Diverse mutational landscapes in human 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], Daniel Leongamornlert[1], Alyssa Cull[11], Francesco Maura[9], Mathijs A. Sanders[1,10], Alex TJ Cagan[1], Craig McDonald[2,3,11], Miriam Belmonte[2,3,11], Mairi S. Shepherd[2,3], Felipe A Vieira Braga[1], 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)

SUMMARY PARAGRAPH

A lymphocyte suffers many threats to its genome, including programmed mutation during differentiation[1], 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 haematopoietic stem cells. All lymphocyte subsets carried more point mutations and structural variants than haematopoietic stem cells, with higher burdens in memory than naive lymphocytes, and with T cells accumulating mutations at a higher rate throughout life. Off-target effects of immunological diversification accounted for approximately half the additional differentiation-associated mutations in lymphocytes. Memory B cells acquired, on average, 18 off-target mutations genome-wide for every one on-target *IGHV* mutation during the germinal centre reaction. Structural variation was 16-fold higher in lymphocytes than stem cells, with ~15% of deletions being attributable to off-target RAG activity. DNA damage from ultraviolet light exposure and other sporadic mutational processes generated hundreds to thousands of mutations in some memory lymphocytes. The mutation burden and signatures of normal B lymphocytes were broadly comparable to those seen in many B-cell cancers, suggesting that malignant transformation of lymphocytes arises from the same mutational processes active across normal ontogeny. 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]. Off-target genome editing in lymphocytes can produce mutations driving lymphoid malignancies, including RAG-mediated deletions in acute lymphoblastic leukaemia[2,3]; AID-mediated somatic hypermutation in diffuse large B-cell lymphoma[4–6]; and class-switch recombination in multiple myeloma[7].

While mutation accumulation in lymphoid malignancies is well characterised, mutation burden of normal lymphocytes has been less comprehensively studied. Patterns of base substitutions in 59 normal, CD19-positive B cells revealed age-related increase in burden, with evidence for off-target somatic hypermutation[8]. More detailed quantification and comparison of the genomic landscape of B versus T cells, naive versus memory lymphocytes, and normal versus malignant lymphocytes is lacking.

Genome sequencing of B and T lymphocytes

Growing single cells into colonies *in vitro* enables accurate identification of all classes of somatic mutation using genome sequencing[9–11]. 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**, **Supplementary Fig. 1**; **Methods**). Culture efficiencies varied by cell type, but were typically 2-5% (**Table S1**), which prompted us to evaluate whether there was evidence for potential bias in culture efficiency among lymphocytes (**Supplementary Note**). Reassuringly, cell surface marker expression was comparable between lymphocytes that succeeded or failed to grow colonies (**Extended Figure 1**). Furthermore, deep sequencing data for one donor showed strong correlation between variant allele fractions in bulk lymphocytes versus colonies (**Extended Figure 2A**) – using bootstrapping, we estimate that any bias in culture efficiency among lineages would amount to just 20% (for example, ranging from 0.04-0.06 for a mean efficiency of 0.05) for both B and T lymphocytes (**Supplementary Note**).

We obtained 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 S2**). All individuals studied were haematopoietically normal and healthy; one subject had a history of inflammatory bowel disease treated with azathioprine and the two tonsil donors had a history of tonsillitis. We focused on four classes of lymphocytes: naive B lymphocytes, memory B lymphocytes, CD4+ and CD8+ naive T lymphocytes, and CD4+ and CD8+ memory T lymphocytes. In one subject we also expanded T-regulatory cells. Five of the subjects reported here were also analysed in a parallel study[12] of haematopoietic stem and progenitor cells (HSPCs), with 39 overlapping HSPC genomes.

We performed whole genome sequencing to an average depth of ~20x. To confirm this provided sufficient depth, we calculated recall statistics for germline heterozygous variants for each colony, generating estimates of sensitivity of 80% at 10x and >98% at 20x depth (**Extended Fig. 2B**). The final dataset comprises 717 whole genomes (**Table S3**).

Mutation burden

The overall burden of both single nucleotide variants (SNVs) and insertion/deletions (indels) per cell varied extensively, influenced predominantly by age and cell type (**Fig. 1B**). The burden of single nucleotide variants (SNVs) 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 base substitutions at ~16 SNVs/cell/year ($CI_{95\%}$=13–19), similar to previous estimates[10,12]. Naive and memory B cells showed broadly similar rates of mutation accumulation (naive B: 15 SNVs/cell/year, $CI_{95\%}$=12–18; memory B cells: 17 SNVs/cell/year, $CI_{95\%}$=6–28). T cells, though, had higher mutation rates (naive T: 22 SNVs/cell/year, $CI_{95\%}$=19–25; memory T cells: 25 SNVs/cell/year, $CI_{95\%}$=17–32). Overall, this suggests that there are clock-like mutational processes adding mutations at constant 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/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 SNVs/cell than HSPCs respectively. 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.

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 donor, 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). This cell-to-cell variability within a donor considerably outweighed the between-person standard deviation, which we estimated at 60 SNVs/cell.

Indels accumulated at an average of 0.7/cell/year in HSPCs ($CI_{95\%}$=0.5–0.9), while lymphocytes had higher indel rates (naive B cells: 0.8/cell/year, $CI_{95\%}$=0.6–1.0; naive T cells: 1.1, $CI_{95\%}$=0.9–1.2; memory B cells: 0.8, $CI_{95\%}$=0.4–1.3; memory T cells: 1.0, $CI_{95\%}$=0.7–1.2; **Extended Fig. 3A**).

Somatic mutations can confer selective advantage on normal cells, driving clonal expansions. Global measures of the strength of positive selection can be obtained by estimating the excess of non-synonymous mutations compared to selectively neutral synonymous mutations[13] (dN/dS ratio, with dN/dS=1 denoting neutrality). Exome-wide, excluding immunoglobulin regions, we estimated the dN/dS ratio in lymphocytes to be 1.12 ($CI_{95\%}$=1.06-1.19). This implies that positive selection shapes clonal competition in lymphocytes, with approximately 11% ($CI_{95\%}$=6-15%) of non-synonymous mutations conferring a selective advantage (**Extended Fig. 3B**). At a single-gene level, *ACTG1* was the only gene significant with false-discovery rate <1% (q=5x10^{-3}) – this gene is recurrently mutated in the plasma cell malignancy, multiple myeloma[14,15].

Mutational signatures

In order to determine whether the excess mutations observed in lymphocyte subsets were due to a specific mutational process, we inferred 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 deamination of methylated cytosines, 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[10,11], which we term 'SBSblood' (**Extended Fig. 4A**). The burden of both signatures correlated linearly with age (**Extended Fig. 4B-C**), 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. 2B**). 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, SBS7a and SBS17a do not, consistent with them being sporadic. SBS7a and SBS17b likely represent exogenous mutational processes, discussed in the next section, while SBS9 is differentiation-associated, discussed thereafter.

Exogenous mutational signatures

SBS7a is the canonical signature of ultraviolet light damage, the predominant mutational process in melanoma[16] and normal skin[17]. The signature we extracted in memory lymphocytes matches the features of SBS7a, with a predominance of C>T substitutions in a dipyrimidine context, transcriptional strand bias and a high rate of CC>TT dinucleotide substitutions (**Fig. 2C**; **Extended Fig. 5**). We found a substantial contribution of SBS7a (>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 SBS7a had significantly shorter telomeres than other memory T cells (p=0.01, Fisher's method; **Extended Fig. 5B**), indicative of increased proliferation. As UVB radiation only penetrates human skin to a depth of 10-50µm[18], the most plausible source of these SBS7a mutations is UV exposure during skin residency.

A second unexpected signature in memory lymphocytes was SBS17. This signature has been observed in cancers of the stomach and oesophagus and occasionally in B and T cell lymphomas[16]. This signature, characterised by T>G mutations in a $\underline{T}$pT 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-fluorouracil chemotherapy in metastatic cancers[19,20], but its occurrence in primary oesophageal 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 gastrointestinal mucosa.

Signatures of the germinal centre

Somatic hypermutation (SHM) at heavy and light chain immunoglobulin regions followed the expected mutational signature (**Fig. 3A**), with the productive rearrangement showing more mutations than non-recombined alleles (**Extended Fig. 6A-C**). However, as reported for lymphoid malignancies[5], off-target mutations in memory B cells, with signature SBS9, had a different spectrum to SHM mutations, characterised by mutations at A:T base-pairs in a $\underline{T}$pW context (**Fig. 3A**), and different distribution across the genome (**Extended Fig. 6D**). SBS9 accounted for 42% mutations (mean, 780 mutations/cell) in memory B cells, at times tripling the baseline mutation burden.

The number of SBS9 mutations genome-wide showed a strong linear correlation with the SHM rate (percentage of the productive *IGHV* gene that was mutated), despite their different spectra (R^2=0.57, p=4x10^{-9}, linear regression; **Fig. 3B**). 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 *IGHV* mutation is accompanied by an average of 18 SBS9 mutations elsewhere in the genome.

Another feature of the germinal centre reaction is increased telomerase activity in B cells[21,22]. We estimated telomere lengths from the genome sequencing data for our dataset. Telomere lengths in HSPCs, T lymphocytes and naive B cells decreased by ~30-50bp/year across life, consistent with cell divisions occurring every 6-24 months[23–25] (**Extended Fig. 7A**). In contrast, 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). Telomere lengths also correlated linearly with the number of SBS9 mutations genome-wide (R^2=0.37, p=3x10^{-8}; **Fig. 3C**). This correlation supports a hypothesis of lengthening telomeres and occurrence of off-target SBS9 mutations during the germinal centre reaction.

A replicative-stress model of SBS9

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[26], which has an error spectrum weighted towards a $\underline{T}$pW context[27]. As has been noted in lymphoid malignancies[5,16], 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 $\underline{T}$pW context (**Fig. 3A**), which makes the role of AID unclear because it specifically targets cytosines. The genome-wide distribution of off-target AID-induced deamination has been measured directly[28], 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 S4**). 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 (**Extended Fig. 7B**). 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 – this is in contradistinction to AID-induced deamination, which occurs in actively transcribed regions[28].

Taken together, our data demonstrate that SBS9 accumulates during the germinal centre 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 hypothesise 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 centre B cells. Normally, mismatch repair and other pathways would accurately correct such lesions, but the high expression of polymerase-eta in germinal centre cells[29] 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[30,31], would be consistent with this as a model of SBS9 mutation.

Epigenetic marks reveal mutation timing

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 fertilised egg[32,33]. We thus hypothesised 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 compared the distribution of somatic mutations across the genome to 149 epigenomes representing 48 distinct blood cell types and differentiation stages. Mutations resulting from the clock-like signature SBSblood in HSPCs correlated best with histone marks from haematopoietic stem cells (p=0.002, Wilcoxon test; **Fig. 3E**), consistent with mutation accumulation in undifferentiated cells. Notably, SBSblood mutational profiles in naive B cells also correlated better with the epigenomes of haematopoietic stem cells than naive B cells (p=0.004; **Fig. 3E**). 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; **Extended Fig. 8**), consistent with a large, long-lived 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. 3E**), 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 centre B cell epigenomes (p=0.049; **Fig. 3E**). This is consistent with our finding of a correlation between SBS9 and other germinal centre-associated processes (SHM and telomere lengthening), providing further evidence that SBS9 arises as a by-product of the germinal centre reaction. For SBS7a, the signature of ultraviolet light exposure seen in memory T cells, the genomic distribution more tightly correlated with epigenomes of differentiated T cells than naive T cells (**Extended Fig. 8**), supporting the hypothesis that SBS7a mutations accumulate in differentiated T cells.

Structural variants

Both V(D)J recombination and class-switch recombination (CSR) are associated with off-target structural variation (SV) in human lymphoid malignancies[2,3,7], but 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 (**Extended Fig. 9**). We identified fewer than the 2 expected on-target V(D)J recombination events per lymphocyte, suggesting that our sensitivity for SVs in these regions is ~62%.

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 (**Fig. 4A**) and cycles of templated insertions[34], most non-Ig/TCR SVs were deletions (49%), several of which affected genes mutated in lymphoid malignancies (**Fig. 4B**, **Table S5**).

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**). Accounting for 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 breakpoints, reaching background levels within ~100bp (**Fig. 4E**). During V(D)J recombination, the TdT protein adds random nucleotides at the dsDNA breaks – this also occurs 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).

Class-switch recombination is achieved through AID cytosine deamination at WGCW clusters, deleting IgH constant region genes and changing the antibody isotype. As expected, on-target CSR was enriched in memory (76%) compared to naive B lymphocytes (12%; **Figure 4F**; **Table S6**). 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 tumours require additional mutational processes to acquire sufficient driver mutations for oncogenic transformation[35]. In many solid tissues, cancers have higher mutation burdens than normal cells from the same organ[36,37], but myeloid leukaemias do not[9]. To address this question in lymphoid malignancies, we compared genomes from normal B and T lymphocytes to 8 blood cancers[38–40], which had similar distributions of effective sequencing coverage (**Extended Fig. 9C**). SNV burdens for follicular lymphoma, diffuse large B-cell lymphoma and multiple myeloma were considerably higher than normal lymphocytes (**Fig. 5A-B**). In contrast, point mutation burdens observed in Burkitt lymphoma, mutated or unmutated chronic lymphocytic leukaemia and acute myeloid leukaemia were well within the range of normal lymphocytes. 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**). Cutaneous T-cell lymphomas carried comparable numbers of mutations attributable to ultraviolet light as the SBS7a-high memory lymphocytes (**Extended Fig. 5C**). These data emphasise that the processes generating point mutations in normal lymphocytes can generate sufficient somatic variants for progression towards many types of lymphoid malignancy.

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.

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 leukaemia had similarly high proportions of RAG-mediated events, but in much higher numbers, as reported previously[2,3] (**Extended Fig. 10A**). 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 (**Extended Fig. 10B**). 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.

Discussion

Positive selection acting on somatic mutations in lymphocytes is more pervasive than negative selection, suggesting that clonal expansions of individual lymphocytes are the evolutionary trade-off for physiological genome editing. Lymphoid cancers are clearly one consequence – that mutation burdens and signatures of normal lymphocytes match those seen in lymphoid malignancies argues that off-target mutagenesis is sufficient to transform occasional lymphocytes. For 50+ years, there has been speculation that driver mutations could underpin autoimmune diseases[41–43], with recent data showing driver mutations in lymphocytes responsible for vasculitis associated with Sjögren's disease[44]. Our data show, first, that mutation rates are high enough to generate considerable genetic diversity among normal lymphocytes, and, second, that selective pressures favour clonal expansion of individual lymphocytes.

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. Given that lymphocytes divide every 3–24 months[45], data supported by our estimates of telomere attrition, mutation rates during these maintenance phases would presumably be ~5–50/cell division. 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, contributing additional mutations. The considerably greater cell-to-cell variation than person-to-person variation suggests that lifelong environmental forces (infections, inflammation, skin residency) are stronger influences on lymphocyte genomes than inherited variation in mutation rates. 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.

REFERENCES

1. Tarlinton, D. & Good-Jacobson, K. Diversity Among Memory B Cells: Origin, Consequences, and Utility. *Science (80-.).* **341**, 1205–1212 (2013).
2. Mullighan, C. G. *et al.* Genomic Analysis of the Clonal Origins of Relapsed Acute Lymphoblastic Leukemia. *Science (80-.).* **322**, 1377–1380 (2008).
3. Papaemmanuil, E. *et al.* RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. *Nat Genet* **46**, 116–125 (2014).
4. Pasqualucci, L. *et al.* Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas. *Nature* **412**, 341–346 (2001).
5. Kasar, S. *et al.* Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* **6**, 8866 (2015).
6. Hadj Khodabakhshi, A. *et al.* Recurrent targets of aberrant somatic hypermutation in

lymphoma. *Oncotarget* **3**, 1308–1319 (2012).

7. Walker, B. A. *et al.* 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).
8. Zhang, L. *et al.* Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. *Proc. Natl. Acad. Sci. U. S. A.* 535906 (2019). doi:10.1073/pnas.1902510116
9. Welch, J. S. *et al.* The Origin and Evolution of Mutations in Acute Myeloid Leukemia. *Cell* **150**, 264–278 (2012).
10. Osorio, F. G. *et al.* Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. *Cell Rep.* **25**, 2308-2316.e4 (2018).
11. Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
12. Mitchell, E. *et al.* Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 343–350 (2022).
13. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029–1041 (2017).
14. Lohr, J. G. *et al.* Widespread genetic heterogeneity in multiple myeloma: Implications for targeted therapy. *Cancer Cell* **25**, 91–101 (2014).
15. Maura, F. *et al.* Genomic landscape and chronological reconstruction of driver events in multiple myeloma. *Nat. Commun.* **10**, (2019).
16. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
17. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science (80-.).* **348**, 880–886 (2015).
18. Meinhardt, M., Krebs, R., Anders, A., Heinrich, U. & Tronnier, H. Wavelength-dependent penetration depths of ultraviolet radiation in human skin. *J. Biomed. Opt.* **13**, 044030 (2008).
19. Pich, O. *et al.* The mutational footprints of cancer therapies. *Nat. Genet.* **51**, 1732–1740 (2019).
20. Christensen, S. *et al.* 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. *Nat. Commun.* **10**, 4571 (2019).
21. Weng, N. P., Granger, L. & Hodes, R. J. Telomere lengthening and telomerase activation during human B cell differentiation. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 10827–32 (1997).
22. Norrback, K. F. *et al.* Telomerase regulation and telomere dynamics in germinal centers. *Eur. J. Haematol.* **67**, 309–317 (2001).
23. Vaziri, H. *et al.* Loss of telomeric DNA during aging of normal and trisomy 21 human lymphocytes. *Am. J. Hum. Genet.* **52**, 661–7 (1993).
24. Weng, N. P., Hathcock, K. S. & Hodes, R. J. Regulation of telomere length and telomerase in T and B cells: a mechanism for maintaining replicative potential. *Immunity* **9**, 151–7 (1998).
25. Weng, N. P., Levine, B. L., June, C. H. & Hodes, R. J. Human naive and memory T lymphocytes differ in telomeric length and replicative potential. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 11091–11094 (1995).

26. Wilson, T. M. *et al.* MSH2-MSH6 stimulates DNA polymerase η, suggesting a role for A:T mutations in antibody genes. *J. Exp. Med.* **201**, 637–645 (2005).

27. Rogozin, I. B., Pavlov, Y. I., Bebenek, K., Matsuda, T. & Kunkel, T. A. Somatic mutation hotspots correlate with DNA polymerase η error spectrum. *Nat. Immunol.* **2**, 530–536 (2001).

28. Álvarez-Prado, Á. F. *et al.* A broad atlas of somatic hypermutation allows prediction of activation-induced deaminase targets. *J. Exp. Med.* **215**, 761–771 (2018).

29. Mcheyzer-Williams, L. J., Milpied, P. J., Okitsu, S. L. & Mcheyzer-Williams, M. G. Class-switched memory B cells remodel BCRs within secondary germinal centers. *Nat. Immunol.* **16**, 296–305 (2015).

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

31. Frigola, J. *et al.* Reduced mutation rate in exons due to differential mismatch repair. *Nat. Genet.* **49**, 1684–1692 (2017).

32. Gonzalez-Perez, A., Sabarinathan, R. & Lopez-Bigas, N. Local Determinants of the Mutational Landscape of the Human Genome. *Cell* **177**, 101–114 (2019).

33. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).

34. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).

35. Tomlinson, I. P., Novelli, M. R. & Bodmer, W. F. The mutation rate and cancer. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 14800–3 (1996).

36. Brunner, S. F. *et al.* Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538–542 (2019).

37. Yoshida, K. *et al.* Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).

38. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, (2020).

39. Maura, F. *et al.* Genomic landscape and chronological reconstruction of driver events in multiple myeloma. *Nat. Commun.* **10**, 3835 (2019).

40. McGirt, L. Y. *et al.* Whole-genome sequencing reveals oncogenic mutations in mycosis fungoides. *Blood* **126**, 508–519 (2015).

41. Dameshek, W. & Schwartz, R. S. Leukemia and auto-immunization- some possible relationships. *Blood* **14**, 1151–1158 (1959).

42. Burnet, F. M. A reassessment of the forbidden clone hypothesis of autoimmune disease. *Aust. J. Exp. Biol. Med. Sci.* **50**, 1–9 (1972).

43. Goodnow, C. C. Multistep Pathogenesis of Autoimmune Disease. *Cell* **130**, 25–35 (2007).

44. Singh, M. *et al.* Lymphoma Driver Mutations in the Pathogenic Evolution of an Iconic Human Autoantibody. *Cell* **180**, 878-894.e19 (2020).

45. Macallan, D. C., Busch, R. & Asquith, B. Current estimates of T cell kinetics in humans. *Curr. Opin. Syst. Biol.* **18**, 77–86 (2019).

FIGURE LEGENDS

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, compared with HSPCs (green points). Each panel has all genomes plotted underneath in white with grey outline. The lines show the fit for the respective populations by linear mixed effects models.

Fig. 2. Mutational processes in lymphocytes. (A) The proportion of SNVs and (B) SNV burden per mutational signature. Each column represents one genome. Per genome, signatures with a 90% CI lower bound of less than 1% are excluded from plotting. (C) Mutational spectra of single colony genomes enriched in the specified mutational signature. The specific genome plotted is identified with the corresponding Roman numeral in panel (B). Trinucleotide contexts on the x-axis represent 16 bars within each substitution class, divided into 4 sets of 4 bars, grouped by the nucleotide 5′ to the mutated base, and within each group by the 3′ nucleotide (in the order A, C, G, then T).

Fig. 3. Correlation of SBS9 with genomic attributes and timing of mutational processes. (A) Mutational spectra of the SBS9 and SHM signatures. Trinucleotide contexts on the x-axis represent 16 bars within each substitution class, divided into 4 sets of 4 bars, grouped by the nucleotide 5′ to the mutated base, and within each group by the 3′ nucleotide. (B) Scatterplot of the number of SBS9 mutations genome-wide and the percentage of bases in *IGHV* mutated in the productive rearrangement of memory B cells. The line represents the linear regression estimate of the correlation; the p-value and R^2 are for this model. (C) Scatterplot of SBS9 versus telomere length per genome, coloured by cell type. The regression line is for memory B cells; the p-value and R^2 are for this model. (D) Explanatory power of each genomic feature significant in the generalised additive model (GAM), expressed as the R^2 of the individual GAM model for predicting number of SBS9 mutations (left) or number of SBSblood/SBS1 mutations (right) per 10kb window. (E) 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 coloured 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 centre; HSC, hematopoietic stem cell; Mem, memory). Mega: megakaryocyte.

Fig. 4. Structural variation burden and off-target RAG-mediated deletion. (A) Chromoplexy cycle (sample PD40667sl, donor KX002). Black points represent corrected read depth along the chromosome; arcs denote structural variants. The final genomic configuration of the four derivative chromosomes is shown as coloured arrows underneath. (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 for Ig/TCR (0.96) and non-Ig-TCR (0.24) regions. Black dashed line represents the genomic background rate of RAG motifs. Error bars represent 95% bootstrap confidence intervals. n = 889 Ig/TCR SVs and 253 non- Ig/TCR SVs. (E) 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.

Fig. 5. Comparison of mutational patterns with malignancy. (A) SNV and (B) SV burden by normal cell type or malignancy. Boxes show the interquartile range and the centre horizontal lines show the median. Whiskers extend to the minimum of either the range or 1.5× the interquartile range. Normal lymphocytes (bold) exclude paediatric samples. (C) Proportion of mutational signatures per genome. Per genome, signatures with a 90% CI lower bound of less than 1% are excluded from plotting. Normal lymphocytes (pink) are from donor AX001. (D) SBS9 burden and (E) proportion by cell/malignancy type. Boxes show the interquartile range and the centre horizontal lines show the median. Whiskers extend to the minimum of either the range or 1.5× the interquartile range. (F,G) Heatmap showing the level of enrichment of (F) SBS9 and (G) SHM signatures nearby frequently mutated genes for that signature compared to the whole genome. Number of SVs per group: B = 145, T = 841, ALL = 523, Burkitt lymphoma = 305, CLL mutated = 252, CLL unmutated = 440, C. T-cell lymphoma = 204, DLBC lymphoma = 3754, follicular lymphoma = 1095. (A,B,D,E) Number of genomes per group: naive B = 68, memory B = 68, naive T = 332, memory T = 87, Burkitt lymphoma = 17, CLL (chronic lymphocytic leukaemia) mutated = 38, CLL unmutated = 45, C. (cutaneous) T-cell lymphoma = 5, DLBC (diffuse large B-cell) lymphoma = 47, follicular lymphoma = 36, multiple myeloma = 30, myeloid-AML (acute myeloid leukaemia) = 10.

METHODS SUMMARY

Samples

Human blood mononuclear cells (MNCs) were obtained from four sources: (1) bone marrow, spleen and peripheral blood taken with written informed consent (provided by next-of-kin) from three deceased 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 taken with written informed consent from one patient (AX001) recruited from Addenbrooke's Hospital (approval 07-MRE05-44), (3) tonsil taken with written informed consent from guardians of two patients (TX001, TX002) recruited from Addenbrooke's Hospital (approval 07-MRE05-44), and (4) one cord blood (CB001) collected with written informed consent from guardian by StemCell Technologies (catalogue #70007) (**Table S2**). All sources were haematopoietically 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$^-$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. Details for the antibody panels used are in **Table S11**. Cells were either single-cell sorted for liquid culture into 96-well plates containing 50µL cell type-specific expansion medium, or (for AX001 HSPCs) bulk-sorted for MethoCult plate-base expansion. Plotting of the FACs data was performed with FlowJo and FCS Express.

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. Detailed step-by-step descriptions of the protocols are provided in **Supplementary Information**. The B cell expansion medium was composed of 5µg/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 1.25µg/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 12.5µL/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). 25µL 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 12 days after sorting (depending on growth).

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 100µL 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 50µL 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 50µl 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 5-10µL 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[46]. 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.

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[47], insertion/deletions (indels) using Pindel[48], structural variants (SVs) using BRASS[49] and copy number variants (CNVs) using ASCAT[50]. 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. 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 artefacts resulting from the low-input library preparation. Detailed description of the artefact filters were provided previously[46] 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 cut-offs, 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 estimated our sensitivity to detect SNVs using germline mutations as a truth set of heterozygous mutations. We called germline mutations by performing a one-sided exact binomial test of the sum of the alternate and sum of the total reads across colonies of an individual for each CaVEMan unfiltered variant (alternate hypothesis of proportion of successes less than 0.5 for autosomes and female X chromosomes, 0.95 for male sex chromosomes). A variant was called as germline on failure to reject the null at a false-discovery rate q-value of 10^{-6}. We calculated sensitivity as the proportion of germline variants detected per colony.

We removed artefacts from the SV calls using AnnotateBRASS with default settings. The full list of statistics calculated and post-hoc filtering strategy was described in detail previously[36]. 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[12]. 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 S7**. 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 maximum likelihood estimation.

Detecting positive selection

In order to estimate an exome-wide rate of selection and to detect selection acting on specific genes we used the dndscv function of the dNdScv R package[13]. This program leverages mutation rate information across genes. As the elevated mutation rate seen with somatic hypermutation may break the assumptions of the test, we excluded the immunoglobulin loci from these analyses (excluded GRCh37 regions: chr14:106304735-107283226, chr2:89160078-90274237, chr22:22385390-23263607). We performed the test for the following subsets of the data: all lymphocytes, naive B, memory B, naive T, memory T, all lymphocytes testing only cancer genes and all lymphocytes excluding cancer genes. Cancer genes were defined as the 566 tier 1 genes from the COSMIC Cancer Gene Census (https://cancer.sanger.ac.uk, downloaded June 6, 2018).

Mutational signature analysis

We characterised 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 leukaemia (mutated), chronic lymphocytic leukaemia (unmutated), and acute myeloid leukaemia[38] and multiple myeloma[39]. We identified mutational signatures present in the data by performing signature extraction with two programs, *SigProfiler*[51] and *hdp* (https://github.com/nicolaroberts/hdp). We used the *SigProfiler* denovo results for the suggested number of extracted signatures. *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[16] SBS1, SBS5, SBS7a, SBS8, SBS9, SBS13, SBS17b, SBS18, SBS19 (version 3).

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[10,11]. 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*[52] to call two signatures across our HSPC genomes, SBS1 and a novel signature, with the novel signature being the final SBSblood (**Extended Fig. 4A**; **Table S8**). SBSblood was highly similar to both the *hdp* and *SigProfiler de novo* 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 S8**).

Ig receptor sequence analysis

In order to identify the immunoglobulin (Ig) rearrangements, productive CDR3 sequences and percent somatic hypermutation for each memory B cell, we ran *IgCaller*[53], 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 identified by *IgCaller* in the productive *IGHV* gene divided by the gene length. For class-switch recombination calling see **Supplementary Information**.

We estimated the number of mutations resulting from on-target (*IGHV* gene) somatic hypermutation compared with those associated with SBS9. We first counted all *IGHV* 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 centre-associated mutations in B cells

We assessed the genomic distribution of the germinal centre-associated mutational signatures, SBS9 and the SHM signature, in memory B cells. We performed per-Mb *de novo* 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 S8**) 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/SBS1.

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 leukaemia, 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 after 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 S4**). 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*[54]. SBS9 was analysed separately from the SBSblood/SBS1 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 S4**.

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') for the 50bp flanking each SV breakpoint using the program FIMO ($p<10^{-4}$)[55]. 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)[56]. 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 leukaemia genomes[3].

Telomere length

We estimated the telomere length for HSPC and lymphocyte genomes (**Table S3**) using the program Telomerecat[57]. Telomere lengths for all genomes for a given donor were estimated as a group.

Timing of mutational processes

Following a procedure described previously[33,58], 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[59–61] (**Table S9**) 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.

DATA AVAILABILITY

Raw sequencing data are available at the European Genome-Phenome Archive (accession number EGAD00001008107: https://ega-archive.org/datasets/EGAD00001008107). All somatic mutation calls and other relevant intermediate datasets are available on the github repository at https://github.com/machadoheather/lymphocyte_somatic_mutation.

CODE AVAILABILITY

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

ADDITIONAL REFERENCES

46. Ellis, P. *et al.* Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat. Protoc.* 1–31 (2020). doi:10.1038/s41596-020-00437-6
47. Jones, D. *et al.* cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. in *Current Protocols in Bioinformatics* **2016**, 15.10.1-15.10.18 (John Wiley & Sons, Inc., 2016).
48. Raine, K. M. *et al.* cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. *Curr. Protoc. Bioinformatics* **52**, 15.7.1-15.7.12 (2015).
49. Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40**, 722–9 (2008).
50. Raine, K. M. *et al.* ascatNgs: Identifying Somatically Acquired Copy-Number Alterations

from Whole-Genome Sequencing Data. in *Current Protocols in Bioinformatics* **2016**, 15.9.1-15.9.17 (John Wiley & Sons, Inc., 2016).

51. Bergstrom, E. N. *et al.* SigProfilerMatrixGenerator: A tool for visualizing and exploring patterns of small mutational events. *BMC Genomics* **20**, 1–12 (2019).
52. Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures. *bioRxiv* 372896 (2020). doi:10.1101/372896
53. Nadeu, F. *et al.* IgCaller for reconstructing immunoglobulin gene rearrangements and oncogenic translocations from whole-genome sequencing in lymphoid neoplasms. *Nat. Commun.* **11**, 1–11 (2020).
54. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 139–44 (2010).
55. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–8 (2011).
56. Bailey, T. L. & Noble, W. S. Searching for statistically significant regulatory modules. *Bioinformatics* **19**, (2003).
57. Farmery, J. H. R. *et al.* Telomerecat: A ploidy-agnostic method for estimating telomere length from whole genome sequencing data. *Sci. Rep.* **8**, 1–17 (2018).
58. Kübler, K. *et al.* Tumor mutational landscape is a record of the pre-malignant state. *bioRxiv* 517565 (2019). doi:10.1101/517565
59. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–329 (2015).
60. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
61. Stunnenberg, H. G. *et al.* The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167**, 1145–1149 (2016).

MAIN TEXT STATEMENTS

Acknowledgments

This work was supported by the WBH Foundation and Wellcome Trust. Investigators at the Sanger Institute are supported by a core grant from the 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 D.G.K. 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.

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.

Author Contributions

H.E.M, P.J.C. and D.K. designed the experiments; P.J.C. and D.K. supervised the project; H.E.M. designed the lymphocyte expansion protocols with advice from D.K., D.H., N.F.O., M.B., and M.S.S; H.E.M. and M.D. performed the lymphocyte cell sorting and colony growth with advice from F.A.V.B, N.F.O., D.H., D.K. and E.L.; H.E.M. and M.D. performed the CellCelector colony picking with advice from C.M.; E.M. and N.F.O. performed the HSPC sorting and colony growth with advice from E.L.; H.E.M. performed the data analyses with advice from M.A.S., R.J.O., I.M., A.R.G., F.M. and P.J.C.; K.M. and K.S.P. collected and processed samples; A.T.J.C. created the artwork for Figure 1A; D.L. performed the class-switch recombination analysis; A.C. analysed the FACs data; K.K. performed the association of epigenetic marks and mutational signatures with advice from G.G. and P.P.; H.E.M and P.J.C. wrote the manuscript; all authors reviewed and edited the manuscript.

Competing interests

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. D.J.H. receives research funding from AstraZeneca and D.K. receives research funding from STRM.bio. All other authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to P.J.C. and D.G.K.

EXTENDED FIGURE LEGENDS

Extended Figure 1. Assessment of culture bias by index flow-sorting. (A) Representative scatterplots of cell surface marker fluorescence intensity measured by flow cytometry (sort AX001 10/05/2018; AX001 13/11/2018 for Treg gate). Cells that successfully seeded colonies are coloured red; cells that did not form colonies are coloured grey. (B) Box-and-whisker plots showing fluorescence intensity for different cell surface markers in the various lymphocyte populations (columns) across different patients and days of flow-sorting (rows). Cells that successfully seeded colonies are shown in teal; cells that did not form colonies in orange. Boxes show the interquartile range and the centre horizontal lines show the median. Whiskers extend to the minimum of either the range or 1.5× the interquartile range. Red asterisks show a statistically significant difference between the fluorescence values of colony forming versus non-colony forming cells (two-sided t-test, false-discovery rate *q<0.05, **q<0.01, ***q<0.001, p-values in **Table S10**). The number of colony and non-colony forming cells per sort per subset can be found in **Table S1**.

Extended Figure 2. Clonal bias and sensitivity correction. (A) To assess clone-to-clone biases in successfully seeded colonies, we reanalysed deep targeted resequencing data of bulk B and T cell lymphocytes from AX001[11]. The figure shows scatterplots of the fraction of lymphocyte colonies reporting a given somatic mutation (x-axis; log scale) with the variant allele fraction of that mutation in the bulk resequencing data (y-axis; log scale). Dashed lines are x=y equality and solid lines show the linear regression fit (B cells, R^2=0.47, p=$1x10^{-18}$; T cells, R^2=0.59, p=$2x10^{-31}$). (B) Estimates of sensitivity for mutation calling as a function of depth for each colony (points in left panels) from each donor (rows; the 5 donors with the highest numbers of colonies are shown). The second column of panels shows uncorrected estimates of mutation burden for HSPCs in each donor, while the third column shows mutation burden estimates after correction for sequencing depth by asymptotic regression. The fourth column shows the corrected mutation burdens for lymphocyte colonies.

Extended Figure 3. Indels and selection pressure. (A) Indel mutation burden per genome for the four main lymphocyte subsets (pink points), compared with HSPCs (green points). Each panel has all genomes plotted underneath in white with grey outline. The lines show the fit by linear mixed effects models for the respective populations. (B) Plots of the estimated dN/dS ratio for mutations genome-wide (excluding immunoglobulin genes) for all lymphocytes, and for the various individual lymphocyte populations. The second row shows the estimated dN/dS ratio for known cancer genes in all lymphocytes. The diamond shows the point estimates, and the lines the 95% confidence intervals. The point estimates / number of variants included in each analysis are as follows: lymphocytes, genome-wide = 1.12 / 7555; lymphocytes, cancer genes = 1.21 / 352; naive B = 1.25 / 671; memory B = 1.10 / 1132; naive T = 1.16 / 4162; memory T = 0.99 / 1414.

Extended Figure 4. Mutational signatures by age. (A) SBSblood signature identified using HSPC genomes and the program *sigfit*. Trinucleotide contexts on the x-axis represent 16 bars within each substitution class, divided into 4 sets of 4 bars, grouped by the nucleotide 5′ to the mutated base, and within each group by the 3′ nucleotide. (B) SNV mutation burden per genome, shown separately for each mutational signature. The lines show the fit by linear mixed effects models for the respective populations. Two outlier cells (PD40667vu and PD40667rx) are excluded from plotting. (C) The rate of mutation accumulation per year (slopes in B) for signatures with strong age effects. Error bars represent the 95% confidence intervals on the slope from the linear mixed effects models.

Extended Figure 5. Ultraviolet light mutational signature (SBS7a) in lymphocytes. (A) Raw mutational spectra shown for all mutation calls from four lymphocyte colonies, two with high contribution of SBS7a (left) and two with a more typical T-cell spectrum (right) from two different donors (rows). For each cell, the top panel shows the SNV spectrum, with trinucleotide contexts on the x-axis representing 16 bars within each substitution class, divided into 4 sets of 4 bars, grouped by the nucleotide 5′ to the mutated base, and within each group by the 3′ nucleotide. The bottom panel shows frequency of dinucleotide substitutions. (B) Telomere lengths for memory T cells with (yellow) and without (grey) high SBS7a signature. A memory T cell with high UV signature is defined as having greater than 9.5% (2 standard deviations above the mean) of its mutations attributable to SBS7a. (C) Proportion of mutations attributable to SBS7a across normal lymphocytes (paediatric samples excluded) and lymphoid malignancies. Boxes show the interquartile range and the centre horizontal lines show the median. Whiskers extend to the minimum of either the range or 1.5× the interquartile range. Number of genomes included per group: naive B: 68, memory B: 68, naive T: 332, memory T SBS7a low: 78, memory T SBS7a high: 9, Burkitt lymphoma: 17, CLL (chronic lymphocytic leukaemia) mutated: 38, CLL unmutated: 45, C. (cutaneous) T-cell lymphoma: 5, DLBC (Diffuse Large B-cell) lymphoma: 47, follicular lymphoma: 36, multiple myeloma: 30, myeloid-AML (acute myeloid leukaemia): 10.

Extended Figure 6. Distribution of mutational signatures across the genome. (A) Estimates of the mutation rate across non-Ig chromosomes and Ig regions for memory (left) and naive B (right) cells. Rates for the Ig regions are calculated separately for the productive (triangles) and non-recombined alleles (circles) and exons (green) versus introns (orange). (B) Estimated mutation rates across different variable segments of the Ig genes for exons (green) versus introns (orange). (C) Number of productive V(D)J rearrangements affecting each variable segment in the dataset.

(D) Proportion of mutations across chromosomes 2, 14 and 22 in each 1Mb window attributed to signatures SBS9, SBSblood and the canonical somatic hypermutation (SHM) signature (rows). Windows spanning the relevant immunoglobulin regions are coloured according to the key.

Extended Figure 7. Telomere lengths and SBS9 versus replication timing. (A) The top left panel includes the tonsil-derived genomes, which have an exceptionally high variance in telomere length. The remaining panels exclude these genomes, and show the estimated telomere lengths (y-axis) for each cell as a function of age (x-axis). Lines show the estimated fit by linear mixed effects models for each cell type, with the slope and 95% confidence intervals quoted in text. (B) Replication timing and number of SBS9 mutations per 10kb window. The line represents the GAM regression prediction. The x-axis is truncated at 5, excluding 0.3% of the data, and points have random noise (-0.5 to 0.5) to facilitate visualisation.

Extended Figure 8. Relationships of signatures to epigenetic marks across haematopoietic cell types. 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 coloured 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 (CS, class switched; GC, germinal centre; HSC, hematopoietic stem cell; Mem, memory).

Extended Figure 9. SV density and patterns in normal and malignant lymphocytes. (A-B) Mutation rates per 1Mb bin across the genome for SNVs (A) and structural variants (B) split by cell type, with chromosomes labelled in the top strip, and Ig/TCR regions marked. Circles (purple) denote bins with more mutations than 2 standard deviations above the mean. (C) Histogram showing the distribution of estimated number of reads per informative chromosome copy for the normal lymphocytes (blue) and lymphoid malignancies from PCAWG (purple). For cancer genomes, purity and ploidy were estimated from the copy number patterns; for lymphocyte colonies, the purity was 1 and ploidy was 2.

Extended Figure 10. RAG-mediated SVs in normal versus malignant lymphocytes. (A) Point estimates and 95% confidence intervals for the proportion of SVs with RSS (RAG) motifs within 50bp of a breakpoint. (B) Number of SVs with RSS (RAG) motifs within 50bp of a breakpoint. Boxes show the interquartile range and the centre horizontal lines show the median. Whiskers extend to the minimum of either the range or 1.5× the interquartile range. Paediatric samples excluded. Number of SVs per group: B = 145, T = 841, ALL = 523, Burkitt lymphoma = 305, CLL mutated =

252, CLL unmutated = 440, C. T-cell lymphoma = 204, DLBC lymphoma = 3754, follicular lymphoma = 1095.

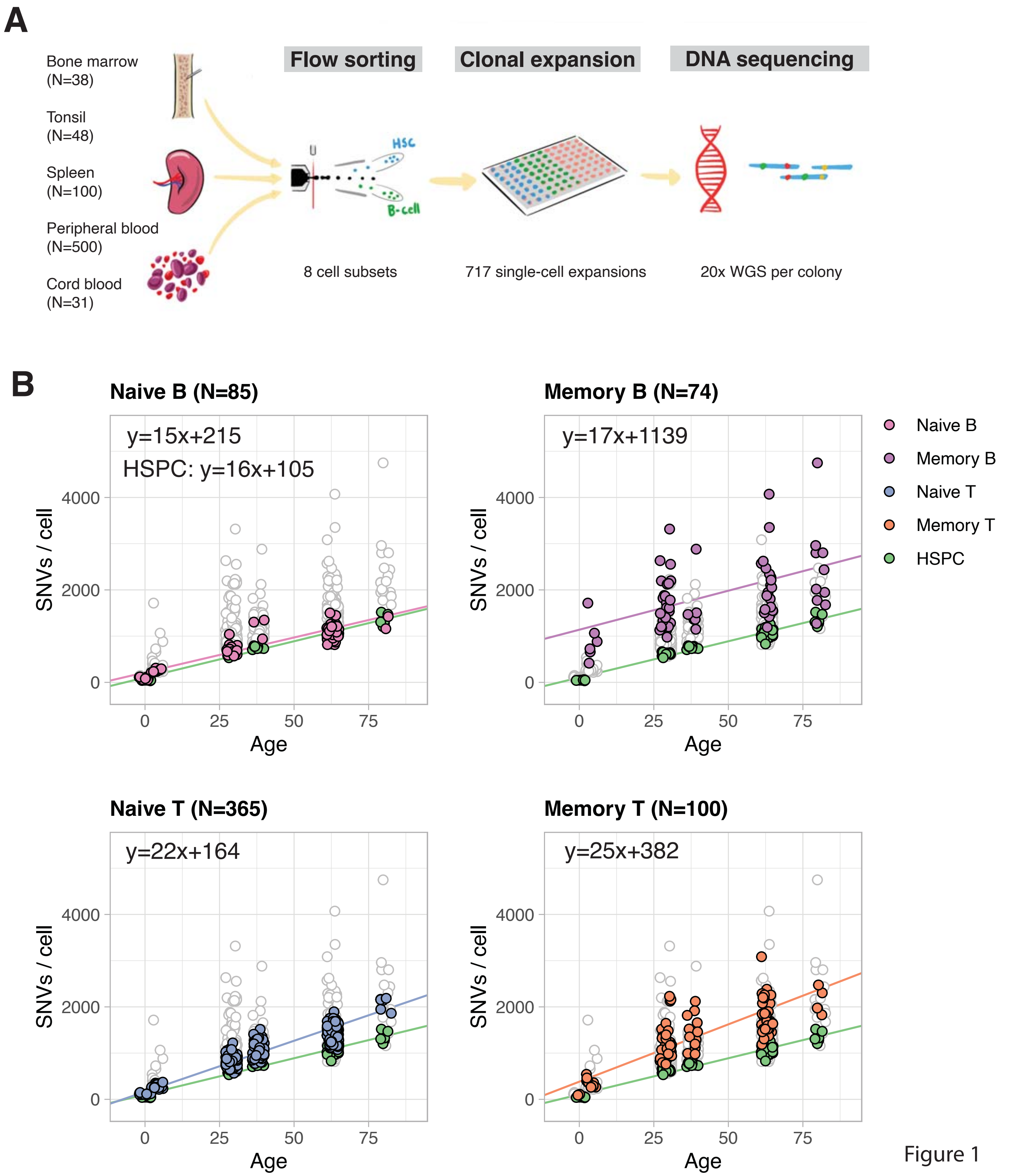
A
Bone marrow
(N=38)
Tonsil
(N=48)
Spleen
(N=100)
Peripheral blood
(N=500)
Cord blood
(N=31)
Flow sorting
Clonal expansion
DNA sequencing
HSC
B-cell
8 cell subsets
717 single-cell expansions
20x WGS per colony
B
Naive B (N=85)
y=15x+215
HSPC: y=16x+105
Memory B (N=74)
y=17x+1139
Naive T (N=365)
y=22x+164
Memory T (N=100)
y=25x+382
SNVs / cell
0
2000
4000
Age
0
25
50
75
● Naive B
● Memory B
● Naive T
● Memory T
● HSPC

Figure 1

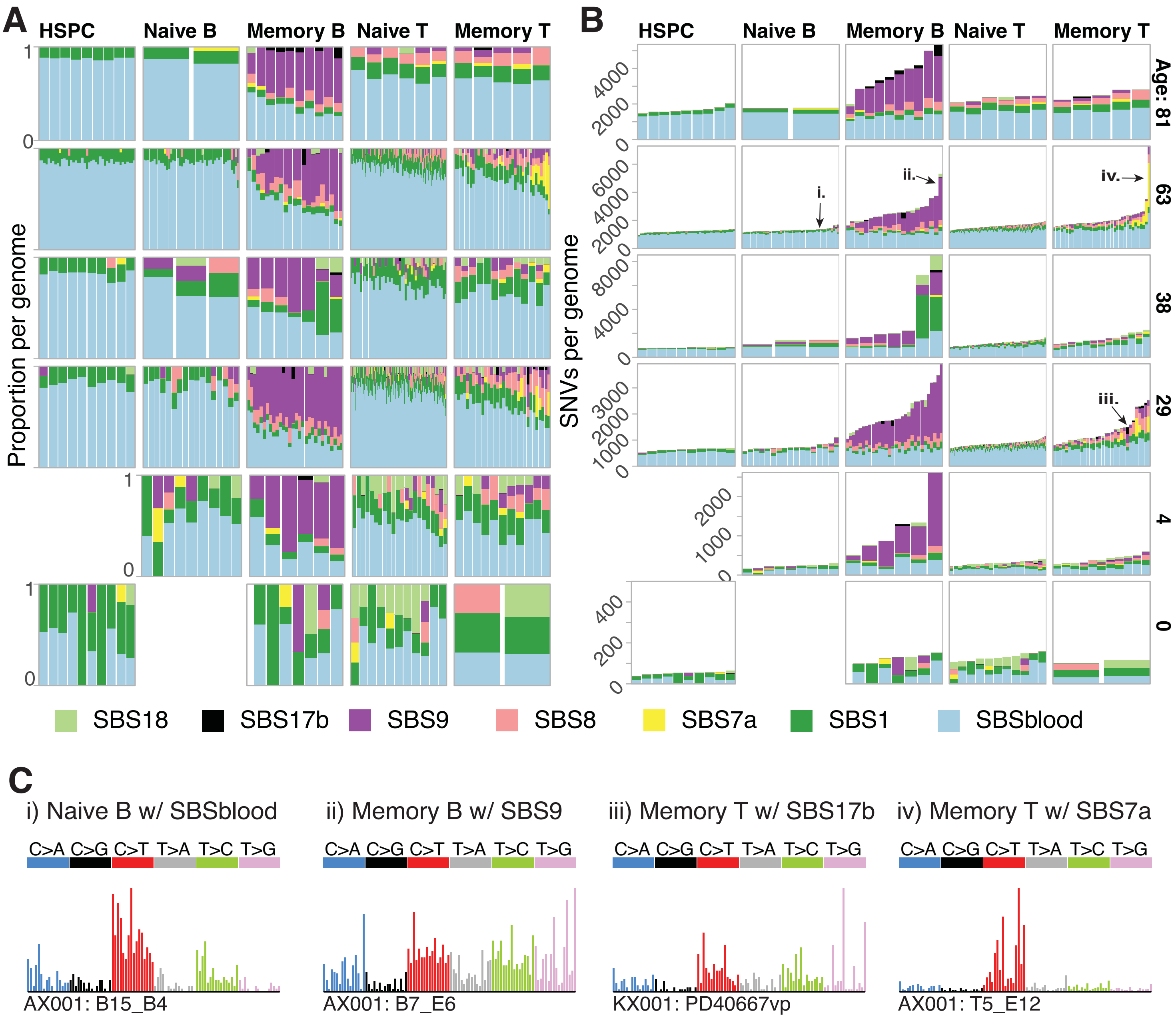
A
HSPC
Naive B
Memory B
Naive T
Memory T
Proportion per genome
1
0
B
SNVs per genome
Age: 81
63
38
29
4
0
2000
4000
6000
8000
3000
1000
200
400
i.
ii.
iii.
iv.
SBS18
SBS17b
SBS9
SBS8
SBS7a
SBS1
SBSblood
C
i) Naive B w/ SBSblood
ii) Memory B w/ SBS9
iii) Memory T w/ SBS17b
iv) Memory T w/ SBS7a
C>A C>G C>T T>A T>C T>G
AX001: B15_B4
AX001: B7_E6
KX001: PD40667vp
AX001: T5_E12

Figure 2

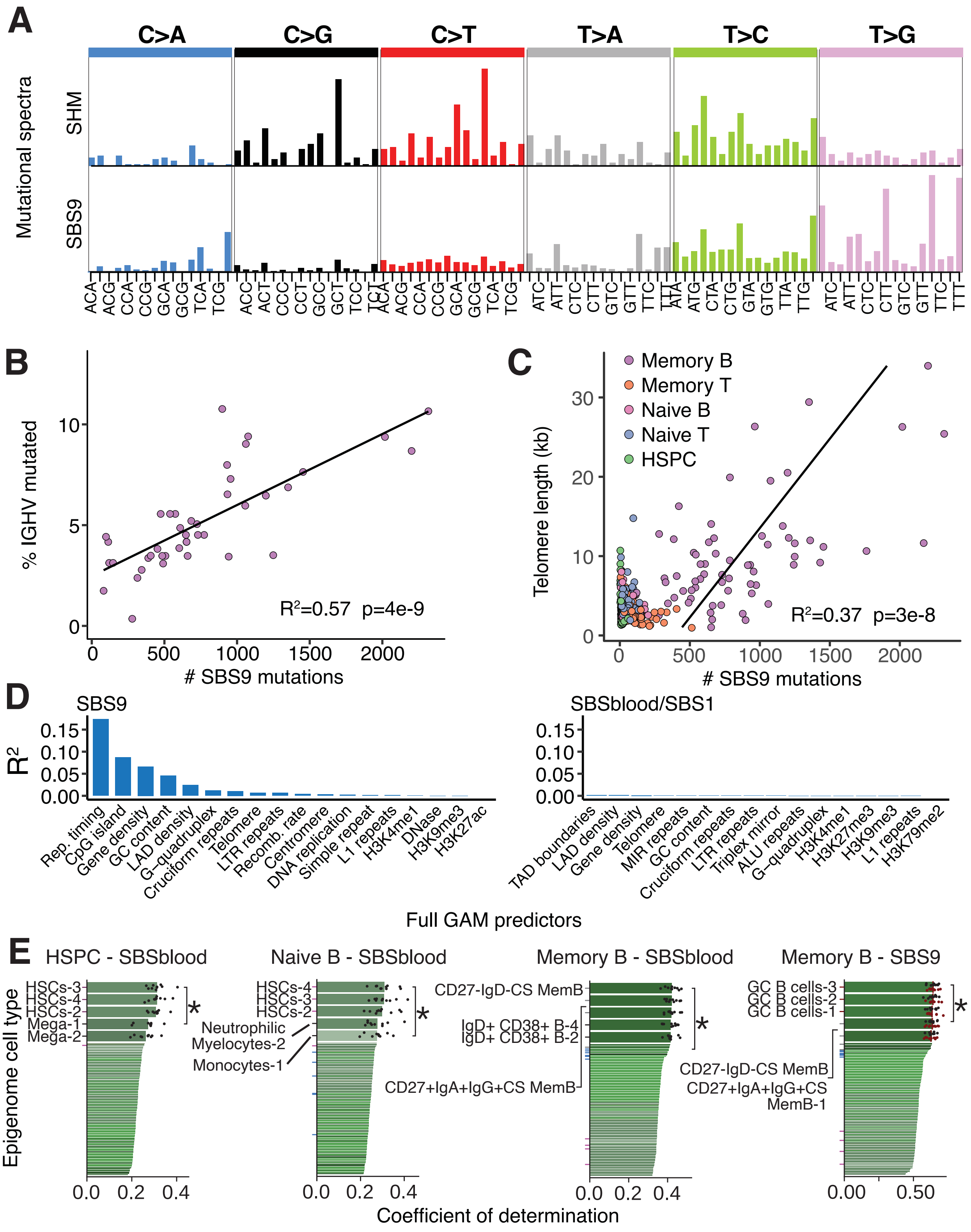
A
C>A
C>G
C>T
T>A
T>C
T>G
Mutational spectra
SHM
SBS9
ACA ACG CCA CCG GCA GCG TCA TCG
ACC ACT CCC CCT GCC GCT TCC TCT
ACA ACG CCA CCG GCA GCG TCA TCG
ATC ATT CTC CTT GTC GTT TTC TTT
ATA ATG CTA CTG GTA GTG TTA TTG
ATC ATT CTC CTT GTC GTT TTC TTT
B
% IGHV mutated
10
5
0
0 500 1000 1500 2000
# SBS9 mutations
R²=0.57 p=4e-9
C
Memory B
Memory T
Naive B
Naive T
HSPC
Telomere length (kb)
30
20
10
0
0 500 1000 1500 2000
# SBS9 mutations
R²=0.37 p=3e-8
D
SBS9
R²
0.15
0.10
0.05
0.00
Rep. timing
CpG island
Gene density
GC content
LAD density
G-quadruplex
Cruciform repeats
Telomere
LTR repeats
Recomb. rate
Centromere
DNA replication
Simple repeat
L1 repeats
H3K4me1
DNase
H3K9me3
H3K27ac
SBSblood/SBS1
0.15
0.10
0.05
0.00
TAD boundaries
LAD density
Gene density
Telomere
MIR repeats
GC content
Cruciform repeats
LTR repeats
Triplex mirror
ALU repeats
G-quadruplex
H3K4me1
H3K27me3
H3K9me3
L1 repeats
H3K79me2
Full GAM predictors
E
HSPC - SBSblood
Naive B - SBSblood
Memory B - SBSblood
Memory B - SBS9
Epigenome cell type
HSCs-3
HSCs-4
HSCs-2
Mega-1
Mega-2
*
HSCs-4
HSCs-3
HSCs-2
Neutrophilic
Myelocytes-2
Monocytes-1
*
CD27-IgD-CS MemB
IgD+ CD38+ B-4
IgD+ CD38+ B-2
CD27+IgA+IgG+CS MemB
*
GC B cells-3
GC B cells-2
GC B cells-1
*
CD27-IgD-CS MemB
CD27+IgA+IgG+CS
MemB-1
0.0 0.2 0.4
0.0 0.2 0.4
0.0 0.2 0.4
0.00 0.50
Coefficient of determination

Figure 3

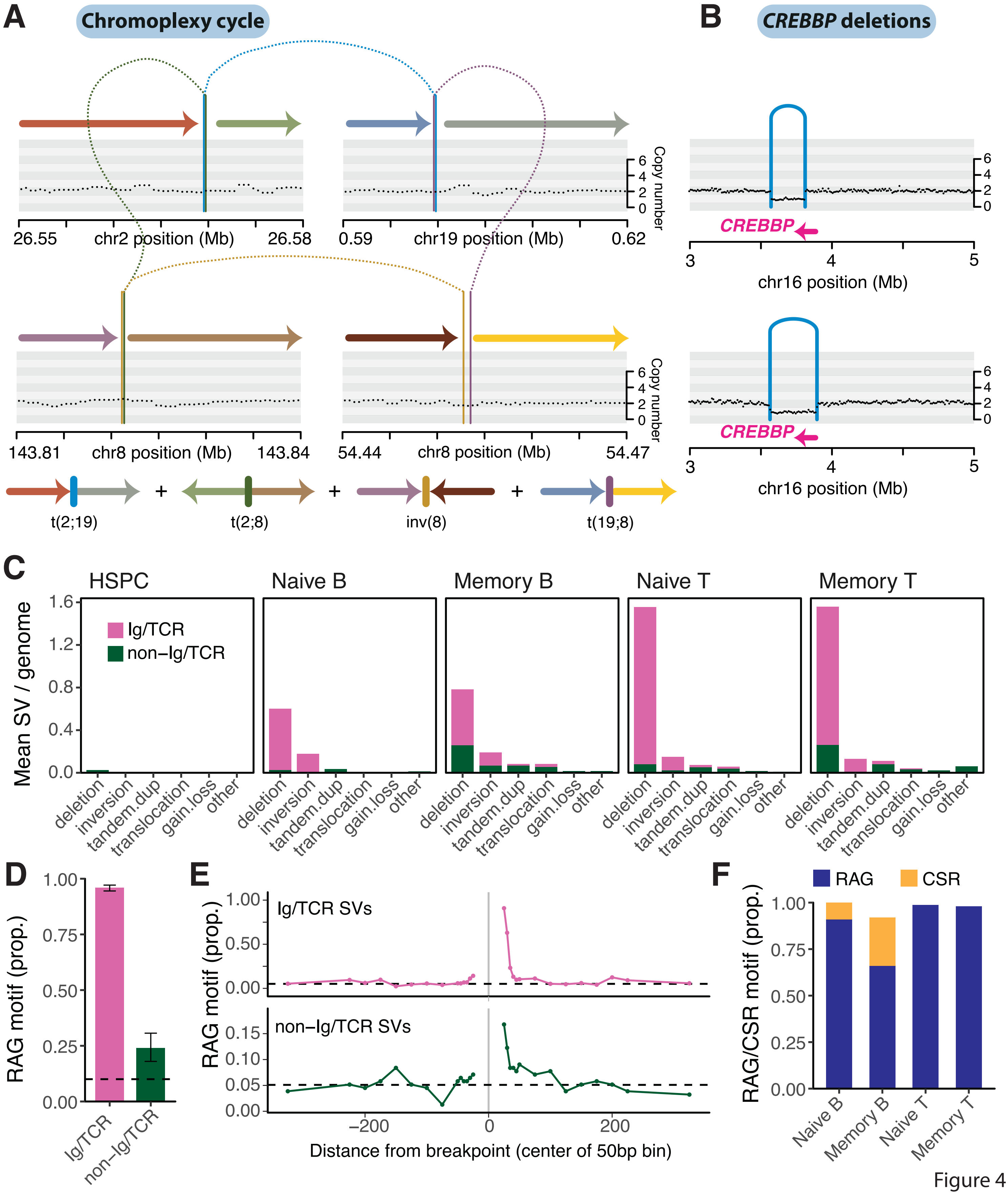
A
Chromoplexy cycle
Copy number
6
4
2
0
26.55
chr2 position (Mb)
26.58
0.59
chr19 position (Mb)
0.62
143.81
chr8 position (Mb)
143.84
54.44
chr8 position (Mb)
54.47
+
t(2;19)
t(2;8)
inv(8)
t(19;8)
B
CREBBP deletions
CREBBP
3
4
5
chr16 position (Mb)
C
HSPC
Naive B
Memory B
Naive T
Memory T
Mean SV / genome
1.6
1.2
0.8
0.4
0.0
Ig/TCR
non−Ig/TCR
deletion
inversion
tandem.dup
translocation
gain.loss
other
D
RAG motif (prop.)
1.00
0.75
0.50
0.25
0.00
Ig/TCR
non−Ig/TCR
E
Ig/TCR SVs
non−Ig/TCR SVs
RAG motif (prop.)
1.00
0.50
0.00
0.15
0.10
0.05
0.00
−200
0
200
Distance from breakpoint (center of 50bp bin)
F
RAG
CSR
RAG/CSR motif (prop.)
1.00
0.75
0.50
0.25
0.00
Naive B
Memory B
Naive T
Memory T

Figure 4

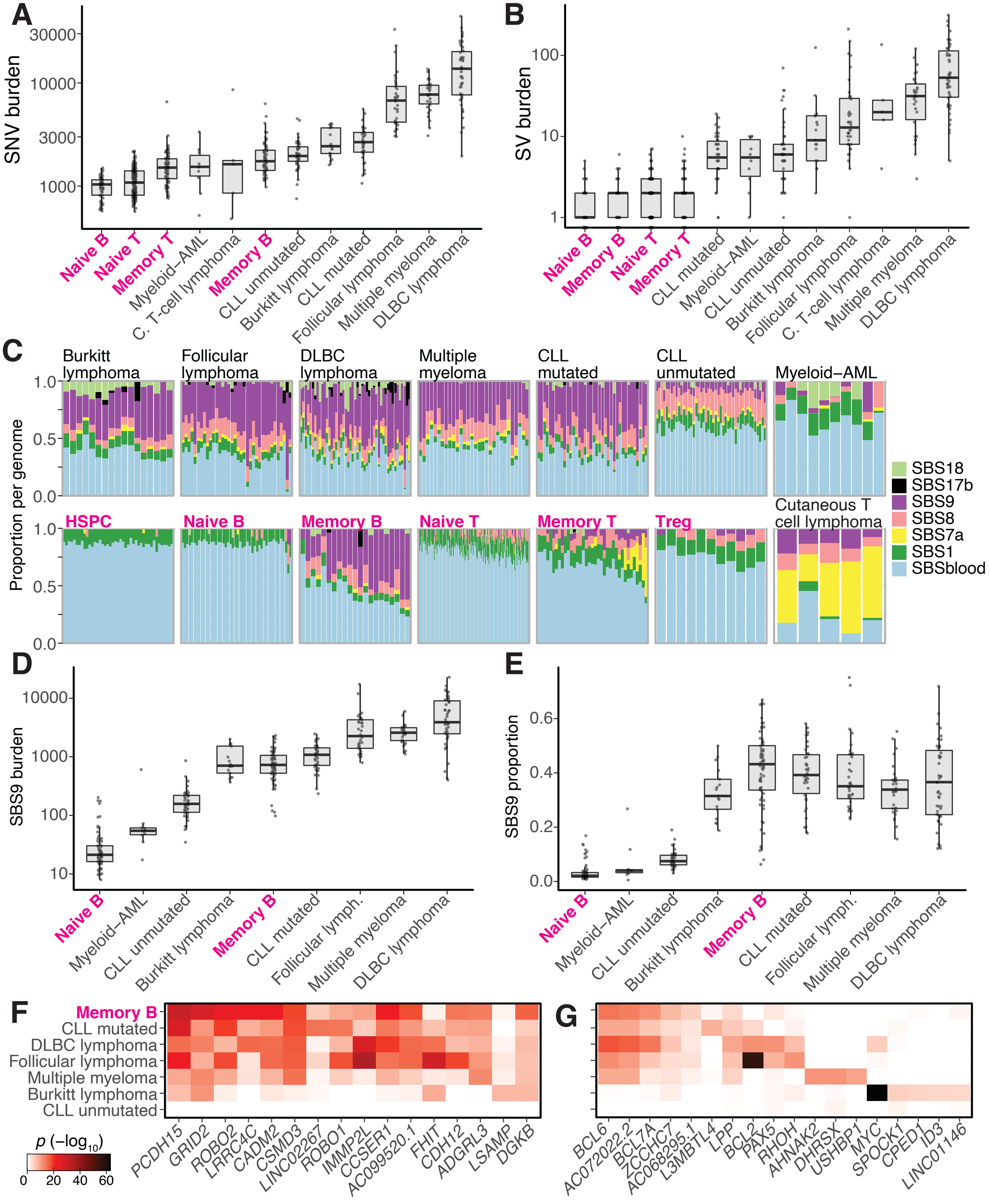
A
SNV burden
30000
10000
3000
1000
Naive B
Naive T
Memory T
Myeloid–AML
C. T-cell lymphoma
Memory B
CLL unmutated
Burkitt lymphoma
CLL mutated
Follicular lymphoma
Multiple myeloma
DLBC lymphoma
B
SV burden
100
10
1
Naive B
Memory B
Naive T
Memory T
CLL mutated
Myeloid–AML
CLL unmutated
Burkitt lymphoma
Follicular lymphoma
C. T-cell lymphoma
Multiple myeloma
DLBC lymphoma
C
Proportion per genome
1.0
0.5
0.0
Burkitt lymphoma
Follicular lymphoma
DLBC lymphoma
Multiple myeloma
CLL mutated
CLL unmutated
Myeloid–AML
HSPC
Naive B
Memory B
Naive T
Memory T
Treg
Cutaneous T cell lymphoma
SBS18
SBS17b
SBS9
SBS8
SBS7a
SBS1
SBSblood
D
SBS9 burden
10000
1000
100
10
Naive B
Myeloid–AML
CLL unmutated
Burkitt lymphoma
Memory B
CLL mutated
Follicular lymph.
Multiple myeloma
DLBC lymphoma
E
SBS9 proportion
0.6
0.4
0.2
0.0
Naive B
Myeloid–AML
CLL unmutated
Burkitt lymphoma
Memory B
CLL mutated
Follicular lymph.
Multiple myeloma
DLBC lymphoma
F
Memory B
CLL mutated
DLBC lymphoma
Follicular lymphoma
Multiple myeloma
Burkitt lymphoma
CLL unmutated
PCDH15
GRID2
ROBO2
LRRC4C
CADM2
CSMD3
LINC02267
ROBO1
IMMP2L
CCSER1
AC099520.1
FHIT
CDH12
ADGRL3
LSAMP
DGKB
p (−log10)
0
20
40
60
G
BCL6
AC072022.2
BCL7A
ZCCHC7
AC068295.1
L3MBTL4
LPP
BCL2
PAX5
RHOH
AHNAK2
DHRSX
USHBP1
MYC
SPOCK1
CPED1
ID3
LINC01146

Figure 5