Highly sensitive and unbiased approach for elucidating antibody repertoires

Sherry G. Lin[a,b,1], Zhaoqing Ba[a,b,1], Zhou Du[b,c,1], Yu Zhang[a,b], Jiazhi Hu[a,b,2], and Frederick W. Alt[a,b,c,2]

[a]Program in Cellular and Molecular Medicine, Boston Children's Hospital, Boston, MA 02115; [b]Department of Genetics, Harvard Medical School, Boston, MA 02115; and [c]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115

Contributed by Frederick W. Alt, May 31, 2016 (sent for review May 17, 2016; reviewed by Jayanta Chaudhuri and Cornelis Murre)

Developing B lymphocytes undergo V(D)J recombination to assemble germ-line V, D, and J gene segments into exons that encode the antigen-binding variable region of Ig heavy (H) and light (L) chains. IgH and IgL chains associate to form the B-cell receptor (BCR), which, upon antigen binding, activates B cells to secrete BCR as an antibody. Each of the huge number of clonally independent B cells expresses a unique set of IgH and IgL variable regions. The ability of V(D)J recombination to generate vast primary B-cell repertoires results from a combinatorial assortment of large numbers of different V, D, and J segments, coupled with diversification of the junctions between them to generate the complementary determining region 3 (CDR3) for antigen contact. Approaches to evaluate in depth the content of primary antibody repertoires and, ultimately, to study how they are further molded by secondary mutation and affinity maturation processes are of great importance to the B-cell development, vaccine, and antibody fields. We now describe an unbiased, sensitive, and readily accessible assay, referred to as high-throughput genome-wide translocation sequencing-adapted repertoire sequencing (HTGTS-Rep-seq), to quantify antibody repertoires. HTGTS-Rep-seq quantitatively identifies the vast majority of IgH and IgL V(D)J exons, including their unique CDR3 sequences, from progenitor and mature mouse B lineage cells via the use of specific J primers. HTGTS-Rep-seq also accurately quantifies DJ_H intermediates and V(D)J exons in either productive or nonproductive configurations. HTGTS-Rep-seq should be useful for studies of human samples, including clonal B-cell expansions, and also for following antibody affinity maturation processes.

antibody repertoires | HTGTS-Rep-seq | V(D)J recombination

The B-lymphocyte antigen receptor (BCR) comprises identical Ig heavy (IgH) and Ig light (IgL) chains. Antibodies are the secreted form of the BCR. The V(D)J recombination process assembles germ-line V, D, and J gene segments into exons that encode the antigen-binding variable region exons of the BCR. The RAG 1 and 2 endonuclease (RAG) initiates V(D)J recombination by generating DNA double-stranded breaks (DSBs) between V, D, and J gene segments and their flanking recombination signal sequences (RSSs) (1). In this process, the V, D, and J coding ends are generated as covalent hairpins that must be opened and that are often further processed, before being joined by classical nonhomologous end joining (2). Processing of V, D, J coding ends can involve generation of deletions or insertions of nucleotides at the junction regions (2), including the frequent de novo addition of nucleotides by the terminal deoxynucleotidyl transferase component of the V(D)J recombination process (3). Notably the V(D)J junctional region encodes a major antigen contact region of the antibody variable region, known as complementarity determining region 3 (CDR3), and thus these junctional diversification processes make a huge contribution to antibody diversity.

The mouse *IgH* locus spans 2.7 megabases (Mb). There are 100s of V_Hs in the several megabase distal portion of the *IgH*, with the number varying substantially in certain mouse strains (4). The V_Hs lie ∼100 kb upstream from a 50-kb region containing 13 D_Hs, which is followed several kilobases downstream by a 2-kb region containing four J_Hs. The IgH constant region (C_H) exons lie downstream of the J_Hs. After assembly of a V_HDJ_H exon, transcription initiates upstream of the V_H and terminates downstream of the C_H exons, with V(D)J and C_H portions being fused into the ultimate *IgH* messenger RNA (mRNA) via splicing of the primary transcript. Due to the random junctional diversification mechanisms, only about 1/3 of assembled *IgH* V(D)J exons are able to generate in-frame splicing events that place the V(D)J and C_H exons in the same reading frame to generate productive (in-frame with functional V_H) rearrangements that encode an IgH polypeptide, with the remainder being nonproductive (out-of-frame, in-frame with a stop codon, or using a pseudo-V_H) (5). IgL chain variable region exons are assembled from just V and J segments but otherwise follow similar basic principles to those of IgH. The mouse *Igκ* light chain locus spans 3.2 Mb with 100s of Vκs in a 3.1-Mb region separated by 20 kb from five Jκs downstream whereas the *Igλ* light chain locus is smaller and less complex (6). RNA splicing again joins assembled VJ_L exons to corresponding C_L exons.

During B-cell development, V(D)J recombination is regulated to ensure specific repertoires and prevent undesired rearrangements. *IgH* V(D)J recombination occurs stage-specifically in progenitor B (pro-B) cells before that of *IgL* loci, which occur in precursor B (pre-B) cells. *IgH* V(D)J recombination is ordered, with D-to-J_H joining occurring, usually on both alleles, before appendage of a V_H to a DJ_H complex (Fig. S1*A*) (2). In addition, the V_H-to-DJ_H step of *IgH* V(D)J recombination is feedback-regulated

Significance

Antibodies are generated by B cells of the adaptive immune system to eliminate various pathogens. A somatic gene rearrangement process, termed V(D)J recombination, assembles antibody gene segments to form sequences encoding the antigen-binding regions of antibodies. Each of the multitude of newly generated B cells produces a different antibody with a unique antigen-binding sequence, which collectively form the primary antibody repertoire of an individual. Given the utility of specific antibodies for treating various human diseases, approaches to elucidate primary antibody repertoires are of great importance. Here, we describe a new method for high-coverage analysis of antibody repertoires termed high-throughput genome-wide translocation sequencing-adapted repertoire sequencing (HTGTS-Rep-seq). We discuss the potential merits of this approach, which is both unbiased and highly sensitive.

Author contributions: S.G.L., Z.B., J.H., and F.W.A. designed research; S.G.L., Z.B., and J.H. performed research; Z.D. and Y.Z. contributed new reagents/analytic tools; S.G.L., Z.B., Z.D., J.H., and F.W.A. analyzed data; and S.G.L., Z.B., J.H., and F.W.A. wrote the paper.

Reviewers: J.C., Memorial Sloan Kettering Cancer Center; and C.M., University of California, San Diego.

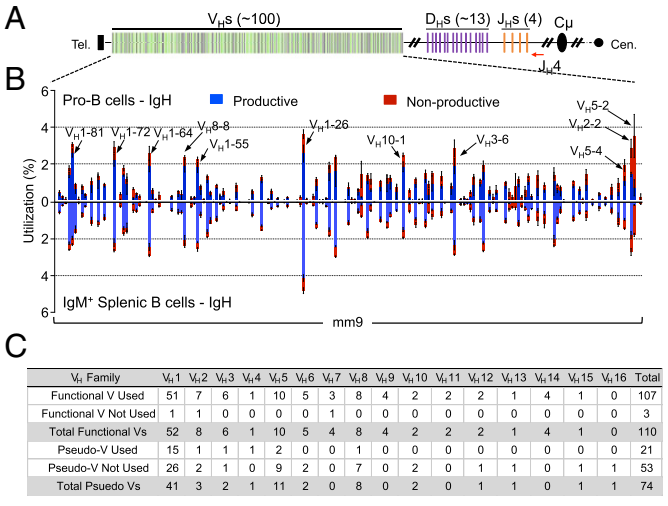
The authors declare no conflict of interest.

Data deposition: The sequencing and processed data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE82126).

[1]S.G.L., Z.B., and Z.D. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: alt@enders.tch.harvard.edu or jiazhi.hu@childrens.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1608649113/-/DCSupplemental.



C

V_H Family	V_H1	V_H2	V_H3	V_H4	V_H5	V_H6	V_H7	V_H8	V_H9	V_H10	V_H11	V_H12	V_H13	V_H14	V_H15	V_H16	Total
Functional V Used	51	7	6	1	10	5	3	8	4	2	2	2	1	4	1	0	107
Functional V Not Used	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	3
Total Functional Vs	52	8	6	1	10	5	4	8	4	2	2	2	1	4	1	0	110
Pseudo-V Used	15	1	1	1	2	0	0	1	0	0	0	0	0	0	0	0	21
Pseudo-V Not Used	26	2	1	0	9	2	0	7	0	2	0	1	1	0	1	1	53
Total Psuedo Vs	41	3	2	1	11	2	0	8	0	2	0	1	1	0	1	1	74

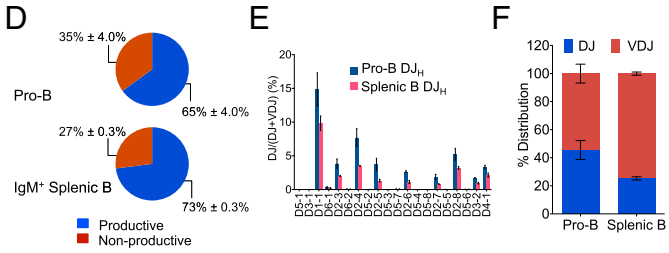


Fig. 1. HTGTS-Rep-seq of V_HDJ_H and DJ_H repertoire in pro-B cells and splenic B cells of C57BL/6 mice. (*A*) Schematic of the murine *IgH* locus showing V_Hs (green, functional; black, pseudo), D_Hs (purple), J_Hs (orange), and C_H region (black). The red arrow indicates the J_H4 coding end bait primer. (*B*) V_H repertoire with productive and nonproductive information from V_HDJ_H joins in pro-B cells (*Upper*) and IgM^+ splenic B cells (*Lower*). Some of the most frequently used V_Hs are highlighted with arrows as indicated. (*C*) Utilization numbers of functional V_Hs and pseudo V_Hs across 16 families in HTGTS-Rep-seq libraries described in *B*. (*D*) Pie chart showing the average overall percentage of productive and nonproductive V_HDJ_H joins from libraries described in *B*. (*E*) D use in V_HDJ_H and DJ_H joins in pro-B cells and IgM^+ splenic B cells as indicated. (*F*) DJ_H:V_HDJ_H ratios in pro-B cells and IgM^+ splenic B cells as indicated. All of the data are showed by mean ± SEM, $n = 3$.

with a productive rearrangement leading to cessation of V(D)J recombination on the other allele if it is still in the DJ_H configuration (2). In contrast, initial nonproductive *IgH* V(D)J rearrangements do not prevent V_H-to-DJ_H rearrangements from occurring on the other allele. Such feedback regulation generally leads to the typical 40/60 ratio of mature B cells, with two *IgH* V(D)J rearrangements (one productive) versus one *IgH* V(D)J plus a DJ_H rearrangement (7). V_H-to-DJ_H rearrangement is also regulated to generate diverse utilization of the 100s of upstream V_Hs. Although proximal V_Hs, notably the most proximal V_H (V_H81X), are somewhat overused in pro-B V(D)J rearrangements, the sequestering of the D_Hs and J_Hs in a separate chromosomal domain from that of the V_Hs (8, 9), coupled with the phenomenon of locus contraction (10, 11), allows even the most distal V_Hs to be used. Subsequently, the somewhat biased primary V_H repertoire in pro-B cells is subjected to cellular selection mechanisms to generate a more normalized primary repertoire in newly generated B cells (12).

Each B cell expresses a unique BCR, and each individual mouse or human has the capacity to generate up to 10^{13} or more distinct BCRs in the primary repertoire (13), with a large fraction of these being generated by junctional diversification of IgH and IgL CDR3s (14). In this regard, the ability to quantitatively identify the *IgH* and *IgL* variable region exons that contribute to the primary antibody repertoire is of great interest in elucidating contributions of this repertoire to immune responses and to immune diseases (15). Several important repertoire sequencing assays that use next-generation sequencing have been developed. These approaches involve the generation of repertoire libraries from either genomic DNA or mRNA (15). Most prior DNA-based approaches rely on use of upstream degenerate V primers, each designed to identify members of particular V_H families, and a downstream degenerate J primer, an approach that covers many, but not necessarily all, V(D)J exons and likely not all equally. RNA-based approaches generally require only one downstream primer (from the J or constant region) and thus obviate biases in prior DNA-based assays, but these approaches can severely underestimate nonproductive rearrangements due to decreased transcript levels (15). In addition, the long length of the 5′ RACE-derived complementary DNAs can also pose a challenge because sequencing technologies cannot always cover the entire length of the V(D)J exons.

We developed linear amplification-mediated high-throughput genome-wide translocation sequencing (LAM-HTGTS) to identify unknown "prey" sequences that join to fixed DSB-associated "bait" sequences (16). LAM-HTGTS, like its predecessor HTGTS (17), employs a single primer for a DSB-associated bait sequence to perform linear amplification across bait–prey junctions to identify all prey sequences joined to the bait DSBs in an unbiased manner (16, 18). We have used various types of DSBs as bait for LAM-HTGTS, including those generated by engineered nucleases and endogenous DSBs (17–22). Because V(D)J recombination generates rearrangements with junctions at borders of V, D, and J segments, we can use primers for any of these gene segments as LAM-HTGTS bait to identify sites of RAG-generated DSBs, both in progenitor or precursor lymphocytes undergoing V(D)J recombination, as well as in mature lymphocytes to retrospectively identify V(D)J recombination events that occurred earlier in development. Notably, LAM-HTGTS using endogenous RAG-generated DSBs identified RAG-generated DJ_H joins, RSS joins in excision circles, and off-target junctions in developing B-lineage cells that were not detected by prior assays (22), illustrating the high sensitivity of the assay. Based on these earlier studies, we now describe an adaptation of LAM-HTGTS as a robust repertoire-sequencing assay that we term "HTGTS-adapted repertoire sequencing" (HTGTS-Rep-seq).

Results

Overview of LAM-HTGTS Adapted Repertoire Sequencing. For HTGTS-Rep-seq libraries, we used bait coding ends of J segments to identify, in unbiased fashion, mouse *IgH* DJ_H repertoires, along with both productive and nonproductive *IgH* V(D)J repertoires from both pro-B and peripheral B cells. Similarly, we also identified mouse productive and nonproductive *Igκ* repertoires from peripheral B cells. For all samples analyzed, genomic DNA isolated from a pool of the given type of B cells was sonicated to generate fragments with an average size of ∼1 kb and that thus would be expected to harbor *IgH* V(D)J or DJ rearrangements, *Igκ* VJ rearrangements, or unrearranged J_Hs or Jκs (Fig. S1*B*). Biotinylated primers that anneal to sequences downstream of the coding end of a particular J_H or Jκ segment will allow linear amplification of any fragments containing the bait J segment(s). Subsequent streptavidin purification, adapter ligation, and library construction steps were carried out as previously described (16) (Fig. S1*B*). To generate longer sequencing reads for more accurate alignment of Vs and Ds, we positioned bait primers closer to the coding ends of bait Js and used MiSeq 2 × 300-bp paired-end sequencing to capture full-length V(D)J sequences in recovered junctions. For bioinformatic analysis, we combined our LAM-HTGTS pipeline with IgBLAST (23) to generate an analysis pipeline that provides comprehensive information on productive or nonproductive junctions and CDR3 sequences (see *Materials and Methods* for details).

For the HTGTS-Rep-seq, we generally kept for analysis all recovered junctions, including all duplicates for reasons described previously (22). To control for experimental variations, we generated three technical repeat HTGTS-Rep-seq libraries from the same

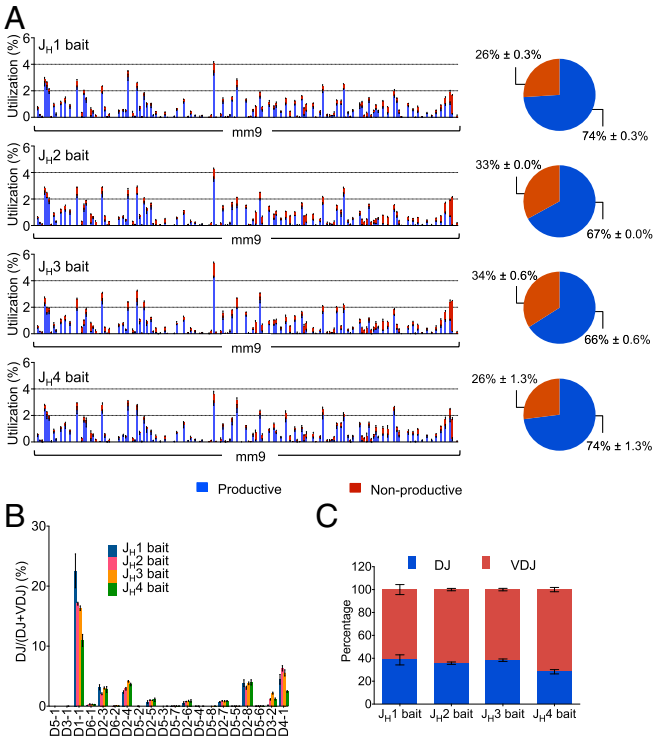


Fig. 2. V_HDJ_H and DJ_H repertoires in IgM^+ splenic B cells across four J_H baits. (*A*) V_H repertoire with productive and nonproductive information from V_HDJ_H joins (*Left*) and pie charts showing the average overall percentage of productive and nonproductive V_HDJ_H joins (*Right*) in IgM^+ splenic B cells using each of the J_H coding end bait primers as indicated. (*B*) Comparison of D use in DJ_H joins in IgM^+ splenic B cells using each of the J_H coding end bait primers. (*C*) Comparison of DJ_H:V_HDJ_H ratios in IgM^+ splenic B cells using each of the J_H coding end bait primers. Mean ± SEM, $n = 3$ for all of the data. Other analysis details are as described for Fig. 1.

splenic B-cell DNA samples, which yielded highly reproducible repertoires with correlation coefficient (r) values of 0.99 (Table S1). Even for biological repeat *IgH* or *IgL* HTGTS-Rep-seq libraries from pro-B or splenic B cells of three different mice, correlation analyses revealed highly reproducible repertoires with r values greater than 0.9 in most of the datasets (Tables S1 and S2). However, as described below, detailed analyses of certain aspects of such libraries, such as the fraction of unique CDR3s in the total repertoire, revealed expected biological variations (Table S1).

HTGTS-Rep-Seq Reveals *IgH* V_HDJ_H and DJ_H Repertoires in Developing and Mature B Cells. To test the ability of HTGTS-Rep-seq to detect differences between primary pro–B-cell *IgH* repertoires versus those of peripheral B lymphocytes, we purified primary $B220^+CD43^+IgM^-$ pro-B cells from the bone marrow and $B220^+IgM^+$ B cells from the spleen of wild-type (WT) C57BL/6 mice. We first used 2 μg of genomic DNA isolated from these cell populations to perform HTGTS-Rep-seq with a J_H4 coding end bait primer to capture V_HDJ_{H4} and DJ_{H4} rearrangements (Fig. 1*A* and Table S1). Libraries from both cell types showed broad use of V_Hs in V_HDJ_{H4} rearrangements throughout the *IgH* variable region locus, with some V_Hs used more frequently (e.g., V_H5-2, V_H2-2, V_H3-6, V_H1-26, V_H1-64, V_H1-72, and V_H1-81) (Fig. 1*B*). The C57BL/6 *IgH* locus has ∼110 potentially functional V_Hs and 74 pseudo V_Hs categorized into 16 families (24). In the *IgH* repertoire libraries generated with a J_H4 coding end bait, we detected in V_HDJ_H exons 107 functional V_Hs from all 16 families, as well as 21 pseudo V_Hs with relatively conserved RSSs (Fig. 1*C*). Notably, the three "functional" V_Hs (V_H1-62-1, V_H2-6-8, and V_H7-2) not detected by HTGTS-Rep-seq also were not found by another high-throughput repertoire sequencing method (25), suggesting that they may actually be nonfunctional with respect to the ability to undergo V(D)J recombination.

V_H-to-DJ_H rearrangements occur at the pro-B stage, with only one in three expected to be in-frame (5). In the V_HDJ_{H4} exons we identified by HTGTS-Rep-seq, on average 65% were productive, and, correspondingly, 35% were nonproductive (Fig. 1*D*). This ratio likely reflects a dynamic differentiation process in which pro-B cells with two nonproductive rearrangements are negatively selected and those with a productive rearrangement on one allele are positively selected (12). Due in large part to feedback mechanisms from productive $V(D)J_H$ rearrangements during pro–B-cell development, ∼40% of splenic B cells displayed V_HDJ_H rearrangements on both alleles (one productive and one nonproductive) and the remaining 60% had one productive V_HDJ_H and one DJ_H rearrangement (5). Thus, a population of splenic B cells theoretically would be expected to have about 71% productive V_HDJ_H exons and 29% nonproductive V_HDJ_H exons. Indeed, we observed a very similar ratio of productive/nonproductive V_HDJ_{H4} exons (73:27) in the HTGTS-Rep-seq libraries from splenic B-cell DNA (Fig. 1*D*). In the DJ_H joins revealed by HTGTS-Rep-seq, D_H1-1 (also known as DFL16.1) was used most frequently in libraries from both pro-B and splenic mature B cells (Fig. 1*E*). Moreover, we observed a much higher percentage of DJ_H exons in pro-B cells compared with that of splenic B cells (45% vs. 25%) (Fig. 1 *E* and *F*), in line with D-to-J_H rearrangement on both alleles preceding V_H-to-DJ_H rearrangement in developing pro-B cells (5, 26, 27).

Biased Proximal V_H Use in 129SVE Mice Revealed by HTGTS-Rep-Seq. The 129SVE mouse strain *IgH* locus contains more V_Hs than the C57BL/6 *IgH* locus with a somewhat different organization (24). Given that 129SVE mice and cell lines have frequently been used in V(D)J recombination studies, we used the same J_H4 bait primers to also generate HTGTS-Rep-seq libraries from 129SVE bone marrow pro-B cells and splenic B cells (Table S2). The 129SVE *IgH* locus V_H sequences are annotated up to ∼1 Mb into the variable V_H region, but V_H sequences lying within the relatively large more distal region of the locus are not completely annotated. Thus, to generate an approximate 129SVE V_HDJ_H repertoire, we ran IgBLAST analyses against a combination of all of the known 129SVE V_H sequences and the annotated distal V_H sequences from the C57BL/6 background starting from V_H8-2 (Fig. S2 *A* and *B*). As with the C57BL/6 libraries, the V_Hs were widely used, and we detected 128 functional V_Hs out of 133 distinct members of the 15 V_H families, plus 34 pseudo V_Hs (Fig. S2*C*).

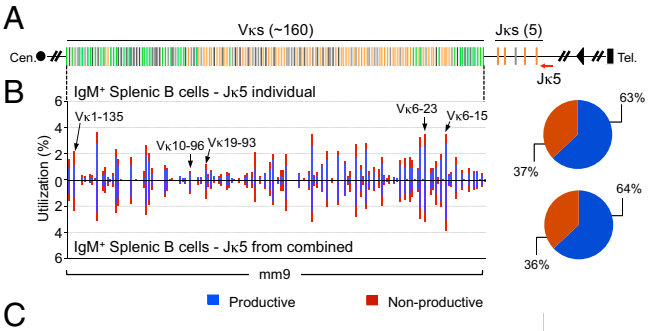
In contrast to the *IgH* V_HDJ_{H4} repertoire in C57BL/6 mice, we found a highly biased use of proximal V_Hs, especially V_H5-2 (also known as V_H81X) and V_H2-2, in 129SVE mice (Fig. 1*B* and Fig. S2*B*). The D-proximal V_H5-2 was used in 9.5% (1.7% productive; 7.7% nonproductive) of all V_HDJ_{H4} exons in pro-B cells and about 4% (0.3% productive; 3.5% nonproductive) of all V_HDJ_{H4} exons in splenic B cells of 129SVE mice (Fig. S2*B*). In contrast, V_H5-2 appeared in only about 3.5% (0.7% productive; 2.8% nonproductive) and about 1.8% (0.15% productive; 1.6% nonproductive) of the V_HDJ_{H4} exons in C57BL/6 pro-B and splenic B cells, respectively (Fig. 1*B*). The majority of V_H5-2–containing V_HDJ_{H4} joins in splenic B cells were nonproductive in both mouse strains, in contrast to other highly used V_Hs throughout both alleles (V_H2-2, V_H5-4, V_H3-6, V_H1-26, V_H1-55, V_H8-8, V_H1-64, V_H1-72, and V_H1-81), consistent with previous reports that most V_H5-2–containing productive rearrangements are selected against due to their autoreactive properties or inability to properly pair with IgL or surrogate IgL chains (28–30). Because the V_H5-2 gene body, associated RSS, and downstream region are conserved in C57BL/6 versus 129SVE mouse strains, the basis for greatly increased V_H5-2 utilization in primary repertoires of the 129SVE strain remains to be determined.

A comparison of V_HDJ_H and DJ_H rearrangements in 129SVE pro–B-cell libraries also revealed a relatively lower ratio of productive/nonproductive V_HDJ_H exons (39:61 in 129SVE vs. 65:35 in C57BL/6), as well as a lower ratio of V_HDJ_H/DJ_H rearrangements (about 45:55 in 129SVE vs. about 55:45 in C57BL/6) (Fig. 1 *D–F* and Fig. S2 *D–F*). V_H5-2 rearrangements did not substantially contribute to these differences. Both pro–B-cell libraries were generated in 4-week-old mice, suggesting that the lower relative proportion of productive V_HDJ_H exons in 129SVE compared with C57BL/6 pro-B cells might be attributed to differential timing of B-cell checkpoint selection in these two mouse strains. For both mouse strains, the splenic B-cell libraries showed comparable productive/nonproductive and VDJ/DJ ratios (Figs. 1 *D–F* and Fig. S2 *D–F*).

IgM$^+$ Splenic B-Cell V_HDJ_H Exons Display Similar V_H Use Profiles Across Different J_Hs. We also designed bait primers to the other three J_Hs in the *IgH* locus and made libraries from splenic B cells of both C57BL/6 and 129SVE mice to compare V_H and D utilization among the different J_Hs. These assays revealed similar V_H and D utilization repertoires for the four different J_Hs, indicating that selection for a particular V_H or D in a V_HDJ_H join did not vary substantially between the J_Hs in both C57BL/6 and 129SVE mice (Fig. 2*A* and Fig. S3*A*). However, we did find higher proportions of nonproductive V_HDJ_H rearrangements using the J_H2 and J_H3 baits, compared with the J_H1 and J_H4 bait libraries (Fig. 2*A* and Fig. S3*A*). In this regard, the stretch of sequence from the J_H coding ends to the highly conserved WGXG-motif that is crucial for a stable antibody structure (24) is shorter in the J_H2 and J_H3 segments relative to the J_H1 and J_H4 segments (Fig. S4*A*). Thus, some V_HDJ_{H2} and V_HDJ_{H3} join sites could lie too close to the WGXG-encoded sequences and be selected against due to unstable antibody structure (Fig. S4*B*). Moreover, we observed moderate differences in the D_H use profiles among the four J_Hs and a larger ratio of V_HDJ_H:DJ_H joins for the J_H4 bait libraries, which potentially could reflect the relative positions of these J_Hs in the recombination center that initiates V(D)J recombination (31) (Fig. 2 *B* and *C* and Fig. S3 *B* and *C*). Finally, we prepared HTGTS-Rep-seq libraries from 129SVE splenic B cells with four sets of J_H HTGTS-Rep-seq primers combined (Fig. S5*A* and Table S2). This approach, which allowed us to detect all $V_HDJ_{H1\text{-}4}$ exons in one HTGTS-Rep-seq library, revealed general V(D)J repertoires similar to those detected with individual J_H primers (Fig. S5 vs. Fig. S3).

HTGTS-Rep-Seq Detects Diverse *Igκ* VJ Rearrangements. In mice, the *Igκ* locus generates the majority of IgL-expressing B cells (32). The Vκ locus organization is distinct from that of the V_H locus. Besides not having D segments and, therefore, undergoing direct Vκ-to-Jκ rearrangements, the Vκ locus contains V segments organized in both direct and inverted orientation relative to the Jκ segments (6) (Fig. 3*A*). Thus, for some Vκs, joining to Jκ occurs deletionally like V_H-to-DJ_H joining, but, for others, it occurs via inversion of the intervening sequence. Direct and inverted Vκs generally occur in distinct clusters but also can be individually interspersed (Fig. 3*A*). To first assess the *Igκ* repertoire, we performed HTGTS-Rep-seq on 1 μg of genomic DNA from C57BL/6 splenic B cells using a Jκ5 coding end bait primer. Similar to the *IgH* locus, we also observed widespread use of Vκs across the entire locus to the Jκs (Fig. 3 *A* and *B*). All of the 100 functional Vκs across 20 Vκ families were detected by HTGTS-Rep-seq, and 11 out of 62 pseudo Vκs were also detected (Fig. 3*C*). We saw productive/nonproductive VJκ joins at a 63:37 ratio in splenic B cells (Fig. 3*B*), which is slightly lower than the predicted 67:33 ratio (33). This small deviation might reflect the presence of nonproductive VJκ joins in Igλ-positive cells (32).

We also generated HTGTS-Rep-seq libraries from splenic B-cell DNAs to capture VJκ joins from the three other functional Jκ segments separately or in a combination of all four Jκ primers. In contrast to *IgH* repertoires with different J_H primers, the *Igκ* repertoires showed apparently different utilization of some Vκs (e.g., Vκ6-15, Vκ6-23, Vκ19-93, Vκ10-96, and Vκ1-135) between different Jκ baits. Moreover, the productive/nonproductive ratios from the other Jκ primer libraries were slightly lower than those observed with the Jκ5 primer (Jκ1, 53:47; Jκ2, 60:40; Jκ4, 53:47; vs. Jκ5, 63:37) (Fig. S6). These differences in utilization and ratios likely reflect the occurrence of sequential VJκ recombination events (34). In this context, alleles containing nonproductive VJκ joins with the three Jκs upstream of Jκ5 have the ability for an unrearranged Vκ upstream of the nonproductive VJκ to join to a remaining Jκ (34). If this secondary rearrangement is inversional, the nonproductive VJκ joins would be retained in the genome and add to the nonproductive fraction of VJκ1, VJκ2, or VJκ4 joins that are detected by HTGTS-Rep-seq. Given this scenario, VJκ5 rearrangements, which are terminal rearrangement events, would be expected to reflect the theoretical productive/nonproductive ratios, as we have found.



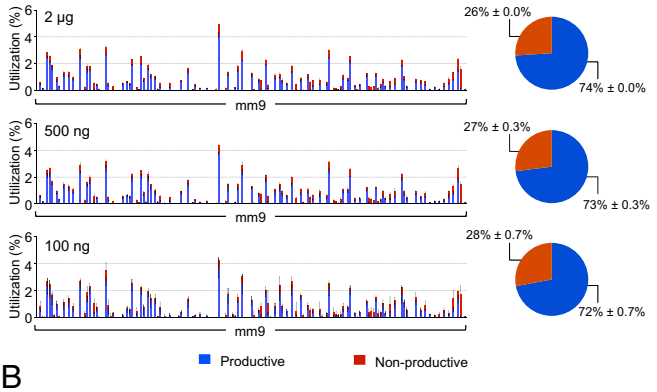
C

Vk family	Vk1	Vk2	Vk3	Vk4	Vk5	Vk6	Vk7	Vk8	Vk9	Vk10	Vk11	Vk12	Vk13	Vk14	Vk15	Vk16	Vk17	Vk18	Vk19	Vk20	Total
Functional V Used	9	3	9	26	5	9	1	11	4	3	1	6	2	4	1	1	2	1	1	1	100
Functional V Not Used	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total Functional Vs	9	3	9	26	5	9	1	11	4	3	1	6	2	4	1	1	2	1	1	1	100
Pseudo-V Used	1	0	1	5	0	0	0	2	0	0	1	1	0	0	0	0	0	0	0	0	11
Pseudo-V Not Used	3	7	3	2	1	0	0	2	0	0	3	5	16	4	4	0	1	0	0	0	51
Total Pseudo-Vs	4	7	4	7	1	0	0	2	2	0	3	6	17	4	4	0	1	0	0	0	62

Fig. 3. HTGTS-Rep-seq of VJκ repertoire in IgM$^+$ splenic B cells of C57BL/6 mice using Jκ5 bait primer. (*A*) Schematic of the murine *Igκ* locus showing Vκs and Jκs. Green and orange bars indicate functional Vκs with convergent and tandem transcriptional orientations, respectively, to the downstream Jκs. Black bars indicate pseudo Vκs. The red arrow indicates the Jκ5 coding end bait primer. (*B*, *Left*) Vκ repertoire with productive and nonproductive information from VJκ joins in IgM$^+$ splenic B cells with Jκ5 bait primer either individually (*Upper*) or from combined Jκ bait primers (*Lower*). Some differentially used Vκs among four different Jκs are highlighted with arrows as indicated (see also Fig. S6). (*Right*) Pie chart showing the overall percentage of productive and nonproductive VJκ joins. Representative results from two repeats are shown. (*C*) Utilization numbers of functional and pseudo Vκs across 20 families in libraries described in *B*.

HTGTS-Rep-Seq Revealed Characteristic CDR3 Properties. We analyzed the CDR3 sequences from productive V_HDJ_H and VJκ rearrangements in pro-B and splenic B cells. The CDR3 of productive V_HDJ_H exons in pro-B and splenic B cells showed a diverse range of lengths from 3 to 24 amino acids (aa) with a peak at 11–15 aa (Fig. S7 *A* and *B*). The consensus CDR3 motifs of these V_HDJ_H exons, made from the unique subset, from unimmunized pro-B and splenic B cells, shared the same V_H contributed and J_H4 contributed amino acid sequences as anticipated (Fig. S7 *A* and *B*). Given that the gene bodies of J_H2 and J_H3 are shorter than those of J_H1 and J_H4, the average lengths of V_HDJ_{H2} and V_HDJ_{H3} exons were shorter than those of V_HDJ_{H1} and V_HDJ_{H4} (median length 11 aa vs. 13 aa) (Fig. S7*C*). In contrast to productive V_HDJ_H exons, ~85% of productive VJκ exons from splenic B cells showed a CDR3 length of 9 aa. The VJκ CDR3 motif also showed the expected flanking cysteine and phenylalanine (Fig. S7*D*). Thus,

A



B

VH Family	V_H1	V_H2	V_H3	V_H4	V_H5	V_H6	V_H7	V_H8	V_H9	V_H10	V_H11	V_H12	V_H13	V_H14	V_H15	V_H16	Total
Functional V Used [2ug]	51	7	6	1	10	5	3	8	4	2	2	2	1	4	1	0	107
Functional V Not Used [2ug]	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	3
Functional V Used [500ng]	51	7	6	1	10	5	3	8	4	2	2	2	1	4	1	0	107
Functional V Not Used [500ng]	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	3
Functional V Used [100ng]	50	7	6	1	9	3	3	7	4	2	2	2	1	4	1	0	102
Functional V Not Used [100ng]	2	1	0	0	1	2	1	1	0	0	0	0	0	0	0	0	8
Total Functional Vs	52	8	6	1	10	5	4	8	4	2	2	2	1	4	1	0	110
Pseudo-V Used [2ug]	12	1	1	1	2	0	0	1	0	0	0	0	0	0	0	0	18
Pseudo-V Not Used [2ug]	29	2	1	0	9	2	0	7	0	2	0	1	1	0	1	1	56
Pseudo-V Used [500ng]	11	0	1	1	1	0	0	1	0	0	0	0	0	0	0	0	15
Pseudo-V Not Used [500ng]	30	3	1	0	10	2	0	7	0	2	0	1	1	0	1	1	59
Pseudo-V Used [100ng]	4	0	1	1	1	0	0	1	0	0	0	0	0	0	0	0	8
Pseudo-V Not Used [100ng]	37	3	1	0	10	2	0	7	0	2	0	1	1	0	1	1	66
Total Psuedo Vs	41	3	2	1	11	2	0	8	0	2	0	1	1	0	1	1	74

Fig. 4. Representative V_HDJ_H repertoire can be generated from small amounts of starting genomic DNA. (*A*) V_H repertoire with productive and nonproductive information from V_HDJ_H joins (*Left*) and pie charts showing the average overall percentage of productive and nonproductive V_HDJ_H joins (*Right*) in IgM^+ splenic B cells cloned from indicated amounts of genomic DNA using J_H4 coding end bait primer. Mean ± SEM, $n = 3$. (*B*) V_H utilization numbers separated by family, organized as in Fig.1*C*.

HTGTS-Rep-seq produces sequences with CDR3 characteristics expected from the various bait loci.

HTGTS-Rep-Seq Can Be Used with Low Amounts of Starting Material. We generated libraries from J_H4 coding end baits with starting DNA amounts of 2 µg, 500 ng, and 100 ng, each purified from the splenic B cells of the same C57BL/6 mouse. Libraries generated from 2 µg and 500 ng of genomic DNA were almost identical ($r >$ 0.97) in V_H use and productive/nonproductive rearrangement ratios (Fig. 4 and Table S1). Even though we saw a slight decrease in the number of detected V_Hs from the libraries generated from 100 ng of genomic DNA, they still displayed a similar repertoire profile ($r \approx 0.8$) and productive/nonproductive ratio (Fig. 4), suggesting that HTGTS-Rep-seq can be used to generate a quite representative V_HDJ_H repertoire library from as little as 20,000 B cells.

We further evaluated $V(D)J_H$ junctional diversities in these titrated libraries by comparing the percentages of unique CDR3 sequences (35). We found that the proportion of V(D)J exons containing unique CDR3 sequences substantially decreased with reduced amounts of starting material (Fig. S8*A*), indicating that higher amounts of DNA starting material allow us to detect a greater fraction of the highly diverse *IgH* CDR3 repertoire. Although sequencing errors might in theory lead to minor overestimation of CDR3 diversity, the biological diversity of CDR3 in these samples was so high that we observed only a very small overlap portion in detected $V(D)J_H$ CDR3 sequences (<1%) between the three technical repeats of 2-µg DNA libraries and even less between 500-ng or 100-ng DNA library repeat subsets (Fig. S8*B*). Thus, 100 ng of DNA is enough to generate a representative $V(D)J_H$ library with respect to V_H use, but even 2 µg of DNA reveals only a very small fraction of the immense diversity of *IgH* CDR3s.

Discussion

HTGTS-Rep-seq is a DNA-based method that requires only a single bait PCR primer, reads out both deletional and inversional V(D)J joins, and can readily be adapted to identify low frequency recombination events invisible to prior repertoire sequencing assays (22). In addition, HTGTS-Rep-seq can be used to comprehensively study productive and nonproductive V exon use. We also can use HTGTS-Rep-seq to developmentally assess the frequency of V(D)J intermediates, most notably by quantitatively identifying the frequency of particular DJ_H rearrangements (22) (Fig. 1 *E* and *F*). HTGTS-Rep-seq also could be adapted for revealing joining patterns of individual Ds or Vs by using them as baits. Thus, this assay, or adaptations of it, could be useful for detecting changes in repertoires that occur during development, or during an immune response. However, use of HTGTS-Rep-seq for assaying certain antigen receptor repertoires, most notably TCRα repertoires, would currently be more limited given the very large number of different Jαs (24).

HTGTS-Rep-seq requires as little as 100 ng of genomic DNA (and potentially less) from mouse splenic B cells to capture a representative profile of V_H use. Thus, this technique can be applied to relatively small numbers of cells and yield accurate repertoire profiles. However, we find that much larger amounts of starting material would be required to capture the full extent of the immense complexity of the CDR3s that we demonstrate to exist in a given population of splenic B cells. Moreover, potential inaccuracies that do arise in quantifying certain rearrangements via HTGTS-Rep-seq, such as productive/nonproductive ratios for the Igκ repertoire, are due to inherent biological events that would be detected in other DNA-based repertoire-sequencing methods, such as nonproductive VJκ rearrangements in the genome in Igλ-expressing cells or sequential rearrangements involving inversional VJκ joining (34) (Fig. S6). This ambiguity in the assay for the Igκ locus could be minimized if desired by adding an initial step to enrich for sonicated DNA fragments containing sequences just downstream of the whole Jκ region.

The ability to use linear amplification with only a single J primer or set of J primers by HTGTS-Rep-seq avoids the necessity of using sets of degenerate V primers (along with J primers) required by prior DNA-based repertoire-sequencing methods, which could lead to variable amplification efficiencies of different V families or Vs within a family (15). Being DNA-based, HTGTS-Rep-seq also bypasses a major limitation of RNA-based methods for certain applications by quantitatively capturing the frequency of Ig rearrangements in a population regardless of their expression level or whether they are productive or nonproductive. Current means to address biases due to multiplex PCR or varying expression levels between cells include the use of universal identifiers (25, 36, 37) or single cell methods (38), but HTGTS-Rep-seq can accurately identify a population repertoire profile without the additional cost or steps of synthesizing primers with random barcodes, or sorting for single cells.

It is striking that, in experiments where we sequenced about 15,000 unique V(D)J rearrangements from each of three technical repeats, we found less than 1% overlap of unique CDR3 sequences, emphasizing the great sensitivity of the approach. This highly sensitive HTGTS-Rep-seq approach should easily be adapted for application to human samples. In that regard, the sensitivity of HTGTS-Rep-seq should provide a low cost and rapid method for identifying clonal rearrangements (even DJ_H rearrangements) that would be diagnostic of clonal B- or T-lymphocyte expansions that occur in the context of certain immune system diseases, including cancers. Finally, in our libraries, approximately one-third of our joined sequences cover the entire length of the ~370-bp V(D)J exons, making HTGTS-Rep-seq applicable to tracking dominant populations of particular V(D)J exons, including particular CDRs, that appear in the B-cell repertoire during antibody affinity maturation in

an immune response. This application may be enhanced as high throughput sequencing technologies are advanced to achieve greater lengths and accuracy.

Materials and Methods

Mice. WT 129SVE and C57BL/6 mice were purchased from Charles River Laboratories International. All animal experiments were performed under protocols approved by the Institutional Animal Care and Use Committee of Boston Children's Hospital.

B-Cell Isolation from Bone Marrow and Spleen. Bone marrow-derived pro-B ($B220^+IgM^-CD43^+$) cells were purified from 129SVE or C57BL/6 mice by sorting and after the depletion of erythrocytes. Single cell suspensions were stained with B220-APC, CD43-PE, and IgM-FITC antibodies. Splenic resting B cells were purified using biotin/streptavidin bead methods (B220-positive selection) (130-049-501; Miltenyi) or EasySep negative B-cell selection (19754; Stem Cell Technologies).

HTGTS-Rep-Seq. HTGTS-Rep-seq was performed as described (16). Primers are listed in Table S3. For the DJ_H joins analysis, we used the standard LAM-HTGTS bioinformatic pipeline (16). For the V_HDJ_H and $V\kappa J\kappa$ identification, we demultiplexed MiSeq reads using the fastq-multx tool in the ea-utils suite (https://code.google.com/archive/p/ea-utils) and trimmed adaptors with cutadapt software (https://cutadapt.readthedocs.io/en/stable/). The paired reads were then joined using the fastq-join tool from the ea-utils suite (overlap region $\geq$10 bp and mismatch rate $\leq$8%). Reads were then grouped as joined reads and unjoined and were analyzed separately in the following analysis. We used IgBLAST (23) using joined reads and unjoined reads against V(D)J gene databases using default parameters. The V(D)J gene sequences were obtained from IMGT (24), manually curated, and used to generate IgBLAST sequence databases. Various stringencies were applied to filter reads that can align to V, D, and J genes (IgBLAST score >150, total alignment length >100, overall mismatch ratio <0.1). In unjoined reads, the top V gene identified in R1 and R2 reads must match. The use of V genes can be computed based on the processed IgBLAST results. A pipeline named "HTGTSrep" was developed to conduct the above-mentioned processing and analyzing and can be downloaded at Bitbucket (https://bitbucket.org/adugduzhou/htgtsrep).

ACKNOWLEDGMENTS. We thank members of the F.W.A. laboratory for stimulating discussions and Dr. Richard Frock for experimental advice. This work is supported by National Institutes of Health Grant R01AI020047 (to F.W.A.) and Grant F31-AI117920 (to S.G.L.). Z.B. is supported by a Cancer Research Institute Irvington Fellowship; J.H. by a Robertson Foundation/Cancer Research Institute Irvington Fellowship; and Y.Z. by a career development fellowship from the Leukemia and Lymphoma Society. F.W.A. is an investigator of the Howard Hughes Medical Institute.

1. Teng G, Schatz DG (2015) Regulation and Evolution of the RAG Recombinase. *Adv Immunol* 128:1–39.
2. Alt FW, Zhang Y, Meng F-L, Guo C, Schwer B (2013) Mechanisms of programmed DNA lesions and genomic instability in the immune system. *Cell* 152(3):417–429.
3. Alt FW, Baltimore D (1982) Joining of immunoglobulin heavy chain gene segments: Implications from a chromosome with evidence of three D-JH fusions. *Proc Natl Acad Sci USA* 79(13):4118–4122.
4. Retter I, et al. (2007) Sequence and characterization of the Ig heavy chain constant and partial variable region of the mouse strain 129S1. *J Immunol* 179(4):2419–2427.
5. Yancopoulos GD, Alt FW (1986) Regulation of the assembly and expression of variable-region genes. *Annu Rev Immunol* 4:339–368.
6. Proudhon C, Hao B, Raviram R, Chaumeil J, Skok JA (2015) Long-range regulation of V(D)J recombination. *Adv Immunol* 128:123–182.
7. Jung D, Giallourakis C, Mostoslavsky R, Alt FW (2006) Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. *Annu Rev Immunol* 24:541–570.
8. Guo C, et al. (2011) CTCF-binding elements mediate control of V(D)J recombination. *Nature* 477(7365):424–430.
9. Lin SG, Guo C, Su A, Zhang Y, Alt FW (2015) CTCF-binding elements 1 and 2 in the Igh intergenic control region cooperatively regulate V(D)J recombination. *Proc Natl Acad Sci USA* 112(6):1815–1820.
10. Fuxa M, et al. (2004) Pax5 induces V-to-DJ rearrangements and locus contraction of the immunoglobulin heavy-chain gene. *Genes Dev* 18(4):411–422.
11. Jhunjhunwala S, et al. (2008) The 3D structure of the immunoglobulin heavy-chain locus: Implications for long-range genomic interactions. *Cell* 133(2):265–279.
12. Melchers F (2015) Checkpoints that control B cell development. *J Clin Invest* 125(6): 2203–2210.
13. Granato A, Chen Y, Wesemann DR (2015) Primary immunoglobulin repertoire development: Time and space matter. *Curr Opin Immunol* 33:126–131.
14. Schroeder HW, Jr, Zemlin M, Khass M, Nguyen HH, Schelonka RL (2010) Genetic control of DH reading frame and its effect on B-cell development and antigen-specifc antibody production. *Crit Rev Immunol* 30(4):327–344.
15. Georgiou G, et al. (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat Biotechnol* 32(2):158–168.
16. Hu J, et al. (2016) Detecting DNA double-stranded breaks in mammalian genomes by linear amplification-mediated high-throughput genome-wide translocation sequencing. *Nat Protoc* 11(5):853–871.
17. Chiarle R, et al. (2011) Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. *Cell* 147(1):107–119.
18. Frock RL, et al. (2015) Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nat Biotechnol* 33(2):179–186.
19. Meng F-L, et al. (2014) Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. *Cell* 159(7):1538–1548.
20. Wei P-C, et al. (2016) Long neural genes harbor recurrent DNA break clusters in neural stem/progenitor cells. *Cell* 164(4):644–655.
21. Dong J, et al. (2015) Orientation-specific joining of AID-initiated DNA breaks promotes antibody class switching. *Nature* 525(7567):134–139.
22. Hu J, et al. (2015) Chromosomal loop domains direct the recombination of antigen receptor genes. *Cell* 163(4):947–959.
23. Ye J, Ma N, Madden TL, Ostell JM (2013) IgBLAST: An immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* 41(Web Server issue):W34–W40.
24. Lefranc M-P, et al. (2015) IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Res* 43(Database issue):D413–D422.
25. Khan TA, et al. (2016) Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci Adv* 2(3):e1501371.
26. Alt FW, et al. (1984) Ordered rearrangement of immunoglobulin heavy chain variable region segments. *EMBO J* 3(6):1209–1219.
27. Daly J, Licence S, Nanou A, Morgan G, Mårtensson I-L (2007) Transcription of productive and nonproductive VDJ-recombined alleles after IgH allelic exclusion. *EMBO J* 26(19):4273–4282.
28. Yancopoulos GD, et al. (1984) Preferential utilization of the most JH-proximal VH gene segments in pre-B-cell lines. *Nature* 311(5988):727–733.
29. Malynn BA, Yancopoulos GD, Barth JE, Bona CA, Alt FW (1990) Biased expression of JH-proximal VH genes occurs in the newly generated repertoire of neonatal and adult mice. *J Exp Med* 171(3):843–859.
30. ten Boekel E, Melchers F, Rolink AG (1997) Changes in the V(H) gene repertoire of developing precursor B lymphocytes in mouse bone marrow mediated by the pre-B cell receptor. *Immunity* 7(3):357–368.
31. Schatz DG, Ji Y (2011) Recombination centres and the orchestration of V(D)J recombination. *Nat Rev Immunol* 11(4):251–263.
32. Gorman JR, Alt FW (1998) Regulation of immunoglobulin light chain isotype expression. *Adv Immunol* 69:113–181.
33. Mostoslavsky R, Alt FW, Rajewsky K (2004) The lingering enigma of the allelic exclusion mechanism. *Cell* 118(5):539–544.
34. Melchers F, ten Boekel E, Yamagami T, Andersson J, Rolink A (1999) The roles of preB and B cell receptors in the stepwise allelic exclusion of mouse IgH and L chain gene loci. *Semin Immunol* 11(5):307–317.
35. Pieper K, Grimbacher B, Eibel H (2013) B-cell biology and development. *J Allergy Clin Immunol* 131(4):959–971.
36. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR (2013) Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc Natl Acad Sci USA* 110(33):13463–13468.
37. Egorov ES, et al. (2015) Quantitative profiling of immune repertoires for minor lymphocyte counts using unique molecular identifiers. *J Immunol* 194(12):6155–6163.
38. Sundling C, et al. (2014) Single-cell and deep sequencing of IgG-switched macaque B cells reveal a diverse Ig repertoire following immunization. *J Immunol* 192(8):3637–3644.