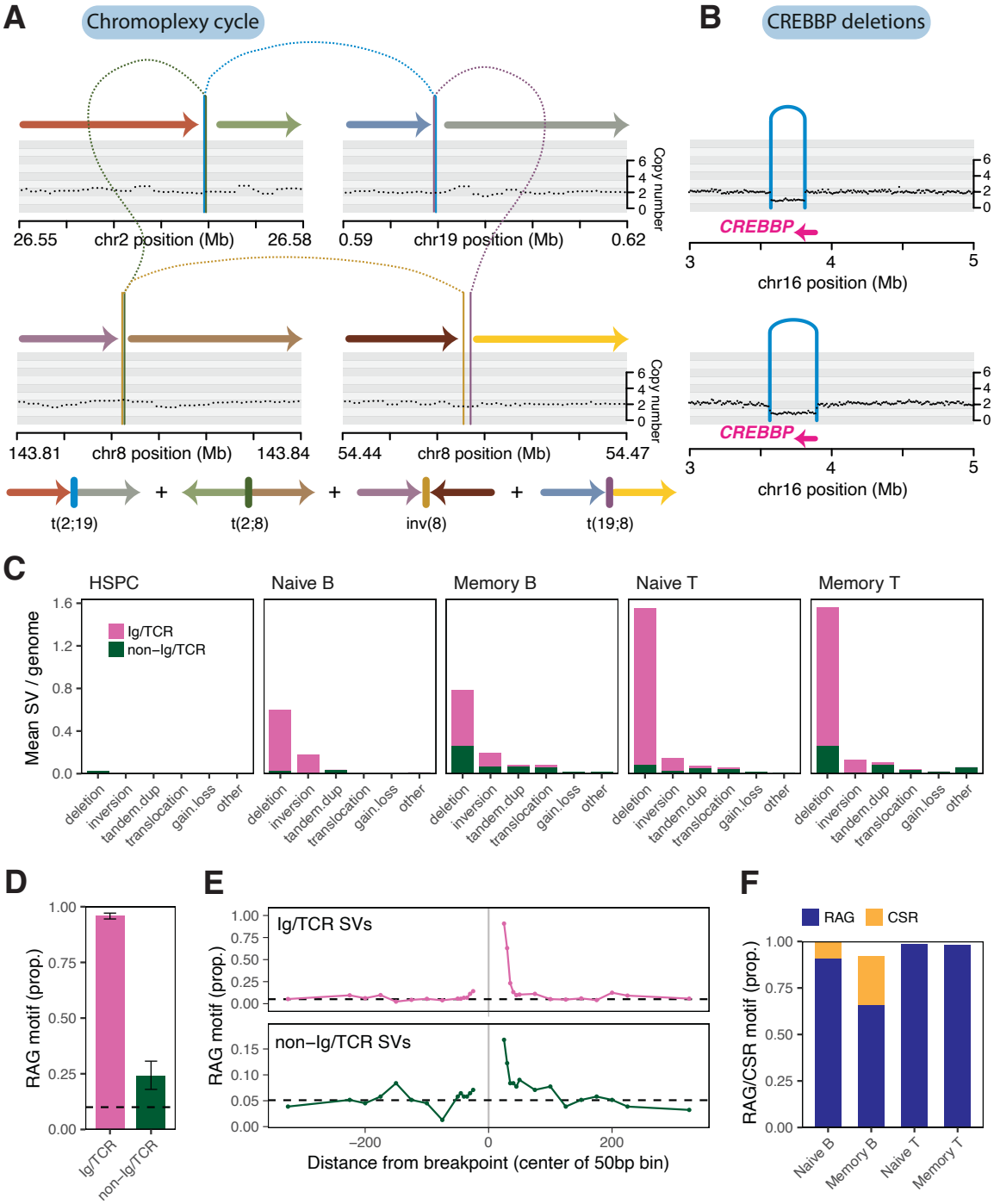

Fig. 4. Structural variation burden and off-target RAG-mediated deletion. (A) Chromoplexy cycle (sample PD40667sl, donor KX002). (B) CREBBP deletions (samples PD40521po, donor KX001 and BMH1_PlateB1_E2, donor AX001). (C) Burden of structural variants per cell type. (D) The proportion of deletions with an RSS (RAG) motif within 50bp of the breakpoint. The black dashed line represents the genomic background rate of RAG motifs. Error bars represent 95% bootstrap confidence intervals. (E) The proportion of deletions with an RSS (RAG) motif as a function of distance from the breakpoint, with a positive distance representing bases interior to the deletion, and a negative value representing bases exterior to the breakpoint. The black dashed line represents the genomic background rate of RAG motifs. (F) Proportion of deletions with an RSS (RAG) or switch (CSR) motif.

breakpoints after the first 50bp window, reaching background levels by ~100bp (**Fig. 4E**). During V(D)J recombination, the TdT protein adds random nucleotides at the dsDNA breaks - this also seems to occur in off-target SVs, with RAG-mediated events enriched for insertions of non-templated sequence at the breakpoint (44% and 88% for non-Ig/TCR and Ig/TCR SVs, respectively, versus 21% of off-target SVs without RSS motif; $p=9x10^{-3}$, Fisher exact test).

CSR is achieved through AID cytosine deamination at WGCW clusters (*68*, *69*), deleting IgH constant region genes and changing the antibody isotype. CSR accounted for the majority of Ig SVs not attributed to RAG; together, RAG and AID accounted for 92% Ig SVs in memory B cells and 100% in naive B cells (**Fig. 4F**). In contrast, none of the non-Ig/TCR SVs had CSR AID motif clusters, suggesting that class-switch recombination is exquisitely targeted.

Comparison with malignancy

A long-standing controversy in cancer modelling is whether tumors require additional mutational processes to acquire sufficient driver mutations for oncogenic transformation (*70*). In many solid tissues, cancers do have higher mutation burdens than normal cells from the same organ (*20*, *71*, *72*), but myeloid leukemias do not (*14*). To address this question in lymphoid malignancies, we compared our normal B and T lymphocytes to 7 blood cancers (*51*, *73*). SNV burdens for follicular lymphoma, DLBC lymphoma and multiple myeloma were considerably higher than normal lymphocytes (**Fig. 5A-B**). In contrast, point mutation burdens observed in Burkitt lymphoma, mutated CLL, unmutated CLL and AML were well within the range of normal lymphocytes. In contrast, all lymphoid malignancies showed higher rates of SV than normal cells.

The elevated point mutation burden could arise from increased activity of mutational processes already present in normal cells, or the emergence of distinct, cancer-specific mutational processes. The vast majority of mutations present across all B-cell malignancies could be attributed to the same mutational processes active in normal memory B cells, and at broadly similar proportions (**Fig. 5C-E**). This suggests that the elevated SNV mutation burdens of DLBC lymphoma, follicular lymphoma and multiple myeloma are due to globally increased activity of the same mutational processes active in normal lymphocytes, rather than enrichment of any single process or novel, cancer-restricted DNA repair defects. This is different to, say, colorectal cancer or breast cancer, where the elevated mutation burden is caused by emergence of cancer-specific mutational processes (*72*, *74*).

A feature of somatic mutations in B-cell lymphomas is clustering of off-target somatic hypermutation in highly expressed genes. For both SBS9 (**Fig. 5F**) and off-target SHM mutations (**Fig. 5G**), we found considerable overlap in genes with elevated mutation rates. For example, *BCL6*, *BCL7A* and *PAX5* had enrichment of mutations with the SHM signature in both normal and post-germinal malignant lymphocytes. Likewise, of the 100 genes most enriched for SBS9 in normal memory B cells, 64% were also SBS9-enriched (top 1%) in ≥3 of the 5 post-germinal malignancies.

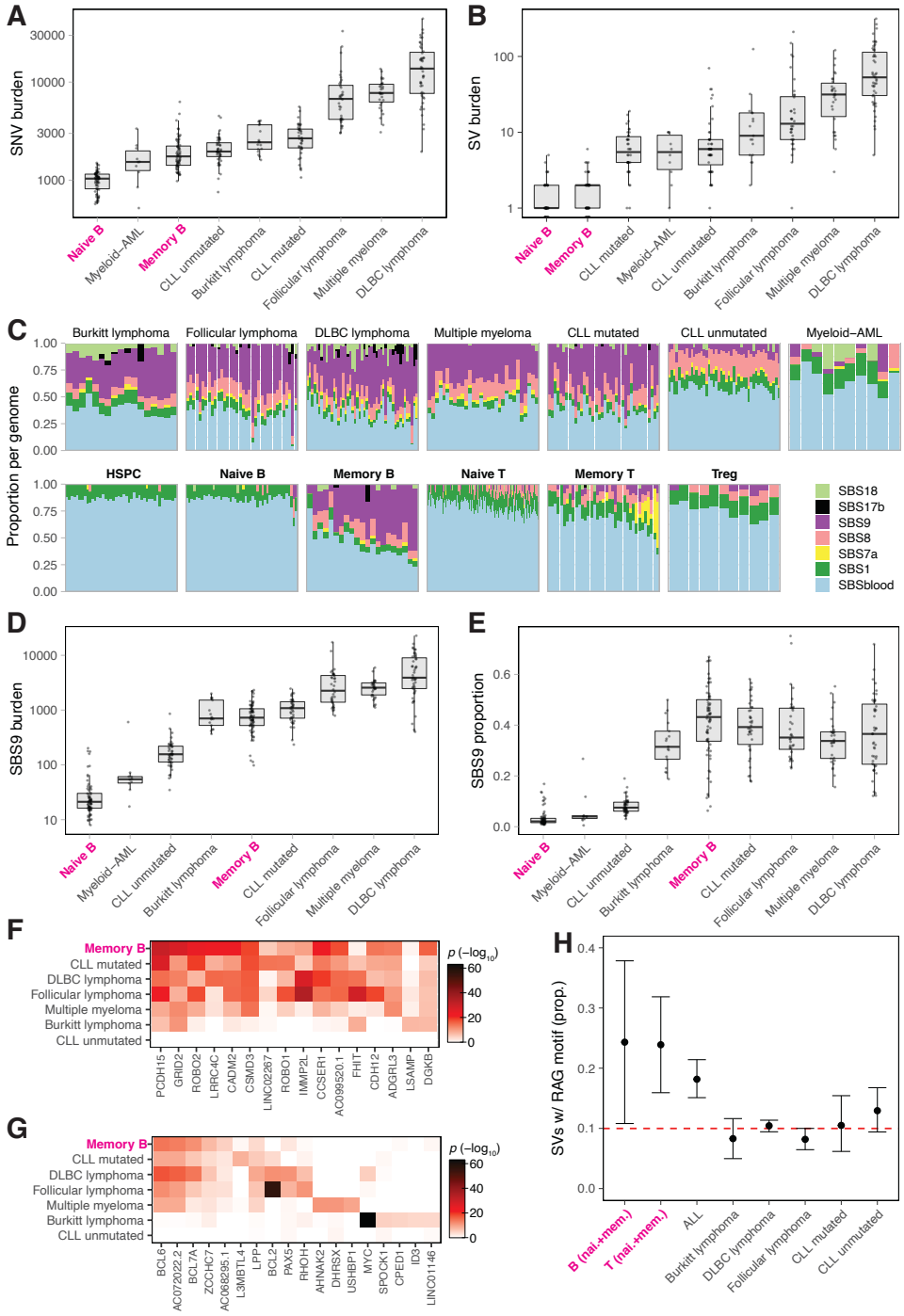

Fig. 5. Comparison of mutational patterns with malignancy. (A) SNV and (B) SV burden by cell/malignancy type. Boxes show the interquartile range and the center horizontal lines show the median. Each genome is plotted as one point. (C) Proportion of mutational signatures per genome. Signatures are identified by the programs *hdp* and *sigprofilier* and attribution per genome is performed by *sigfit*. Per genome, signatures with a 90% CI lower bound of less than 1% are excluded from plotting. (D) SBS9 burden and (E) proportion by cell/malignancy type. (F) SBS9 and G) SHM signature enrichment per gene. (H) Proportion of SV's with RSS (RAG) motifs within 50bp of a breakpoint. (A-E,H) Normal lymphocytes (bold) exclude pediatric samples.

About 10% of normal lymphocytes have a non-Ig/TCR RAG-mediated SV, accounting for 24% of off-target rearrangements. Across lymphoid malignancies, acute lymphoblastic leukemias (*4*) had similarly high proportions of RAG-mediated events, but in much higher numbers, as reported previously (*4*, *5*) (**Fig. 5H**). For other lymphoid malignancies, although the proportions were low, the absolute numbers of RAG-mediated SVs (≥0.5/lymphoma) were broadly comparable to those seen in normal lymphocytes (**Fig. S10**). This suggests that malignant transformation of lymphocytes is associated with the emergence of cancer-specific genomic instability, generating a genome with considerably more large-scale rearrangement.

Conclusions

Unique among human cell types, a lymphocyte experiences long periods of its life in diverse microenvironments, be it bone marrow, thymus, lymph node, skin or mucosa. These stages are interspersed with short-lived bursts of differentiation, each of which is associated with proliferation and/or programmed genome engineering to improve antigen recognition. Our data show that each phase of lymphocyte differentiation contributes an additional burden of mutations - naive B and T lymphocytes have ~100 more mutations than hematopoietic stem cells; memory lymphocytes have hundreds to thousands more than naive lymphocytes. The signatures of these mutations reflect both the unintended by-products of immunological diversification and exposure to exogenous mutagens; their genomic distribution reflects the chromatin landscape of the cell at the time the mutational process was active. The rare lymphocyte that transforms to cancer draws its stock of somatic mutations from the same mutational processes that are active in normal lymphocytes.

Code availability

An exhaustive repository of code for statistical analyses reported in this manuscript is available at https://github.com/machadoheather/lymphocyte_somatic_mutation

References

1. F. W. Alt, Y. Zhang, F.-L. Meng, C. Guo, B. Schwer, Mechanisms of Programmed DNA Lesions and Genomic Instability in the Immune System. *Cell*. **152**, 417–429 (2013).
2. D. Tarlinton, K. Good-Jacobson, Diversity Among Memory B Cells: Origin, Consequences, and Utility. *Science*. **341**, 1205–1211 (2013).
3. M. Nishana, S. C. Raghavan, Role of recombination activating genes in the generation of antigen receptor diversity and beyond. *Immunology*. **137**, 271–281 (2012).
4. E. Papaemmanuil, I. Rapado, Y. Li, N. E. Potter, D. C. Wedge, J. Tubio, L. B. Alexandrov, P. Van Loo, S. L. Cooke, J. Marshall, I. Martincorena, J. Hinton, G. Gundem, F. W. van Delft, S. Nik-Zainal, D. R. Jones, M. Ramakrishna, I. Titley, L. Stebbings, C. Leroy, A. Menzies, J. Gamble, B. Robinson, L. Mudie, K. Raine, S. O'Meara, J. W. Teague, A. P. Butler, G. Cazzaniga, A. Biondi, J. Zuna, H. Kempski, M. Muschen, A. M. Ford, M. R. Stratton, M. Greaves, P. J. Campbell, RAG-mediated recombination is the predominant driver of oncogenic rearrangement in *ETV6-RUNX1* acute lymphoblastic leukemia. *Nat. Genet.* **46**, 116–125 (2014).
5. C. G. Mullighan, L. A. Phillips, X. Su, J. Ma, C. B. Miller, S. A. Shurtleff, J. R. Downing, Genomic Analysis of the Clonal Origins of Relapsed Acute Lymphoblastic Leukemia. *Science*. **322**, 1377–1380 (2008).

6. L. Pasqualucci, P. Neumeister, T. Goossens, G. Nanjangud, R. S. K. Chaganti, R. Küppers, R. Dalla-Favera, Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas. *Nature*. **412**, 341–346 (2001).
7. X. S. Puente, M. Pinyol, V. Quesada, L. Conde, G. R. Ordóñez, N. Villamor, G. Escaramis, P. Jares, S. Beà, M. González-Díaz, L. Bassaganyas, T. Baumann, M. Juan, M. López-Guerra, D. Colomer, J. M. C. Tubío, C. López, A. Navarro, C. Tornador, M. Aymerich, M. Rozman, J. M. Hernández, D. A. Puente, J. M. P. Freije, G. Velasco, A. Gutiérrez-Fernández, D. Costa, A. Carrió, S. Guijarro, A. Enjuanes, L. Hernández, J. Yagüe, P. Nicolás, C. M. Romeo-Casabona, H. Himmelbauer, E. Castillo, J. C. Dohm, S. de Sanjosé, M. A. Piris, E. de Alava, J. S. Miguel, R. Royo, J. L. Gelpí, D. Torrents, M. Orozco, D. G. Pisano, A. Valencia, R. Guigó, M. Bayés, S. Heath, M. Gut, P. Klatt, J. Marshall, K. Raine, L. A. Stebbings, P. A. Futreal, M. R. Stratton, P. J. Campbell, I. Gut, A. López-Guillermo, X. Estivill, E. Montserrat, C. López-Otín, E. Campo, Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature*. **475**, 101–105 (2011).
8. S. Kasar, J. Kim, R. Improgo, G. Tiao, P. Polak, N. Haradhvala, M. S. Lawrence, A. Kiezun, S. M. Fernandes, S. Bahl, C. Sougnez, S. Gabriel, E. S. Lander, H. T. Kim, G. Getz, J. R. Brown, Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* **6**, 8866 (2015).
9. A. H. Khodabakhshi, R. D. Morin, A. P. Fejes, A. J. Mungall, K. L. Mungall, M. Bolger-Munro, N. A. Johnson, J. M. Connors, R. D. Gascoyne, M. A. Marra, I. Birol, S. J. Jones, Recurrent targets of aberrant somatic hypermutation in lymphoma. *Oncotarget*. **3**, 1308–1319 (2012).
10. B. A. Walker, C. P. Wardell, D. C. Johnson, M. F. Kaiser, D. B. Begum, N. B. Dahir, F. M. Ross, F. E. Davies, D. Gonzalez, G. J. Morgan, Characterization of IGH locus breakpoints in multiple myeloma indicates a subset of translocations appear to occur in pregerminal center B cells. *Blood*. **121**, 3413–3419 (2013).
11. G. J. Liu, L. Cimmino, J. G. Jude, Y. Hu, M. T. Witkowski, M. D. McKenzie, M. Kartal-Kaess, S. A. Best, L. Tuohey, Y. Liao, W. Shi, C. G. Mullighan, M. A. Farrar, S. L. Nutt, G. K. Smyth, J. Zuber, R. A. Dickins, Pax5 loss imposes a reversible differentiation block in B-progenitor acute lymphoblastic leukemia. *Genes Dev.* **28**, 1337–1350 (2014).
12. M. Caganova, C. Carrisi, G. Varano, F. Mainoldi, F. Zanardi, P.-L. Germain, L. George, F. Alberghini, L. Ferrarini, A. K. Talukder, M. Ponzoni, G. Testa, T. Nojima, C. Doglioni, D. Kitamura, K.-M. Toellner, I. -hsin Su, S. Casola, Germinal center dysregulation by histone methyltransferase EZH2 promotes lymphomagenesis. *J. Clin. Invest.* **124**, 1869–1869 (2014).
13. L. Zhang, X. Dong, M. Lee, A. Y. Maslov, T. Wang, J. Vijg, Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. *Proc. Natl. Acad. Sci.* **116**, 9014–9019 (2019).
14. J. S. Welch, T. J. Ley, D. C. Link, C. A. Miller, D. E. Larson, D. C. Koboldt, L. D. Wartman, T. L. Lamprecht, F. Liu, J. Xia, C. Kandoth, R. S. Fulton, M. D. McLellan, D. J. Dooling, J. W. Wallis, K. Chen, C. C. Harris, H. K. Schmidt, J. M. Kalicki-Veizer, C. Lu, Q. Zhang, L. Lin, M. D. O'Laughlin, J. F. McMichael, K. D. Delehaunty, L. A. Fulton, V. J. Magrini, S. D. McGrath, R. T. Demeter, T. L. Vickery, J. Hundal, L. L. Cook, G. W. Swift, J. P. Reed, P. A. Alldredge, T. N. Wylie, J. R. Walker, M. A. Watson, S. E. Heath, W. D. Shannon, N. Varghese, R. Nagarajan, J. E. Payton, J. D. Baty, S. Kulkarni, J. M. Klco, M. H. Tomasson, P. Westervelt, M. J. Walter, T. A. Graubert, J. F. DiPersio, L. Ding, E. R. Mardis, R. K. Wilson, The Origin and Evolution of Mutations in Acute Myeloid Leukemia. *Cell*. **150**, 264–278 (2012).
15. H. Lee-Six, N. F. Øbro, M. S. Shepherd, S. Grossmann, K. Dawson, M. Belmonte, R. J. Osborne, B. J. P. Huntly, I. Martincorena, E. Anderson, L. O'Neill, M. R. Stratton, E. Laurenti, A. R. Green, D. G. Kent, P. J. Campbell, Population dynamics of normal human blood inferred from somatic mutations. *Nature*. **561**, 473 (2018).
16. F. G. Osorio, A. Rosendahl Huber, R. Oka, M. Verheul, S. H. Patel, K. Hasaart, L. de la Fonteijne, I. Varela, F. D. Camargo, R. van Boxtel, Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. *Cell Rep.* **25**, 2308-2316.e4 (2018).
17. K. Yizhak, F. Aguet, J. Kim, J. M. Hess, K. Kübler, J. Grimsby, R. Frazer, H. Zhang, N. J. Haradhvala, D. Rosebrock, D. Livitz, X. Li, E. Arich-Landkof, N. Shoresh, C. Stewart, A. V. Segrè, P. A. Branton, P. Polak, K. G. Ardlie, G. Getz, RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science*. **364**, eaaw0726 (2019).
18. I. Martincorena, A. Roshan, M. Gerstung, P. Ellis, P. Van Loo, S. McLaren, D. C. Wedge, A. Fullam, L. B. Alexandrov, J. M. Tubio, others, High burden and pervasive positive selection of somatic mutations in normal

human skin. *Science*. **348**, 880–886 (2015).

19. A. Yokoyama, N. Kakiuchi, T. Yoshizato, Y. Nannya, H. Suzuki, Y. Takeuchi, Y. Shiozawa, Y. Sato, K. Aoki, S. K. Kim, Y. Fujii, K. Yoshida, K. Kataoka, M. M. Nakagawa, Y. Inoue, T. Hirano, Y. Shiraishi, K. Chiba, H. Tanaka, M. Sanada, Y. Nishikawa, Y. Amanuma, S. Ohashi, I. Aoyama, T. Horimatsu, S. Miyamoto, S. Tsunoda, Y. Sakai, M. Narahara, J. B. Brown, Y. Sato, G. Sawada, K. Mimori, S. Minamiguchi, H. Haga, H. Seno, S. Miyano, H. Makishima, M. Muto, S. Ogawa, Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature*. **565**, 312–317 (2019).
20. K. Yoshida, K. H. C. Gowers, H. Lee-Six, D. P. Chandrasekharan, T. Coorens, E. F. Maughan, K. Beal, A. Menzies, F. R. Millar, E. Anderson, S. E. Clarke, A. Pennycuick, R. M. Thakrar, C. R. Butler, N. Kakiuchi, T. Hirano, R. E. Hynds, M. R. Stratton, I. Martincorena, S. M. Janes, P. J. Campbell, Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature*. **578**, 266–272 (2020).
21. C. J. Watson, A. L. Papula, G. Y. P. Poon, W. H. Wong, A. L. Young, T. E. Druley, D. S. Fisher, J. R. Blundell, The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science*. **367**, 1449–1454 (2020).
22. S. Jaiswal, P. Fontanillas, J. Flannick, A. Manning, P. V. Grauman, B. G. Mar, R. C. Lindsley, C. H. Mermel, N. Burtt, A. Chavez, J. M. Higgins, V. Moltchanov, F. C. Kuo, M. J. Kluk, B. Henderson, L. Kinnunen, H. A. Koistinen, C. Ladenvall, G. Getz, A. Correa, B. F. Banahan, S. Gabriel, S. Kathiresan, H. M. Stringham, M. I. McCarthy, M. Boehnke, J. Tuomilehto, C. Haiman, L. Groop, G. Atzmon, J. G. Wilson, D. Neuberg, D. Altshuler, B. L. Ebert, Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *http://dx.doi.org/10.1056/NEJMoa1408617* (2014), , doi:10.1056/NEJMoa1408617.
23. M. Xie, C. Lu, J. Wang, M. D. McLellan, K. J. Johnson, M. C. Wendl, J. F. McMichael, H. K. Schmidt, V. Yellapantula, C. A. Miller, B. A. Ozenberger, J. S. Welch, D. C. Link, M. J. Walter, E. R. Mardis, J. F. Dipersio, F. Chen, R. K. Wilson, T. J. Ley, L. Ding, Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* **20**, 1472–1478 (2014).
24. G. Genovese, A. K. Kähler, R. E. Handsaker, J. Lindberg, S. A. Rose, S. F. Bakhoum, K. Chambert, E. Mick, B. M. Neale, M. Fromer, S. M. Purcell, O. Svantesson, M. Landén, M. Höglund, S. Lehmann, S. B. Gabriel, J. L. Moran, E. S. Lander, P. F. Sullivan, P. Sklar, H. Grönberg, C. M. Hultman, S. A. McCarroll, Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
25. L. Busque, J. P. Patel, M. E. Figueroa, A. Vasanthakumar, S. Provost, Z. Hamilou, L. Mollica, J. Li, A. Viale, A. Heguy, M. Hassimi, N. Socci, P. K. Bhatt, M. Gonen, C. E. Mason, A. Melnick, L. A. Godley, C. W. Brennan, O. Abdel-Wahab, R. L. Levine, Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. *Nat. Genet.* **44**, 1179–1181 (2012).
26. G. P. Pfeifer, Mutagenesis at methylated CpG sequences. *Curr. Top. Microbiol. Immunol.* **301**, 259–281 (2006).
27. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. J. R. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjörd, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinsk, N. Jäger, D. T. W. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. López-Otín, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. J. Tutt, R. Valdés-Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, J. Zucman-Rossi, P. Andrew Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, M. R. Stratton, Signatures of mutational processes in human cancer. *Nature*. **500**, 415–421 (2013).
28. C. Jhappan, F. P. Noonan, G. Merlino, Ultraviolet radiation and cutaneous malignant melanoma. *Oncogene*. **22**, 3099–3112 (2003).
29. L. Wei, S. R. Christensen, M. E. Fitzgerald, J. Graham, N. D. Hutson, C. Zhang, Z. Huang, Q. Hu, F. Zhan, J. Xie, J. Zhang, S. Liu, E. Remenyik, E. Gellen, O. R. Colegio, M. Bax, J. Xu, H. Lin, W. J. Huss, B. A. Foster, G. Paragh, Ultradeep sequencing differentiates patterns of skin clonal mutations associated with sun-exposure status and skin cancer burden. *Sci. Adv.* **7** (2021), doi:10.1126/sciadv.abd7703.
30. L. Y. McGirt, P. Jia, D. A. Baerenwald, R. J. Duszynski, K. B. Dahlman, J. A. Zic, J. P. Zwerner, D. Hucks, U. Dave, Z. Zhao, C. M. Eischen, Whole-genome sequencing reveals oncogenic mutations in mycosis fungoides. *Blood*. **126**, 508–519 (2015).
31. C. L. Jones, A. Degasperi, V. Grandi, T. D. Amarante, T. J. Mitchell, S. Nik-Zainal, S. J. Whittaker, Spectrum of mutational signatures in T-cell lymphoma reveals a key role for UV radiation in cutaneous T-cell lymphoma.

Sci. Rep. **11**, 3962 (2021).
32. J. J. Campbell, R. A. Clark, R. Watanabe, T. S. Kupper, Sézary syndrome and mycosis fungoides arise from distinct T-cell subsets: a biologic rationale for their distinct clinical behaviors. *Blood*. **116**, 767–771 (2010).
33. M. Meinhardt, R. Krebs, A. Anders, U. Heinrich, H. Tronnier, Wavelength-dependent penetration depths of ultraviolet radiation in human skin. *J. Biomed. Opt.* **13**, 044030 (2008).
34. R. A. Clark, B. F. Chong, N. Mirchandani, K.-I. Yamanaka, G. F. Murphy, R. K. Dowgiert, T. S. Kupper, A Novel Method for the Isolation of Skin Resident T Cells from Normal and Diseased Human Skin. *J. Invest. Dermatol.* **126**, 1059–1070 (2006).
35. R. A. Clark, B. Chong, N. Mirchandani, N. K. Brinster, K. Yamanaka, R. K. Dowgiert, T. S. Kupper, The Vast Majority of CLA+ T Cells Are Resident in Normal Skin. *J. Immunol.* **176**, 4431–4439 (2006).
36. L. B. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. W. Tian Ng, Y. Wu, A. Boot, K. R. Covington, D. A. Gordenin, E. N. Bergstrom, S. M. A. Islam, N. Lopez-Bigas, L. J. Klimczak, J. R. McPherson, S. Morganella, R. Sabarinathan, D. A. Wheeler, V. Mustonen, G. Getz, S. G. Rozen, M. R. Stratton, The repertoire of mutational signatures in human cancer. *Nature*. **578**, 94–101 (2020).
37. F. Maura, A. Dodero, C. Carniti, N. Bolli, M. Magni, V. Monti, A. Cabras, D. Leongamornlert, F. Abascal, B. Diamond, B. Rodriguez-Martin, J. Zamora, A. Butler, I. Martincorena, J. M. C. Tubio, P. J. Campbell, A. Chiappella, G. Pruneri, P. Corradini, *CDKN2A* deletion is a frequent event associated with poor outcome in patients with peripheral T-cell lymphoma not otherwise specified (PTCL-NOS). *Haematologica*. **Online ahead of print** (2020), doi:10.3324/haematol.2020.262659.
38. O. Pich, F. Muiños, M. P. Lolkema, N. Steeghs, A. Gonzalez-Perez, N. Lopez-Bigas, The mutational footprints of cancer therapies. *Nat. Genet.* **51**, 1732–1740 (2019).
39. S. Christensen, B. Van der Roest, N. Besselink, R. Janssen, S. Boymans, J. W. M. Martens, M.-L. Yaspo, P. Priestley, E. Kuijk, E. Cuppen, A. Van Hoeck, 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. *Nat. Commun.* **10**, 4571 (2019).
40. F. Maura, A. Degasperi, F. Nadeu, D. Leongamornlert, H. Davies, L. Moore, R. Royo, B. Ziccheddu, X. S. Puente, H. Avet-Loiseau, P. J. Campbell, S. Nik-Zainal, E. Campo, N. Munshi, N. Bolli, A practical guide for mutational signature analysis in hematological malignancies. *Nat. Commun.* **10**, 2969 (2019).
41. N. Bolli, F. Maura, S. Minvielle, D. Gloznik, R. Szalat, A. Fullam, I. Martincorena, K. J. Dawson, M. K. Samur, J. Zamora, P. Tarpey, H. Davies, M. Fulciniti, M. A. Shammas, Y. T. Tai, F. Magrangeas, P. Moreau, P. Corradini, K. Anderson, L. Alexandrov, D. C. Wedge, H. Avet-Loiseau, P. Campbell, N. Munshi, Genomic patterns of progression in smoldering multiple myeloma. *Nat. Commun.* **9** (2018), doi:10.1038/s41467-018-05058-y.
42. N. Weng, L. Granger, R. J. Hodes, Telomere lengthening and telomerase activation during human B cell differentiation. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 10827–10832 (1997).
43. K.-F. Norrback, M. Hultdin, K. Dahlenborg, P. Osterman, R. Carlsson, G. Roos, Telomerase regulation and telomere dynamics in germinal centers. *Eur. J. Haematol.* **67**, 309–317 (2001).
44. F. Supek, B. Lehner, Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. *Cell*. **170**, 534-547.e23 (2017).
45. V. H. Odegard, D. G. Schatz, Targeting of somatic hypermutation. *Nat. Rev. Immunol.* **6**, 573–583 (2006).
46. B. Pilzecker, H. Jacobs, Mutating for Good: DNA Damage Responses During Somatic Hypermutation. *Front. Immunol.* **10** (2019), doi:10.3389/fimmu.2019.00438.
47. T. M. Wilson, A. Vaisman, S. A. Martomo, P. Sullivan, L. Lan, F. Hanaoka, A. Yasui, R. Woodgate, P. J. Gearhart, MSH2–MSH6 stimulates DNA polymerase η, suggesting a role for A:T mutations in antibody genes. *J. Exp. Med.* **201**, 637–645 (2005).
48. J. M. Di Noia, M. S. Neuberger, Molecular Mechanisms of Antibody Somatic Hypermutation. *Annu. Rev. Biochem.* **76**, 1–22 (2007).
49. Y. Chen, T. Sugiyama, NGS-based analysis of base-substitution signatures created by yeast DNA polymerase eta and zeta on undamaged and abasic DNA templates in vitro. *DNA Repair*. **59**, 34–43 (2017).
50. T. Matsuda, K. Bebenek, C. Masutani, F. Hanaoka, T. A. Kunkel, Low fidelity DNA synthesis by human DNA polymerase-η. *Nature*. **404**, 1011–1013 (2000).
51. F. Maura, N. Bolli, N. Angelopoulos, K. J. Dawson, D. Leongamornlert, I. Martincorena, T. J. Mitchell, A. Fullam, S. Gonzalez, R. Szalat, F. Abascal, B. Rodriguez-Martin, M. K. Samur, D. Glodzik, M. Roncador, M. Fulciniti, Y. T. Tai, S. Minvielle, F. Magrangeas, P. Moreau, P. Corradini, K. C. Anderson, J. M. C. Tubio, D. C. Wedge, M. Gerstung, H. Avet-Loiseau, N. Munshi, P. J. Campbell, Genomic landscape and chronological reconstruction of

driver events in multiple myeloma. *Nat. Commun.* **10**, 3835 (2019).

52. Á. F. Álvarez-Prado, P. Pérez-Durán, A. Pérez-García, A. Benguria, C. Torroja, V. G. de Yébenes, A. R. Ramiro, A broad atlas of somatic hypermutation allows prediction of activation-induced deaminase targets. *J. Exp. Med.* **215**, 761–771 (2018).
53. L. J. McHeyzer-Williams, P. J. Milpied, S. L. Okitsu, M. G. McHeyzer-Williams, Class-switched memory B cells remodel BCRs within secondary germinal centers. *Nat. Immunol.* **16**, 296–305 (2015).
54. F. Supek, B. Lehner, Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature*. **521**, 81–84 (2015).
55. J. Frigola, R. Sabarinathan, L. Mularoni, F. Muiños, A. Gonzalez-Perez, N. López-Bigas, Reduced mutation rate in exons due to differential mismatch repair. *Nat. Genet.* **49**, 1684–1692 (2017).
56. G. Yaari, J. Vander Heiden, M. Uduman, D. Gadala-Maria, N. Gupta, J. N. H. Stern, K. O'Connor, D. Hafler, U. Laserson, F. Vigneault, S. Kleinstein, Models of Somatic Hypermutation Targeting and Substitution Based on Synonymous Mutations from High-Throughput Immunoglobulin Sequencing Data. *Front. Immunol.* **4** (2013), doi:10.3389/fimmu.2013.00358.
57. A. Koren, P. Polak, J. Nemesh, J. J. Michaelson, J. Sebat, S. R. Sunyaev, S. A. McCarroll, Differential Relationship of DNA Replication Timing to Different Forms of Human Mutation and Variation. *Am. J. Hum. Genet.* **91**, 1033–1040 (2012).
58. A. Gonzalez-Perez, R. Sabarinathan, N. Lopez-Bigas, Local Determinants of the Mutational Landscape of the Human Genome. *Cell*. **177**, 101–114 (2019).
59. P. Polak, R. Karlić, A. Koren, R. Thurman, R. Sandstrom, M. S. Lawrence, A. Reynolds, E. Rynes, K. Vlahoviček, J. A. Stamatoyannopoulos, S. R. Sunyaev, Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature*. **518**, 360–364 (2015).
60. K. Kübler, R. Karlić, N. J. Haradhvala, K. Ha, J. Kim, M. Kuzman, W. Jiao, S. Gakkhar, K. W. Mouw, L. Z. Braunstein, O. Elemento, A. V. Biankin, I. Rooman, M. Miller, W. R. Karthaus, C. D. Nogiec, E. Juvenson, E. Curry, M. M.- Kenudson, L. W. Ellisen, R. Brown, A. Gusev, C. Tomasetti, M. P. Lolkema, N. Steeghs, C. van Herpen, H.-G. Kim, H. Lee, K. Vlahoviček, B. E. Bernstein, C. L. Sawyers, K. A. Hoadley, E. Cuppen, A. Koren, P. F. Arndt, D. N. Louis, L. D. Stein, W. D. Foulkes, P. Polak, G. Getz, and the I. P.-C. A. of W. G. N. on behalf of the PCAWG Pathology and Clinical Correlates Working Group, Tumor mutational landscape is a record of the pre-malignant state. *bioRxiv*, 517565 (2019).
61. A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Yaping Liu, C. Coarfa, R. Alan Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. David Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. Scott Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. D. Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, M. Kellis, Integrative analysis of 111 reference human epigenomes. *Nature*. **518**, 317–330 (2015).
62. C. A. Davis, B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U. K. Baymuradov, A. K. Narayanan, K. C. Onate, K. Graham, S. R. Miyasato, T. R. Dreszer, J. S. Strattan, O. Jolanki, F. Y. Tanaka, J. M. Cherry, The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
63. H. G. Stunnenberg, S. Abrignani, D. Adams, M. de Almeida, L. Altucci, V. Amin, I. Amit, S. E. Antonarakis, S. Aparicio, T. Arima, L. Arrigoni, R. Arts, V. Asnafi, M. Esteller, J.-B. Bae, K. Bassler, S. Beck, B. Berkman, B. E. Bernstein, M. Bilenky, A. Bird, C. Bock, B. Boehm, G. Bourque, C. E. Breeze, B. Brors, D. Bujold, O. Burren, M. J. Bussemakers, A. Butterworth, E. Campo, E. Carrillo-de-Santa-Pau, L. Chadwick, K. M. Chan, W. Chen, T. H. Cheung, L. Chiapperino, N. H. Choi, H.-R. Chung, L. Clarke, J. M. Connors, P. Cronet, J. Danesh, M. Dermitzakis, G. Drewes, P. Durek, S. Dyke, T. Dylag, C. J. Eaves, P. Ebert, R. Eils, J. Eils, C. A. Ennis, T. Enver, E. A. Feingold, B. Felder, A. Ferguson-Smith, J. Fitzgibbon, P. Flicek, R. S.-Y. Foo, P. Fraser, M. Frontini, E. Furlong, S. Gakkhar, N. Gasparoni, G. Gasparoni, D. H. Geschwind, P. Glažar, T. Graf, F. Grosveld, X.-Y. Guan, R. Guigo, I. G. Gut, A.

Hamann, B.-G. Han, R. A. Harris, S. Heath, K. Helin, J. G. Hengstler, A. Heravi-Moussavi, K. Herrup, S. Hill, J. A. Hilton, B. C. Hitz, B. Horsthemke, M. Hu, J.-Y. Hwang, N. Y. Ip, T. Ito, B.-M. Javierre, S. Jenko, T. Jenuwein, Y. Joly, S. J. M. Jones, Y. Kanai, H. G. Kang, A. Karsan, A. K. Kiemer, S. C. Kim, B.-J. Kim, H.-H. Kim, H. Kimura, S. Kinkley, F. Klironomos, I.-U. Koh, M. Kostadima, C. Kressler, R. Kreuzhuber, A. Kundaje, R. Küppers, C. Larabell, P. Lasko, M. Lathrop, D. H. S. Lee, S. Lee, H. Lehrach, E. Leitão, T. Lengauer, Å. Lernmark, R. D. Leslie, G. K. K. Leung, D. Leung, M. Loeffler, Y. Ma, A. Mai, T. Manke, E. R. Marcotte, M. A. Marra, J. H. A. Martens, J. I. Martin-Subero, K. Maschke, C. Merten, A. Milosavljevic, S. Minucci, T. Mitsuyama, R. A. Moore, F. Müller, A. J. Mungall, M. G. Netea, K. Nordström, I. Norstedt, H. Okae, V. Onuchic, F. Ouellette, W. Ouwehand, M. Pagani, V. Pancaldi, T. Pap, T. Pastinen, R. Patel, D. S. Paul, M. J. Pazin, P. G. Pelicci, A. G. Phillips, J. Polansky, B. Porse, J. A. Pospisilik, S. Prabhakar, D. C. Procaccini, A. Radbruch, N. Rajewsky, V. Rakyan, W. Reik, B. Ren, D. Richardson, A. Richter, D. Rico, D. J. Roberts, P. Rosenstiel, M. Rothstein, A. Salhab, H. Sasaki, J. S. Satterlee, S. Sauer, C. Schacht, F. Schmidt, G. Schmitz, S. Schreiber, C. Schröder, D. Schübeler, J. L. Schultze, R. P. Schulyer, M. Schulz, M. Seifert, K. Shirahige, R. Siebert, T. Sierocinski, L. Siminoff, A. Sinha, N. Soranzo, S. Spicuglia, M. Spivakov, C. Steidl, J. S. Strattan, M. Stratton, P. Südbeck, H. Sun, N. Suzuki, Y. Suzuki, A. Tanay, D. Torrents, F. L. Tyson, T. Ulas, S. Ullrich, T. Ushijima, A. Valencia, E. Vellenga, M. Vingron, C. Wallace, S. Wallner, J. Walter, H. Wang, S. Weber, N. Weiler, A. Weller, A. Weng, S. Wilder, S. M. Wiseman, A. R. Wu, Z. Wu, J. Xiong, Y. Yamashita, X. Yang, D. Y. Yap, K. Y. Yip, S. Yip, J.-I. Yoo, D. Zerbino, G. Zipprich, M. Hirst, The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell*. **167**, 1145–1149 (2016).

64. Y. Li, N. D. Roberts, J. A. Wala, O. Shapira, S. E. Schumacher, K. Kumar, E. Khurana, S. Waszak, J. O. Korbel, J. E. Haber, M. Imielinski, J. Weischenfeldt, R. Beroukhim, P. J. Campbell, Patterns of somatic structural variation in human cancer genomes. *Nature*. **578**, 112–121 (2020).
65. J. Hu, Y. Zhang, L. Zhao, R. L. Frock, Z. Du, R. M. Meyers, F. Meng, D. G. Schatz, F. W. Alt, Chromosomal Loop Domains Direct the Recombination of Antigen Receptor Genes. *Cell*. **163**, 947–959 (2015).
66. G. Teng, Y. Maman, W. Resch, M. Kim, A. Yamane, J. Qian, K.-R. Kieffer-Kwon, M. Mandal, Y. Ji, E. Meffre, M. R. Clark, L. G. Cowell, R. Casellas, D. G. Schatz, RAG Represents a Widespread Threat to the Lymphocyte Genome. *Cell*. **162**, 751–765 (2015).
67. S. C. Baca, D. Prandi, M. S. Lawrence, J. M. Mosquera, A. Romanel, Y. Drier, K. Park, N. Kitabayashi, T. Y. MacDonald, M. Ghandi, E. Van Allen, G. V. Kryukov, A. Sboner, J.-P. Theurillat, T. D. Soong, E. Nickerson, D. Auclair, A. Tewari, H. Beltran, R. C. Onofrio, G. Boysen, C. Guiducci, C. E. Barbieri, K. Cibulskis, A. Sivachenko, S. L. Carter, G. Saksena, D. Voet, A. H. Ramos, W. Winckler, M. Cipicchio, K. Ardlie, P. W. Kantoff, M. F. Berger, S. B. Gabriel, T. R. Golub, M. Meyerson, E. S. Lander, O. Elemento, G. Getz, F. Demichelis, M. A. Rubin, L. A. Garraway, Punctuated Evolution of Prostate Cancer Genomes. *Cell*. **153**, 666–677 (2013).
68. M. Muramatsu, K. Kinoshita, S. Fagarasan, S. Yamada, Y. Shinkai, T. Honjo, Class Switch Recombination and Hypermutation Require Activation-Induced Cytidine Deaminase (AID), a Potential RNA Editing Enzyme. *Cell*. **102**, 553–563 (2000).
69. P. Revy, T. Muto, Y. Levy, F. Geissmann, A. Plebani, O. Sanal, N. Catalan, M. Forveille, R. Dufourcq-Lagelouse, A. Gennery, I. Tezcan, F. Ersoy, H. Kayserili, A. G. Ugazio, N. Brousse, M. Muramatsu, L. D. Notarangelo, K. Kinoshita, T. Honjo, A. Fischer, A. Durandy, Activation-Induced Cytidine Deaminase (AID) Deficiency Causes the Autosomal Recessive Form of the Hyper-IgM Syndrome (HIGM2). *Cell*. **102**, 565–575 (2000).
70. I. P. M. Tomlinson, M. R. Novelli, W. F. Bodmer, The mutation rate and cancer. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 14800–14803 (1996).
71. S. F. Brunner, N. D. Roberts, L. A. Wylie, L. Moore, S. J. Aitken, S. E. Davies, M. A. Sanders, P. Ellis, C. Alder, Y. Hooks, F. Abascal, M. R. Stratton, I. Martincorena, M. Hoare, P. J. Campbell, Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature*. **574**, 538–542 (2019).
72. S. F. Roerink, N. Sasaki, H. Lee-Six, M. D. Young, L. B. Alexandrov, S. Behjati, T. J. Mitchell, S. Grossmann, H. Lightfoot, D. A. Egan, A. Pronk, N. Smakman, J. van Gorp, E. Anderson, S. J. Gamble, C. Alder, M. van de Wetering, P. J. Campbell, M. R. Stratton, H. Clevers, Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature*. **556**, 457–462 (2018).
73. P. J. Campbell, G. Getz, J. O. Korbel, J. M. Stuart, J. L. Jennings, L. D. Stein, M. D. Perry, H. K. Nahal-Bose, B. F. F. Ouellette, C. H. Li, E. Rheinbay, G. P. Nielsen, D. C. Sgroi, C.-L. Wu, W. C. Faquin, V. Deshpande, P. C. Boutros, A. J. Lazar, K. A. Hoadley, D. N. Louis, L. J. Dursi, C. K. Yung, M. H. Bailey, G. Saksena, K. M. Raine, I. Buchhalter, K. Kleinheinz, M. Schlesner, J. Zhang, W. Wang, D. A. Wheeler, L. Ding, J. T. Simpson, B. D. O'Connor, S.

Yakneen, K. Ellrott, N. Miyoshi, A. P. Butler, R. Royo, S. I. Shorser, M. Vazquez, T. Rausch, G. Tiao, S. M. Waszak, B. Rodriguez-Martin, S. Shringarpure, D.-Y. Wu, G. M. Demidov, O. Delaneau, S. Hayashi, S. Imoto, N. Habermann, A. V. Segre, E. Garrison, A. Cafferkey, E. G. Alvarez, J. M. Heredia-Genestar, F. Muyas, O. Drechsel, A. L. Bruzos, J. Temes, J. Zamora, A. Baez-Ortega, H.-L. Kim, R. J. Mashl, K. Ye, A. DiBiase, K. Huang, I. Letunic, M. D. McLellan, S. J. Newhouse, T. Shmaya, S. Kumar, D. C. Wedge, M. H. Wright, V. D. Yellapantula, M. Gerstein, E. Khurana, T. Marques-Bonet, A. Navarro, C. D. Bustamante, R. Siebert, H. Nakagawa, D. F. Easton, S. Ossowski, J. M. C. Tubio, F. M. De La Vega, X. Estivill, D. Yuen, G. L. Mihaiescu, L. Omberg, V. Ferretti, R. Sabarinathan, O. Pich, A. Gonzalez-Perez, A. Taylor-Weiner, M. W. Fittall, J. Demeulemeester, M. Tarabichi, N. D. Roberts, P. Van Loo, I. Cortés-Ciriano, L. Urban, P. Park, B. Zhu, E. Pitkänen, Y. Li, N. Saini, L. J. Klimczak, J. Weischenfeldt, N. Sidiropoulos, L. B. Alexandrov, R. Rabionet, G. Escaramis, M. Bosio, A. Z. Holik, H. Susak, A. Prasad, S. Erkek, C. Calabrese, B. Raeder, E. Harrington, S. Mayes, D. Turner, S. Juul, S. A. Roberts, L. Song, R. Koster, L. Mirabello, X. Hua, T. J. Tanskanen, M. Tojo, J. Chen, L. A. Aaltonen, G. Rätsch, R. F. Schwarz, A. J. Butte, A. Brazma, S. J. Chanock, N. Chatterjee, O. Stegle, O. Harismendy, G. S. Bova, D. A. Gordenin, D. Haan, L. Sieverling, L. Feuerbach, D. Chalmers, Y. Joly, B. Knoppers, F. Molnár-Gábor, M. Phillips, A. Thorogood, D. Townend, M. Goldman, N. A. Fonseca, Q. Xiang, B. Craft, E. Piñeiro-Yáñez, A. Muñoz, R. Petryszak, A. Füllgrabe, F. Al-Shahrour, M. Keays, D. Haussler, J. Weinstein, W. Huber, A. Valencia, I. Papatheodorou, J. Zhu, Y. Fan, D. Torrents, M. Bieg, K. Chen, Z. Chong, K. Cibulskis, R. Eils, R. S. Fulton, J. L. Gelpi, S. Gonzalez, I. G. Gut, F. Hach, M. Heinold, T. Hu, V. Huang, B. Hutter, N. Jäger, J. Jung, Y. Kumar, C. Lalansingh, I. Leshchiner, D. Livitz, E. Z. Ma, Y. E. Maruvka, A. Milovanovic, M. M. Nielsen, N. Paramasivam, J. S. Pedersen, M. Puiggròs, S. C. Sahinalp, I. Sarrafi, C. Stewart, M. D. Stobbe, J. A. Wala, J. Wang, M. Wendl, J. Werner, Z. Wu, H. Xue, T. N. Yamaguchi, V. Yellapantula, B. N. Davis-Dusenbery, R. L. Grossman, Y. Kim, M. C. Heinold, J. Hinton, D. R. Jones, A. Menzies, L. Stebbings, J. M. Hess, M. Rosenberg, A. J. Dunford, M. Gupta, M. Imielinski, M. Meyerson, R. Beroukhim, J. Reimand, P. Dhingra, F. Favero, S. Dentro, J. Wintersinger, V. Rudneva, J. W. Park, E. P. Hong, S. G. Heo, A. Kahles, K.-V. Lehmann, C. M. Soulette, Y. Shiraishi, F. Liu, Y. He, D. Demircioğlu, N. R. Davidson, L. Greger, S. Li, D. Liu, S. G. Stark, F. Zhang, S. B. Amin, P. Bailey, A. Chateigner, M. Frenkel-Morgenstern, Y. Hou, M. R. Huska, H. Kilpinen, F. C. Lamaze, C. Li, X. Li, X. Li, X. Liu, M. G. Marin, J. Markowski, T. Nandi, A. I. Ojesina, Q. Pan-Hammarström, P. J. Park, C. S. Pedamallu, H. Su, P. Tan, B. T. Teh, J. Wang, H. Xiong, C. Ye, C. Yung, X. Zhang, L. Zheng, S. Zhu, P. Awadalla, C. J. Creighton, K. Wu, H. Yang, J. Göke, Z. Zhang, A. N. Brooks, M. W. Fittall, I. Martincorena, C. Rubio-Perez, M. Juul, S. Schumacher, O. Shapira, D. Tamborero, L. Mularoni, H. Hornshøj, J. Deu-Pons, F. Muiños, J. Bertl, Q. Guo, A. Gonzalez-Perez, Q. Xiang, The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, Pan-cancer analysis of whole genomes. *Nature*. **578**, 82–93 (2020).

74. S. Nik-Zainal, P. Van Loo, D. C. Wedge, L. B. Alexandrov, C. D. Greenman, K. W. Lau, K. Raine, D. Jones, J. Marshall, M. Ramakrishna, A. Shlien, S. L. Cooke, J. Hinton, A. Menzies, L. A. Stebbings, C. Leroy, M. Jia, R. Rance, L. J. Mudie, S. J. Gamble, P. J. Stephens, S. McLaren, P. S. Tarpey, E. Papaemmanuil, H. R. Davies, I. Varela, D. J. McBride, G. R. Bignell, K. Leung, A. P. Butler, J. W. Teague, S. Martin, G. Jönsson, O. Mariani, S. Boyault, P. Miron, A. Fatima, A. Langerød, S. A. J. R. Aparicio, A. Tutt, A. M. Sieuwerts, Å. Borg, G. Thomas, A. V. Salomon, A. L. Richardson, A.-L. Børresen-Dale, P. A. Futreal, M. R. Stratton, P. J. Campbell, The Life History of 21 Breast Cancers. *Cell*. **149**, 994–1007 (2012).
75. P. Ellis, L. Moore, M. A. Sanders, T. M. Butler, S. F. Brunner, H. Lee-Six, R. Osborne, B. Farr, T. H. H. Coorens, A. R. J. Lawson, A. Cagan, M. R. Stratton, I. Martincorena, P. J. Campbell, Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat. Protoc.* **16**, 841–871 (2021).
76. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).
77. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* (2013) (available at http://arxiv.org/abs/1303.3997).
78. D. Jones, K. M. Raine, H. Davies, P. S. Tarpey, A. P. Butler, J. W. Teague, S. Nik-Zainal, P. J. Campbell, *Curr. Protoc. Bioinforma.*, in press, doi:https://doi.org/10.1002/cpbi.20.
79. K. Ye, M. H. Schulz, Q. Long, R. Apweiler, Z. Ning, Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinforma. Oxf. Engl.* **25**, 2865–2871 (2009).
80. S. Nik-Zainal, H. Davies, J. Staaf, M. Ramakrishna, D. Glodzik, X. Zou, I. Martincorena, L. B. Alexandrov, S.

Martin, D. C. Wedge, P. Van Loo, Y. S. Ju, M. Smid, A. B. Brinkman, S. Morganella, M. R. Aure, O. C. Lingjærde, A. Langerød, M. Ringnér, S.-M. Ahn, S. Boyault, J. E. Brock, A. Broeks, A. Butler, C. Desmedt, L. Dirix, S. Dronov, A. Fatima, J. A. Foekens, M. Gerstung, G. K. J. Hooijer, S. J. Jang, D. R. Jones, H.-Y. Kim, T. A. King, S. Krishnamurthy, H. J. Lee, J.-Y. Lee, Y. Li, S. McLaren, A. Menzies, V. Mustonen, S. O'Meara, I. Pauporté, X. Pivot, C. A. Purdie, K. Raine, K. Ramakrishnan, F. G. Rodríguez-González, G. Romieu, A. M. Sieuwerts, P. T. Simpson, R. Shepherd, L. Stebbings, O. A. Stefansson, J. Teague, S. Tommasi, I. Treilleux, G. G. Van den Eynden, P. Vermeulen, A. Vincent-Salomon, L. Yates, C. Caldas, L. van't Veer, A. Tutt, S. Knappskog, B. K. T. Tan, J. Jonkers, Å. Borg, N. T. Ueno, C. Sotiriou, A. Viari, P. A. Futreal, P. J. Campbell, P. N. Span, S. Van Laere, S. R. Lakhani, J. E. Eyfjord, A. M. Thompson, E. Birney, H. G. Stunnenberg, M. J. van de Vijver, J. W. M. Martens, A.-L. Børresen-Dale, A. L. Richardson, G. Kong, G. Thomas, M. R. Stratton, Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*. **534**, 47–54 (2016).

81. P. V. Loo, S. H. Nordgard, O. C. Lingjærde, H. G. Russnes, I. H. Rye, W. Sun, V. J. Weigman, P. Marynen, A. Zetterberg, B. Naume, C. M. Perou, A.-L. Børresen-Dale, V. N. Kristensen, Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci.* **107**, 16910–16915 (2010).
82. T. H. H. Coorens, T. R. W. Oliver, R. Sanghvi, U. Sovio, E. Cook, R. Vento-Tormo, M. Haniffa, M. D. Young, R. Rahbari, N. Sebire, P. J. Campbell, D. S. Charnock-Jones, G. C. S. Smith, S. Behjati, Inherent mosaicism and extensive mutation of human placentas. *Nature*. **592**, 80–85 (2021).
83. L. Moore, D. Leongamornlert, T. H. H. Coorens, M. A. Sanders, P. Ellis, S. C. Dentro, K. J. Dawson, T. Butler, R. Rahbari, T. J. Mitchell, F. Maura, J. Nangalia, P. S. Tarpey, S. F. Brunner, H. Lee-Six, Y. Hooks, S. Moody, K. T. Mahbubani, M. Jimenez-Linan, J. J. Brosens, C. A. Iacobuzio-Donahue, I. Martincorena, K. Saeb-Parsy, P. J. Campbell, M. R. Stratton, The mutational landscape of normal human endometrial epithelium. *Nature*. **580**, 640–646 (2020).
84. K. Gori, A. Baez-Ortega, sigfit: flexible Bayesian inference of mutational signatures. *bioRxiv*, 372896 (2020).
85. F. Nadeu, R. Mas-de-les-Valls, A. Navarro, R. Royo, S. Martín, N. Villamor, H. Suárez-Cisneros, R. Mares, J. Lu, A. Enjuanes, A. Rivas-Delgado, M. Aymerich, T. Baumann, D. Colomer, J. Delgado, R. D. Morin, T. Zenz, X. S. Puente, P. J. Campbell, S. Beà, F. Maura, E. Campo, IgCaller for reconstructing immunoglobulin gene rearrangements and oncogenic translocations from whole-genome sequencing in lymphoid neoplasms. *Nat. Commun.* **11**, 3390 (2020).
86. R. S. Hansen, S. Thomas, R. Sandstrom, T. K. Canfield, R. E. Thurman, M. Weaver, M. O. Dorschner, S. M. Gartler, J. A. Stamatoyannopoulos, Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci.* **107**, 139–144 (2010).
87. J. E. Hesse, M. R. Lieber, K. Mizuuchi, M. Gellert, V(D)J recombination: a functional definition of the joining signals. *Genes Dev.* **3**, 1053–1061 (1989).
88. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: scanning for occurrences of a given motif. *Bioinforma. Oxf. Engl.* **27**, 1017–1018 (2011).
89. T. L. Bailey, W. S. Noble, Searching for statistically significant regulatory modules. *Bioinforma. Oxf. Engl.* **19 Suppl 2**, ii16-25 (2003).
90. J. H. R. Farmery, M. L. Smith, NIHR BioResource - Rare Diseases, A. G. Lynch, Telomerecat: A ploidy-agnostic method for estimating telomere length from whole genome sequencing data. *Sci. Rep.* **8**, 1300 (2018).

Acknowledgments:

The authors would like to thank Federico Abascal, Tim Coorens, Timothy Butler and Simon Brunner for valuable guidance in data analysis, the CASM lab, including Laura O'Neill and Calli Latimer, for sample and data management, and CASM IT for technical support. This research was supported by the Cambridge NIHR BRC Cell Phenotyping Hub and staff, including Esther Perez and Natalia Savinykh who provided advice and support in flow cytometry and cell sorting. We are especially grateful to the tissue donors and their families and to the Cambridge Biorepository for Translational Medicine for the gift of tissue from transplant organ donors.

Funding:

This work was supported by the Harrison Foundation and Wellcome Trust. M.S.S. was the recipient of a Biotechnology and Biological Sciences Research Council Industrial Collaborative Awards in Science and Engineering PhD Studentship. The DGK laboratory is supported by a Blood Cancer UK Bennett Fellowship (15008), an ERC Starting Grant (ERC-2016-STG–715371), a CR-UK Programme Foundation award (DCRPGF\100008) and an MRC-AMED joint award (MR/V005502/1). DGK, EL, and ARG are supported by a core support grant to the Wellcome MRC Cambridge Stem Cell Institute, Blood Cancer UK, the NIHR Cambridge Biomedical Research Centre, and the CRUK Cambridge Cancer Centre.

E.L. is supported by a Sir Henry Dale fellowship from Wellcome/Royal Society (107630/Z/15/Z), BBSRC (BB/P002293/1), and core support grants by Wellcome and MRC to the Wellcome-MRC Cambridge Stem Cell Institute (203151/Z/16/Z).

K.K. and G.G. are supported by a GDAN grant (grant number U24CA210999). G.G. is partly supported by the Paul C. Zamecnik Chair in Oncology at the Massachusetts General Hospital Cancer Center.

Competing interests:

P.J.C. is a founder, consultant and director for Mu Genomics Ltd. G.G. receives research funds from Pharmacyclics and IBM. G.G. is an inventor on multiple patents related to bioinformatics methods (MuTect, MutSig, ABSOLUTE, MSMutSig, MSMuTect, POLYSOLVER and TensorQTL). G.G. is a founder, consultant and holds privately held equity in Scorpion Therapeutics.

Materials and Methods

Samples

Human blood mononuclear cells (MNCs) were obtained from four sources: 1) bone marrow, spleen and peripheral blood from three transplant organ donors (KX001, KX002, KX003) recruited from Cambridge University Hospitals NHS Trust, Addenbrooke's Hospital (by Cambridge Biorepository for Translational Medicine, Research Ethics Committee approval 15/EE/0152), 2) peripheral blood from one patient (AX001) recruited from Addenbrooke's Hospital (approval 07-MRE05-44), 3) tonsil from two patients (TX001, TX002) recruited from Addenbrooke's Hospital (approval 07-MRE05-44), and 4) one cord blood (CB001) collected with informed consent by StemCell Technologies (catalog #70007) (Table S1). All sources were hematopoietically normal and healthy. Donor KX002 had a history of Crohn's disease and treatment with Azathioprine. Patients TX001 and TX002 had a history of tonsillitis. MNCs from (1), (2) and (3) were extracted using Lymphoprep (Axis-Shield), depleted of red blood cells using RBC lysis buffer (BioLegend) and frozen viable in 10% DMSO. Cord blood MNCs (4) were received frozen and then $CD34^+$ selected using the EasySep human whole blood CD34 positive selection kit (Stemcell Technologies) as per the manufacturer's instructions, with the $CD34^+$ fraction used for hematopoietic stem and progenitor cell (HSPC) cultures and the $CD34^-$ fraction used for lymphocyte cultures. Additional peripheral blood MNCs from (1) also underwent CD34 positive selection and was used for HSPC cultures.

Flow cytometry

MNC samples were sorted by flow cytometry at the NIHR Cambridge BRC Cell Phenotyping Hub on AriaIII or Aria-Fusion cell sorters into naive B lymphocytes ($CD3^-CD19^+CD20^+CD27^-CD38^-IgD^+$), memory B lymphocytes ($CD3^-CD19^+CD20^+CD27^+CD38^-IgD^-$), naive T lymphocytes ($CD3^+CD4/CD8^+CCR7^+CD45RA^{high}$), memory T lymphocytes ($CD3^+CD4/CD8^+CD45RA^-$), regulatory T cells (Tregs: $CD3^+CD4^+CD25^{high}CD127^-$) and HSPCs ($CD3^-CD19^-CD34^+CD38^-CD90^+CD45RA^-$) (Fig. S1). HSPCs from AX001 included HSCs ($CD34^+CD38^-$) and progenitors ($CD34^+CD38^+CD10^{-/dim}$). The antibody panels used are as follows: lymphocytes (excluding Tregs): CD3-APC, CD4-BV785, CD8-BV650, CD14-BV605, CD19-AF700, CD20-PEDazzle, CD27-BV421, CD34-APC-Cy7, CD38-FITC, CD45RA-PerCP-Cy5.5, CD56-PE, CCR7-BV711, IgD-PECy7, Zombie-Aqua; Tregs: CD3-APC, CD4-BV785, CD8-BV650, CD19-APC-Cy7, CD45RA-PerCP-Cy5.5, CD56-PE, CCR7-FITC, CD25-PECy5, CD127-PECy7, CD69-AF700, CD103-BV421, CCR9-PE, Zombie-Aqua; HSPCs (excluding AX001): CD3-FITC, CD90-PE, CD49f-PECy5, CD38-PECy7, CD33-APC, CD19-A700, CD34-APC-Cy7, CD45RA-BV421, Zombie-Aqua; HSPCs (AX001): CD38-FITC, CD135-PE, CD34-PE-Cy7, CD90-APC, CD10-APC-Cy7, CD45RA-V450, Zombie-Aqua. Cells were either single-cell sorted for liquid culture into 96-well plates containing 50ul cell type-specific expansion medium, or (for AX001 HSPCs) bulk-sorted for MethoCult plate-base expansion.

In vitro liquid culture expansion

We designed novel protocols to expand B and T lymphocytes from single cells into colonies of at least 30 cells. The B cell expansion medium was composed of 5ug/ml Anti-IgM (Stratech Scientific Ltd), 100ng/ml IL-2, 20ng/ml IL-4, and 50ng/ml IL-21 (PeproTech EC Ltd), 2.5ng/ml CD40L-HA (Bio-Techne Ltd) and 1.25ug/ml HA Tag (Bio-Techne Ltd), in Advanced RPMI 1640 Medium (ThermoFisher Scientific) with 10% fetal bovine serum (ThermoFisher Scientific), 1% penicillin/streptomycin (Sigma-Aldrich), and 1% L-glutamine (Sigma-Aldrich). The T cell expansion medium was composed of 12.5ul/ml ImmunoCult CD3/CD28 (STEMCELL Technologies) and 100ng/ml IL-2 and 5ng/ml IL-15 (PeproTech EC Ltd), in ImmunoCult-XF T Cell Expansion Medium (STEMCELL Technologies) with 5% fetal bovine serum (ThermoFisher Scientific) and 0.5% penicillin/streptomycin (Sigma-Aldrich). 25ul of fresh

expansion medium was added to each culture every 3-4 days. Colonies (30-2000 cells per colony) were harvested either manually or robotically using a CellCelector (Automated Lab Solutions) approximately 14 days after sorting.

Sorted HSPCs from donors KX001, KX002, KX003 and CB001 were expanded from single cells into colonies of 200-100,000+ cells in Nunc 96 well flat-bottomed TC plates (ThermoFisher Scientific) containing 100uL of supplemented StemPro media (Stem Cell Technologies) (MEM media). MEM media contained StemPro Nutrients (0.035%) (Stem Cell Technologies), L-Glutamine (1%) (ThermoFisher Scientific), Penicillin-Streptomycin (1%) (ThermoFisher Scientific) and cytokines (SCF: 100ng/ml; FLT3: 20ng/ml; TPO: 100ng/ml; EPO: 3ng/ml; IL-6: 50ng/ml; IL-3: 10ng/ml; IL-11: 50ng/ml; GM-CSF: 20ng/ml; IL-2: 10ng/ml; IL-7: 20ng/ml; lipids: 50ng/ml) to promote differentiation towards Myeloid/Erythroid/Megakaryocyte (MEM) and NK lineages. Manual assessment of colony growth was made at 14 days. Colonies were topped up with an additional 50ul of MEM media on day 15 if the colony was ≥1/4 size of well. Following 21 +/- 2 days in culture, colonies were selected by size criteria. Colonies ≥ 3000 cells in size were harvested into a U bottomed 96 well plate (ThermoFisher Scientific). Plates were then centrifuged (500g/5min), media was discarded, and the cells were resuspended in 50ul PBS prior to freezing at -80C. Colonies less than 3000 cells but greater than 200 cells in size were harvested into 96 well skirted Lo Bind plates (Eppendorf) and centrifuged (800g/5min). Supernatant was removed to 5-10ul using an aspirator prior to DNA extraction on the fresh cell pellet. Sorted HSPCs from donor AX001 were plated onto CFC media MethoCult H4435 (STEMCELL Technologies) and colonies were picked following 24 days in culture.

Whole genome sequencing of colonies

DNA was extracted from 717 colonies with Arcturus PicoPure DNA Extraction Kit (ThermoFisher Scientific), with the exception of larger HSPC colonies which were extracted using the DNeasy 96 blood and tissue plate kit (Qiagen) and then diluted to 1-5ng. DNA was used to make Illumina sequencing libraries using a custom low input protocol (*75*). We performed whole genome sequencing using 150bp paired-end sequencing reads on an Illumina XTen platform, to an average depth of 20x per colony. Sequence data were mapped to the human genome reference GRCh37d5 using the BWA-MEM algorithm (*76*, *77*).

Variant calling

We called all classes of variants using validated pipelines at the Wellcome Sanger Institute. Single nucleotide variants (SNVs) were called using the program CaVEMan (*78*), insertion/deletions (indels) using Pindel (*79*), structural variants (SVs) using BRASS (*80*) and copy number variants (CNVs) using ASCAT (*81*). In order to recover all mutations, including high frequency ones, we used an *in silico* sample produced from the reference genome rather than use a matched normal for the CaVEMan, Pindel, and BRASS analyses. Germline mutations were removed after variant calling (see below). For the ASCAT analysis we elected one colony (arbitrarily chosen) to serve as the matched normal.

Variants were filtered to remove false positives and germline variants. First, variants with a mean VAF greater than 40% across colonies of an individual were likely germline variants and were removed. To remove remaining germline variants and false positives, we exploited the fact that we have several, highly clonal samples per individual. We performed a beta-binomial test per variant per individual, retaining only SNVs and indels that were highly over-dispersed within an individual (*82*). For SNVs we also required that the variants be identified as significantly subclonal within an individual using the program Shearwater, and applied filters to remove artifacts resulting from the low-input library preparation. Detailed description of the artifact filters were provided previously (*75*) and the complete filtering pipeline is made available on GitHub (https://github.com/MathijsSanders/SangerLCMFiltering). For both the beta-binomial filter and the Shearwater filter we observed bimodal distributions separating the data into low and high confidence variants. We made use of this feature, using a valley-finding algorithm (R package *quantmod*) to determine the p-value cutoffs, per

individual. We genotyped each colony for the set of filtered somatic SNVs and indels (per respective individual), calling a variant present if it had a minimum VAF of 20% and a minimum of two alternate reads in that colony.

We removed artifacts from the SV calls using AnnotateBRASS (https://github.com/MathijsSanders/AnnotateBRASS) with default settings. The full list of statistics calculated and post-hoc filtering strategy was described in detail previously (*83*). Somatic SVs were identified as those shared by less than 25% of the colonies within an individual. SVs and CNVs were both subsequently manually curated by visual inspection.

Mutation burden analysis

We found that sequencing depth was a strong predictor of mutation burden in our samples. Therefore, in order to more accurately estimate the mutation burden for each colony, we corrected the number of SNVs or indels (corrected separately) by fitting an asymptotic regression (function *NLSstAsymptotic*, R package *stats*) to mutation burden as a function of sequencing depth per colony. For this correction we used HSPC genomes (excepting the tonsil samples, for which naive B and T cells were used), as lymphocyte genomes are more variable in mutation burden, and included additional unpublished HSPC genomes to increase the reliability of the model (Mitchell *et al.* in prep). Genomes with a mean sequencing depth of greater than 50x were omitted. The model parameters b0, b1, and lrc for each dataset for the model y = b0 + b1*(1-exp(-exp(lrc) * x)) are in Table S4. Mutation burden per colony was adjusted to a sequencing depth of 30.

We used a linear mixed effects model (function *lme*, R package *nlme*) to test for a significant linear relationship between mutation burden and age, and for an effect of cell subset on this relationship (separately for SNVs and indels). Number of mutations per colony was regressed on age of donor and cell type as fixed effects, with interaction between age and cell type, donor by cell type as a random effect, weighted by cell type, and with log-likelihood maximization.

Mutational signature analysis

We characterized per-colony mutational profiles by estimating the proportion of known and novel mutational signatures present in each colony. For comparison, we included in the analysis 223 genomes from 7 blood cancer types: Burkitt lymphoma, follicular lymphoma, diffuse large B cell lymphoma, chronic lymphocytic leukemia (mutated), chronic lymphocytic leukemia (unmutated), and acute myeloid leukemia (*73*) and multiple myeloma (*51*). We identified mutational signatures present in the data by performing signature extraction with two programs, *SigProfiler* (*36*) and *hdp* (https://github.com/nicolaroberts/hdp). We used the *SigProfiler* denovo results for the suggested number of extracted signatures (12). *hdp* was run without any signatures as prior, with no specified grouping of the data. These programs identified the presence of 9 mutational signatures with strong similarity (cosine similarity >= 0.85) to Cosmic signatures SBS1, SBS5, SBS7a, SBS8, SBS9, SBS13, SBS17b, SBS18, SBS19 (version 3, (*36*)).

Both *SigProfiler* and *hdp* also identified the same novel signature (cosine similarity = 0.93), which we term the 'blood signature' or 'SBSblood'. This signature is very similar to the mutational profile seen previously in HSPCs (*15*, *16*). As the signature SBSblood co-occurs with SBS1 in HSPCs, leading to the potential for these signatures being merged into one signature, we further purified SBSblood by using the program *sigfit* (*84*) to call two signatures across our HSPC genomes - SBS1 and a novel signature - with the novel signature being the final SBSblood (Supplemental Figure 3, Table S5). SBSblood was highly similar to both the *hdp* and *SigProfiler* denovo extracted signatures (cosine similarity of 0.95 and 0.94, respectively) and had similarity to the Cosmic v3 SBS5 signature

(cosine similarity = 0.87). One hypothesis is that SBSblood is the manifestation of SBS5 mutational processes in the blood cell environment.

We estimated the proportion of each of the 10 identified mutational signatures using the program *sigfit*. From these results we identified three signatures (SBS5, SBS13, SBS19) that were at nominal frequencies in the HSPC and lymphocyte genomes (less than 10% in each genome)- these were excluded from the analysis and the signature proportions were re-estimated in *sigfit* using the remaining 7 signatures: SBSblood, SBS1, SBS7a, SBS8, SBS9, SBS17b, SBS18 (Table S5).

Ig receptor sequence analysis

In order to identify the immunoglobulin (Ig) rearrangements, productive CDR3 sequences, class-switch recombination and percent somatic hypermutation for each memory B cell, we ran *IgCaller* (*85*), using a genome from the same donor (HSPC or T cell) as a matched normal for germline variant removal. We considered the somatic hypermutation rate to be the number of variants in the productive IGHV gene divided by the gene length.

We estimated the number of mutations resulting from on-target (IGV gene) somatic hypermutation compared with those associated with SBS9. We first counted all IGV variants identified by Caveman pre-filtering, as we found that standard filtering removes many somatic hypermutation variants. We then estimated SBS9 burden as the proportion of SBS9 mutations per genome multiplied by the SNV burden. The SBS9 mutation rate per genome was the SBS9 burden divided by the 'callable genome' (genome size of 3.1Gb minus an average of 383Kb excluded from variant calling).

Distribution of germinal center-associated mutations in B cells

We assessed the genomic distribution of the germinal center-associated mutational signatures, SBS9 and the SHM signature, in memory B cells. We performed per-Mb denovo signature analyses with *hdp* (no *a priori* signatures), treating mutations across all normal memory B cells within a given Mb window as a sample. The extracted 'SHM' signature (Table S5) had a cosine similarity of 0.96 to the spectrum of memory B cell mutations in the immunoglobulin gene regions, supporting the assumption that it is indeed the signature of SHM. In this analysis, SBSblood and SBS1 resolved as a single combined signature that we refer to in the genomic feature regression (below) as SBSblood/1.

We estimated the per-gene enrichment of SBS9 and SHM signatures across normal memory B and malignant B cell genomes (Burkitt lymphoma, follicular lymphoma, diffuse large B-cell lymphoma, chronic lymphocytic leukemia, and multiple myeloma). We first used *sigfit* to perform signature attribution of the signatures found in memory B cells (from the main signature analysis; SBSblood, SBS1, SBS8, SBS9, SBS17b, SBS18) and the extracted SHM signature from the above 1Mb *hdp* analysis, considering each 1Mb bin a sample. We subsequently calculated a signature attribution per variant. Gene coordinates were downloaded from UCSC (gencode.v30lift37.basic.annotation.geneonly.genename.bed). We calculated the mean attribution of variants in a given gene, representing the proportion of variants attributable to a given signature. We estimated the enrichment of SBS9 or SHM over genomic background per gene per cell type as the *p*-value of individual t-tests. While for this down-sampled dataset few genes were significant after multiple testing correction, analysis of full

datasets with larger sample sizes show statistically significant enrichment in most presented genes (Figure 5) post-multiple testing correction (data not shown).

Regression of SBS9 and genomic features

The *hdp* per-Mb memory B cell mutational signature results above were used to identify genomic features associated with the location of mutations attributable to a particular mutational signature. To achieve a finer-scale genomic resolution, each Mb bin was further divided up into 10Kb bins, and the proportion of each mutational signature in a Mb bin was used to calculate a signature attribution per 10Kb bin, based on the type and trinucleotide context of mutations in the 10Kb bin.

The number of mutations attributable to a particular mutational signature, per 10Kb window, was regressed on each of 36 genomic features (Table S3) (*64*). Noise was further removed from the replication timing data, using the GM12878 blood cell line data, and filtering the Wave Signal data by removing low Sum Signal (<95) regions, per Hansen *et al.* 2010 (*86*). SBS9 was analyzed separately from the SBSblood/1 combined signature. The number of mutations per signature per bin was calculated as the sum of the per-nucleotide probabilities per signature within a given bin. For the analysis of a given signature, a bin was only included if the average contribution of that signature was greater than 50%. This step ameliorates the problem of artificially high numbers of mutations being ascribed to a bin due to the combination of a trivially small attribution but a high overall mutation rate. This can occur in high SHM or SBS9 regions. This left 26,151 bins for SBS9 and 25,202 bins for SBSblood, out of 91,343 bins with mutations and 279,094 bins genome-wide. We also included a random sample of zero-mutation bins to equal 10% of the total bins.

We performed lasso-penalized general additive model regressions of the number of mutations per bin with the value of the genomic features. We used the *gamsel* function in R (package *gamsel*), with the lambda estimated from a 5-fold cross-validation of training data (2/3 the data). To estimate individual effect sizes, we performed general additive model regressions per genomic feature using the function *gam* (R package *mgcv*). The same analysis was also performed on HSPC mutations. The results for the full and individual regression models for each of SBS9 and SBSblood/1 in memory B cells and for all HSPC mutations can be found in Table S3.

RAG and CSR motif analysis

We assessed the enrichment of V(D)J recombination (mediated by RAG) and class switch recombination (CSR, mediated by AID) associated motifs in regions proximal to lymphocyte SVs. We identified the presence of full length and heptamer RSS motifs associated with RAG binding and endonuclease activity ('RAG motifs') (*87*) for the 50bp flanking each SV breakpoint using the program FIMO ($p<10^{-4}$) (*88*). Clusters of AGCT and TGCA repeats, associated with AID cytosine deamination and CSR ('CSR motifs'), were identified in the 1000bp flanking each SV breakpoint using the program MCAST (p<0.1, max gap=100, *E*<10,000) (*89*). In order to estimate a genomic background rate of these motifs, we generated 100 genomic controls sets, randomly selected from regions of the genome not excluded from variant calling, and performed both the RAG and CSR motif analyses on these sets. The genomic background rate presented is the median of the 100 control datasets for each motif analysis. Both the

RAG and CSR motif analyses were also performed for SVs from the PCAWG cancer genomes included in the mutational signatures analysis and for acute lymphoblastic leukemia genomes (*4*).

Telomere length

We estimated the telomere length for HSPC and lymphocyte genomes (Table S2) using the program Telomerecat (*90*). Telomere lengths for all genomes for a given donor were estimated as a group.

Timing of mutational processes

Following a procedure described previously (*59*, *60*), we modelled the distribution of somatic mutations along the genome from the density of ChIP-sequencing reads using Random Forest regression in a 10-fold cross-validation setting and the LogCosh distance between observed and predicted profiles. Each mutation was attributed to the signature that most likely generated it and aggregated into 2,128 windows of 1Mb spanning ~2.1Gb of DNA. Signatures with an average number of mutations per window <1 were not evaluated due to lack of power. We determined the difference between models using a paired two-sided Wilcoxon test on the values from the ten-fold cross-validation. Epigenetic data were gathered from different sources (Table S6) (*61*–*63*) and consisted of 149 epigenomes representing 48 distinct blood cell types and differentiation stages and their replicates. Histone marks used included H3K27me3, H3K36me3, H3K4me1 and H3K9me3. To evaluate the specificity of SBS9 mutational profiles in memory B cells, we took the same number of mutations as in SBSblood with the highest association with SBS9 and compared models with an unpaired two-sided Wilcoxon test.

Supplementary Figures

Figure S1

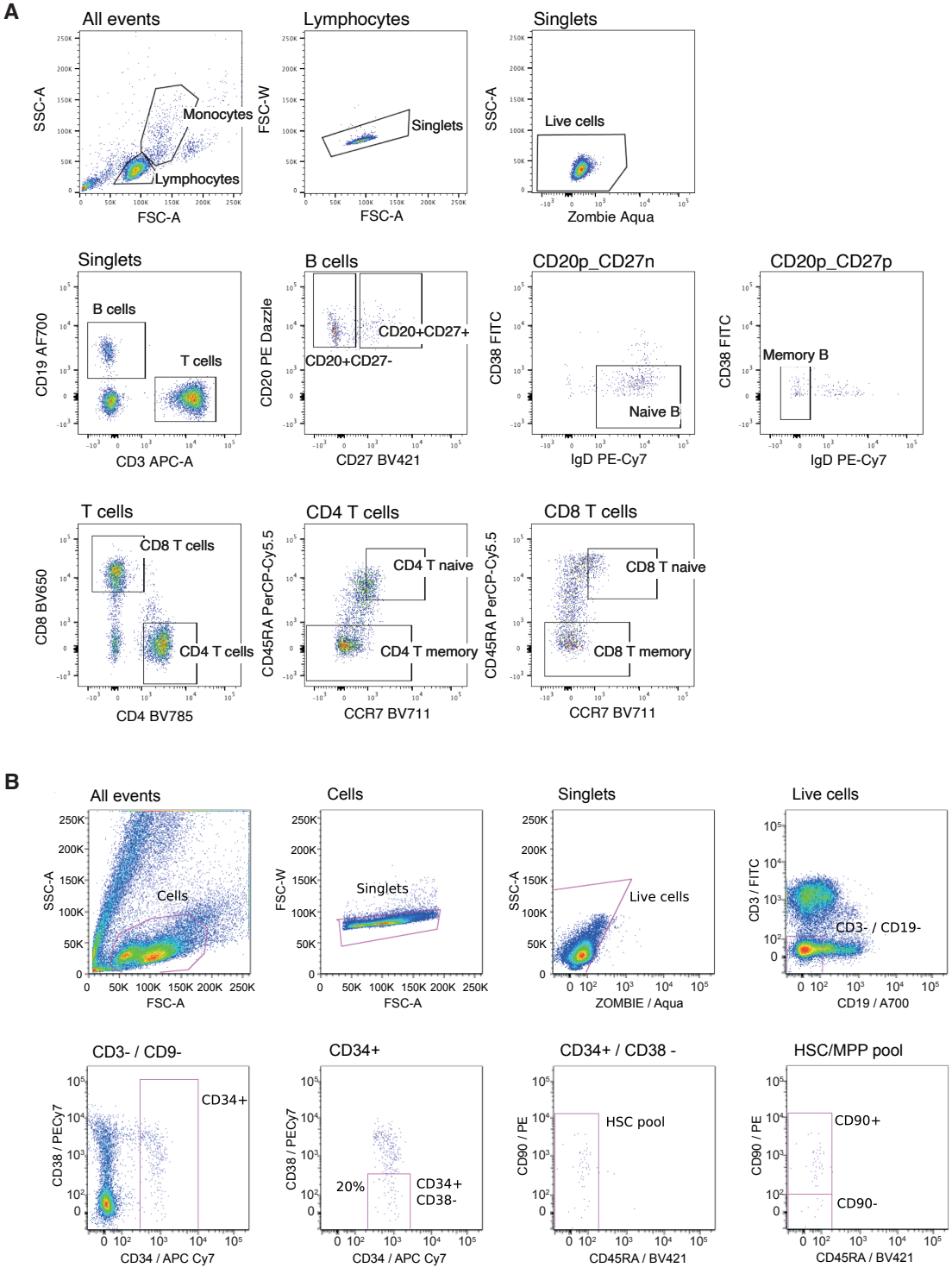

Fig. S1. Flow cytometry gating for A) lymphocytes and B) HSPCs.

Figure S2

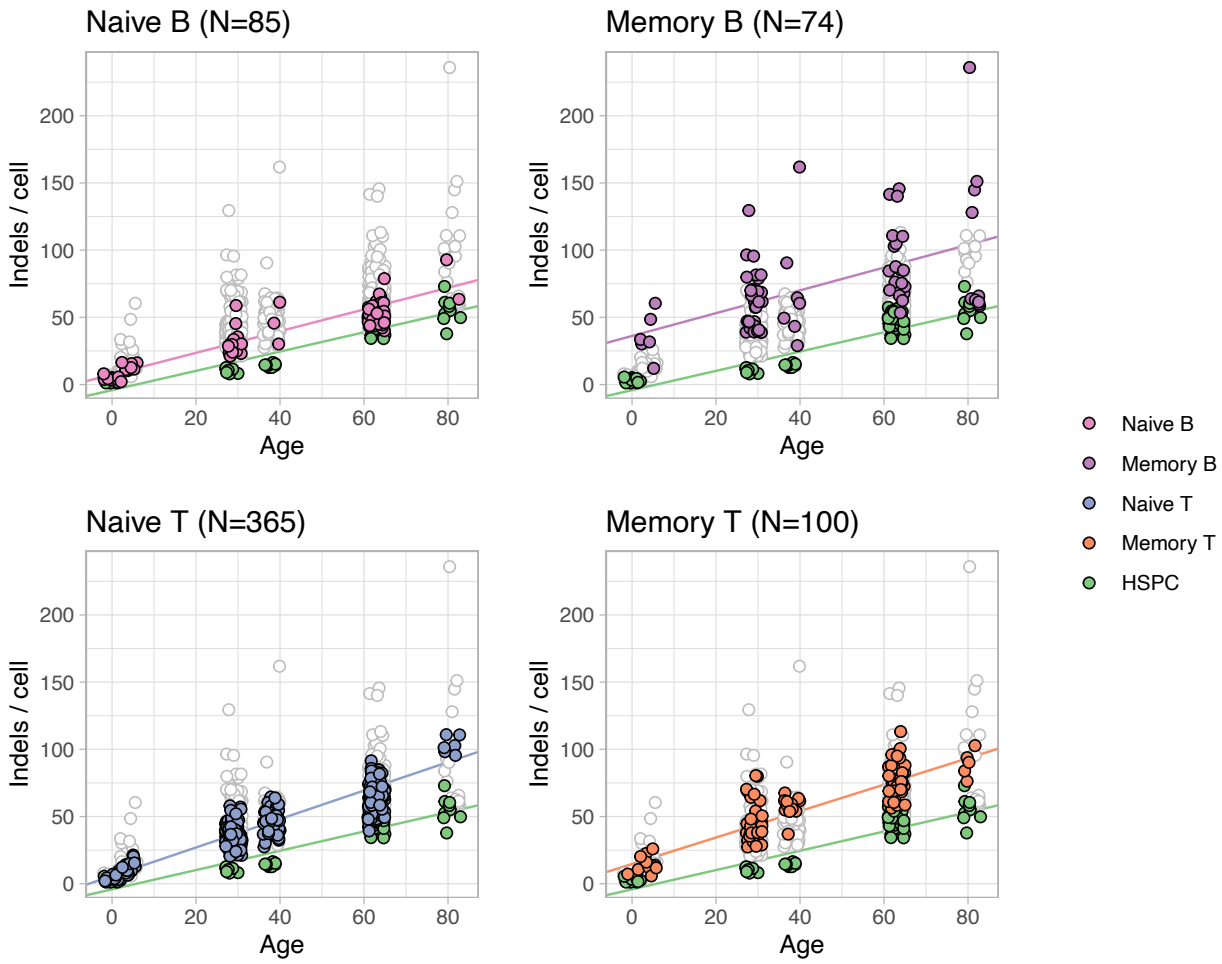

Fig. S2. Indel mutation burden per genome for the four main lymphocyte subsets (colored points), compared with HSPCs (green points). Each panel has all genomes plotted underneath in white with grey outline.

Figure S3

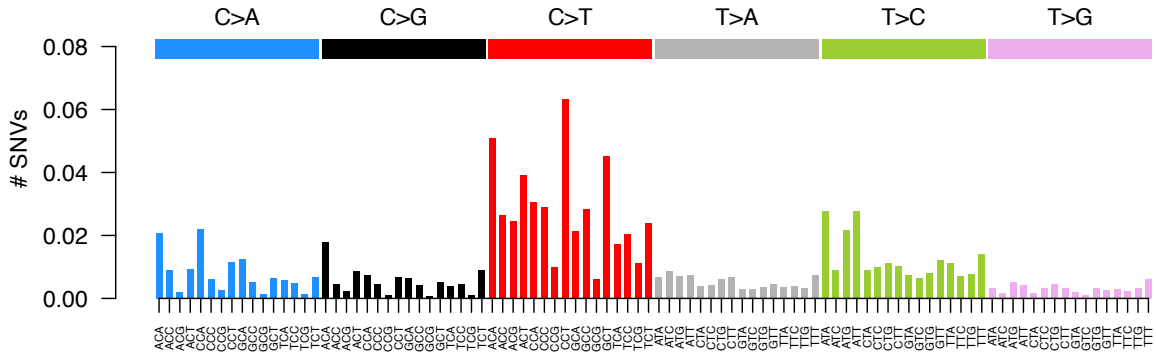

Fig. S3. SBSblood signature identified using HSPC genomes and the program *sigfit* (excludes SBS1).

Figure S4

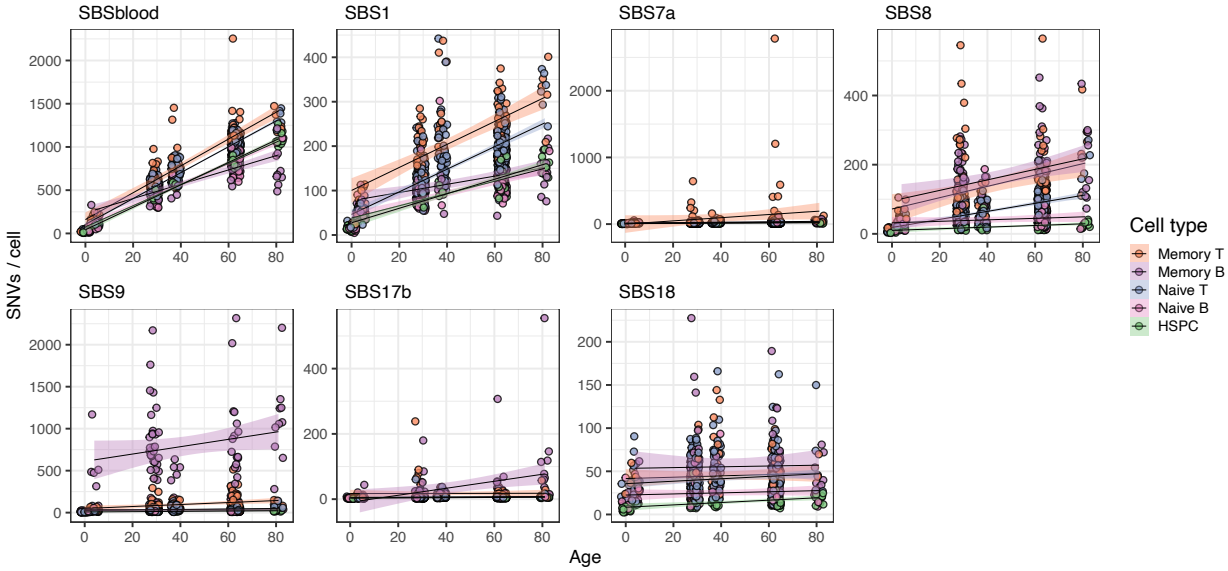

Figure S4. Mutation burden with age, per signature.

Figure S5

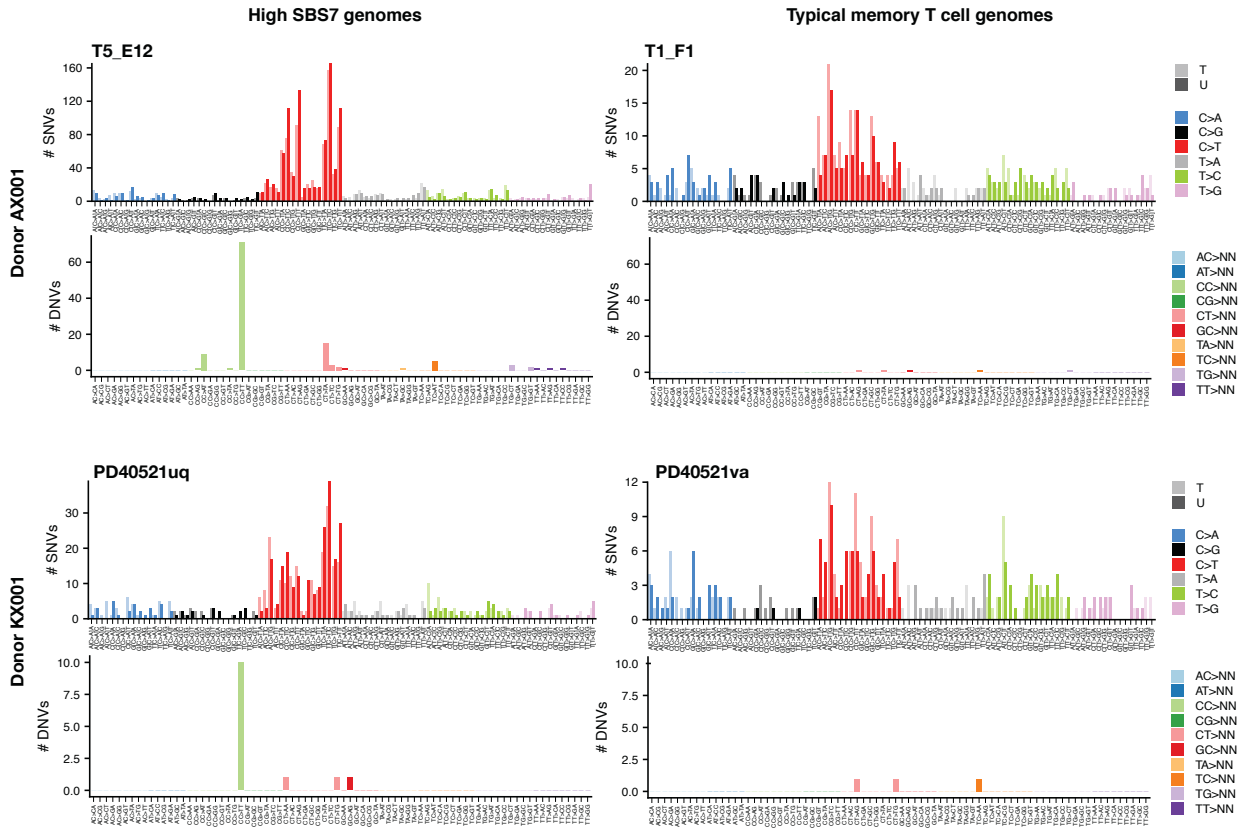

Fig. S5. Single and dinucleotide mutational profiles for high UV cells. The SNV profile shows the transcriptional strand bias, with mutations on transcribed strands in a lighter shaded bar to the left and mutations on untranscribed strands in a darker shaded bar to the right. The two sets of plots on the left are of genomes with high levels of SBS7 (samples T5_E12 and PD40521uq), while the plots of the right are of genomes with the more common relative absence of SBS7.

Figure S6

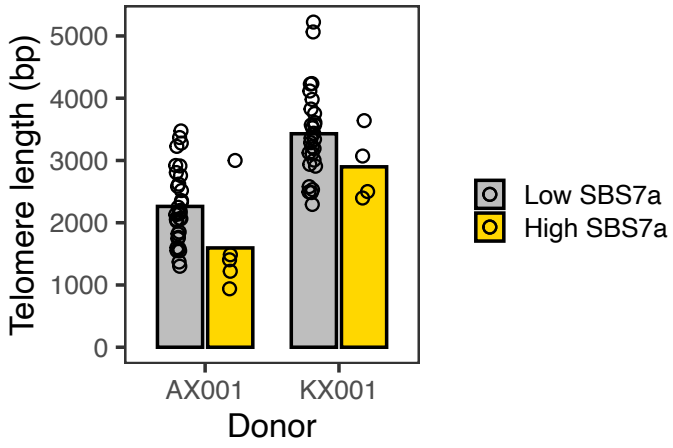

Fig. S6. Telomere length in high UV signature memory T cells. A high UV signature memory T cell is defined as having greater than 9.5% (2 standard deviations above the mean) of its mutations attributable to SBS7a.

Figure S7

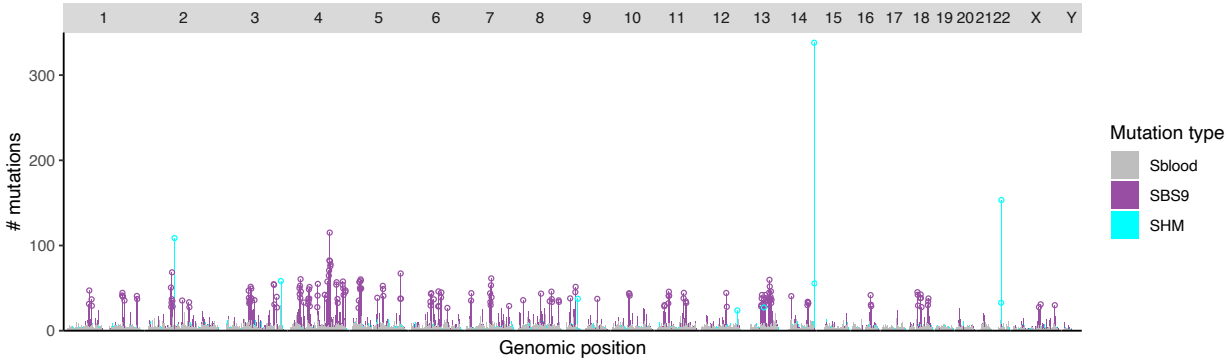

Fig. S7. SBS9 and SHM signature genomic distribution in memory B cells. Signature extraction is performed per 1Mb genomic bin. Open circles denote bins with more mutations than 2 standard deviations above the mean.

Figure S8

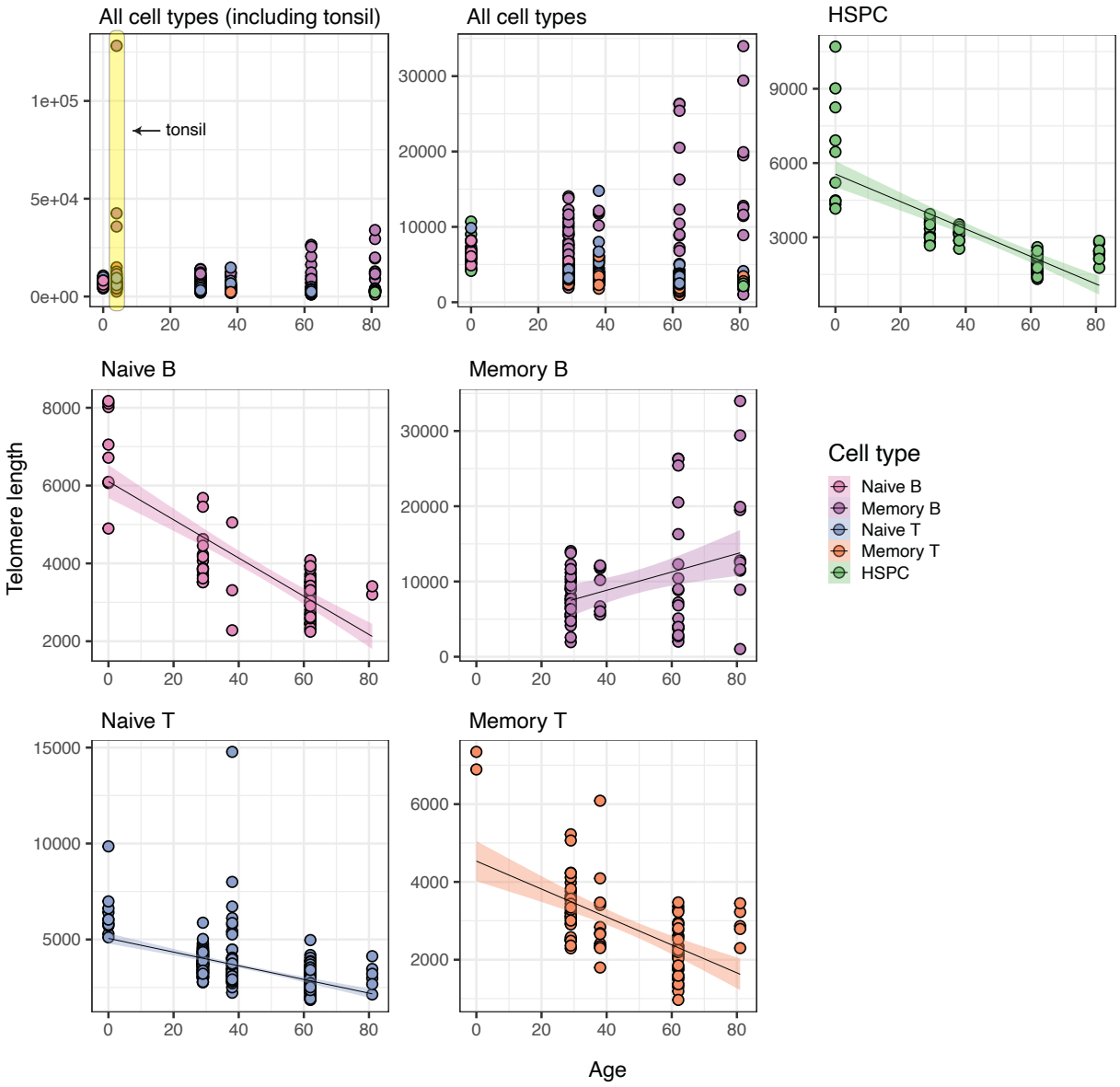

Fig. S8. Telomere length as a function of age. The top left panel includes the tonsil-derived genomes, which have an exceptionally high variance in telomere length. The remaining panels exclude these genomes.

Figure S9

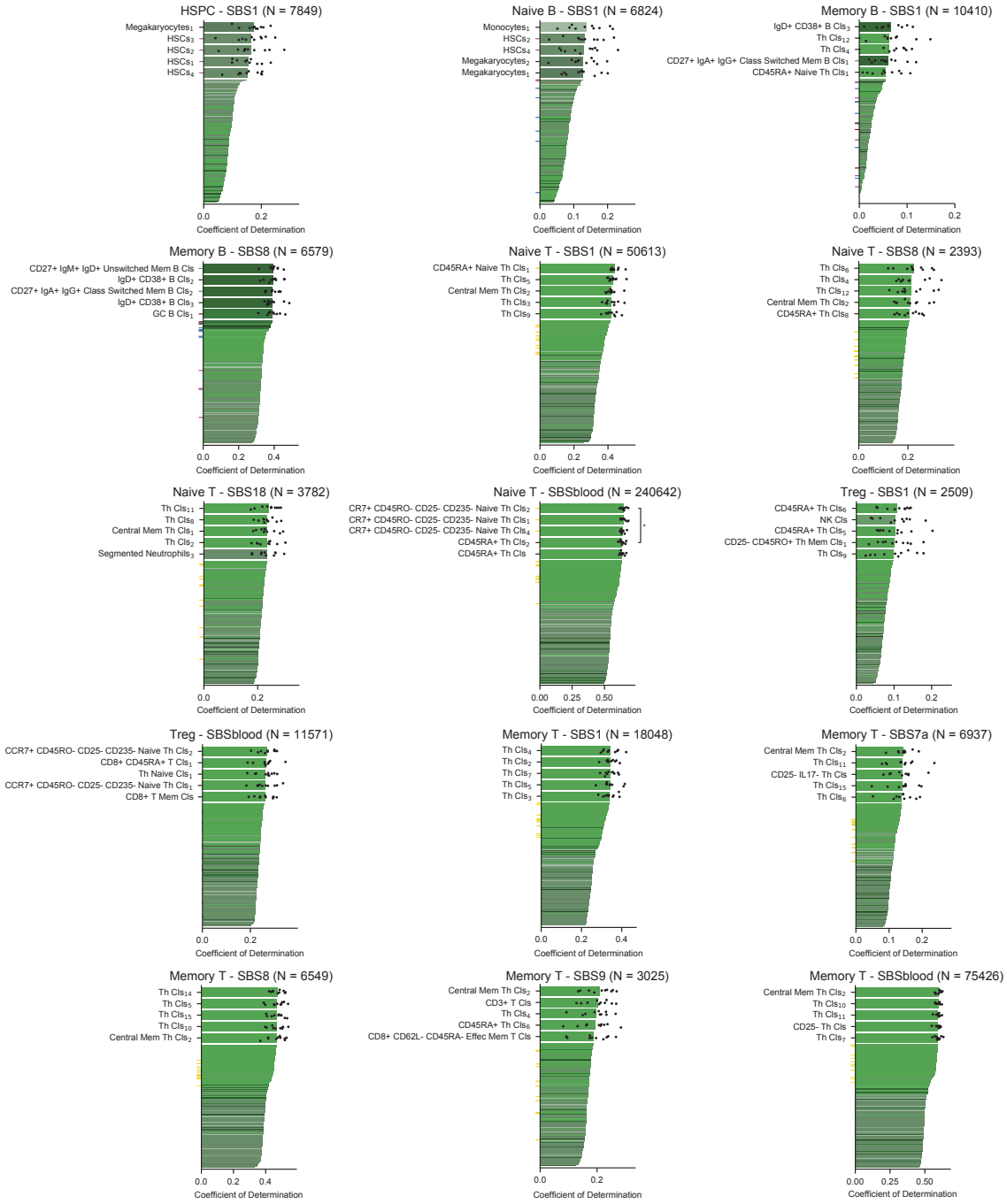

Fig. S9. Performance of prediction of genome-wide mutational profiles attributable to particular mutational signatures from histone marks of 149 epigenomes representing distinct blood cell types and different phases of development (subscripts indicate replicates); ticks are colored according to the epigenetic cell type (purple, HSC; blue, naive B cell; grey, memory B cell; maroon, GC B cell); black points depict values from ten-fold cross validation; p-values were obtained for the comparison of the 10-fold cross validation values using the two-sided Wilcoxon test (Cls, cells; CS, class switched; GC, germinal center; HSC, hematopoietic stem cell; Mem, memory).

Figure S10

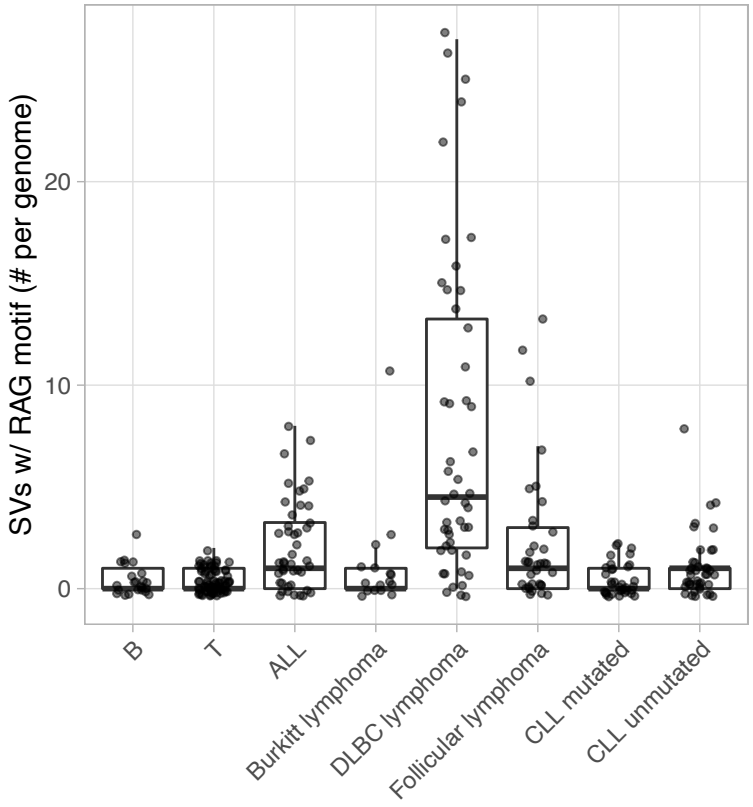

Fig. S10. Number of SV's with RSS (RAG) motifs within 50bp of a breakpoint.