PNAS

Supporting Information for

Different classes of genomic inserts contribute to human antibody diversity

Mikhail Lebedin[a,b,c,1], Mathilde Foglierini[d,e,1], Svetlana Khorkova[a,b,f], Clara Vázquez García[a,c], Christoph Ratswohl[a,g], Alexey N. Davydov[h], Maria A. Turchaninova[b,f], Claudia Daubenberger[i], Dmitriy M. Chudakov[b,f,h], Antonio Lanzavecchia[d], and Kathrin de la Rosa[a,c,j,2]

[a]Max Delbrück Center for Molecular Medicine in the Helmholtz Association, 13125 Berlin, Germany;

[b]Department of Genomics of Adaptive Immunity, Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, 117997 Moscow, Russian Federation;

[c]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin, Charitéplatz 1, 10117 Berlin, Germany;

[d]Institute for Research in Biomedicine, Università della Svizzera Italiana, Via Francesco Chiesa 5, 6500 Bellinzona, Switzerland;

[e]Service of Immunology and Allergy, Department of Medicine, Lausanne University Hospital and University of Lausanne, 1011 Lausanne, Switzerland;

[f]Department of Molecular Technologies, Pirogov Russian National Research Medical University, 117997 Moscow, Russian Federation;

[g]Department of Biology, Chemistry and Pharmacy, Free University of Berlin, 14195 Berlin, Germany;

[h]Central European Institute of Technology, Masaryk University, 601 77 Brno, Czech Republic;

[i]Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, 4123 Allschwil, Switzerland; and

[j]Berlin Institute of Health at Charité, 10117 Berlin, Germany

[1]These authors contributed equally to this work

[2]To whom correspondence may be addressed: Kathrin de la Rosa, Email: kathrin.delarosa@mdc-berlin.de

This PDF file includes:

- Supporting methods
- SI References
- Figures and legends S1 to S12

Supporting Methods

Generation of in silico controls

For generation of in silico controls, it needs to be considered that inserts derive from genic and intergenic regions with different frequencies between distinct insert classes. For instance, 87.8% J-CH1 insertions compared to 71.1% of V-DJ insertions originate from genic regions. To account for this imbalance, for each real insert we create a matched in silico insertion, which was randomly originating from either genic or intergenic regions across all chromosomes. Gap regions without mapping, such as microsatellites, were excluded from the randomization process. After random assignment of the start coordinate, the end coordinate was calculated based on the length of the original insertion. Thus, in silico control data preserves the original length distribution and genic/intergenic bias. GRCh38 genome assembly was used throughout the study. The randomization was repeated 100 times and the resulting data were pooled.

Expression level analysis

For expression level analysis, TPM values of B cell RNA-seq data were used (1). Naïve and memory B cell TPM values were merged, as TPM values were compared to all genic inserts from aggregated data of all B cells that were analyzed in this study. TPM values were filtered according to the threshold described in Monaco et al., briefly, only genes with raw counts ≥4 in at least three samples were considered. Statistical significance was calculated with Wilcoxon rank-sum test in R.

Calculation of the distance to and overlap with documented genomic regions

DRIP-seq (2), CFSs (3), ERFSs (4), LINEs, and SINE/Alu (5) regions were retrieved from the corresponding publications. AID off-target sites were extracted from various publications (6-13). RAG off-target sites were extracted from Mijušković et. al. (14). In order to analyze the overlap with insert origin sites, mouse genomic coordinates were lifted over to GRCh38 human genome assembly via UCSC Genome Browser (15). To calculate the distance of inserts to genomic regions covering the sites listed above and chromosome ends (15), we used the central coordinates of the corresponding genomic region. Overlaps were determined by identification of shared region coordinates. For DRIP-seq data analysis, mitochondrial insertions were excluded.

GC content was computed with R using in-house scripts (https://github.com/lebmih/LAIR) and compared to GC content of the human genome (16). Statistical significance was calculated with Wilcoxon rank-sum test in R.

Calculation of the estimated genomic length

Gene lengths of insert donor genes were extracted from UCSC Table Browser. The estimated genomic length of inserts comprising multiple exons was computed using the 5´ and 3´ end-genomic coordinates.

Cryptic RSS site analysis

Cryptic RSS site analysis was performed on the 100 nt flanking regions of all VDJ-inserts and in silico controls by extracting the sequences using an in-house R script and computing the recombination information content (RIC) by RSS database (https://www.itb.cnr.it/rss/index.html; 17, 18)

Hotspot analysis

For hotspot analysis, we mapped the inserts detected in IGH transcripts to the human genome, resulting in a dataset of insert origin coordinates. Next, we created a control dataset by randomizing coordinates. To identify hotspots, we determined the distance of each insert to the closest neighboring insert in the experimental dataset and the control dataset. As expected, the randomized control showed a unimodal distribution of the inter-loci distance with a peak at 63 Mb, with < 1% loci overlap. By contrast, the experimental data demonstrated an additional peak at ~800 kb, with > 2% loci overlap. This analysis suggests that certain regions, < 1 Mb in size, are prone to donate inserts at a higher frequency compared to the rest of the genome (Fig. S5). Based on this distribution analysis we defined a threshold of 1 Mb to identify clusters with an enrichment of inserts. We ranked the top ten clusters (hotspots) based on the number of donors, from which we could identify hotspot-derived insert transcripts.

SI References

1. G. Monaco, *et al.*, RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Reports* 26, 1627-1640.e7 (2019).
2. L. A. Sanz, *et al.*, Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. *Molecular cell* 63(1), 167-178 (2016).
3. R. Kumar, *et al.*, HumCFS: a database of fragile sites in human chromosomes. *BMC genomics*, 19(Suppl 9), 985 (2019).
4. J. H. Barlow, *et al.*, Identification of Early Replicating Fragile Sites that Contribute to Genome Instability. *Cell* 152, 620–632 (2013).
5. D. Karolchik, *et al.*, The UCSC Table Browser data retrieval tool. *Nucleic acids research*, 32(Database issue), D493–D496 (2004).
6. F.-L. Meng, et al., Convergent Transcription at Intragenic Super-Enhancers Targets AID-Initiated Genomic Instability. Cell 159, 1538–1548 (2014).
7. I. A. Klein, et al., Translocation-Capture Sequencing Reveals the Extent and Nature of Chromosomal Rearrangements in B Lymphocytes. Cell 147, 95–106 (2011).
8. R. Chiarle, et al., Genome-wide Translocation Sequencing Reveals Mechanisms of Chromosome Breaks and Rearrangements in B Cells. Cell 147, 107–119 (2011).
9. J. Qian, et al., B Cell Super-Enhancers and Regulatory Clusters Recruit AID Tumorigenic Activity. Cell 159, 1524–1537 (2014).
10. O. Staszewski, et al., Activation-Induced Cytidine Deaminase Induces Reproducible DNA Breaks at Many Non-Ig Loci in Activated B Cells. Mol Cell 41, 232–242 (2011).
11. L. Khair, R. E. Baker, E. K. Linehan, C. E. Schrader, J. Stavnezer, Nbs1 ChIP-Seq Identifies Off-Target DNA Double-Strand Breaks Induced by AID in Activated Splenic B Cells. Plos Genet 11, e1005438 (2015).
12. Á. F. Álvarez-Prado, et al., A broad atlas of somatic hypermutation allows prediction of activation-induced deaminase targets. J Exp Medicine 215, 761–771 (2018).
13. A. Yamane, et al., Deep-sequencing identification of the genomic targets of the cytidine deaminase AID and its cofactor RPA in B lymphocytes. Nat Immunol 12, 62–69 (2011).
14. M. Mijušković, et al., Off-Target V(D)J Recombination Drives Lymphomagenesis and Is Escalated by Loss of the Rag2 C Terminus. Cell Reports 12, 1842–1852 (2015).
15. W. J. Kent, *et al.*, The human genome browser at UCSC. *Genome research*, 12(6), 996–1006 (2002).
16. A. Piovesan, *et al.*, On the length, weight and GC content of the human genome. *BMC research notes*, 12(1), 106 (2019).
17. L. G. Cowell, *et al.*, Identification and utilization of arbitrary correlations in models of recombination signal sequences. *Genome biology*, 3(12), RESEARCH0072 (2002).
18. I. Merelli, *et al.*, RSSsite: a reference database and prediction tool for the identification of cryptic Recombination Signal Sequences in human and murine genomes, *Nucleic Acids Research*, 38-2 262-267 (2010).

Supplementary Figures

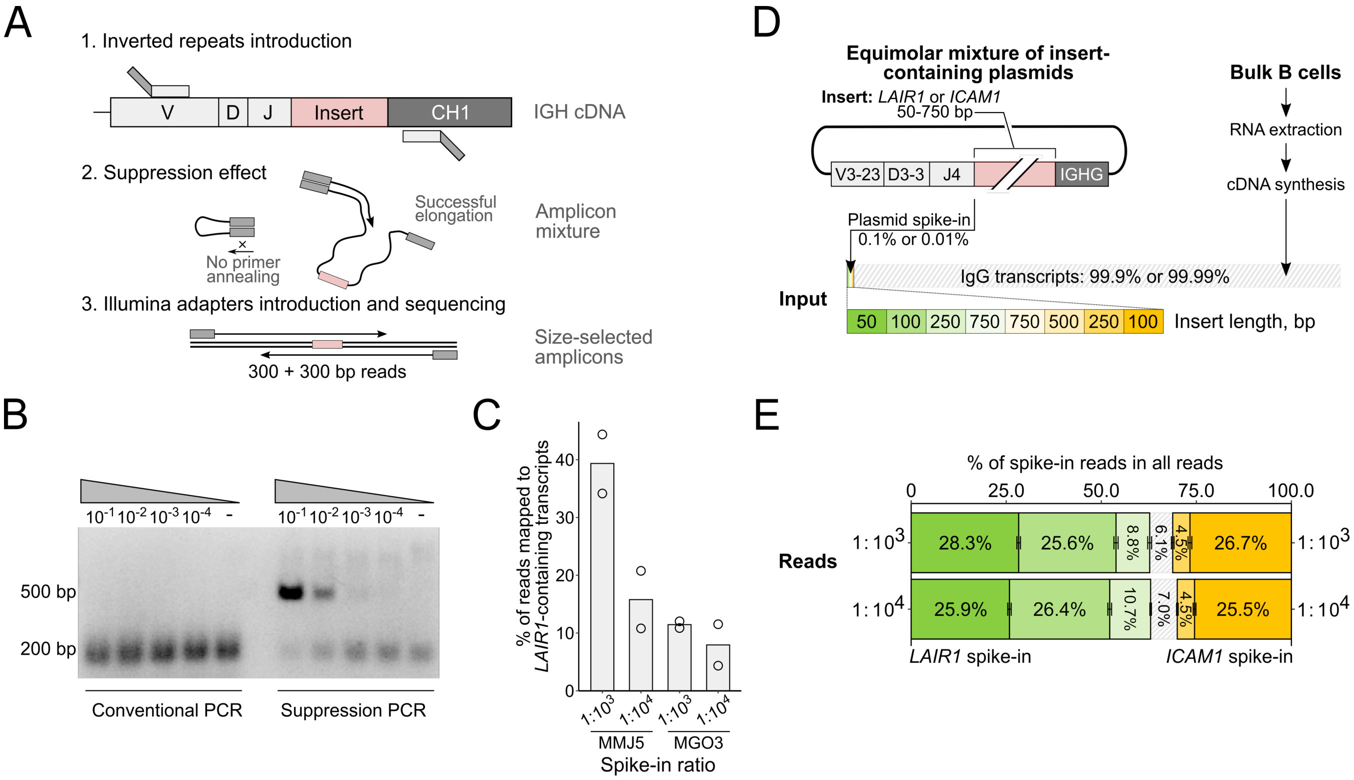
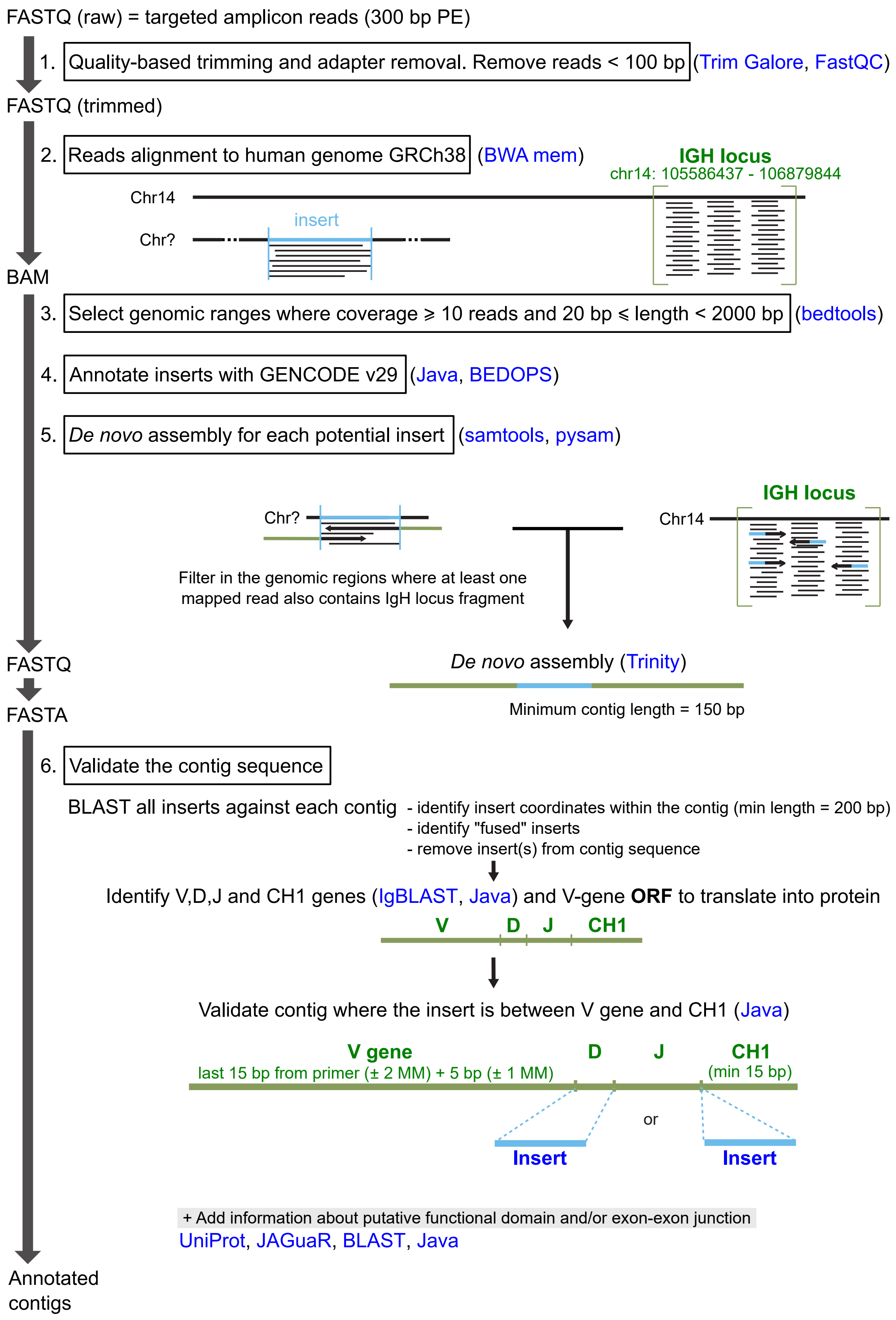


Fig. S1. Suppression PCR enables in-depth analysis of insert-containing antibody transcripts. (*A*) Scheme of the suppression PCR approach. Inverted repeats (grey rectangles) are introduced to mediate intramolecular hairpins in amplicons of the consecutive PCR. Final products are sequenced by Illumina MiSeq. (*B*) Natural LAIR1-transcripts (long) isolated from a monoclonal cell line were added to polyclonal IgM transcripts (short) at serial dilutions, amplified by conventional versus suppression PCR, and analyzed by agarose gel electrophoresis. (*C*) Percent of Illumina reads mapping to LAIR1-transcripts amplified from different dilutions. Monoclonal cell lines expressing LAIR1-antibodies (MMJ5, IgM; MGO3, IgG) were spiked into memory B cells at ratios 1:10^3 and 1:10^4. n = 2 technical replicates. (*D*) Scheme and (*E*) readout of spike-in experiments performed with artificial insert-containing plasmid products of indicated size and genetic origin mixed at 1:10^3 and 1:10^4 with cDNA of polyclonal B cells. Median ± SD for 3 independent biological replicates.



FASTQ (raw) = targeted amplicon reads (300 bp PE)

1. Quality-based trimming and adapter removal. Remove reads < 100 bp (Trim Galore, FastQC)

FASTQ (trimmed)

2. Reads alignment to human genome GRCh38 (BWA mem)

BAM

3. Select genomic ranges where coverage ⩾ 10 reads and 20 bp ⩽ length < 2000 bp (bedtools)

4. Annotate inserts with GENCODE v29 (Java, BEDOPS)

5. *De novo* assembly for each potential insert (samtools, pysam)

Filter in the genomic regions where at least one mapped read also contains IgH locus fragment

FASTQ

FASTA

De novo assembly (Trinity)

Minimum contig length = 150 bp

6. Validate the contig sequence

BLAST all inserts against each contig
- identify insert coordinates within the contig (min length = 200 bp)
- identify "fused" inserts
- remove insert(s) from contig sequence

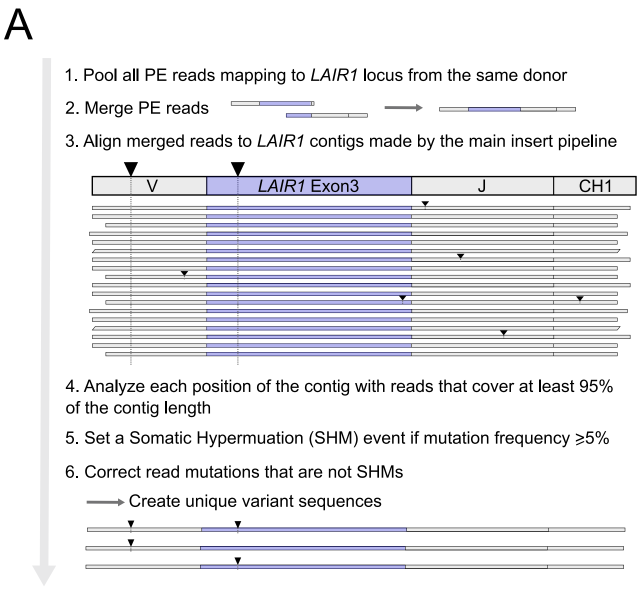
Identify V,D,J and CH1 genes (IgBLAST, Java) and V-gene **ORF** to translate into protein

Validate contig where the insert is between V gene and CH1 (Java)

+ Add information about putative functional domain and/or exon-exon junction
UniProt, JAGuaR, BLAST, Java

Annotated contigs

Fig. S2. Insert-containing sequence extraction and annotation. See detailed description in Methods. Briefly, demultiplexed reads in FASTQ format derived from suppression PCR sequencing were trimmed with Trim Galore and filtered based on read quality by FastQC. Short reads below 100 nt were removed. Filtered reads were aligned to the human genome and selected if they map to the 20-2000 bp loci with >10 reads coverage. Reads were annotated with GENCODE and *de novo* assembled with Trinity. Contig sequences were annotated to determine the insert position, detect the "fused" insert and extract contig sequence devoid of the insert in order to perform IgBLAST. MM - mismatch, ORF - open reading frame.


A
1. Pool all PE reads mapping to LAIR1 locus from the same donor
2. Merge PE reads
3. Align merged reads to LAIR1 contigs made by the main insert pipeline
V
LAIR1 Exon3
J
CH1
4. Analyze each position of the contig with reads that cover at least 95% of the contig length
5. Set a Somatic Hypermuation (SHM) event if mutation frequency ≥5%
6. Correct read mutations that are not SHMs
Create unique variant sequences

B

		Reads covering 95% of the *LAIR1* contig	Contig nucleotide variants	Contig amino acid variants	*LAIR1* nucleotide variants	*LAIR1* amino acid variants
Polyclonals	070	300'390	175	103	16	16
	215	2'164	73	54	56	43
Monoclonals	MMJ5	355'852	2	2	2	2
	MGO3	159'378	1	1	1	1

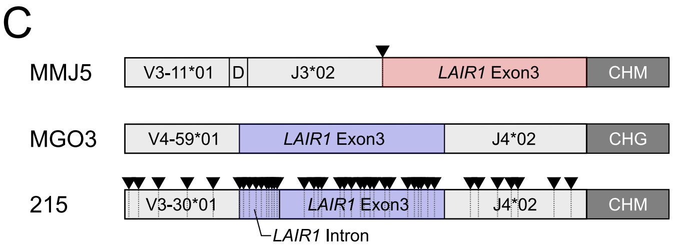


Fig. S3. Nucleotide and amino acid variant analysis of LAIR1-containing transcripts. (*A*) Outline of data processing. Paired-end reads mapping to *LAIR1* locus were extracted from samples and merged. Resulting reads were aligned to the consensus contig created by the main pipeline. Somatic hypermutations were assigned if a mismatch was present in more than 5% of reads. (*B*) Variants of *LAIR1* transcripts profiled for two donors of which one displayed LAIR1-containing antibodies in the serum, and another donor identified by suppression PCR (215). Two monoclonal cell lines (MMJ5, MGO3) expressing LAIR1-containing antibodies were used as controls. (*C*) Schematic representation of transcripts derived from monoclonal cell lines and donor 215. Black arrows indicate positions of nucleotide exchanges. The arrow in MMJ5 indicates a splice variant.

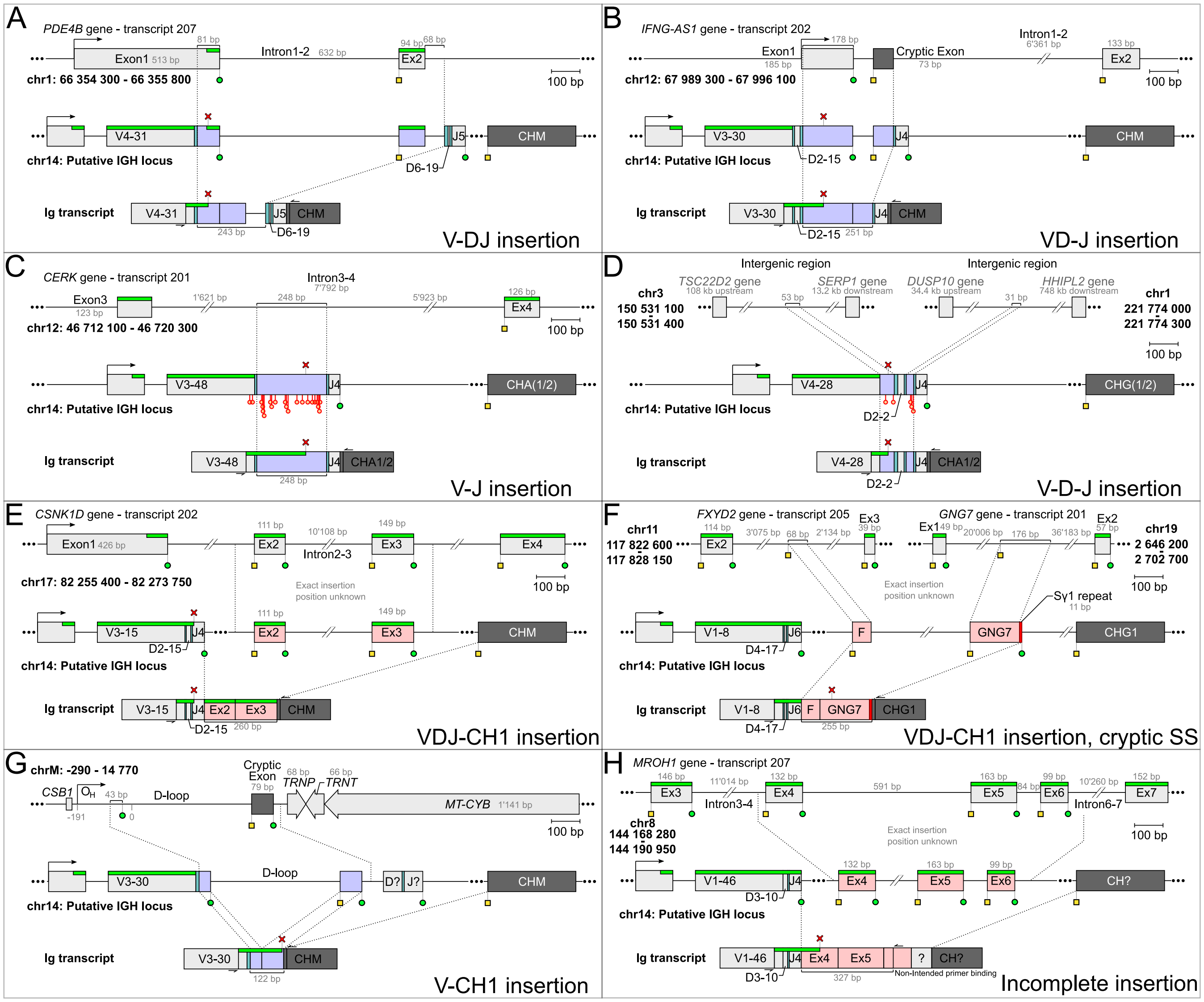


Fig. S4. Insert-containing transcripts with reconstruction of the putative genomic insertion and splicing. (*A*) Insert between V and D segment of 243 bp covering partial exon1, full exon2, and its downstream intron of the *PDE4B* gene. (*B*) Insert between D and J segment of 251 bp of *IFNG-AS1* long-noncoding RNA gene. The inserted fragment contains cryptic splice sites. (*C*) Insert between V and J segments of a 248 bp intron sequence originating from the *CERK* gene. (*D*) Two distinct intergenic sequences from different chromosomes incorporated up- and downstream of a D segment. (*E*) Insert between J and CH1 region of IGHM (CHM) comprising full exon 2 and 3 of *CSNK1D.* (*F*) Two distinct intron fragments inserted between J and CH1 of IGHG1 (CHG1). Cryptic splice sites were identified in the gamma 1 switch region (red rectangle) and both inserted fragments. (*G*) Insert between V and CH1 of 574 bp originating from mtDNA D-loop containing cryptic RSS sites. (*H*) Example of an incomplete contig with a putative insert deriving from *MROH1* gene covering full exons 3, 4, and part of exon 6. A primer binding site in exon 6 was identified. For all panels: green bar: open reading frame; red cross: premature termination codon. Green circles and yellow squares: splice donor and acceptor sites. Red circles: somatic hypermutations. Transcript numbers correspond to the ENSEMBL nomenclature. Regions length is to scale.

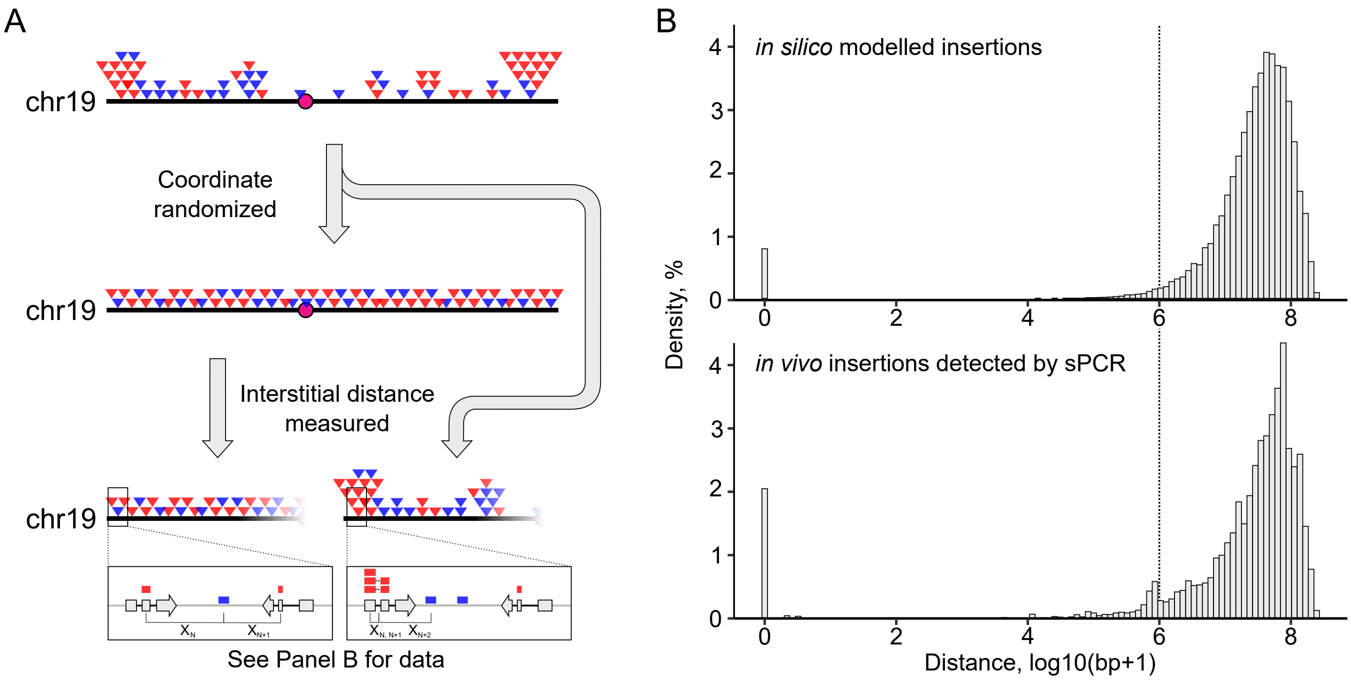


Fig. S5. Insert hotspot (hs) analysis. (A) Schematic outline of the analysis with a chromosome 19 hotspot as an example. Inserts occurring in vivo (upper panel) and after in silico coordinate randomization (middle panel) were analyzed for their distance to the closest insert (bottom panel). (B) Inter-insert distances are represented for in silico and in vivo data. A threshold (dashed line) for hotspots was set at 1 Mb according to the peak that was unique to insertions detected in vivo.

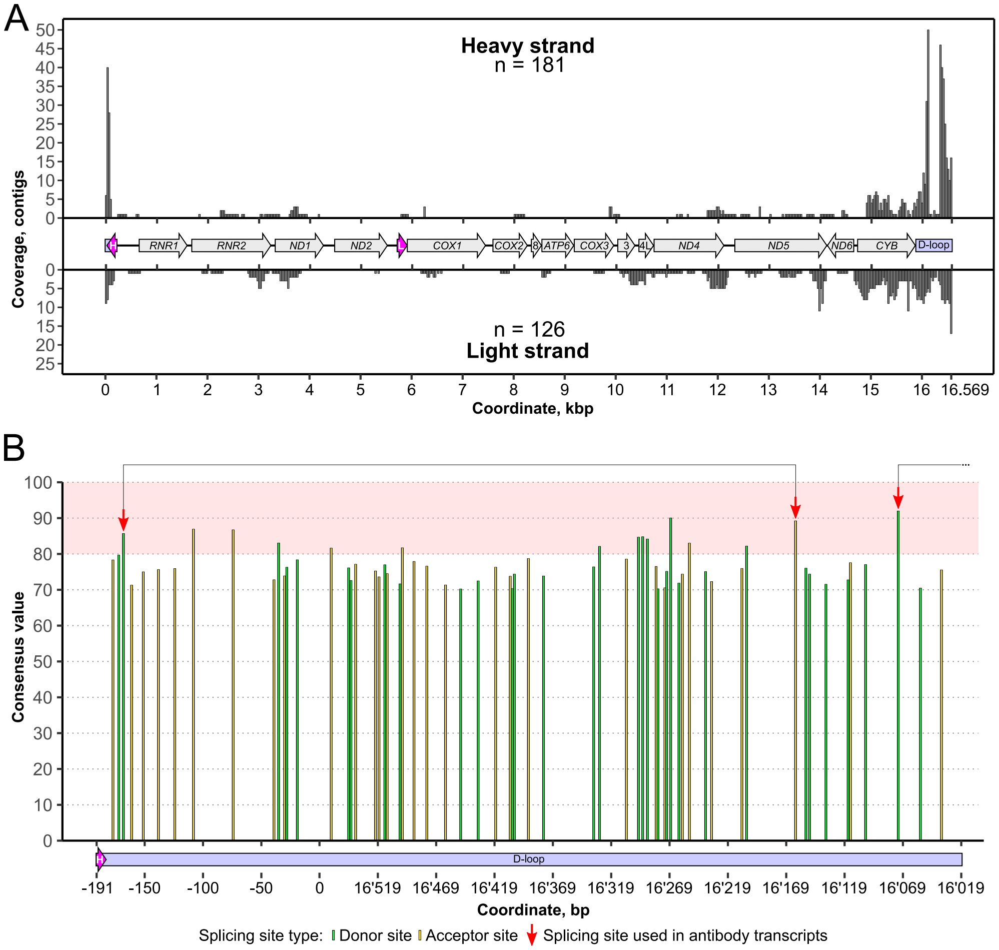


Fig. S6. mtDNA inserts originate from the D-loop region and may undergo splicing. (*A*) Mapping of inserts to the heavy (n = 181, upper panel) and light strand (n = 126, bottom panel) of the mitochondrial chromosome. Heavy and light strand origins are depicted as purple arrows. D-loop is shown in blue, protein-coding sequences, and ribosomal RNAs are shown in grey. *ATP8*, *ND3,* and *ND4L* genes are denoted as "8", "3" and "4L". (*B*) The D-loop region was analyzed by the Human Splice Finder algorithm. Acceptor sites are shown as yellow bars, donor sites as green bars. Only splicing sites of consensus value above 80 (red area) are shown as considered potent.

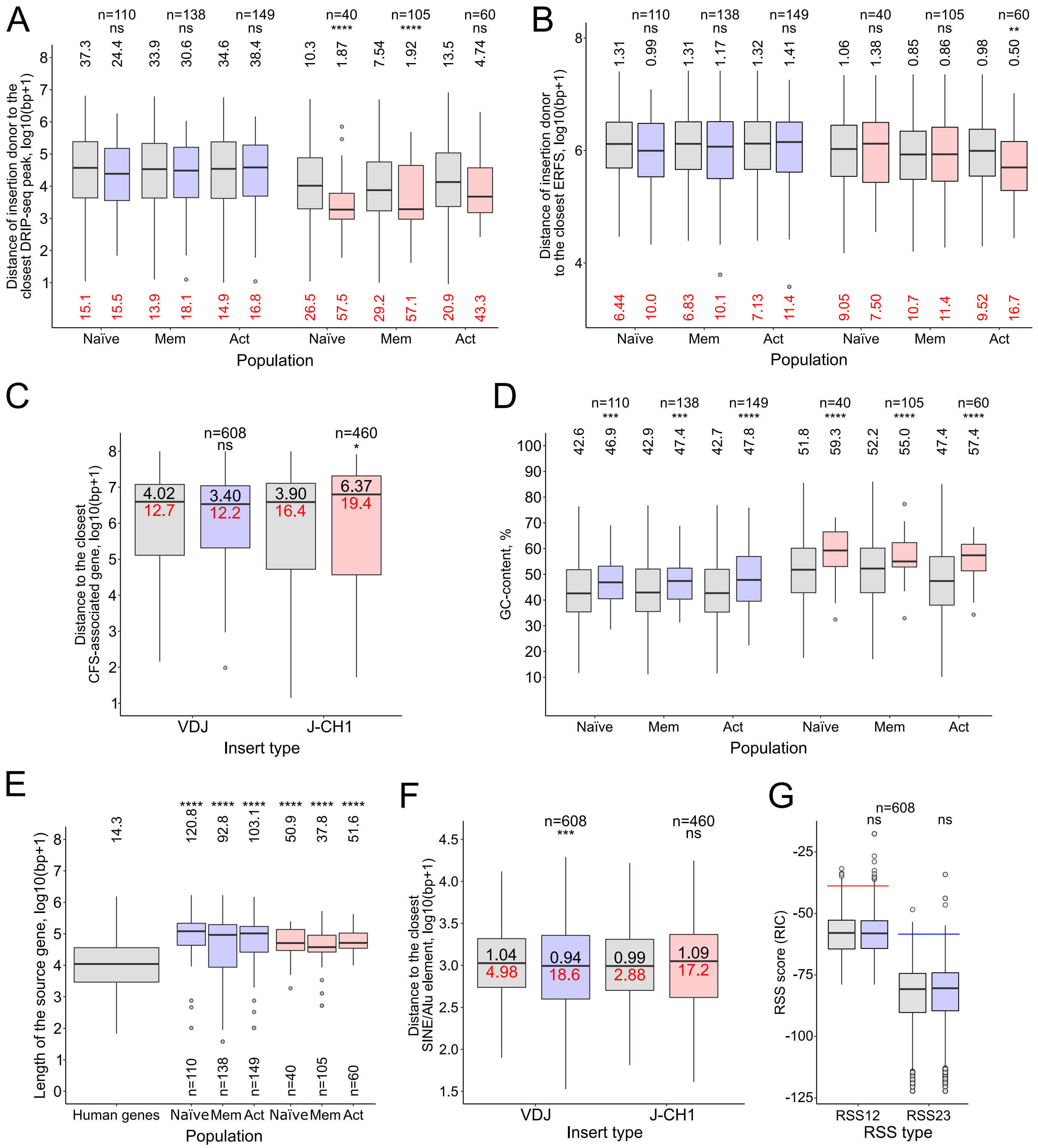


Fig. S7. Genomic features of insert donor sites. Insert donor sites, which were detected in naïve, memory (Mem) and in vitro activated (Act) B cells, were analyzed for their distance to genomic sites containing (*A*) R-loops (in kb) determined by DRIP-seq and (*B*) ERFSs (in Mb). (*C*) Inserts detected in all samples studied were analyzed for proximity to CFS-associated genes (in Mb). (*D*) GC content in %, and (*E*) length of donor genes (in kb) of inserts detected in distinct B cell populations. (*F*) All inserts identified in this study were analyzed for their proximity (in kb)

to neighboring SINE/Alu elements. (*G*) Presence of cryptic RSS sites in the flanking regions of VDJ-inserts was analyzed by determining recombination information content (RIC) for 100 nt insert flanking regions. Red and blue lines denote the threshold used in Cowell et al. (17) to identify putative cryptic RSS. For all panels: Data for VDJ-inserts are represented in blue, J-CH1 inserts in red, in silico controls in grey. n = number of insertions. Black numbers: Median values. Red numbers: overlap in %. The black lines in the boxplot represent the median, the top and the bottom of the boxplots represent 25th and 75th percentile, and the whiskers spread from the borders of the box for 1.5 x interquartile range. ns – $P \geq 0.05$, * – $P<0.05$, ** – $P<0.01$, *** – $P<0.001$, and **** – $P<0.0001$.

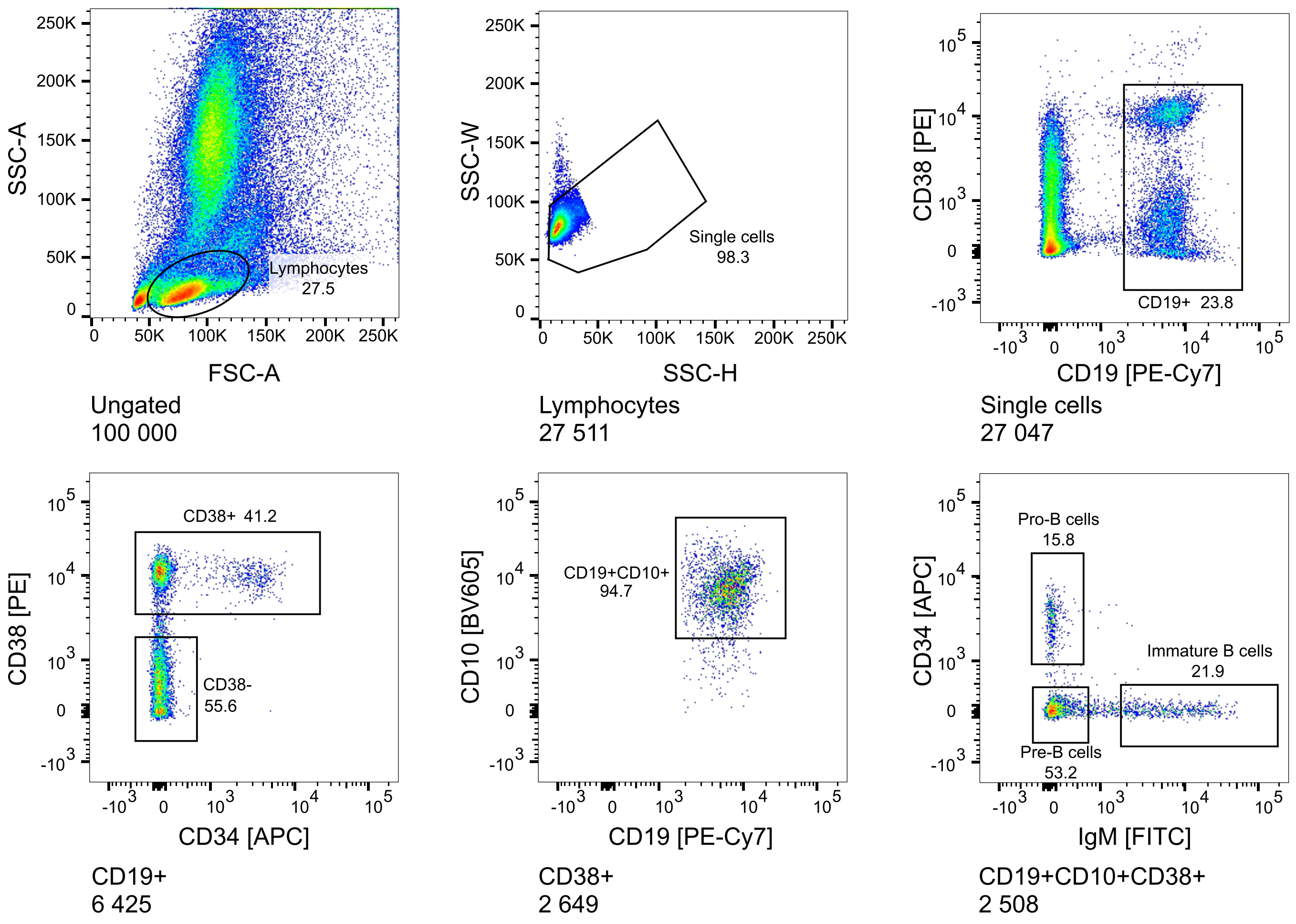


Fig. S8. Sorting strategy to isolate pro- and pre-B cells from human bone marrow. Single cells were gated through SSC and FSC channels. Pro-B cells were defined as CD19^{+}CD38^{+}CD10^{+}CD34^{+}μ-chain^{-}; pre-B cells as CD19^{+}CD38^{+}CD10^{+}CD34^{-}μ-chain^{-}; immature B cells as CD19^{+}CD38^{+}CD10^{+}CD34^{-}μ-chain^{+}.

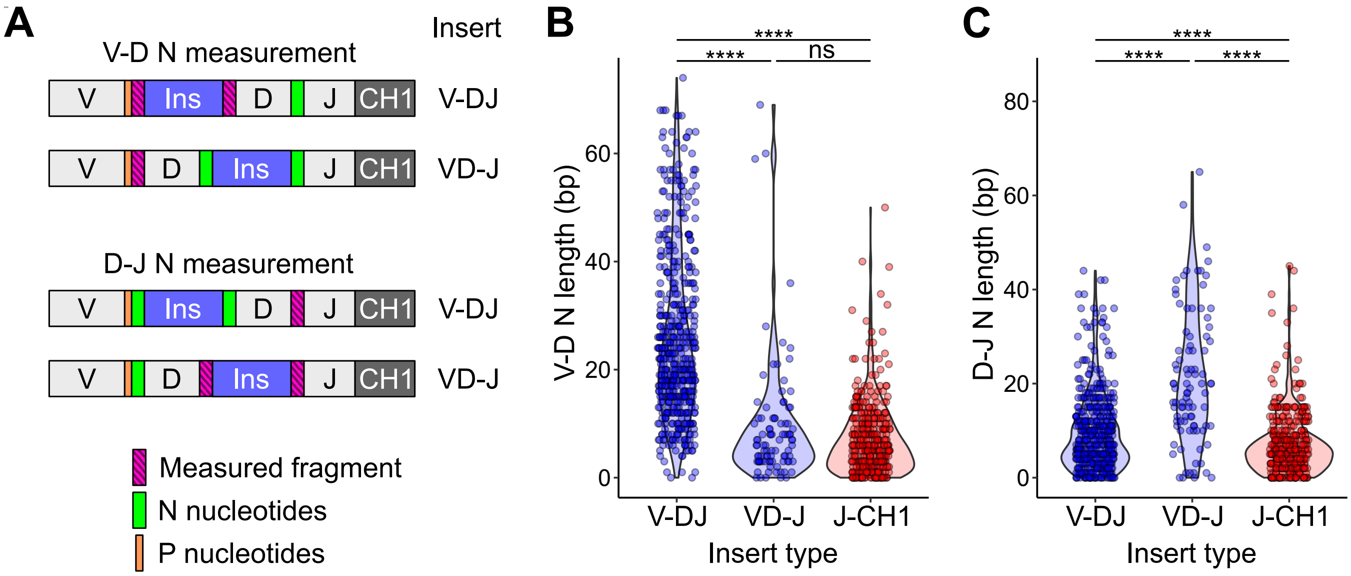


Fig. S9. Joint characteristics of insertion-containing transcripts. (*A*) Schematic representation of total length of N nucleotides (N-length) measurement, implying that if an insert is incorporated between V and D, it will sum the N-length between V and the insertion and between the insertion and D. (*B*) The shift in N-length for V-D and (*C*) D-J regions is firmly connected to the insert position. Wilcoxon test was used to determine the significance level. ns – p >= 0.05, **** - p < 0.0001.

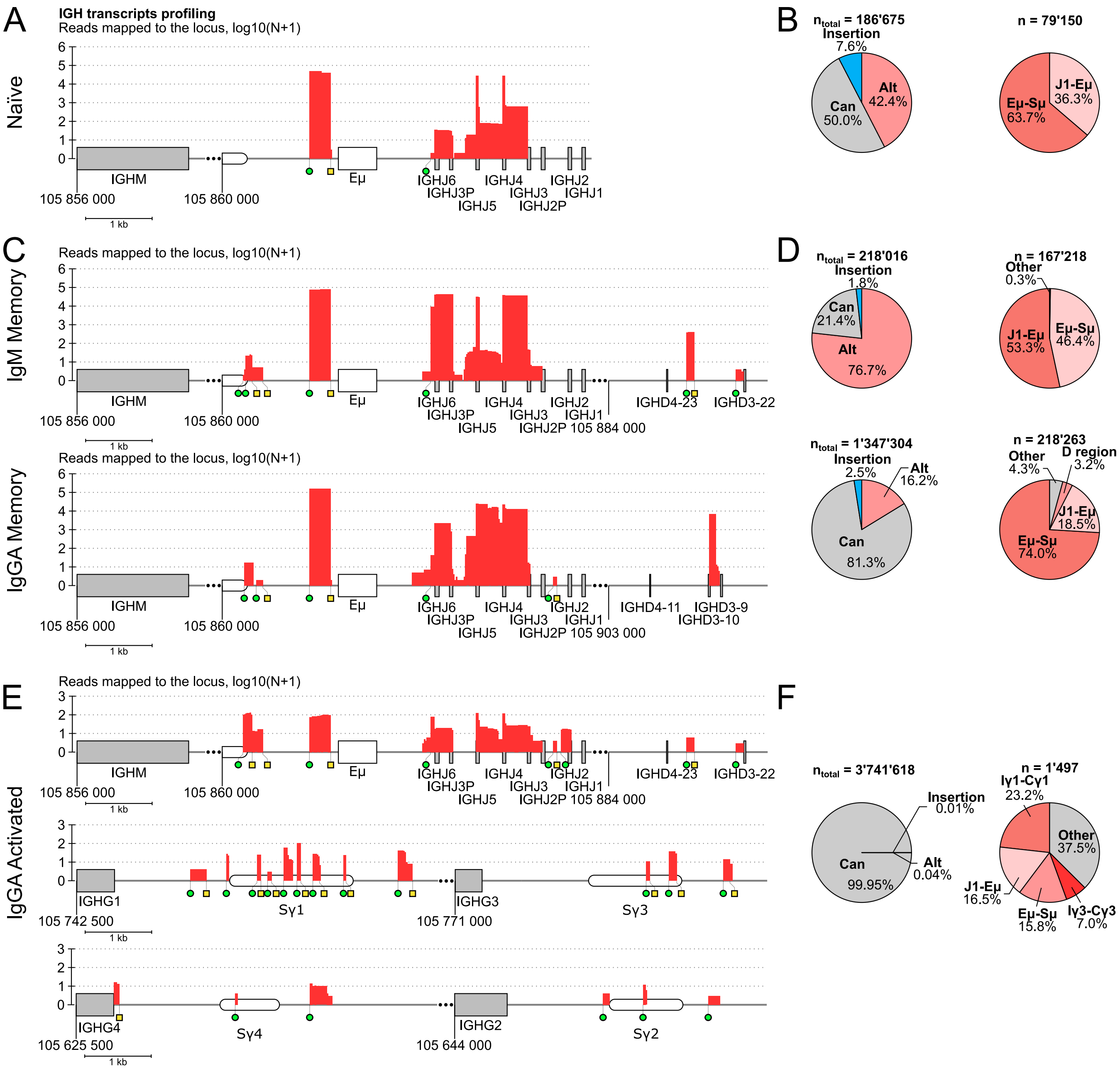


Fig. S10. The integration of alternative exons decreases upon B cell stimulation. Illumina reads obtained from (*A, B*) naïve B cells, (*C, D*) IgM and IgG/A memory B cells, and (*E, F*) IgG/A in vitro activated B cells were mapped to *IGHM* and *IGHG* loci (*A, C, E*). Pie charts (*B, D, F*) show proportions of canonical transcripts without inserts (Can, grey), alternatively spliced transcripts containing alternative and cryptic exons of the *IGH* locus (Alt, red) and ectopic inserts (blue). The pie charts on the right side represent the subset of alternatively spliced, cryptic *IGH* exons (Alt), indicating their proportions and origins. Data are represented for 4 independent biological replicates. n_{total} = total number of analyzed reads. n = number of reads mapped to alternative exons plotted in *A, C,* and *E*.

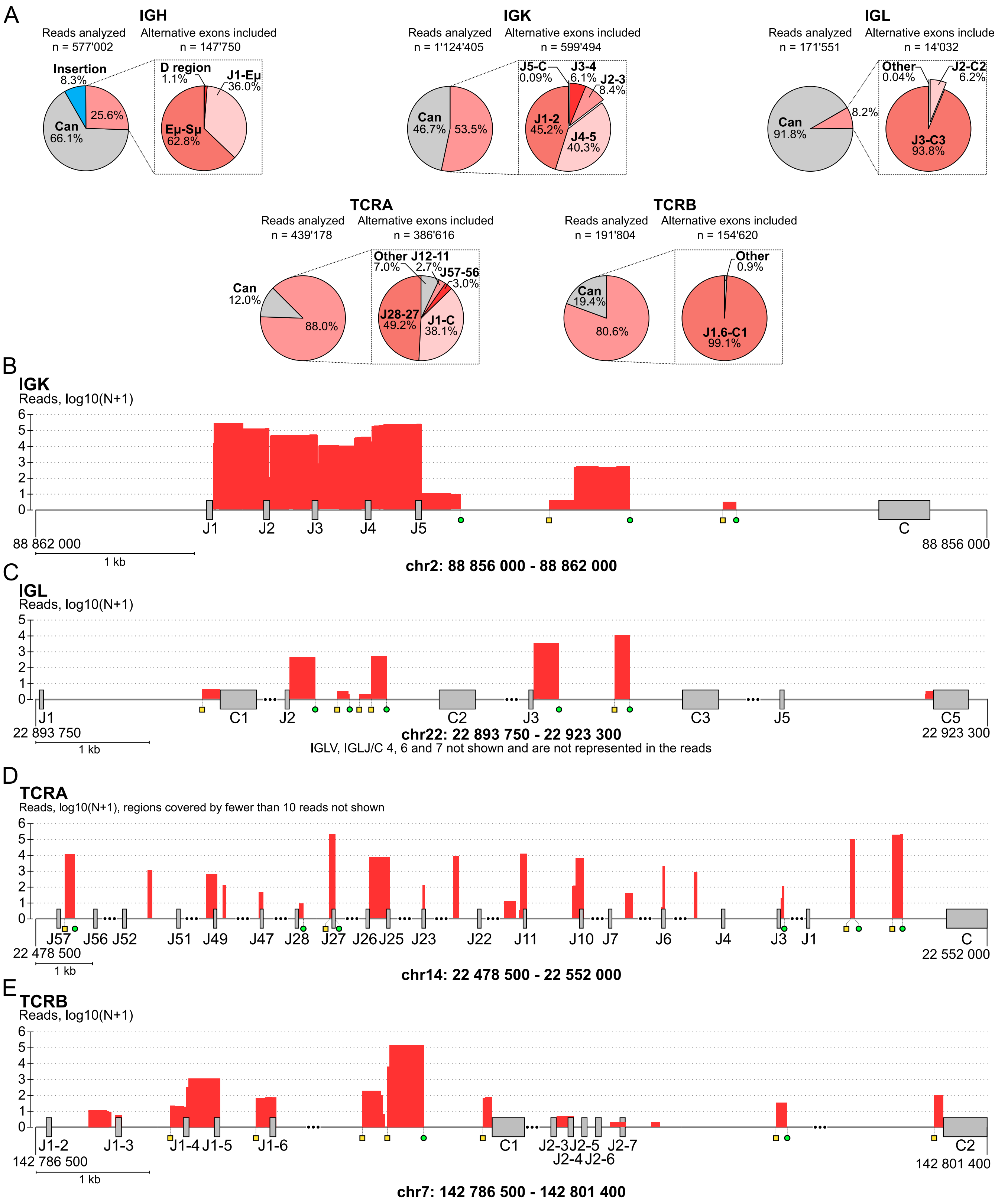


Fig. S11. Suppression PCR reveals alternatively spliced exons and inter-J fragments but no inserts in light chain and TCR transcripts. Pie charts represent Illumina reads binned into three categories: canonical (Can, grey), insert-containing (blue), and alternatively spliced transcripts (red) with subsequent distinction of alternatively spliced

transcripts by origin. (*A*) *IGH* profiling of naïve B cells, *IGK* (27 samples from 9 donors) and *IGL* (6 samples from one donor) profiling of bulk B cells, *TCRA* (2 samples from 2 donors) and *TCRB* (2 samples from 2 donors) profiling of bulk T cells. n = number of reads analyzed. "Other" refers to reads mapped to the regions within a locus not represented by a pie chart section. Histograms show Illumina reads mapping to (*B*) kappa, (*C*) lambda light chain, (*D*) *TCR* alpha, and (*E*) *TCR* beta loci. Yellow squares and green circles depict cryptic splice acceptor and donor sites. Regions are to scale, coverage is shown in logarithmical scale.

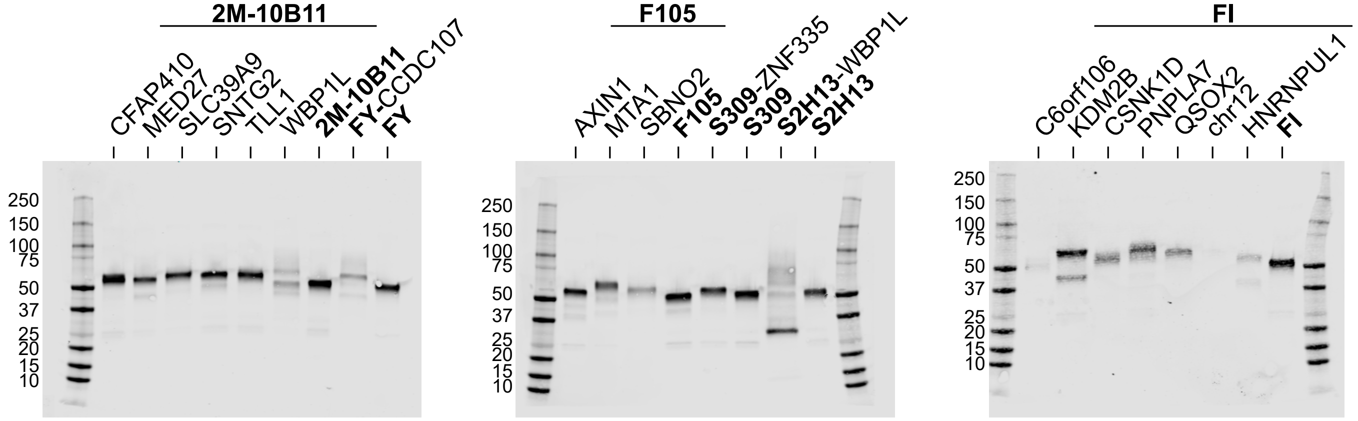


Fig. S12. Western Blot analysis of grafted insertion-containing antibodies. 500 ng of a purified antibody was loaded; for SBNO2, C6orf106, chr12, and HNRNPUL1 20 μl of the protein G pull-down eluate was loaded.