PNAS RESEARCH ARTICLE | GENETICS OPEN ACCESS

# 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]

Edited by David Schatz, Yale University School of Medicine, New Haven, CT; received March 29, 2022; accepted August 1, 2022

**Recombination of antibody genes in B cells can involve distant genomic loci and contribute a foreign antigen-binding element to form hybrid antibodies with broad reactivity for *Plasmodium falciparum*. So far, antibodies containing the extracellular domain of the LAIR1 and LILRB1 receptors represent unique examples of cross-chromosomal antibody diversification. Here, we devise a technique to profile non-VDJ elements from distant genes in antibody transcripts. Independent of the preexposure of donors to malaria parasites, non-VDJ inserts were detected in 80% of individuals at frequencies of 1 in $10^4$ to $10^5$ B cells. We detected insertions in heavy, but not in light chain or T cell receptor transcripts. We classify the insertions into four types depending on the insert origin and destination: 1) mitochondrial and 2) nuclear DNA inserts integrated at VDJ junctions; 3) inserts originating from telomere proximal genes; and 4) fragile sites incorporated between J-to-constant junctions. The latter class of inserts was exclusively found in memory and in in vitro activated B cells, while all other classes were already detected in naïve B cells. More than 10% of inserts preserved the reading frame, including transcripts with signs of antigen-driven affinity maturation. Collectively, our study unravels a mechanism of antibody diversification that is layered on the classical V(D)J and switch recombination.**

B cell diversity | antibody repertoire | insert

## Significance

The enormous diversity of antibodies is a key element to combat infections. Antibodies containing pathogen receptors were a surprising discovery that contrasted antibody diversification through classic recombination events. However, such insert-containing antibodies were thus far exclusively detected in African individuals exposed to malaria parasites and were identified as screening byproducts or through hypothesis-driven search. The prevalence and complexity of insertion events remained elusive. In this study, we devise an unbiased, systematic approach to identify inserts in the human antibody repertoire. We show that inserts from distant genomic regions occur in the majority of donors and are independent of *Plasmodium falciparum* preexposure. Our findings suggest that four distinct classes of insertion events contribute diversity to the human antibody repertoire.

The generation of B cell diversity relies on two main mechanisms. The primary repertoire emerges in early B cell development by activation of recombination-activating gene (RAG) enzymes. Random and imprecise joining of numerous variable, diversity, and joining (V, D, J) gene segments assemble a V(D)J exon that encodes the antibody variable region (1). Upon antigen encounter, B cells express activation-induced cytidine deaminase (AID) to further diversify the antibody repertoire in secondary lymphoid organs (2). AID mediates DNA nicks and double-strand breaks (DSB) in the variable as well as in the switch region, thereby initiating somatic hypermutation (SHM) and class switch recombination (CSR) (3). SHMs adjust the antibody affinity, while CSR replaces the constant domain of antibodies by switching from IgM to IgG, IgE, or IgA conferring different immune effector functions.

Using an antigen-based screening, we previously identified antibodies that gain *Plasmodium* parasite reactivity through integration of the extracellular immunoglobulin (Ig)-like domains of the leukocyte-associated immunoglobulin-like receptor 1 (LAIR1) (4, 5) or of the leukocyte immunoglobulin-like receptor 1 (LILRB1) (6). In six of nine donors, the >300 bp LAIR1 insert was positioned between V and D/J segments, while in the remaining three donors a LAIR1 exon with flanking introns integrated into the switch region and was spliced into the J-to-constant junction of the mRNA. In three donors with *LILRB1* inserts, two extra exons were exclusively detected in the J-to-constant junction. *LAIR1* and *LILRB1* are originally encoded on chromosome (chr) 19.

The recombination process to integrate inserts into the antibody locus on chr 14 relies on the generation of a DNA break acceptor site and the availability of an insert substrate. RAG and AID are enzymes known to cut at specific sites in Ig loci and are therefore likely to provide the acceptor site. For example, transfected DNA is integrated into VDJ and switch regions with a 7- and 100-fold higher frequency than into average genomic sites (7). In contrast, the mechanism that may generate insert substrates is less clear.

Both RAG and AID were previously shown to excise pieces of DNA that can be reinserted into the genome (8). Likewise, the RAG machinery was shown to insert recombination signal sequence (RSS)-containing Ig gene fragments into non-Ig sites in vitro (9–15). Reciprocally, in two human cases of follicular lymphoma, the *BCL2* gene with cryptic RSS was excised from chr 18 and inserted into the Ig locus (16). In another study, microRNA-125b-1 was found inserted at the rearranged Ig locus in a case of B cell acute lymphoblastic leukemia (17). In contrast to *LAIR1* insertions, *BCL2* and

Author contributions: K.d.l.R. designed research; D.M.C. and K.d.l.R. conceived experiments; M.L., M.F., S.K., C.V.G., C.R., A.N.D., M.A.T., and K.d.l.R. performed research; C.D. and A.L. contributed reagents and to discussions; M.L., M.F., S.K., and K.d.l.R. analyzed data; and M.L. and K.d.l.R. wrote the paper.

Competing interest statement: A.L. is an employee of Vir Biotechnology and holds shares in Vir Biotechnology.

This article is a PNAS Direct Submission.



[1]M.L. and M.F. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: Kathrin.delaRosa@mdc-berlin.de.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2205470119/-/DCSupplemental.

Published August 29, 2022.

*MIR125B1* (microRNA-125b) were accompanied by a deletion from their original loci, suggesting RAG-mediated cut-and-pasting. Instead, endogenous *LAIR1* alleles in B cells carrying the insertion remained intact, suggesting a copy-and-paste mechanism (4).

Examples of large sequences inserted at DSBs have been reported in different experimental systems and in vivo. In yeast, the absence of the Dna2 nuclease promotes duplicates of genomic DNA fragments that are captured at DSBs (18). In human nonlymphoid cells, natural DSBs can be repaired by large templated DNA patches deriving from duplication of retrotransposons and reversely transcribed RNA (19, 20). In murine pro-B cells deficient for RAG2, inserts deriving from highly transcribed genes and early replicating fragile sites (ERFSs) integrated at an I-SceI restriction site (21).

A distinct form of genomic aberrations is chromosomal translocations found in certain cancers (22, 23). Intriguingly, individuals endemically exposed to malaria are at higher risk to develop endemic Burkitt lymphoma arising from germinal center B cells (24, 25). In mice, *Plasmodium chabaudi* infection leads to chronically stimulated germinal centers with high levels of AID, thereby predisposing B cells for genomic instability and translocations (26).

LAIR1-containing antibodies, despite their prevalence in about 10% of Africans exposed to malaria, have not been detected in a cohort of more than 800 European individuals (5). It remains to be established whether malaria plays an exclusive role in selection of LAIR1 antibodies or also contributes to their generation. So far, large genomic DNA insertions have been observed in the absence of malaria in the genomic switch region of plasmacytoma (27), as well as primary human B cells of healthy European donors (5).

Here, we apply an unbiased, systematic approach to identify ectopic inserts in human antibody transcripts and address the general relevance of inserts to antibody diversity. Our methodology overcomes technical difficulties to enrich and screen for large insert-containing antibody transcripts. We characterize numerous antibody insertions in different B cell subsets, thereby shedding light on the molecular mechanisms involved. Importantly, we show that inserts in Ig transcripts occur in the vast majority of donors, independent of *Plasmodium falciparum* preexposure. Contrasting the classic recombination by predefined segments and addition of random nucleotides, our data suggest that ectopic inserts can contribute another layer of diversity forming the human antibody repertoire.

## Results

**Suppression PCR and a Data-Processing Pipeline Identify Insert-Containing Antibody Transcripts.** Previous repertoire studies have systematically omitted insert-containing antibodies because: 1) PCR amplifies preferentially shorter products, 2) size selection steps remove long PCR amplicons, and 3) limited read lengths prevent detection of chimeric sequences. To overcome these limitations, we developed an approach based on suppression PCR (28) to achieve selective amplification of rare, long antibody transcripts (*SI Appendix*, Fig. S1*A*). In the first amplification step, we introduce inverted repeats that allow the formation of intramolecular hairpins. During downstream amplification, short amplicons are disfavored due to circularization, while the ends of long amplicons remain accessible for primer annealing. To enable reliable detection of V segments and isotype, we designed PCR primers that bind in framework region (FR) 3 at least 30 bp upstream of the CDR3 and at least 14 bp downstream in CH1 exons. While a conventional PCR disabled amplification of a LAIR1-containing IgM transcript already at a 1:10 dilution, our approach detected insert-transcripts in the presence of a $10^3$- to $10^4$-fold excess of polyclonal IgM transcripts (*SI Appendix*, Fig. S1 *B* and *C*).

As insert-containing antibody transcripts may be of low abundance and different size, we designed spike-in probes of J-to-constant inserts of defined length (50, 100, 250, 500, and 750 bp) using *LAIR1* and *ICAM1* genes as templates. Probes of each insert size were mixed equimolarly and spiked into primary B cell cDNA to represent 0.1% or 0.01% of total IgG sequences (*SI Appendix*, Fig. S1 *D* and *E*). A dedicated data-processing pipeline (*SI Appendix*, Fig. S2) was developed and applied, identifying insertions from non-*IGH* loci ranging from 50 to 250 bp. Insertions of 500 bp and above were excluded, which was expected as the suppression effect was optimized to capture LAIR1-like events about 300 bp in length.

Finally, detection of the natural diversity of insert-containing transcripts was confirmed with monoclonal cell lines and primary blood samples of donors containing LAIR1-antibodies, revealing 1 versus 16 and 43 somatically hypermutated clones, respectively (*SI Appendix*, Fig. S3).

Collectively, these results validate the suppression PCR method and demonstrate that this unbiased approach can be used to identify antibody mRNAs containing inserts of sizes ranging from 50 to 300 bp.

**Inserts Are Found in Donors of Different Origin and at Various Junctions of Heavy Chain Segments.** To determine the prevalence of insert-containing Ig transcripts in individuals and their possible relationship to exposure to malaria, we screened 56 healthy individuals living in Europe and Africa by suppression PCR, from which we isolated a total of $57.6 \times 10^6$ peripheral blood B cells (Dataset S1). Overall, we identified 1,822 antibody transcripts containing inserts that mapped to regions outside of the *IGH* locus (Dataset S2). Insert-containing IGH transcripts were detected in the majority of individuals (50 of 56 donors, 89.3%) (Fig. 1*A*). Frequencies were comparable between European and African individuals ranging from $1.67 \times 10^{-6}$ to $1.74 \times 10^{-4}$ per B cell. The absence of insert-containing transcripts in six monoclonal B cell lines excludes PCR recombination artifacts, confirming the specificity of the suppression PCR approach (Fig. 1*A*).

The insertions were classified according to the insert position in the antibody transcript (Fig. 1 *B* and *C*). In the first group of transcripts, denoted as VDJ inserts (*SI Appendix*, Fig. S4 *A–D*), fragments were inserted in the antibody CDR3 region: between V and DJ (V-DJ, 33.4%,), VD and J (VD-J, 5.5%), or, when no D segment could be assigned, between V and J (V-J, 17.7%). Rare transcripts contained two distinct inserts, both between V-D and D-J segments (V-D-J, 0.7%). In the second group of transcripts, insertions were located between the VDJ and the constant region (J-CH1 inserts, 26.2%) (*SI Appendix*, Fig. S4 *E–G*). Of J-CH1 inserts, 45% spanned entire exons with the 5′- and 3′-ends precisely matching the original splice sites (Fig. 1*D*), while for a fraction of the remaining J-CH1 inserts we could detect cryptic splice sites. These findings suggest that J-CH1 inserts are the product of a genomic insert comprising an exon or a cryptic exon with flanking introns that is spliced. We also identified transcripts missing the J-segment with inserts positioned between a V segment and constant domain (V-CH1, 11.8%). This insert type may represent CDR3 region integrations that are accompanied by the deletion of the J segment via genomic loss or alternative splicing.

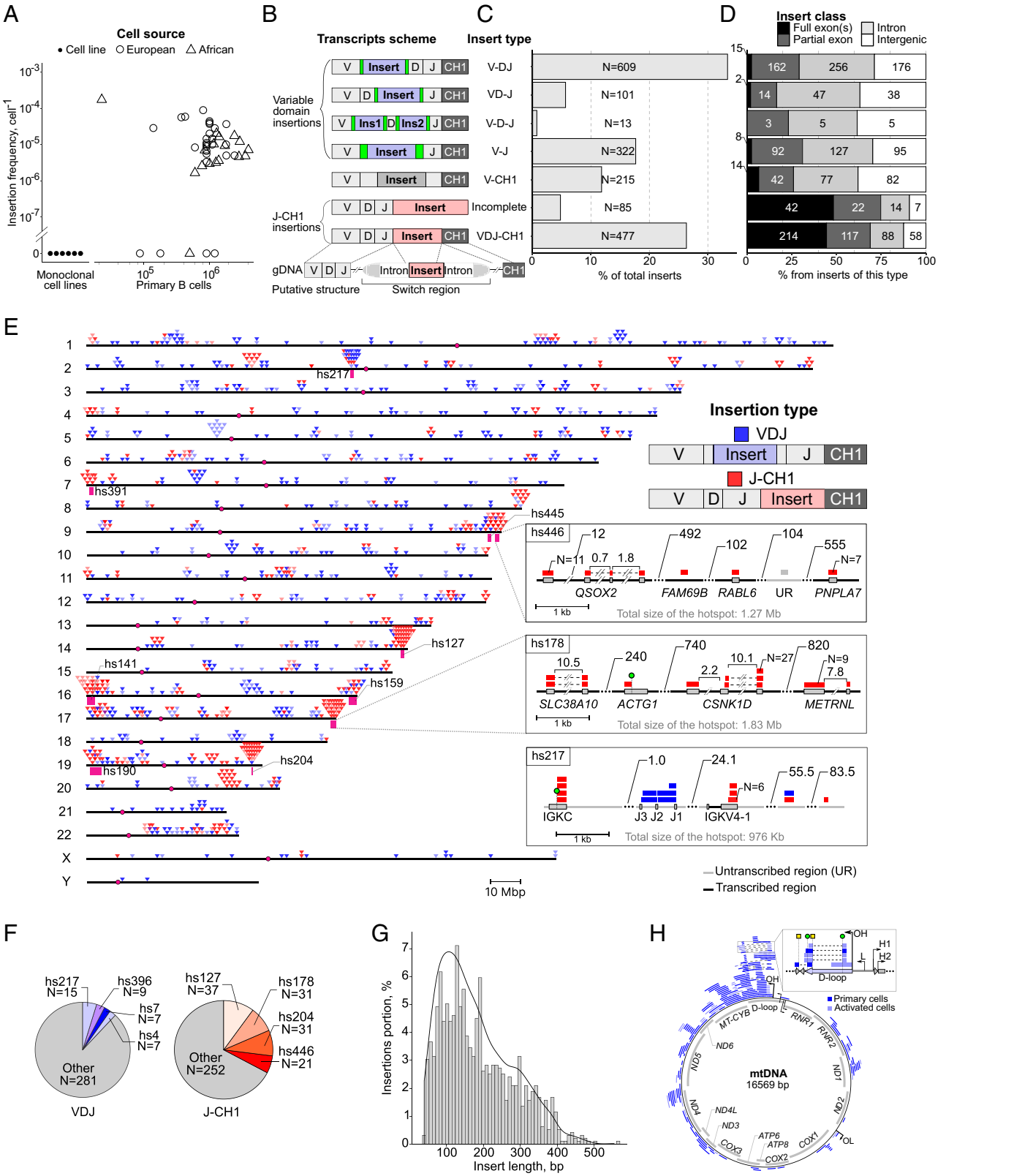


**Fig. 1.** Molecular characteristics of insertions in antibody transcripts. (*A*) Minimum frequencies of unique inserts and analyzed cell numbers of polyclonal B cells isolated from blood of European ($n$ = 29) and African ($n$ = 17) donors, and monoclonal cell lines ($n$ = 6). Frequency calculations are based on three ($n$ = 45), two ($n$ = 7), and one ($n$ = 4) biological replicates for each donor. (*B*) Insert types classified by the position of the insert between antibody V, D, J, and CH1 (constant) segments. Blue: variable (V-D-J) insert; gray: nonclassified; red: insert between VDJ and CH1 domain (J-CH1); green: P/N-nucleotides. Bottom scheme depicts the putative genomic structure of a switch region insert allowing exon splicing between J and CH1. (*C*) Frequencies, n = numbers, and (*D*) classification of inserts by origin from genes: introns, exons, or intergenic regions. (*E*) Mapping of inserts identified in blood derived B cells to nuclear chromosomes. Blue: VDJ-inserts; red: J-CH1 inserts; pale colors: inserts detected in in vitro-activated B cells. Numbers indicate genomic distance in kilobases if not stated otherwise. Dark-pink rectangles mark the 10 most common hotspots (hs). Boxes show top three hotspot donor sites. Gray rectangles: Exons; red or blue rectangles: inserts. (*F*) Portion of top four hotspot donating VDJ and J-CH1 inserts. (*G*) Length distribution of detected insertions. 352 biological replicates from 56 donors. (*H*) Inserts mapped to mtDNA. Zoom-in box on D-loop region with origins of heavy (OH) and light strand replication sites (OL). L, H1, H2: light, heavy-1, heavy-2 strand promoters. Yellow squares and green circles depict cryptic splice acceptor and donor sites.

In conclusion, within the size limitation, our experimental approach revealed numerous unique inserts at frequencies ranging from $10^{-4}$ to $10^{-6}$ B cells in most of the individuals analyzed, irrespective of their origin and preexposure to the malaria parasite.

**Insert Substrates Originate from Nuclear and Mitochondrial DNA.** We found that 85.7% of all inserts derived from the nuclear genome and map to all chromosomes (Fig. 1*E*). The majority of insertions were unique, but certain inserts were detected multiple times. Of all nuclear insertions, 18.3% originated from 10 prominent regions (Fig. 1*E* and *SI Appendix*, Fig. S5). Such hotspots were primarily associated with J-CH1 insert events (Fig. 1*F* and Dataset S3). Certain genes were frequently detected, such as *CSNK1D*, *QSOX2*, and *PNPLA7* (boxes in Fig. 1*E*). The detected insert length ranged from 29 to 563 bp, with a median equal to 160 bp (Fig. 1*G*).

Of all inserts, 14.3% originated from mitochondrial (mt) DNA (Fig. 1*H*), many of which derived from hotspots. The most prominent hotspot donated 44.1% of inserts and overlapped with the D-loop region containing transcription and replication initiation sites (Fig. 1*H* and *SI Appendix*, Fig. S6*A*). *MT-CYB*, *MT-ND5*, and *MT-ND4* genes donated 36.4% of mtDNA inserts. Of note, mitochondrial inserts were exclusively found in VDJ junctions. We hypothesized that the bacterial ancestry of mtDNA with its lack of exon-intron structures might prevent an insert from splicing between J and CH1 segments. To investigate this hypothesis, we reanalyzed our published genomic dataset of the antibody switch regions sequenced by MinION (5) and found that none of 232 genomic switch inserts detected in six European individuals was of mtDNA origin. We conclude that mtDNA is an exclusive donor for inserts that are incorporated in VDJ regions. Certain mtDNA inserts (3.4%) carried small deletions, which can be attributed to splicing at cryptic sites (*SI Appendix*, Fig. S6*B*) and points to the processing of such transcripts by the nuclear splicing machinery within the nucleus of a cell.

Collectively, the data show that inserts deriving from mtDNA exclusively integrate between V-D-J segments, while fragments of nuclear DNA are found at both V-D-J and J-CH1 junctions. Furthermore, our analysis reveals distinct hotspots in the nuclear genome and in the D-loop region of mtDNA that provide a majority of inserted templates.

**ERFS and R-Loops Contribute to J-CH1 Inserts.** We further defined the donor regions of nuclear DNA and observed that 70.0% and 88.4% of VDJ and J-CH1 inserts originated from mRNA-encoding regions, suggesting a possible link between insert source and transcription. To quantify transcription of insert donor regions, we used a published RNA sequencing dataset for human immune cells (29). Both VDJ and J-CH1 inserts originated from genes that are highly transcribed in human peripheral blood B cells (Fig. 2*A*). Previous studies have linked transcription to DNA damage (30, 31), possibly by formation of RNA/DNA hybrids called R-loops (32). To test whether inserts originate from such structures, we compared our dataset to an R-loops immunoprecipitation sequencing (DRIP-seq) database derived from human leukemic cell line K562 (33). As a control, we generated in silico datasets simulated to originate from randomized positions in the genome (*Materials and Methods*). J-CH1 insertions, but not VDJ insertions, displayed proximity to R-loop regions, with 45.4% of inserts overlapping with R-loops. In contrast, only 24.4% overlap was found for in silico controls (Fig. 2*B* and *SI Appendix*, Fig. S7*A*).

In malaria infection-driven germinal centers, DNA damage in replicating B cells preferentially occurs at ERFSs but not at common fragile sites (CFS) (26), which are genomic regions prone to break during late DNA replication (34, 35). DNA breakage at ERFSs is likely to be induced by replicative stress and is independent of AID activity (36). Insert origins showed

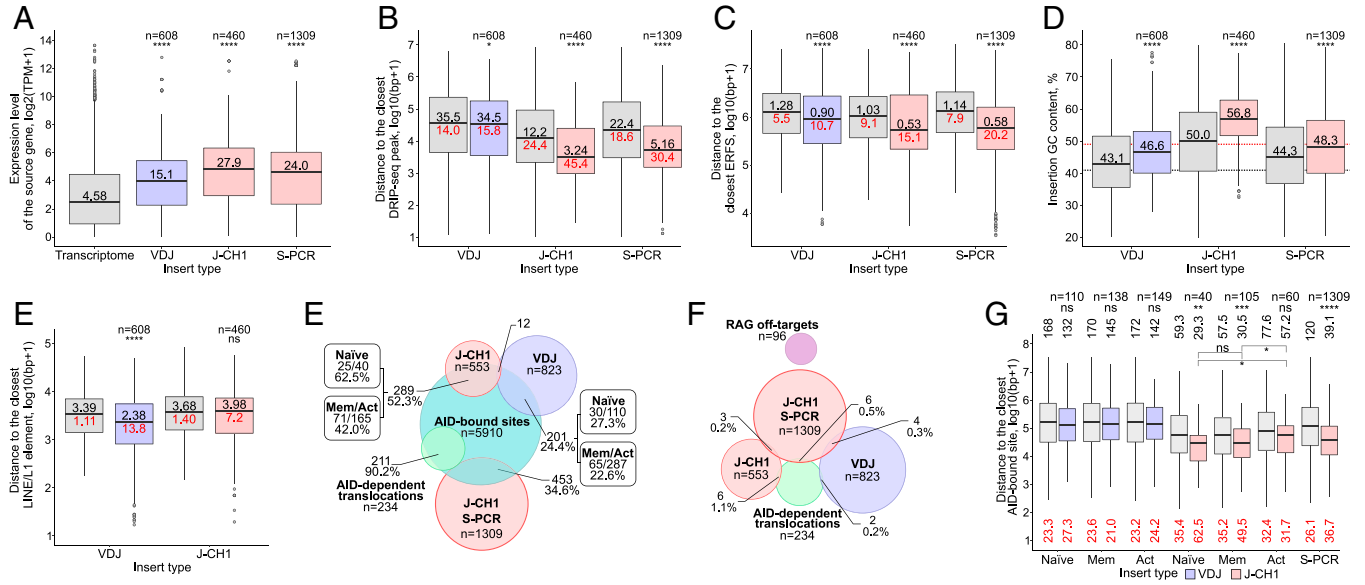


**Fig. 2.** Features of insert substrates deriving from nuclear DNA. (*A*) Expression level (transcripts per million, TPM) of VDJ and J-CH1 compared to S-PCR (previously detected genomic inserts in the switch region) (5) insert donor genes and the transcriptome in B cells. (*B*) Distance (kb) of the insert donor site to the closest R-loop determined by DRIP-seq. (*C*) Distance (Mb) to ERFS sites retrieved by lift-over of murine data to human. (*D*) GC content in percent of the inserted sequence. (*E*) Distance (kb) of the insert donor site to closest LINE element. (*F*) Euler diagram depicting overlap of insert donor sites with AID-bound sites. (*G*) Overlap of insert donor sites with RAG-mediated off-targets and AID translocations. (*H*) Median insert donor site proximity (kb) to the AID-bound sites. For all panels: red, J-CH1 inserts; blue, VDJ-inserts; gray, in silico-generated control data (182,200 artificial inserts, 100 technical replicates). *n* = number of inserts. Black numbers: median. The black lines in the boxplot represent the median, the top and the bottom of the boxplots represent 25th and 75th percentile. Red numbers: overlap in percent. Normality tested by Shapiro–Wilk test, significance computed by Wilcoxon signed-rank test. ns, $P \geq 0.05$, *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, and ****$P < 0.0001$.

significant proximity to ERFSs previously described for murine B cells (Fig. 2*C* and *SI Appendix*, Fig. S7*B*), while no proximity was observed for documented human CFSs (*SI Appendix*, Fig. S7*C*). Inserts showed an elevated GC-content (Fig. 2*D* and *SI Appendix*, Fig. S7*D*), sharing this feature with ERFS (36), while CFSs were shown to be AT-rich (37–39). In addition, inserts originated from genes that are significantly longer than the average human gene (mean 36.7 kb for J-CH1 and 100.7 kb for VDJ inserts) (*SI Appendix*, Fig. S7*E*).

As a long interspersed retrotransposable element (LINE) retrotransposon-mediated DNA repair was observed in mammalian cells and human genetic diseases (40), we tested if inserts may originate from these elements. Reported LINEs yielded a significant overlap with VDJ inserts but not J-CH1 inserts (Fig. 2*E* and *SI Appendix*, Fig. S7*F*), while no overlap was found for short interspersed retrotransposable element (SINEs) (*SI Appendix*, Fig. S7*F*).

To test if AID off-target activity may contribute to the described events, the donor sites were analyzed for overlaps with AID off-target regions using a chromatin immunoprecipitation-sequencing (ChIP-seq) database of murine B cells (41). While a random control displayed 22.8 to 32% overlap, 52.3% of J-CH1 inserts, and 24.4% of VDJ inserts derived from regions that were shown to bind AID (Fig. 2*F* and Dataset S4). As AID binding is not sufficient for the induction of a DSB (41), we compared the locations of insert origins with available datasets of AID-mediated DSBs, translocations, or mutagenesis (31, 42–47). We found a moderate overlap between J-CH1 inserts detected in memory B cells with AID-mutated sites (47) (14.3% of the inserts, 1.47% of the control dataset) and with AID-associated translocations characterized by high-throughput genomic translocation sequences (31) (5.7% of the inserts, 0.58% for the control dataset). From 234 documented AID translocation hotspots that derived from translocation-capture sequencing (43) and were converted from mice to human syntenic regions, only five genes were detected in eight unique insert-containing transcripts (Fig. 2*G*). We detected significant proximity to AID target sites for J-CH1 inserts in both naïve and memory B cells (Fig. 2*H*) but not in in vitro-activated B cells. While the overlaps in naïve B cells argue against AID off-target activity generating inserted fragments, a lack in activated cells may be explained by a limited AID activity in vitro. Instead, in vitro activation significantly increased acquisition of J-CH1 inserts deriving from ERFSs (*SI Appendix*, Fig. S7*B*), suggesting that during B cell activation and proliferation inserts are provided by replication stress rather than AID off-targeting. Inserts did not overlap with described RAG off-targets (48) (Fig. 2*G*) and, importantly, cryptic RSS flanking the insert were as frequently detected as in random controls (*SI Appendix*, Fig. S7*G*).

To match inserts detected in transcripts with events observed in genomic DNA, we made use of our previously published dataset of switch region insertions (5). As expected, more than any other insertion type, J-CH1 insertions shared similarities with fragments detected by switch PCR (S-PCR), including their origin from highly expressed genes, overlap with R-loops, and proximity to ERFS and AID-bound sites (Fig. 2). Only the GC-content of genomic switch inserts was lower, which can be explained by the presence of introns that are less GC-rich compared to exons (Fig. 2*D*).

In conclusion, our data suggest that distinct molecular mechanisms contribute to insert substrates. While generation of the insert fragments appears to be independent of RAG activity, our results suggest that a portion of VDJ inserts derive from LINE elements. J-CH1 inserts may originate from ERFSs and R-loop proximal regions, and a contribution of AID off-targeting may be moderate and cannot be excluded.

**Acceptor Break Sites Occur at Distinct B Cell Developmental Stages.** VDJ inserts were detected in all populations analyzed at a similar frequency, namely pre-B cells (*SI Appendix*, Fig. S8), naïve, and memory B cells (Dataset S2), suggesting the involvement of RAG recombinase in providing acceptor sites during heavy chain rearrangement. Interestingly, the analysis of N nucleotides unraveled that an insert donor template may be subject to TdT-mediated N nucleotide addition (*SI Appendix*, Fig. S9). We expected that J-CH1 inserts mainly occur in memory B cells as acceptor sites can be provided by AID-mediated DSBs in the genomic switch region (Fig. 1 *B*, bottom scheme). However, J-CH1 inserts were detected in naïve B cells, suggesting that certain J-CH1 inserts can be acquired in the absence of AID activity. Since inserts deriving from ERFSs were detected only after CD40L/interleukin (IL)-4 in vitro stimulation (*SI Appendix*, Fig. S7*B*), we expected that B cell activation increases insert frequency. However, suppression PCR analysis revealed no significant change in frequency, which might be explained by a lower efficiency of splicing after in vitro stimulation (*SI Appendix*, Fig. S10). Nevertheless, we observed profound qualitative differences, as J-CH1 inserts detected in naïve B cells derived from telomere-proximal regions (Fig. 3*A*), while ERFS proximity was exclusive for activated cells. Of note, insertions detected by genomic S-PCR (5) are not prone to originate from the subtelomeric regions (Fig. 3*A*), which is a feature they share with the J-CH1 inserts detected in in vitro activated cells.

Taken together, our results suggest that acceptor sites for VDJ inserts occur during RAG recombination and a portion of J-CH1 inserts may be acquired during AID targeting. In addition, a particular J-CH1 insert class of telomere-proximal origin was detected in naïve B cells.

**J-CH1 Inserts of Telomere-Proximal Origin Span Multiple Exons and Are Shared between Donors.** We observed that 10 to 20% of inserts comprise multiple consecutive exons of one donor gene (Fig. 3*B*), with introns spanning up to multiple kilobases. For example, the *CSNK1D* gene donated an insert with two exons that are 10.1 kb apart in the donor gene (Fig. 3*C*), and the *QSOX2* gene donated up to three exons covering a genomic distance of 2.9 kb (Fig. 1 *E*, upper box). J-CH1 inserts deriving from telomere-proximal regions are thus supposed to span >1 kb genomic distances (Fig. 3*D*). Inserts deriving from the *CSNK1D* gene were detected in 8 of 56 donors. In three of them, the *CSNK1D* insert was observed in two to five distinct antibodies with unique CDR3 junctions (Fig. 3*E*). Similarly, *QSOX2* inserts were detected in 12 antibodies of 7 donors (Dataset S2).

In conclusion, J-CH1 inserts in naïve B cells belong to a particular insert class likely resulting from splicing of multiexon genomic insertions that originate from telomere-proximal hotspots.

**Ectopic Inserts Were Not Undetectable in Light Chain or T Cell Receptor Transcripts.** To address if V-J junctions of light chains serve as acceptor sites for inserts, we applied the suppression PCR technique to Ig-κ (IGK) transcripts. In 27 samples from 9 donors, no light chain inserts were detected (*SI Appendix*, Fig. S11*A*). This finding was somewhat unexpected, as our data suggested that RAG might mediate insertions during heavy chain rearrangement. Mapping the reads to the *IGK* locus revealed frequent inclusions of inter-J sequences and alternatively spliced exons in the transcripts (*SI Appendix*, Fig. S11*B*), confirming

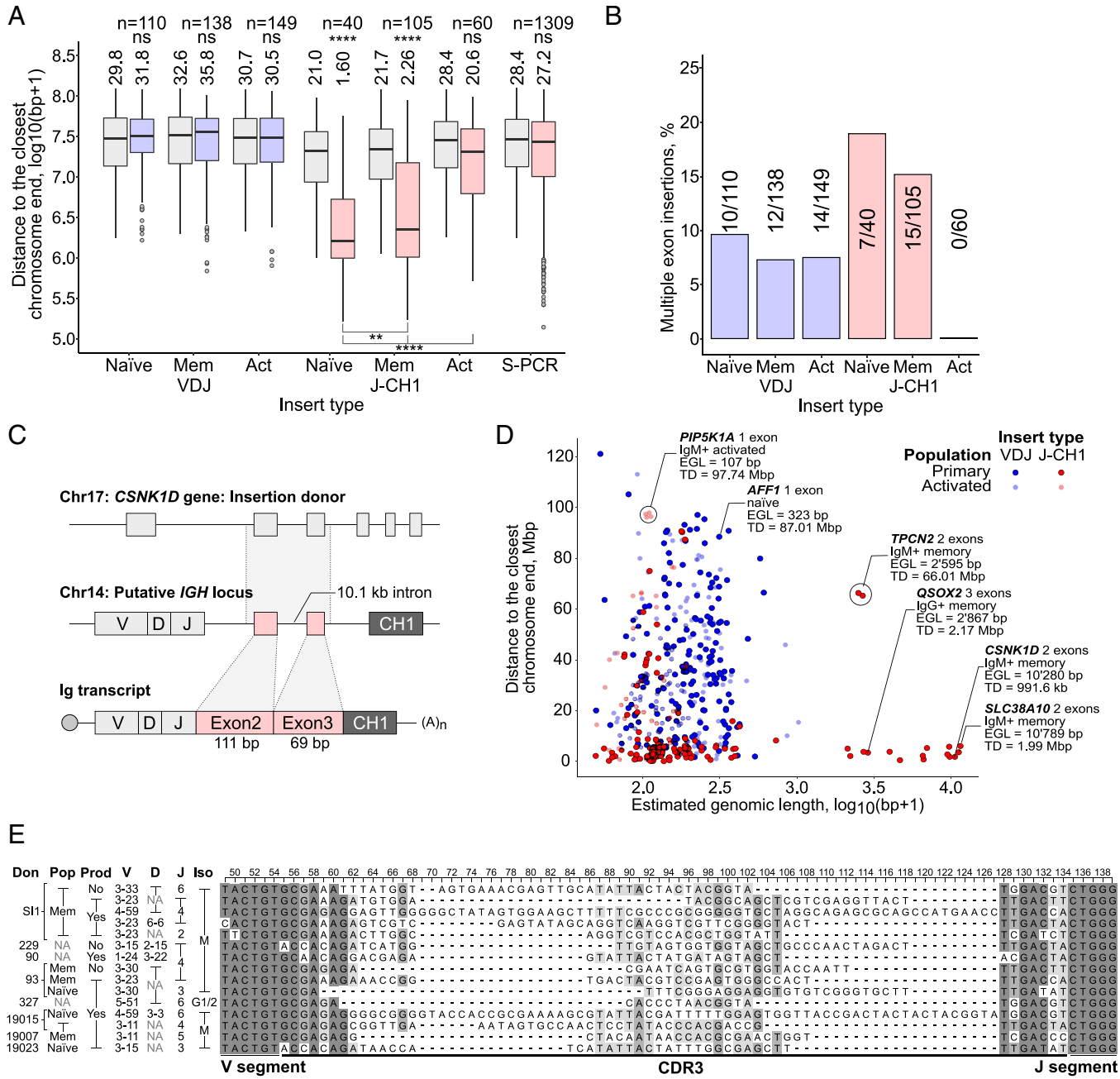


**Fig. 3.** Telomere-proximal J-CH1 inserts span multiple exons and are shared between distinct donors. (*A*) Distance of inserts to chromosome ends detected by suppression PCR for naïve, Mem, Act B cells, and by S-PCR (previously detected genomic inserts in the switch region) (5); in silico controls in gray. Black text above the boxes: median in Mb. *n* = number of inserts. (*B*) Percentage and numbers of multiple exon inserts. (*C*) Schematic representation of a *CSNK1D* insert with the original insert donor locus, the putative IGH structure, and the detected transcript. (*D*) Telomer proximity and estimated genomic length calculated by 5′- and 3′-end coordinates of insert flanks of VDJ and J-CH1 inserts detected in primary and Act B cells (semitransparent dots). (*E*) CDR3 region alignments for CSNK1D-containing insert transcripts. Donor names (Don; *n* = 8), B cell population from which inserts were isolated (Pop), frame preservation (Prod), VDJ segment usage, and isotype (Iso) are shown. Gray shade density represents homology. The black lines in the boxplot represent the median, the top and the bottom of the boxplots represent 25th and 75th percentile. ns, $P \geq 0.05$, *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, and ****$P < 0.0001$.

enrichment of long natural transcripts by suppression PCR. To clarify if alternative IGH transcripts, such as inter-J splice variants, could mask insert detection, we studied the *IGL* locus with consecutive J-C cassettes that would exclude integration of inter-J sequences. No inserts were found in λ light chain transcripts in six samples from a single donor (*SI Appendix*, Fig. S11 *A* and *C*). Finally, the interlaced structure of the *IGL* locus does not prevent the inclusion of cryptic J-C exons, which dominated the size-selected λ light chain amplification. Similar to the light chains, suppression PCR in two donors targeting T cell receptor (TCR)A and TCRB transcripts did not reveal any ectopic insert, but pointed to a significant incorporation of inter-J fragments and J-C alternative exons (*SI Appendix*, Fig. S11 *A*, *D*, and *E*). We conclude that suppression PCR did not extract any ectopic inserts in TCR and light chain transcripts, suggesting that the *IGH* locus is most permissive for insert acquisition.

**In-Frame Insert Transcripts Somatically Hypermutate and Class Switch.** To address the potential contribution of inserts to the functional antibody repertoire, we analyzed the frequency of in-frame transcripts and signatures of antigen-driven affinity maturation. A total of 13.2% of inserts were in-frame,

which were more frequently detected between J-CH1 junctions (32.8% of in-frame inserts; 11.1% of total inserts) (Fig. 4 *A* and *B*). In contrast, 3.1% of all VDJ inserts (1.8% of total inserts) preserved the frame. To investigate if inserts could contribute to immune responses, we analyzed in-frame insert-containing transcripts isolated from memory B cells. Of in-frame VDJ inserts, 4.2% and 16.7% derived from IgM memory and IgG/A class-switched memory B cells, respectively; and 21.2% and 51% of J-CH1 inserts derived from IgM memory and IgG/A class-switched memory B cells, respectively (Fig. 4*B*). SHMs were detected in VDJ in-frame inserts of memory but not naïve or bone marrow-derived pre-B cells (Fig. 4*C*). Our data thus provide evidence that B cells expressing in-frame insert-containing transcripts underwent antigen-driven affinity maturation.

To assess the effect of ectopic insertions on the stability and binding properties of antibodies by in vitro expression, we cloned 20 of the most frequently detected in-frame fragments (8 J-CH1 and 12 VDJ) into available recombinant antibodies that share VDJ-homology with the insertion-carrier transcript (*Materials and Methods*, Fig. 4*D*, and Dataset S5). One VDJ insert (WBP1L) was cloned into four antibody backbones. In total, 11 of 15 VDJ and all J-CH1 insert-containing antibodies (further called VDJ Abs and J-CH1 Abs) were produced, detected by ELISA in culture supernatants, and purified by protein G (Fig. 4*E*). Except for WBP1L, FI-KDM2B, and FI-HNRNPUL1 insertions, single bands of expected molecular weight compared to the backbone antibodies were observed (*SI Appendix*, Fig. S12). Next, we determined antibody binding toward the antigen recognized by the backbone antibody (Fig. 4*F* and Dataset S5). As expected, 9 of 14 purified VDJ Abs lost reactivity to the original target. Interestingly, 2 of 14 VDJ Abs presented with moderate target binding, but at the same time gained weak reactivity to irrelevant control antigens. Three of 14 VDJ Abs gained unspecific binding properties to all tested targets. Six of eight purified J-CH1 Abs maintained binding and in part gained weak to moderate unspecific binding. We conclude that the antibody structure is permissible for an insertion of 35 to 76 amino acids between VDJ and CH1 domains and for 33 to 86 amino acid insertions in the heavy chain CDR3 region.

## Discussion

**Frequency of Inserts in the Human Antibody Repertoire.** By developing a target-independent approach, we identified insert-containing antibody transcripts in the vast majority of individuals of a genetically diverse cohort. Inserts were detected at frequencies of $10^{-4}$ to $10^{-6}$ B cells, of which about 90% were out-of-frame. This finding is consistent with a limited contribution of inserts to total B cell diversity. We have to consider, however, that the antibody repertoire is dynamic and includes spatial and temporal diversity. Naïve B cells are continuously renewed and, at a given time, the number of circulating clones in all body compartments may range between $10^9$ (49) and $10^{11}$ (50) naïve and total B cells, respectively. LAIR1 inserts, despite being undetectable in nonexposed subjects, were readily selected in at least 5% of malaria-exposed individuals (5), demonstrating that inserts can contribute to the antibody response. This may especially apply when the insert is a pathogen receptor, but it is also possible that random insertions may generate new antigen-binding domains that may further diversify through somatic mutation.

Insert detection was limited not only by the number of B cells analyzed but also by the fact that suppression PCR is size-selective. As PCR and screening conditions limit the detectable range, the obtained numbers need to be interpreted as minimum insert frequencies. Therefore, the contribution of inserts to diversity might be higher than estimated in this study.

**Four Insert Classes Defined by Acceptor Break Sites and Insert Origins.** Our results suggest a classification of inserts into four classes based on two parameters: the site of insertion and the source of the insert (Table 1). V-D-J inserts derived either from nuclear genes (nucVDJ) or mitochondrial DNA (mtVDJ) and J-CH1 inserts derived from telomere proximal regions (nucJC) or ERFS (telJC).

We assume that nucVDJ and mtVDJ inserts occur prior to antigen encounter as a consequence of repair at RAG-mediated breaks. The origins of nucVDJ inserts in part overlapped with LINE elements and large genes. The latter are known to be prone to DNA damage due to a clash of transcription with replication (51, 52). Although cellular divisions of B cell progenitors are limited, it may be possible that in rare cases single-stranded DNA may be released from stalled replication forks (53) serving as patches for the repair of DNA breaks.

We observed that mtVDJ inserts derived from the mitochondrial chromosome but not from genomic mitochondrial pseudogenes (54). NUMTogenesis—the transfer of mtDNA to the nucleus—has been suggested to play a role in cancer development (55) and 19% of templated sequence insertion polymorphisms found in the human genome derive from mtDNA (56). Intriguingly, mtDNA nuclear insertion was detected in a zygote or early after fertilization but not in experimentally induced I-SceI DSBs in a leukemia cell line (20). These results point to either a different mechanism exploited by mtDNA or to a distinct cellular state at which mtDNA donates inserts.

NucJC inserts were mainly present in memory and in vitro class-switched B cells, suggesting that acceptor break sites are generated by AID. Substrate sources, instead, overlapped with ERFSs that are independent of AID activity (36). The observed overlap of some detected inserts with AID-bound sites may result from transcription and open chromatin contributing to both AID binding and fragment generation.

Naïve B cells harbored telJC inserts that comprised multiple exons. It remains unclear how acceptor sites between the J-to-constant region are generated, since AID may only rarely be expressed at the early stages of B cell development (57). TelJC insert substrates and acceptor sites may, for example, be mediated by RAG proteins that excise DNA over large genomic distances and have the potential to function as a transposase that targets GC-rich sequences and hairpins (12, 14, 58). Considering that telJC inserts may span multiple kilobases in the genome, conventional techniques in the past may have erroneously classified telJC inserts as translocation events. Linear amplification methods (59) could in the future be combined with suppression PCR to study the genomic architecture and deliver further mechanistic insights into the nature of the phenomenon. However, a dual-sided approach may be needed to clearly distinguish a multikilobase genomic insert from a translocation.

Finally, the *LAIR1* gene, which donates inserts in malaria-exposed individuals (5), shares characteristics with inserts identified in this study. The *LAIR1* gene is 19 kb long, which is almost double the average gene size, and is proximal (at 4.26 Mb distance) to the chromosome end, 8.6 kb apart from the closest R-loop. *LAIR1* is highly expressed in progenitor and naïve B cells, as well as in a fraction of memory B cells. *LAIR1* does not overlap with ERFSs; however, it contains a CFS. Together, these features hint at *LAIR1* inserts resembling nucVDJ and telJC insert classes.

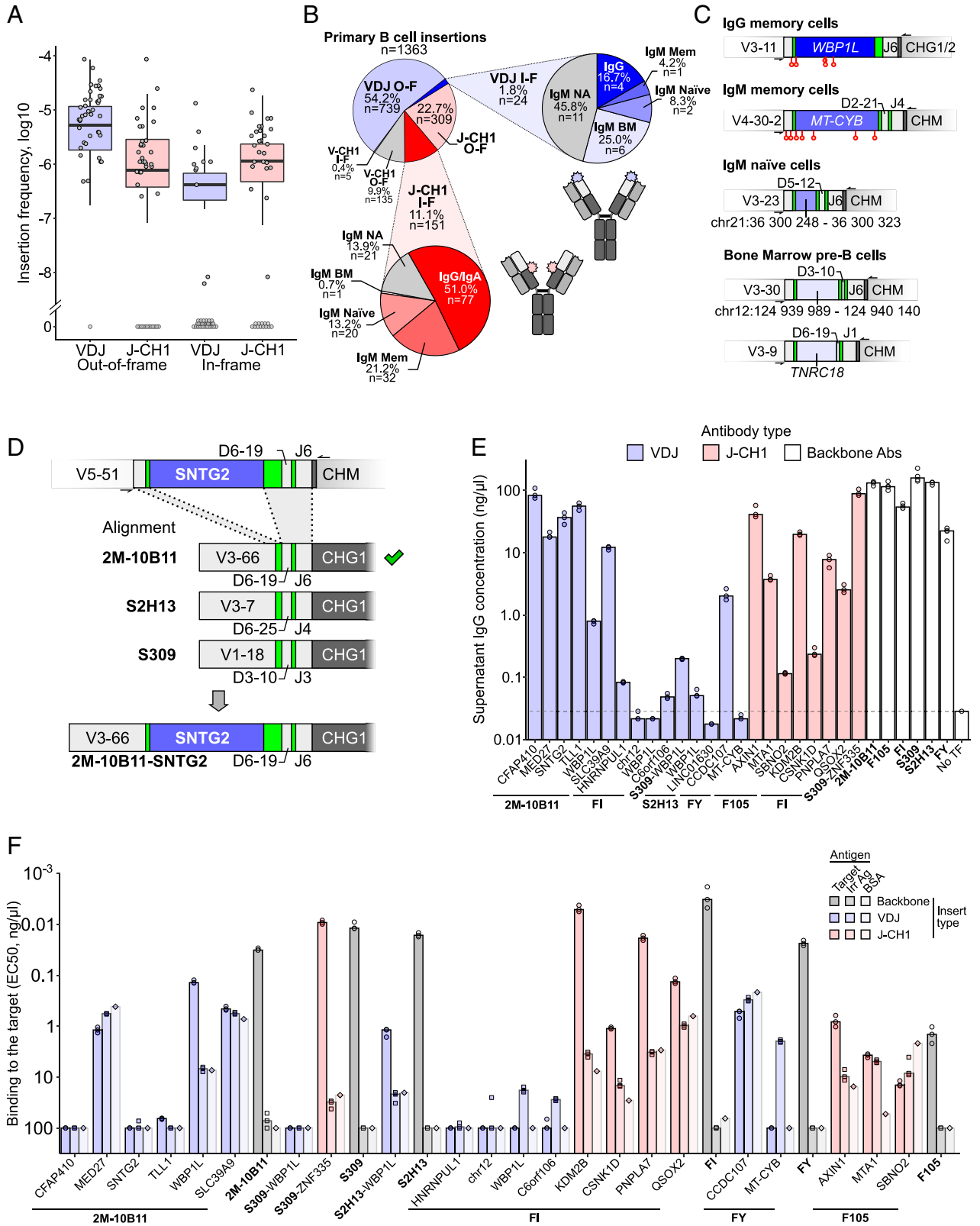


**Fig. 4.** In-frame inserts show signs of antigen-driven affinity maturation in memory B cells and are compatible with expression when grafted into recombinant antibodies. (*A*) In- and out-of-frame insert frequencies for different insert types. Each point represents an insert frequency of 1 of 56 donors analyzed (352 biological replicates). The black lines in the boxplot represent the median, the top and the bottom of the boxplots represent 25th and 75th percentile. (*B*) Pie chart depicts out-of-frame (O-F) and in-frame (I-F) insertion frequencies with two subpies deconvoluting contributing B cell subpopulations. Schemes represent insert position of J-CH1 (red) and VDJ (blue) inserts in the antibody protein. *n* = insert numbers. NA = IgM transcripts derived from bulk sorting. (*C*) Schematic representation of in-frame VDJ insert-containing transcripts detected in IgG/IgM memory cells, naïve B cells, and pre-B cells. Red circles mark the SHMs. (*D*) Scheme depicting the selection of recombinant backbone antibodies with known specificity for grafting of in-frame inserts, which was based on a most favorable alignment of VDJ joints. SNTG2 is shown as an example. (*E*) Indicated constructs (15 VDJ Abs and 8 J-CH1 Abs) were expressed in Expi293F cells and the concentration of IgG in culture supernatants was determined by ELISA. (*F*) Protein G purified insert-antibody grafts were analyzed for their ability to bind the target antigen of the backbone antibody, an irrelevant antigen (Irr Ag), and bovine serum albumin (BSA). Green rectangles in *C* and *D* denote the N nucleotides. In *A* and *B*, in vitro-activated cells were excluded from analysis.

**Table 1. Four distinct insert classes: nucVDJ, mtVDJ, nucJC, and telJC were defined**

| | nucVDJ | mtVDJ | nucJC | telJC |
|---|---|---|---|---|
| Acceptor site | RAG mediated | RAG-mediated | AID-mediated | Switch region fragility, RAG or AID |
| Insert source | Nuclear DNA: LINE elements, very large genes (mean 100 kb), slightly GC-rich | mtDNA: D-loop | Nuclear DNA: ERFS proximal, highly GC-rich, large genes (mean 50 kb) | Nuclear DNA: telomere-proximal, R-loop proximal, highly GC-rich |
| Occurrence | Prior to antigen encounter | Prior to antigen encounter | After B cell activation | Prior to antigen encounter |

**Insert Contribution to the Antibody Repertoire.** VDJ inserts and J-CH1 inserts have to meet different requirements to be expressed as a protein. VDJ inserts integrate into the CDR3 region and thus need to preserve the open reading frame to allow a positive B cell selection. Instead, J-CH1 insertions can be alternatively spliced giving rise to B cell receptors (BCRs) with and without an insert. As VDJ inserts are likely to arise during B cell development, they may be subject to selection. Despite rare exceptions (60), the vast majority of B cells is committed to express a single antibody heavy chain due to allelic exclusion (61–64). Therefore, detection of heavy chain transcripts containing hypermutated, in-frame VDJ inserts suggest that inserts contribute to functional diversity.

J-CH1 inserts comprise entire exons and are susceptible to alternative splicing. Despite having a direct impact on specificity, the addition of an insert between VDJ and CH1 domains could also affect protein conformation and binding, as well as clustering of BCRs, which was shown to depend on the spatial organization in the plasma membrane (65). Despite detecting several insert transcripts compatible with expression, we did not detect another *LAIR1*-like example. Two main factors may be crucial: gain of a particular specificity by the integration of a pathogen receptor and chronic stimulation by that pathogen. The latter would increase the likelihood to activate a rare B cell clone. In the future, an exclusive screening of memory B cells of donors exposed to chronic infections may enable isolation of other functional insert antibodies. Finally, our work not only provides a tool to unravel an additional layer of antibody diversity, but also provides molecular insights into the mechanisms of insert acquisition that might be of particular relevance in chronic infectious diseases.

## Materials and Methods

**Human Specimens.** Blood from healthy individuals was obtained from the German and Swiss Red Cross. In all cases, written informed consents were obtained and samples were anonymized. All healthy donor samples were tested negative for HIV, HBV, HCV. Peripheral blood mononuclear cells (PBMCs) from malaria preexposed volunteers were obtained from the prevaccination period of the P27A vaccine phase Ib trial (ClinicalTrials.gov Identifier: NCT01949909, Pan African Clinical Trial Registry identifier: PACTR201310000683408). This study was conducted with approval from the Tanzanian Food and Drug Administration (TFDA; Dar-es-Salaam, TFDA13/CTR/004/03), National Institute for Medical Research (NIMR; Dar-es-Salaam, NIMR/HQ/R8a/Vol.IX/1742), and ethical review boards at Ifakara Health Institute and the University of Lausanne, and all volunteers provided informed consent before blood donation. Bone marrow specimens were obtained from AllCells.

**Extraction of Human PBMCs.** Blood samples were diluted 1:1 with Dulbecco's phosphate-buffered saline (DPBS, Sigma, Cat#D8537-500ML) containing 2 mM EDTA and separated using Ficoll density gradient centrifugation (Roth, Cat#0642.2). PBMCs were either immediately processed or resuspended in 90% FBS with 10% DMSO, frozen in freezing chambers at −80 °C, and transferred to liquid nitrogen. Frozen PBMC or cells extracted from bone marrow material were thawed at 37 °C for 4 min and gradually diluted in 10-fold excess of B cell medium (composition is described below in *B Cell Activation*). Thawed cells were centrifuged for 5 min at 350 × *g*, resuspended in B cell medium, and used for flow cytometry, in vitro activation, or RNA extraction.

**Cell Enrichment and Flow Cytometry.** PBMCs were enriched for $CD19^+$ cells by magnetic cell separation (MACS) according to the manufacturer's manual (Miltenyi Biotec, Cat#130-050-301). For T cell analysis, PBMCs were enriched by MACS using anti-CD3 microbeads (Miltenyi Biotec, Cat#130-050-101). For FACS analysis and cell sorting, $CD19^+$ B cells were incubated with antibodies specific for CD19, CD27, CD38, IgM, IgG, IgD, IgA, and κ light chains. $CD3^+$ T cells were FACS-sorted using anti-CD3, anti-CD4, and anti-CD8 antibodies. Bone marrow samples were thawed, washed, and incubated with antibodies specific for CD10, CD34, CD38, CD19, and the IgM μ-chain. After sorting, collected cells were immediately processed or frozen in 90% FBS with 10% DMSO. Cell populations were sorted according to the following phenotypes: 1) naïve B cells: $CD19^+CD27^-IgM^+IgD^+$; 2) IgM memory B and Activated IgM cells: $CD19^+CD27^+IgM^+$; 3) $IgG/A^+$ switched memory and Activated $IgG/A^+$ cells: $CD19^+CD27^+IgG/A^+$; 4) $IgK^+$ B cells: $CD19^+IgK^+$; 5) $IgL^+$ B cells: $CD19^+IgK^-$; 6) CD4 T cells: $CD3^+CD4^+$; 7) CD8 T cells: $CD3^+CD8^+$; 8) pro-B cells: $CD38^+CD19^+CD10^+CD34^+$μ-$chain^-$; 9) pre-B cells: $CD38^+CD19^+CD10^+CD34^-$μ-$chain^-$; 10) immature B cells: $CD38^+CD19^+CD10^+CD34^-$ μ-$chain^+$.

**B Cell Activation.** Sorted naïve B cells were seeded at $3 \times 10^4/cm^2$ and cocultured with irradiated cells expressing CD40L at a 10:1 ratio. Cells were cultured in B cell medium: complete RPMI plus 10% FBS, 1% sodium pyruvate, 1% nonessential amino acids, 1% penicillin-streptomycin (Thermo Fisher), 1% GlutaMAX, 0.1% 2-mercaptoethanol, 0.02 ng/mL transferrin (Sigma), and 0.1 mg/mL kanamycin (Serva). Cells were activated with 25 ng/mL IL-4. Every 3 d cells were restimulated by adding 25 ng/mL IL-4. On the seventh day, cells were reseeded at a density of $10^6$ cells/mL, provided with fresh CD40L expressing cells, and activated with 25 ng/mL IL-4 every 3 d. On the 14th day, cells were collected, sorted as described above and immediately processed by either extracting total RNA or genomic DNA.

**Suppression PCR.** Total RNA was isolated from up to $10^5$ cells with RNeasy Mini Kit (Qiagen) according to the manufacturer's protocol and either stored at −80 °C or processed immediately. cDNA synthesis was performed using SuperScript IV Reverse Transcriptase (Thermo Fisher) with 1 μM final primer concentration (hIGM or an equimolar mixture of hIGA and hIGG; for primer sequence, see Dataset S6) keeping reaction at a minimum 55 °C to avoid off-target priming. Reactions were purified with 1.4× AMPure XP beads (Beckman Coulter), according to manufacturer's recommendations and eluted in TE buffer (Tris-HCl 10 mM, EDTA 1 mM). All PCRs described below are performed with Q5 Hot Start Polymerase (New England Biolabs) in the recommended buffer and 0.25 mM of each dNTP and 0.2 μM of each primer if not stated otherwise. PCR-I: purified cDNA was amplified with a mixture of suppression primers annealing to V gene segments and constant regions (Dataset S6) using the following thermocycling settings: 98 °C 30 s, (98 °C 10 s, 55 °C 10 s, 72 °C 10 s) × 10, 72° 60 s. Amplicons were purified with 0.8× AMPure XP beads. PCR-II: Tagged amplicons with inverted repeats on both ends were amplified with distal_22 primer at the final concentration 0.5 μM and the following thermocycling settings: 98 °C

30 s, (98 °C 10 s, 52.7 °C 10 s, 72 °C 10 s) × 37, 72 °C 60 s. Products were purified with 0.8× AMPure XP beads. PCR-III: to prepare for Illumina adapters introduction, suppression products were amplified with U1- and U2-overhang primers, with the thermocycling conditions: 98 °C 30 s, (98 °C 10 s, 55 °C 10 s, 72 °C 10 s) × 10, 72°60 s. Products were purified with 0.8× AMPure XP beads. iPCR: unique indices and flow cell adapters were introduced with FC1-i5X-U1 and FC2-i7X-U2 primers (X = index identifier) by the following program: 98 °C 30 s, (98 °C 10 s, 55 °C 10 s, 72 °C 10 s) × 5, 72 °C 60 s. PCR products were analyzed by agarose gel electrophoresis, purified with 0.8× beads. Concentration was measured either by DS-11 spectrophotometer (DeNovix) or Qubit HS DNA kit (Thermo Fisher). The size distribution of the products was estimated by gel image processing with GelAnalyzer software (66) and used to calculate the molarity. Indexed samples were mixed equimolarly, repurified by 0.6× AMPure XP beads, measured with BioAnalyzer High Sensitivity DNA kit (Agilent), and prepared for sequencing with 2 × 300 MiSeq Reagents kits v3 (Illumina) according to manufacturer's manual. Libraries for light chains and T-cell receptor chains were performed following the same procedure with distinct primers (Dataset S6).

**Molecular Cloning of Spike-In Plasmids.** Control plasmids were generated by introducing LAIR1 or ICAM1 inserts into antibody vector encoding IgG heavy chain. Fragments of the designed constructs were amplified from total PBMC cDNA with the corresponding primers and included: 1) IGH variable domain (amplified with VH3-23_Fwd_NcoI and IGHJ4_Rev_BamHI); 2) IGHG1 constant gene fragment (IGHG_Fwd_NheI, IGHG_Rev_SalI); and 3) *LAIR1* and *ICAM1* fragments of variable length (LAIR1_Fwd_BamHI, ICAM1_Fwd_BamHI, and LAIR1_Rev_X_NheI or ICAM1_Rev_X_NheI, where X - amplified fragment length). The purified PCR products were digested with corresponding restriction enzymes and ligated into NcoI/SalI-digested pAL2-T vector (Evrogen) with T4 DNA ligase (Thermo Fisher Scientific). To prepare spike-in samples, IgG transcript copy number was assessed by qPCR, plasmids were diluted to either 200 or 20 fg/μL each and spiked into cDNA of bulk B cells at 1:1,000 and 1:10,000 ratios. The mixtures were used for suppression PCR.

**Selection of In-Frame Inserts and Cloning into Recombinant Antibodies.** Suppression PCR does not allow us to obtain the full-length V segment sequence or the sequence of the light chain. To express the insertion-containing antibodies in the most favorable VDJ context, we aligned the available VDJ sequence to our in-house antibody library and selected the backbone for each insertion with the highest VDJ homology. The WBP1L insertion was cloned into four different backbones to assess the backbone influence, resulting in 15 total VDJ insertion-containing antibodies (Dataset S5). The insertions were synthesized either by overlap-extension PCR or amplified from the human genomic PCR and introduced in the backbone by Gibson assembly (primer sequences in Dataset S6). Successful cloning was confirmed by Sanger sequencing.

**Expression and Purification of Grafted Insert Antibodies.** Heavy chain constructs containing the selected insertions were mixed with the corresponding light chain plasmids to the final amount of 1.2 μg/mL of production. The plasmids were mixed with polyethylenimine (PEI, #23966-1, Polysciences Europe; 4 μg/mL of production) in OptiMEM (#31985047, Life Technologies; 100 μL/mL of production). The DNA-PEI mixture was incubated at room temperature for 20 min and added dropwise to the Expi293F culture (Thermo Scientific) seeded at 2 million per milliliter in Expi293 expression medium (#A1435102, Thermo Scientific) on the previous day. The supernatant was harvested, filtered through a 0.45-μm membrane, and analyzed after 72 h posttransfection. Antibodies were purified through protein G pull-down using a Ab SpinTrap kit (#28408347, Cytiva). Protein concentration in the eluates was calculated through 280-nm absorbance measured by DS-11 spectrophotometer (DeNovix).

**Insert Antibodies Concentration and Binding Analysis.** The concentration of the antibodies in the Expi293F supernatant was determined by the ELISA. Half-area high binding polystyrene plates (#7626991, Greiner) were coated with 25 μL of 10 ng/μL PBS dilution of goat anti-human IgG (#2040-01, Southern Biotechnologies) overnight at +4 °C. The next day, plates were blocked with 1% BSA in PBS (PBS-BSA) for 1 h at room temperature, washed three times with 0.1% Tween in PBS (PBST), and incubated with the serial dilutions of the Expi293F supernatant (72 h posttransfection) in PBS-BSA for 1 h at room temperature. Human IgG preparation (#0150-01, Southern Biotechnologies) was used as a standard. Washed plates were incubated with 1:500 PBS-BSA dilution of goat anti-human IgG conjugated with alkaline phosphatase (#2040-04, Southern Biotechnologies) for 1 h at room temperature, and the washing procedure was repeated. Plates were incubated with 250 μg/mL solution of 4-nitrophenyl phosphate (#S0942, Sigma Aldrich) for 30 min and analyzed for 405 nm absorbance with Cytation 5 device (Biotek).

For all antibodies expressed and purified at a detectable level, we determine the binding affinity toward the backbone target (SARS-CoV-2 RBD for 2M-10B11, S309, and S2H13; gp140 for F105; Influenza HA for FI and FY), the irrelevant viral antigen (gp140 for 2M-10B11, S309, and S2H13; SARS-CoV-2 RBD for F105, FI, and FY), and BSA via ELISA. The procedure is performed according to the protocol described above with the following changes: 1) the plates were coated with the corresponding antigens listed above and 2) protein G-purified antibodies were used instead of Expi293F supernatant. Corresponding backbone antibodies were used as a standard. The optical density (OD) measurements were analyzed by in-house R scripts, effective dilution/concentration 50 ($ED_{50}$/$EC_{50}$) was calculated using the four-parametric curve fitting (Dataset S5).

**Cell Lines.** EBV-immortalized cell lines derived from African donors were produced in an earlier study (4).

**Suppression PCR Bioinformatics Analysis.** For suppression of PCR bioinformatics analysis, 300-nt paired-end (PE) reads were trimmed to remove adapters and poor-quality base calls using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/, v0.4.3, parameters –nextera –paired -q 20 –length 100). Quality score per base position was assessed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Trimmed reads were aligned in paired-end mode to the GRCh38.p12 human genome assembly using Burrows-Wheeler alignment tool (v0.7.17, parameters: bwa mem) (67). The *IGH* locus (V, D, J, and constant genes) was defined on GRCh38, with the following coordinates: chr14 105586437–106879844 (heavy). Light chain and *TCR* loci were mapped to: chr2 88848000–90362000 (κ) and chr22 22026076–22922913 (λ), chr14 21621904–22552132 (TCR-α), chr14 142299011–142813287 (TCR-β). Corresponding target locus (IGH/IGK/IGL/TCRA/TCRB) was excluded from the insertion mapping to distinguish between the genomic insertion and a cryptic exon splicing.

To find the insertion within the VDJ rearranged sequences, we selected genomic ranges when sequence coverage was above 10 reads, with a sequence length between 20 and 2,000 bp. Genome coverage was calculated using BEDTools (https://bedtools.readthedocs.io/en/latest/index.html, v2.27.1) (68) and a dedicated python script using pysam (https://github.com/pysam-developers/pysam) was written to identify potential inserts. Potential inserts were assigned if they fulfill the following criteria: 1) one mapping read is chimeric with the IGH region; 2) it has no secondary alignment; 3) its mapping quality is equal or higher than 5; and 4) its mate is mapped. The list of potential inserts was annotated using GENCODE v291 (69) and BEDOPS tools (70). mtDNA origin of the insertions was confirmed by comparing the E-values of the alignment against mitochondrial and nuclear genomes. Original reads were retrieved, pooled, and used as input for de novo assembly using Trinity software (71). BLAST was used to define the insert coordinates within the contig (command-line version, v2.7.1+). Contig sequence bordering the insert was annotated by IgBLAST (72). Finally, we validated contig sequences that: 1) contain an insert embedded by V-J genes and CH1 gene or by V gene and J-CH1 genes; and 2) contain the last 15 bp of one of the V primers (two mismatches allowed) + 5 bp belonging to the V gene (1 mismatch allowed). To identify exon–exon junctions, we used JAGuaR to generate a junction database of the human transcriptome (100-bp junction length) and then blast the contig sequences against this database (parameters: -task megablast -perc_identity 95). Inserts covering more than 70% of the junction length (70 bp) were tagged as having an exon–exon junction.

**Insert Frequencies and Feature Analysis.** Frequencies of the insertions were determined for each donor as well as distinct B cell populations by dividing the number of detected unique inserts by the number of analyzed B cells for a given sample. Insertions were mapped to nuclear and mtDNA using in-house R scripts (73) (https://github.com/lebmih/LAIR). Details on in silico controls generation and calculation of the distance and overlap with previously reported genomic regions such as R-loops, ERFS, CFS, LINEs, SINEs, AID, and RAG off-target

sites, are provided in *SI Appendix*. CDR3 sequence alignment was performed with the MUSCLE algorithm using UGENE software (74).

**Alternative Splicing.** Illumina reads received from suppression PCR libraries of IGH, IGL, and IGK chain profiling were processed as follows: reads were paired by PEAR software (75), adapters were removed with Trimmomatic (76), and constant genes mapped by an in-house script in R (73). Variable domain was annotated by IgBLAST (77), the alternative exons quantified and extracted by an in-house script in R. Extracted sequences were mapped to the corresponding locus by BLAST+ (78) and visualized as the histogram in R. Cryptic splice sites were annotated by Human Splicing Finder web service (79).

**P/N Nucleotide Analysis.** To measure the P/N nucleotide length, we analyzed the transcript sequence surrounding the inserted fragment. To extract the junction sequence, we detected the conserved second framework cysteine residue through TT[AG][CATG]TGT[GA][CT] pattern search and the J-segment conserved tryptophane with CTGGGGC[CA][AG][ATG]GG[AGC]AC[AC][AC][CT] pattern search. The nucleotide sequence between the conserved Cys and Trp residues was characterized with IMGT junction analysis (https://imgt.org/IMGTindex/IMGTJunctionAnalysis.php).

**Statistical Analysis.** Sampling size was not predetermined and was limited by the availability of the donor material. Researchers were not blinded to the studied groups. Data were analyzed for normality with Shapiro–Wilk test. Colocalization and overlap data were not distributed normally, Wilcoxon rank-sum test was used with the null hypothesis stating the absence of a significant difference (80). Plots were created using ggplot2 and gridExtra packages in R 3.6.2 (73). In all boxplots, the black lines 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× interquartile range. Significance level showed according to the legend: ns, $P \geq 0.05$, *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, and ****$P < 0.0001$. Schemes and loci depictions are generated with Inkscape software (not-in-scale schemes) or with ggplot2 (in-scale maps).

**Data, Materials, and Software Availability.** Sequencing data have been deposited in the National Center for Biotechnology Sequence Read Archive (SRA) database accession PRJNA638005 (81). The code for the main data-processing pipeline is available at https://bitbucket.org/mathildefog/vdjinsertillumina/src/master/ (82). The code to analyze insert-containing antibodies (*SI Appendix*, Fig. S3) is also available at https://bitbucket.org/mathildefog/lair1vdjinsertillumina/src/master/ (83).

**ACKNOWLEDGMENTS.** We thank Tomasz Zemojtel and Dieter Beule of the Berlin Institute of Health Core Genomics Facility team for next-generation sequencing. The work was funded by the Deutsche Forschungsgemeinschaft Grant 394523286 (to K.d.l.R.), by the Helmhotz Association, and the Berlin Institute of Health at Charité (K.d.l.R.). Furthermore, this project has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme Grants 948464 (to K.d.l.R.) and 670955 (to A.L.). The work was further supported by the grant Ministry of Science and Higher Education 075-15-2020-807 (to D.M.C.) and by the Ministry of Education, Youth and Sports of the Czech Republic project CEITEC 2020 LQ1601 (to A.N.D).

Author affiliations: [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. D. Jung, C. Giallourakis, R. Mostoslavsky, F. W. Alt, Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. *Annu. Rev. Immunol.* **24**, 541–570 (2006).
2. V. H. Odegard, D. G. Schatz, Targeting of somatic hypermutation. *Nat. Rev. Immunol.* **6**, 573–583 (2006).
3. M. Muramatsu *et al.*, Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **102**, 553–563 (2000).
4. J. Tan *et al.*, A LAIR1 insertion generates broadly reactive antibodies against malaria variant antigens. *Nature* **529**, 105–109 (2016).
5. K. Pieper *et al.*, Public antibodies to malaria antigens generated by two LAIR1 insertion modalities. *Nature* **548**, 597–601 (2017).
6. Y. Chen *et al.*, Structural basis of malaria RIFIN binding by LILRB1-containing antibodies. *Nature* **592**, 639–643 (2021).
7. J. Baar, M. J. Shulman, The Ig heavy chain switch region is a hotspot for insertion of transfected DNA. *J. Immunol.* **155**, 1911–1920 (1995).
8. A. Gabrea, P. L. Bergsagel, M. Chesi, Y. Shou, W. M. Kuehl, Insertion of excised IgH switch sequences causes overexpression of cyclin D1 in a myeloma tumor cell. *Mol. Cell* **3**, 119–123 (1999).
9. A. Agrawal, Q. M. Eastman, D. G. Schatz, Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature* **394**, 744–751 (1998).
10. K. Hiom, M. Melek, M. Gellert, DNA transposition by the RAG1 and RAG2 proteins: A possible source of oncogenic translocations. *Cell* **94**, 463–470 (1998).
11. M. Melek, M. Gellert, RAG1/2-mediated resolution of transposition intermediates: Two pathways and possible consequences. *Cell* **101**, 625–633 (2000).
12. G. S. Lee, M. B. Neiditch, R. R. Sinden, D. B. Roth, Targeted transposition by the V(D)J recombinase. *Mol. Cell. Biol.* **22**, 2068–2077 (2002).
13. M. B. Neiditch, G. S. Lee, L. E. Huye, V. L. Brandt, D. B. Roth, The V(D)J recombinase efficiently cleaves and transposes signal joints. *Mol. Cell* **9**, 871–878 (2002).
14. C. L. Tsai, M. Chatterji, D. G. Schatz, DNA mismatches and GC-rich motifs target transposition by the RAG1/RAG2 transposase. *Nucleic Acids Res.* **31**, 6180–6190 (2003).
15. Y. V. R. Reddy, E. J. Perkins, D. A. Ramsden, Genomic instability due to V(D)J recombination-associated transposition. *Genes Dev.* **20**, 1575–1582 (2006).
16. J. W. Vaandrager, E. Schuuring, K. Philippo, P. M. Kluin, V(D)J recombinase-mediated transposition of the BCL2 gene to the IGH locus in follicular lymphoma. *Blood* **96**, 1947–1952 (2000).
17. T. Sonoki, E. Iwanaga, H. Mitsuya, N. Asou, Insertion of microRNA-125b-1, a human homologue of lin-4, into a rearranged immunoglobulin heavy chain gene locus in a patient with precursor B-cell acute lymphoblastic leukemia. *Leukemia* **19**, 2009–2010 (2005).
18. Y. Yu *et al.*, Dna2 nuclease deficiency results in large and complex DNA insertions at chromosomal breaks. *Nature* **564**, 287–290 (2018).
19. M. Onozawa, P. D. Aplan, Templated sequence insertion polymorphisms in the human genome. *Front Chem.* **4**, 43 (2016).
20. M. Onozawa *et al.*, Repair of DNA double-strand breaks by templated nucleotide sequence insertions derived from distant regions of the genome. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 7729–7734 (2014).
21. P. C. Rommel, T. Y. Oliveira, M. C. Nussenzweig, D. F. Robbiani, RAG1/2 induces genomic insertions by mobilizing DNA into RAG1/2-independent breaks. *J. Exp. Med.* **214**, 815–831 (2017).
22. F. W. Alt, Y. Zhang, F.-L. Meng, C. Guo, B. Schwer, Mechanisms of programmed DNA lesions and genomic instability in the immune system. *Cell* **152**, 417–429 (2013).
23. A. Nussenzweig, M. C. Nussenzweig, Origin of chromosomal translocations in lymphoid cancer. *Cell* **141**, 27–38 (2010).
24. A. Geser, G. Brubaker, C. C. Draper, Effect of a malaria suppression program on the incidence of African Burkitt's lymphoma. *Am. J. Epidemiol.* **129**, 740–752 (1989).
25. A. M. Moormann *et al.*, Exposure to holoendemic malaria results in elevated Epstein-Barr virus loads in children. *J. Infect. Dis.* **191**, 1233–1238 (2005).
26. D. F. Robbiani *et al.*, Plasmodium infection promotes genomic instability and AID-dependent B cell lymphoma. *Cell* **162**, 727–737 (2015).
27. L. A. Elenich, W. A. Dunnick, Sequence at insertion site of E.Tn retrotransposon into an immunoglobulin switch region suggests a role for switch recombinase. *Nucleic Acids Res.* **19**, 396 (1991).
28. K. A. Lukyanov, G. A. Launer, V. S. Tarabykin, A. G. Zaraisky, S. A. Lukyanov, Inverted terminal repeats permit the average length of amplified DNA fragments to be regulated during preparation of cDNA libraries by polymerase chain reaction. *Anal. Biochem.* **229**, 198–202 (1995).
29. G. Monaco *et al.*, RNA-seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. *Cell Rep.* **26**, 1627–1640.e7 (2019).
30. A. Aguilera, The connection between transcription and genomic instability. *EMBO J.* **21**, 195–201 (2002).
31. R. Chiarle *et al.*, Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. *Cell* **147**, 107–119 (2011).
32. A. Aguilera, T. García-Muse, R loops: From transcription byproducts to threats to genome stability. *Mol. Cell* **46**, 115–124 (2012).
33. L. A. Sanz *et al.*, Prevalent, dynamic, and conserved R-loop structures associate with specific epigenomic signatures in mammals. *Mol. Cell* **63**, 167–178 (2016).
34. R. S. Hansen *et al.*, A variable domain of delayed replication in FRAXA fragile X chromosomes: X inactivation-like spread of late replication. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 4587–4592 (1997).
35. M. M. Le Beau *et al.*, Replication of a common fragile site, FRA3B, occurs late in S phase and is delayed further upon induction: Implications for the mechanism of fragile site induction. *Hum. Mol. Genet.* **7**, 755–761 (1998).
36. J. H. Barlow *et al.*, Identification of early replicating fragile sites that contribute to genome instability. *Cell* **152**, 620–632 (2013).
37. H. Zhang, C. H. Freudenreich, An AT-rich sequence in human common fragile site FRA16D causes fork stalling and chromosome breakage in S. cerevisiae. *Mol. Cell* **27**, 367–379 (2007).
38. L. W. Dillon, A. A. Burrow, Y.-H. Wang, DNA instability at chromosomal fragile sites in cancer. *Curr. Genomics* **11**, 326–337 (2010).
39. A. Fungtammasan, E. Walsh, F. Chiaromonte, K. A. Eckert, K. D. Makova, A genome-wide analysis of common fragile sites: What features determine chromosomal instability in the human genome? *Genome Res.* **22**, 993–1005 (2012).
40. L. M. Payer, K. H. Burns, Transposable elements in human genetic disease. *Nat. Rev. Genet.* **20**, 760–772 (2019).

41. 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).
42. F.-L. Meng *et al.*, Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. *Cell* **159**, 1538–1548 (2014).
43. I. A. Klein *et al.*, Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. *Cell* **147**, 95–106 (2011).
44. J. Qian *et al.*, B cell super-enhancers and regulatory clusters recruit AID tumorigenic activity. *Cell* **159**, 1524–1537 (2014).
45. 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).
46. 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).
47. Á. F. Álvarez-Prado *et al.*, A broad atlas of somatic hypermutation allows prediction of activation-induced deaminase targets. *J. Exp. Med.* **215**, 761–771 (2018).
48. M. Mijušković *et al.*, Off-target V(D)J recombination drives lymphomagenesis and is escalated by loss of the Rag2 C terminus. *Cell Rep.* **12**, 1842–1852 (2015).
49. S. D. Boyd, S. A. Joshi, High-throughput DNA sequencing analysis of antibody repertoires. *Microbiol. Spectr.* **2**, 10.1128/microbiolspec.AID-0017-2014 (2014).
50. H. Morbach, E. M. Eichhorn, J. G. Liese, H. J. Girschick, Reference values for B cell subpopulations from infancy to adulthood. *Clin. Exp. Immunol.* **162**, 271–279 (2010).
51. A. Helmrich, M. Ballarino, L. Tora, Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. *Mol. Cell* **44**, 966–977 (2011).
52. T. E. Wilson *et al.*, Large transcription units unify copy number variants and common fragile sites arising under replication stress. *Genome Res.* **25**, 189–200 (2015).
53. F. Coquel *et al.*, SAMHD1 acts at stalled replication forks to prevent interferon induction. *Nature* **557**, 57–61 (2018).
54. J. E. Willett-Brozick, S. A. Savul, L. E. Richey, B. E. Baysal, Germ line insertion of mtDNA at the breakpoint junction of a reciprocal constitutional translocation. *Hum. Genet.* **109**, 216–223 (2001).
55. K. K. Singh, A. R. Choudhury, H. K. Tiwari, Numtogenesis as a mechanism for development of cancer. *Semin. Cancer Biol.* **47**, 101–109 (2017).
56. M. Onozawa, L. Goldberg, P. D. Aplan, Landscape of insertion polymorphisms in the human genome. *Genome Biol. Evol.* **7**, 960–968 (2015).
57. T. Cantaert *et al.*, Activation-induced cytidine deaminase expression in human B cell precursors is essential for central B cell tolerance. *Immunity* **43**, 884–895 (2015).
58. J. E. Posey, M. J. Pytlos, R. R. Sinden, D. B. Roth, Target DNA structure plays a critical role in RAG transposition. *PLoS Biol.* **4**, e350 (2006).
59. J. Hu *et al.*, Detecting DNA double-stranded breaks in mammalian genomes by linear amplification-mediated high-throughput genome-wide translocation sequencing. *Nat. Protoc.* **11**, 853–871 (2016).
60. E. ten Boekel, F. Melchers, A. G. Rolink, Precursor B cells showing H chain allelic inclusion display allelic exclusion at the level of pre-B cell receptor surface expression. *Immunity* **8**, 199–207 (1998).
61. M. C. Nussenzweig *et al.*, Allelic exclusion in transgenic mice that express the membrane form of immunoglobulin μ. *Science* **236**, 816–819 (1987).
62. M. Reth, E. Petrac, P. Wiese, L. Lobel, F. W. Alt, Activation of V kappa gene rearrangement in pre-B cells follows the expression of membrane-bound immunoglobulin heavy chains. *EMBO J.* **6**, 3299–3305 (1987).
63. D. Kitamura, K. Rajewsky, Targeted disruption of mu chain membrane exon causes loss of heavy-chain allelic exclusion. *Nature* **356**, 154–156 (1992).
64. J. Manz, K. Denis, O. Witte, R. Brinster, U. Storb, Feedback inhibition of immunoglobulin gene rearrangement by membrane mu, but not by secreted mu heavy chains. *J. Exp. Med.* **168**, 1363–1381 (1988).
65. M. R. Gold, M. G. Reth, Antigen receptor function in the context of the nanoscale organization of the B cell membrane. *Annu. Rev. Immunol.* **37**, 97–123 (2019).
66. I. Lazar, I. Zwecker-Lazar, R. Lazar, Gel Analyzer 2010a: Freeware 1D gel electrophoresis image analysis software (2010). https://www.scienceopen.com/document?vid=416d30b4-53d9-4565-a37d-e9d8709afd93. Accessed 22 August 2022.
67. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
68. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
69. J. Harrow *et al.*, GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
70. S. Neph *et al.*, BEDOPS: High-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920 (2012).
71. B. J. Haas *et al.*, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
72. M.-P. Lefranc, IMGT, the international ImMunoGeneTics information system: A standardized approach for immunogenetics and immunoinformatics. *Immunome Res.* **1**, 3 (2005).
73. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2019)
74. K. Okonechnikov, O. Golosova, M. Fursov; UGENE team, Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics* **28**, 1166–1167 (2012).
75. J. Zhang, K. Kobert, T. Flouri, A. Stamatakis, PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).
76. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
77. J. Ye, N. Ma, T. L. Madden, J. M. Ostell, IgBLAST: An immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* **41**, W34–W40 (2013).
78. C. Camacho *et al.*, BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
79. F.-O. Desmet *et al.*, Human Splicing Finder: An online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* **37**, e67 (2009).
80. L. G. Cowell, M. Davila, T. B. Kepler, G. Kelsoe, Identification and utilization of arbitrary correlations in models of recombination signal sequences. *Genome Biol.* **3**, RESEARCH0072 (2002).
81. M. Lebedin *et al.*, Sequencing of non-VDJ insertions in human antibody transcripts. NCBI SRA. https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA638005. Accessed 22 August 2022.
82. M. Foglierini Perez, vdjinsertIllumina, Pipeline and scripts to process PCR suppression products to find insert into V(D)J transcripts. Bitbucket. https://bitbucket.org/mathildefog/vdjinsertillumina/src/master/. Deposited 16 June 2020.
83. M. Foglierini Perez, LAIR1vdjinsertIllumina, In-depth analysis of insert-containing antibody transcripts after running vdjinsertIllumina pipeline. Bitbucket. https://bitbucket.org/mathildefog/lair1vdjinsertillumina/src/master/. Deposited 16 June 2020.