The usage of six human IGHJ genes follows a particular nonrandom selection: The recombination signal sequence affects the usage frequency of six human IGHJ genes

Bin Shi[1,3,4,†], Xiaoheng Dong[1,†], Qingqing Ma[1,†], Suhong Sun[2,#], Long Ma[1,#], Jiang Yu[1], Xiaomei Wang[1], Juan Pan[1], Xiaoyan He[1], Danhua Su[1], Xinsheng Yao[1,*].

[1]Department of Immunology, Center of Immunomolecular Engineering, Innovation & Practice Base for Graduate Students Education, Zunyi Medical University, Zunyi, China, [2]Departmentof Breast Surgery, The first Affiliated Hospital of ZunYi Medical University, Zunyi, China, [3]Department of Laboratory Medicine, Affiliated Hospital of Zunyi Medical University, Zunyi, China, [4]School of Laboratory Medicine, Zunyi Medical University, Zunyi, China

[*]To whom correspondence should be addressed. Tel: +86 13985671591; Email: immunology01@126.com

[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

[#] These two authors should be regarded as Co-Corresponding author.

ABSTRACT

The formation of the B cell receptor (BCR) heavy chain variable region is derived from the germline V(D)J gene rearrangement according to the "12/23" rule and the "beyond 12/23" rule. The usage frequency of each V(D)J gene in the peripheral BCR repertoires is related to the initial recombination, self-tolerance selection, and the clonal proliferative response. However, their specific differences and possible mechanisms are still unknown. We analyzed in-frame and out-of-frame BCR-H repertoires from human samples with physiological and various pathological conditions by high-throughput sequencing. Our results showed that IGHJ gene frequency follows a similar pattern where IGHJ4 is used at high frequency (>40%), IGHJ6/IGHJ3/IGHJ5 is used at medium frequencies (10%~20%), and IGH2/IGHJ1 is used at low frequency (<4%) under whether physiological or various pathological conditions. Furthermore, analysis of the recombination signal sequences suggested that the conserved nonamer and heptamer and certain 23 bp spacer length may affect the initial IGHD-IGHJ recombination, which results in different frequencies of IGHJ genes among the initial BCR-H repertoire. Based on this "background repertoire", we recommend that re-evaluation and further investigation are needed when analyzing the significance and mechanism of IGHJ gene frequency in self-tolerance selection and the clonal proliferative response.

INTRODUCTION

The diversity of the initial vertebrate B cell receptor (BCR) originates from the recombination of multiple germline genes (V(D)J) and insertion and deletion during the recombination process. There is a consensus recombination signal sequence (RSS) (1) at the 5' or 3' end of each V(D)J gene segment that participates in recombination according to the "12/23" rule (2, 3, 4) and the "beyond 12/23" rule (5). In addition, recombination-activating gene (RAG) enzymes, terminal deoxynucleotidyl transferase (TDT), heterodimer-KU70/KU80, DNA-dependent protein kinase (DNA-PK/Artemis), DNA ligase IV (XRCC4) and other proteins are involved in the complex V(D)J recombination process (4, 6).

Theoretically, the usage frequency of V(D)J gene segments is random in the pro-B cell or pre-B cell recombination process (before autoantigen selection). However, in vitro experiments in B cell lines confirmed that V(D)J gene segments contribute unequally to the primary repertoire, and the consensus heptamer and nonamer sequences of the RSSs are considered the major factor (7). The contributing factors may relate to the usage frequency of V(D)J gene segments. The usage of proximal and distal gene segments in recombination is not random; for example, the JH-proximal VH gene of pre-B cell lines has a preferential usage (8), and VH near Cu may be preferred during early rearrangement (9). During pre-B cell differentiation and development, the initial DH-JH rearrangements employ more 3' (JH-proximal) DH segments (10); however, Feeney AJ et al found that there is no apparent preference for the more JK-proximal over the more JK-distal genes in the proximal region (11). In addition, compared with RSSs with one or more base mutations, the corresponding gene subfamily of RSSs with a consensus heptamer/nonamer (conserved) has preferred usage (3, 12, 13, 14, 15). Moreover, the usage frequency of the corresponding gene segment will be affected when the lengths of the 12 bp spacer/23 bp spacer in RSSs increase or decrease (12, 13, 15) and when the base sequences of the 12 bp spacer/23 bp spacer in RSS change (16,17,18).

However, these results are derived from experiments based on B cell lines in vitro, and whether RSSs influence the V(D)J usage frequency of initial repertoires in vivo is unclear. The difference in each V(D)J usage frequency in the peripheral B cell repertoires is mainly derived from the selection of self-tolerance and the response of clonal proliferation (8, 19, 20, 21). How the difference in usage frequency of each V(D)J gene segment in initial repertoires influences the peripheral repertoire has not been clarified and has received little attention.

With the development of high-throughput sequencing (HTS) analysis for V(D)J tracking, analyzing each V(D)J usage frequency in individual BCR-H repertoires is now possible. We have broadly analyzed the composition characteristics of BCR-H repertoires by HTS since 2013 and found that the human IGHJ4 gene subfamily has the highest usage frequency in physiological and various pathological conditions, followed by IGHJ6, IGHJ3 and IGHJ5 with medium usage frequency and by IGHJ1 and IGHJ2 with significantly low usage frequency. Additionally, the usage frequency of 6 IGHJ gene families shows amazing consistency by analyzing the BCR-H sequences of public databases (IMGT, etc.) and published articles (HTS data) from subjects with physiological or various pathological conditions. Moreover, we analyzed the composition characteristics of the RSSs in human IGHJ genes. Our results suggest that the consensus nonamer and heptamer, the standard spacer length (23 bp), and the mutation site of RSSs may affect the usage frequency of 6 IGHJ gene segments (nonrandom selection), and this specific primary repertoire may result in the lack of significant changes in the usage frequency of 6 IGHJ genes in the peripheral repertoire under physiological and various pathological conditions.

MATERIALS AND METHODS

Subjects and sample preparation

The subjects included six healthy volunteers (6 samples: H-1, H-2, H-3, H-4, H-5 and H-6) (22), two volunteers with systemic lupus erythematosus (SLE) (including 6 total samples pretreatment, during treatment and after treatment, namely, S1-1, S1-2, S1-3, S2-1, S2-2 and S2-3) (23), three volunteers with breast cancer (including 9 total samples pretreatment, during treatment and after treatment, namely, B3-1, B2-1, B1-1, B3-2, B2-2, B1-2, B3-3, B2-3, and B1-3), two volunteers with a high titer of HBsAb (2 samples: HBsAb-1, HBsAb-2) (24) and three volunteers with samples before and after immunization with the HBV vaccine (6 IgM samples (V1-BM, V1-AM, V2-BM, V2-AM, V3-BM, V3-AM) and 6 IgG samples (V1-BG, V1-AG, V2-BG, V2-AG, V3-BG, and V3-AG)) (25). The peripheral blood samples were obtained from the Affiliated Hospital of Zunyi Medical University. All the volunteers were informed of the purpose of peripheral blood collection and were under a protocol approved by The Committee on the Ethics of Human Experiments of Zunyi Medical University, and all the experiments were performed in accordance with the guidelines of the committee. Peripheral blood mononuclear cells (PBMCs) were obtained using Ficoll 1640 (Biochrom AG, Berlin, Germany) density centrifugation.

Total RNA/DNA extraction and cDNA synthesis

Total RNA was extracted from the PBMCs in three volunteers with immunization with HBV vaccine according to the manufacturer's protocol for the total RNA extraction kit (OmegaBio-Tek). The total RNA was then reverse transcribed into cDNA using Oligo dT18 according to the manufacturer's protocol for the reverse transcription kit (MBI, Fermentas). The genomic DNA from PBMCs in other samples was obtained using the QIAamp DNA Mini Kit (QIAGEN, CA) and was stored in a QIAsafe DNA tube (QIAGEN).

High-throughput sequencing

All the DNA samples were sent to Adaptive Biotechnologies Corp (http://www.adaptivebiotech.com) for multiplex PCR amplification of human BCR-HCDR3 regions. Error from bias in this multiplex PCR assay was controlled using synthetic templates (26), and the HCDR3 sequences were acquired by HTS on the ImmunoSEQ platform (http://www.adaptivebiotech.com) (23). All the PCR products of cDNA samples after PCR amplification were sent to Tongji-SCBIT Biotechnology Corporation for HTS, and detailed experimental procedures have been described in our previous article (25). The HCDR3 regions were identified within the sequencing reads according to the definition established by the International ImMunoGeneTics (IMGT) collaboration. A standard algorithm was used to identify which V(D)J segments contributed to each HCDR3 sequence.

Public data

We used 9,340 unique in-frame BCR-H sequences (non-HTS data in different pathological states) derived from the IMGT/LIGM-DB to analyze the IGHJ gene frequency by IMGT/HighV-QUEST (27). In addition, 50,290 BCR-H sequences of memory B cells and 48,167 HCDR3 sequences of naive B cells from a public database (HTS data from 3 healthy volunteers) were used for this study (28). The unique in-frame BCR-H sequences (n=84,804) and out-of-frame sequences (n=13,653) were compared and analyzed by IMGT/HighV-QUEST software in this study.

Sequence analysis

The raw sequences in FASTA format were analyzed with IMGT/HighV-QUEST online software (version 1.3.1, http://www.imgt.org). Using the IMGT summary document, the sequences not meeting the following criteria were filtered out: (1) no results (sequences for which IMGT/HighV-QUEST did not return any result) and (2) unknown (sequences for which no functionality was detected. This category corresponds to the sequences for which the junction could not be identified (no evidence of rearrangement, no evidence of junction anchors).). In-frame and out-of-frame unique sequences remaining after filtering were used for IGHJ gene frequency, D-J pairing, and nucleotide insertion and deletion analyses.

RSS composition analysis

According to the accession numbers of these human IGHJ and IGHD genes in IMGT and GenBank, we obtained detailed annotations of complete human genome sequences for RSS composition analysis, including sequence characteristics of nonamers and heptamers, length characteristics of 12 bp and 23 bp spacers, and the IGHJ gene segment (amino acid) composition of code end.

Software and statistics

IMGT/HighV-QUEST (version 1.3.1) was used for identification of sequences (JH and DH), evaluation of functionality and statistical analysis of the sequence data; IMGT/V-QUEST (version 3.3.1) was used for identification of nonamers, heptamers, 12 bp and 23 bp spacers, and IGHJ gene segments of the coding end; Microsoft Office Excel (version 365) was used for storage, filtering and statistical analysis of the sequences. The resulting sequences were graphed using Prism 8 software (GraphPad). IGHJ gene usages were compared using a χ2 test. Insertions and deletions of the nucleotides were compared using one-way ANOVA with Bonferroni correction. All statistically significant differences are indicated as *=p<0.05; **=p<0.01, and ***=p<0.001.

RESULTS

The IGHJ gene frequency follows a similar pattern and is rarely influenced by antigen selection

The number of BCR-H sequences from 6 healthy volunteer samples ranged from 250,000 to 1,250,000 (Supplementary Table 1). The order of frequency of IGHJ genes (in-frame) was IGHJ4>IGHJ6>IGHJ3>IGHJ5>IGHJ2>IGHJ1, while out-of-frame sequences followed an order of IGHJ4>IGHJ6> IGHJ5>IGHJ3>IGHJ1>IGHJ2 (Figure 1A). For these two groups, the frequency of IGHJ4 was significantly higher than that of each IGHJ gene, while IGHJ1 and IGHJ2 were significantly less frequently used (Figure 1A). Supplementary Table 2 shows the data of the naive B cell repertoire (primary repertoire, n=48,167) and the memory B cell repertoire (n=50,290). The order of IGHJ gene usage (in-frame) was IGHJ4>IGHJ6>IGHJ3>IGHJ5>IGHJ2>IGHJ1. Sequences (n=9,340) from the IMGT/LIGM-DB also followed this pattern (Supplementary Table 3 and Figure 1C), while the usage of IGHJ genes (out-of-frame) followed IGHJ4>IGHJ6> IGHJ5>IGHJ3>IGHJ1>IGHJ2 (Figure 1B). Similarly, IGHJ4 was significantly used, while the IGHJ1 or IGHJ2 frequency was significantly lower than those of other IGHJ genes.

A similar pattern of IGHJ gene frequency was found not only under physiological conditions but also under pathological conditions. IgM and IgG sequences from three volunteers before and after HBV vaccine are shown in Supplementary Table 4**.** IgM in-frame sequences presented as IGHJ4> IGHJ6> IGHJ3>IGHJ5>IGHJ2>IGHJ1, while IgM out-of-frame sequences showed IGHJ4>IGHJ3> IGHJ5> IGHJ6>IGHJ1>IGHJ2 (Figure 1D). For IgG sequences, IGHJ4>IGHJ6>IGHJ5>IGHJ3> IGHJ2>IGHJ1 was found in the in-frame sequences, while out-of-frame sequences showed IGHJ4>IGHJ5>IGHJ3> IGHJ6> IGHJ1>IGHJ2 (Figure 1E). The BCR-H sequences from 6 SLE samples ranged from 170,000 to 610,000 sequences (Supplementary Table 5). The usage frequency of 6 IGHJ genes (in-frame) followed IGHJ4>IGHJ6>IGHJ3>IGHJ5>IGHJ2>IGHJ1, while the order of usage frequency of 6 IGHJ genes (out-of-frame) was IGHJ4>IGHJ6>IGHJ5>IGHJ3> IGHJ1>IGHJ2 (Figure 1F). The BCR-H sequence number from breast cancer samples was approximately 70,000~160,000 for each sample (Supplementary Table 6), and the sequence number from two volunteers with a high titer of HBsAb was 760,000 and 880,000 (Table 7). Interestingly, in-frame and out-of-frame sequences from these two groups consistently presented as IGHJ4>IGHJ6>IGHJ5>IGHJ3>IGHJ2>IGHJ1 (Figure 1G and H).

In addition, we analyzed the ratio of unique to total sequences of each IGHJ gene (in-frame and out-of-frame) and found no differences in 6 IGHJ gene families (Supplementary Table 1 and 2, and Figure 2), which suggests that the multiplex PCR library and the experimental system of HTS did not show obvious bias. Taken together, these results indicate that IGHJ gene frequency follows a similar pattern where IGHJ4 is used at high frequency (>40%), IGHJ6/IGHJ3/IGHJ5 is used at medium frequencies (10%~20%), and IGH2/IGHJ1 is used at low frequencies (<4%). Therefore, the pattern shows high consistency in physiological and various pathological conditions, which suggests that the recombination selection of each IGHJ gene is nonrandom and rarely influenced by antigen selection.

IGHJ-IGHD pairing and trimming and insertion between IGHD and IGHJ

Six IGHJ gene families have different initial BCR-H repertoires, which may be related to nonrandom selection of D-J recombination, thus prompting us to investigate IGHJ-IGHD pairing (Figure 3) and trimming and insertion between IGHD and IGHJ (Figure 4). Most of the 27 IGHD gene subfamilies showed a higher proportion of IGHJ4 pairing (Figure 3). However, whether they were in frame or out of frame, the paired IGHD genes of different IGHJ genes at high or low frequencies were similar. For 6 IGHJ gene families, the IGHD genes paired at high frequency included IGHD6-13, IGHD6-19, IGHD3-22, IGHD3-10, and IGHD2-15, while the low frequency parings included IGHD1-20, IGHD1-7, IGHD4-11, IGHD6-25, and IGHD7-27 (Figure 3).

Trimming and insertion between IGHD and IGHJ mainly presented as 3'D trimmed, 5'J trimmed, and N2 insertion (Figure 4). We found that the mean length of 5'J trimmed showed significant differences among different IGHJ genes under some conditions, while 3'D trimmed and N2 insertion did not show significant differences (data not shown). For IGHJ1 and IGHJ2, the 5'J trimmed length of IGHJ1 (in-frame sequences) showed significant differences compared with the other IGHJ subfamilies in the SLE and IgM with HBV vaccine groups (one-way ANOVA with Bonferroni correction, p<0.001). A similar situation occurred on 5'J trimmed of IGHJ2 in the breast cancer group. The mean length of 5'J trimmed of the IGHJ4 (in-frame or out-of-frame sequences) showed significant differences compared with the other IGHJ genes in the SLE group (one-way ANOVA with Bonferroni correction, each p<0.001). In all groups, IGHJ4 (high usage) showed significant differences compared with IGH1 and IGHJ2 (low usage) (one-way ANOVA with Bonferroni correction, each p<0.001). The mean length of 5'J trimmed from IGHJ6/IGHJ5/IGHJ3 (in-frame sequences) showed significant differences compared with that of the other 5 IGHJ subfamilies in different groups (one-way ANOVA with Bonferroni correction, each p<0.001). These results suggest that the composition of the IGHJ front end (5'J trimmed) may have an impact on the usage and efficiency of the D-J recombination, especially for the IGHJ genes with high or low usage.

The usage frequency of 6 IGHJ families in the BCR-H repertoires from 19 published articles

We analyzed the usage frequency of the 6 IGHJ gene families in BCR-H repertoires from 19 published articles (29-47) (Supplementary Table 8). Overall, subjects included healthy volunteers of different ages (2 months to 87 years) and patients with different pathological conditions, including SLE, primary biliary cholangitis (PBC), colorectal adenoma and carcinoma (CRC), celiac disease (CD), congenital heart disease, atopic dermatitis, hepatitis C virus infection, rheumatoid arthritis, and primary immune thrombocytopenia, as well as in humanized NOD-scid-IL2R gamma (null) mice. The sample sources included peripheral blood, PBMC (DNA), PBMC (RNA), cord blood, biopsies (RNA), humanized mouse spleen, bone marrow, mucosal tissues, small intestine, lung, stomach, lymph node, tonsil, and thymus. The B cell subsets included B cells, pre-B cells, immature B cells, transitional B cells, naive B cells, normal B cells with IGHV1-69-DJ-C rearrangements, memory B cells, plasmacytes, etc.

The usage frequency of the IGHJ4 gene subfamily was higher than that of other IGHJ genes, suggesting that IGHJ4 had the highest frequency in the initial rearrangement and showed high consistency in peripheral repertoires (after self-tolerance selection or the clonal proliferation response). The usage frequencies of IGHJ1 and IGHJ2 were significantly lower than those of the other IGHJ genes, suggesting that IGHJ1 and IGHJ2 may be partially restricted in the initial rearrangement and that they showed consistency in the peripheral repertoires. IGHJ6, IGHJ3, and IGHJ5 have a medium usage frequency, and the usage frequency of IGHJ6 was higher than that of IGH3 and IGHJ5, except for articles 2, 7, 8, 13 and 19. Additional results showed that IGHJ3 usage was higher than IGHJ5. Regardless of the physiological or pathological conditions, the usage frequencies of the 6 IGHJ gene families in our results are almost identical to those in the 19 published articles. The overall results indicate the nonrandomness of the 6 human IGHJ gene usages during the initial rearrangement process.

IGHD-IGHJ recombination may affect IGHJ gene usage through the RSS composition

Recombination of IGHJ-IGHD can be divided into two phases. The first phase involves recognition and cleavage of the DNA, and the second phase involves resolution and joining (4, 6). In the evolutionary process, the human IGHJ nonamer sequence is 5'-GGTTTTTTT-3' (the complementary sequences, CCAAAAACA), and the IGHD nonamer sequence is 5'-ACAAACC-3' (the complementary sequences, TGTTTTTGG). This evolutionary IGHD-IGHJ “double-stranded complementary pairing” relationship may play a role in the efficiency of D-J recombination. The IGHJ-IGHD recombination schematic diagram is shown in Figure 5A.

To investigate whether RSSs affect IGHJ usage, we obtained human IGHJ gene sequences (X97051, X86356, M25625, J00256, AJ879487 from the IMGT and GenBank) for RSS composition analysis. The composition and characteristics of the human IGHJ RSSs (nonamer--spacer--heptamer (9-23-7)), J region sequence and AA are shown in Supplementary Tables 9-11. IGHJ4 and IGHJ6 have the consensus nonamer sequences “5’-GGTTTTTGT-3’” (the complementary sequence is “CCAAAAACA”). However, the nonamer had one or two base mutations in other IGHJ families. Position 4 of IGHJ1 mutated from A to G, position 9 of IGHJ2 mutated from C to A, position 4 of IGHJ3 mutated from A to C, position 6 of IGHJ5 mutated from A to G, and position 8 of IGHJ5 mutated from C to A (Supplementary Table 9 and Figure 5B). The consensus heptamer is CACAGTG/GTGTCAC. Position 5 of IGHJ4 and IGHJ5 mutated from G to T (IGHJ6 mutated to A), position 4 of IGHJ1 mutated from A to G, position 6 of IGHJ2 mutated from T&G to C, and position 6 of IGHJ3 mutated from T to G. In addition, IGHJ4 and IGHJ3 have a consensus spacer length (23 bp), while the spacer length is reduced by 1 or 2 bases in other IGHJ gene families (IGHJ1-22 bp, IGHJ2-22 bp, IGHJ5-21 or 22 bp, and IGHJ6-22 bp) (Figure 5C).

Overall, compared to the conserved RSS, the IGHJ4 gene subfamily is roughly consistent, the spacer lengths are changed in IGHJ6, the nonamer and heptamer are altered in IGHJ3, the spacer lengths and the nonamer are changed in IGHJ5, and the nonamer, heptamer, and spacer lengths are changed simultaneously in IGHJ1 and IGHJ2. There were different code end sequences (AA) in the IGHJ genes IGHJ4 (15AA), IGHJ1 and IGHJ2 (17AA), IGHJ3 and IGHJ5 (16AA), and IGHJ6 (20AA).

DISCUSSION

The V(D)J gene family of the human BCR heavy chain variable region contains 56 functional V genes with 3' ends of 7-23-9 RSS, 27 functional D genes with 3' ends of 9-12-7 RSS and 5' ends of 7-12-9 RSS, and 6 functional J genes with 5' ends of 9-23-7 RSS. The recombination starts with recombination of the 3' end of the D gene and the 5' end of the J gene, and then the 3' end of the V gene is recombined with the 5' end of the D gene (D-J recombination). In the peripheral BCR-H repertoires, the usage frequency of each V(D)J gene is related to the preferred usage in the initial rearrangement, the selection of self-tolerance and the response of peripheral clonal proliferation. However, the mechanism and significance of differential selection among V(D)J gene subfamilies have not been fully elucidated (4, 6, 48).

We investigated the usage frequency of the 6 IGHJ genes in unique BCR-H repertoires (in-frame and out-of-frame) by HTS under physiological and various pathological conditions. In addition, we analyzed non-HTS-derived BCR-H sequences from the IMGT database, the HTS-derived BCR-H sequences from the public database (other laboratory), and the usage frequency data of 6 IGHJ genes from 19 published articles. The results indicate that IGHJ4 has a significantly high usage frequency in all subjects, various tissues, and different B cell subset samples. IGHJ6, IGHJ3, and IGHJ5 have medium usage frequencies, and IGHJ1 and IGHJ2 have significantly low usage frequencies. Taken together, these results suggest that the recombination selection of each human IGHJ gene is nonrandom and rarely influenced by antigen selection, which is different from the traditional understanding.

The IGHJ nonamer and combination frequency

Early studies suggested that the composition characteristics of human IGHJ RSSs may affect the usage frequency of IGHJ in the initial rearrangement. In 1987, Akira S et al found that two sets of heptamer (CACTGTG) and nonamer (GGTTTTTGT) sequences were enough to initiate the V(D)J joining if the 12-bp and 23-bp spacer rule is satisfied in the recombination-competent pre-B cell line (49). A point mutation in the heptamer sequence or a change in the combination of the two spacer lengths (21 bp 22 bp24 bp/11 bp13 bp) would drastically reduce the recombination frequency.

Variation from the conserved sequences in the heptamer and nonamer of the RSSs is considered a major factor affecting the relative representation of gene segments in the primary repertoire. The mechanism of RSSs on gene recombination is mainly related to the interaction efficiency of RAG protein (recombinase) (50-54). Based on the composition of human IGHJ gene families, we found differences in RSSs among 6 IGHJ gene families (Supplementary Table 9 and Figure 5), which suggests that these differences may affect the usage frequency (nonrandom) of IGHJ gene families.

The nonamer of human IGHJ4 and IGHJ6 is the conserved sequence 5'-GGTTTTTGT-3' or 5'-CCAAAACA-3', while the other IGHJ nonamers have one or two base mutations. Experiments in vitro based on B cell lines showed that the mutation of nonamer had a significant effect on the corresponding gene recombination. Ramsden DA et al found that the nonamers were probably the most important element in initial RAG protein binding (12). A single base mutation of the nonamer resulted in a reduction in overall cleavage levels when the heptamer was retained, but the entire nonamer was substituted with random sequence. Both nicks and hairpins were still found, but overall cleavage was reduced fold. Kowalski D et al found (55, 56) that A-rich core sequences of the nonamer may be important to facilitate strand dissociation during the process of recombination.

The presence of three consecutive A residues was necessary for efficient recombination in the nonamer; furthermore, the nucleotides flanking the A-rich core needed to be other than one residue. The mechanism may be that the recombinase must measure the distance between the heptamer and the nonamer to satisfy the 12/23-bp spacer rule (3, 12, 13, 14, 15). Akamatsu Y et al found that the A residue at position 5 (nonamer A-rich core) was most crucial in their recombination assay (13). However, Hesse JE et al considered that the "A residue" at position 6 (nonamer) was most crucial in their recombination assay (3). Regarding the effect of nonamer A-rich core mutation and corresponding gene usage, Akamatsu Y et al found that recombination frequency decreased to 27.3% of the control with the mutant 9-4G (position 4 was changed to G) (13). A mutant at position 9-5C gave the lowest recombination frequency (10.4%). With the double mutant at positions 9-3G and 9-4G, the joining rate dropped only to 19.3% (9-6G and 9-7G was 26.0%). According to the results from cell line experiments, human IGH4 and IGHJ6 gene subfamilies appear to have a "complete A-rich core" in the nonamer (conserved), which may play an important role in their high usage selection. However, 9-4A of human IGHJ1 is mutated to 9-4G, 9-4A of IGHJ3 is mutated to 9-4C, 9-6A of IGHJ5 is mutated to 9-6G, and 9-8C is mutated to 9-8A, which is a possible cause of their disfavored usages.

In addition, Akamatsu Y et al found that the nonamer 9-2C was changed to 9-2A, and the recombination frequency was reduced to 2.7% of the control level; 9-2C was changed to 9-2T, and the frequency was reduced to 12.9%; and 9-2C was changed to 9-2G, and the frequency remained at 61.3% (13). When 9-8C/9-9C were changed to 9-8N/9-9A, the recombination frequency dramatically dropped to less than 0.1%, which suggested that the C residue plays an important role when the recombinase measures the distance between the heptamer and the nonamer sequences. In this study, one factor for the low usage frequency of the human IGHJ2 gene may be its 9-9C mutation to 9-9A.

The IGHJ heptamer and combination frequency

Human IGHJ4 and IGHJ5 genes have the same heptamer sequence (CAATGTG/GTTACAC). Position 7-3C is mutated to 7-3A compared to the conserved heptamer, and 7-3C is mutated to 7-3T in IGHJ6, while the heptamer sequences of the IGHJ4/IGHJ5/IGHJ6 gene subfamilies are uniform on the double strand. Position 7-4A is mutated to 7-4G in the IGHJ1 gene, position 7-6T/7-7G is mutated to 7-6C/7-7C in the IGHJ2 gene, and position 7-6T is mutated to 7-6G in the IGHJ3 gene.

The relationship between the heptamer and the recombination frequency of the corresponding gene family has been confirmed by several laboratories. Both studies found that the mutation of the entire heptamer resulted in low levels of nicking distributed across several sites, the mechanism of heptamer affecting recombination was related to the formation of hairpins, and the nicks and hairpins were reduced 2-fold when the sequence of the last four positions of the heptamer was changed (12, 57). In addition, nicking formation depended on the heptamer for the generation of double strand breaks (DSBs) by RAG1 and RAG2, and the nonamer at the correct distance would improve heptamer efficiency in the natural RSSs. The first three nucleotide positions were nearly 100% conserved (CAC/GTG) in the BCR gene. The mutations were in the first three positions, and cleavage was impaired either at the nicking step or the hairpin formation site. No rearrangement was detected with the mutant at position l (7-1G). Mutations at position 2 (7-2T) and position 3 (7-3G) produced detectable levels of recombination, 0.5% and 0.6%, respectively. The G residue at position 5 was changed to C (7-5C), and the recombination frequency dropped to 5.9% of the control level. For the rest of the residues in the heptamer, mutation effects were moderate, ranging from 28.5 to 52.0%. Akamatsu Y et al found that no rearrangement was detected with the mutant at position l (7-1G), and mutations at position 2 (7-2T) and position 3 (7-3G) produced detectable levels of recombination, 0.5% and 0.6%, respectively. The recombination frequency dropped to 5.9% of the control level when the G residue at position 5 was changed to C (7-5C); for the rest of the residues in the heptamer, mutation effects were moderate, ranging from 28.5 to 52.0% (13).

The first three positions of the 6 human IGHJ gene subfamily heptamers are a conserved CAC/GTG sequence. Based on the results of Akamatsu Y et al, position 7-4A of human IGHJ1 mutated to 7-4G, and 7-6T/7-7G of IGHJ2 mutated to 7-6C/7-7C, which may be one important factor causing their low usage frequency. In addition, the 7-5G mutation (IGHJ3, IGHJ4, IGHJ5, and IGHJ6) may have a moderate effect on their usage frequency. The effect of mutations in the human IGHJ heptamer on usage frequency needs to be further explored.

The RSS spacer and combination frequency

The length of the spacer is also a determining factor contributing to the usage frequency of V(D)J rearrangement. Human IGHJ4 and IGHJ3 gene subfamilies have a conserved 23 bp length; however, the IGHJ1, IGHJ2, IGHJ5, and IGH6 gene subfamilies have 21 bp or 22 bp spacer lengths.

Akamatsu Y et al found that the recombination frequency dropped to 7.7% with the 11-bp RSSs when one C residue was added to the 12 bp RSSs (13 bp spacer) (11.0% joining rate); when two C residues were added (14 bp spacer), recombination dropped below the detection level, indicating that RSS spacer length was critical for combination frequency (13). Nadel B et al found that the effect of the spacer on the recombination rate of various human Vk gene segments in the peripheral repertoire correlated with their frequency in pre-B cells (in vivo) (16). Steen SB et al found that changing the spacer length by one nucleotide (23 bp1 bp only moderately reduced DSB formation, altering the spacer length by greater than one nucleotide (23 bp-2 bp and 23 bp-3 bp), severely reduced cleavage to a lesser degree (15). If each RSS contains a severe mutation (12 bp-3 bp/23 bp-3 bp), no DSBs were observed. According to the above research, the length of the 23 bp spacer of the human IGHJ4 and IGHJ3 gene subfamilies is an important factor in the higher usage frequency, and the length reduction of the 23 bp spacer in the IGHJ1, IGHJ2, IGHJ5, and IGHJ6 genes reduces their recombination usage.

The sequences of RSSs may affect the usage frequency of V(D)J gene recombination. Fanning L et al found that when the Igk 12 bp spacer of the natural sequence CTAC "A" GACTGGA was changed to CTAC "C" GACTGGA but the corresponding 23RSSs-GTAGTACTCCACTG TCTGGCTGT were not changed, the mutant proximal RSSs were consistently used less frequently (17). In addition, the recombination efficiency was 63.0% of the control level when the 12 bp spacer was changed to an artificial sequence GATCGATCGATC (13, 57, 58). Larijani M et al found that the frequency of recombination decreased by approximately 5-fold when the V81x spacer (AGCAAAAGTTACTGTGAGCTCAA) was replaced by that of VA1 (TTGTAA CCACATCCTGAGTGTGT) (14). Montalbano A et al found that single base pair changes in the spacer sequence can significantly affect recombination efficiency (18). Nadel B et al confirmed that natural variation in spacer sequences could contribute to the nonrandom use of human V genes observed in vivo and that a randomly generated variant of a human V spacer was significantly worse in recombination efficiency (16). These results suggest that the spacer sequence plays an important role in recombination efficiency. Our results show that the ratio of AT and CG in 23 bp spacer sequences of 6 human IGHJ gene families is inconsistent (Supplementary Table 9). Base C has the highest ratio in IGHJ4. Is this the reason for the higher usage frequency in the recombination of the IGHJ4 gene subfamily? Whether the base composition of spacer sequences such as the nonamer has the key "A-rich core" structure need to be further explored.

Distance and combination frequency

It has been confirmed that the proximal gene has preferred usage in the initial rearrangement (8-10). Malynn, B.A., et al. believe that the difference in IGHV gene usage in adult spleen B cells is mainly due to the selection of the initial rearrangement rather than the changes in expression frequency after rearrangement (59). The "proximal and distal" studies of BCR recombinant genes are mainly focused on the V gene. "Proximal and distal" differences in the J gene have not been reported. In our results, we did not find the "proximal" phenomenon in the 6 IGHJ gene families with high usage frequencies.

Other factors and combination frequency

Ramsden DA et al found that the sequence of the coding end may be related to the usage frequency of gene combination (12). We found that there are differences in the amino acid length and the coding flank sequences of the human 6 IGHJ families (Supplementary Table 9). The IGHJ4 gene has the shortest 16 amino acid components. The sequence of the coding end and AA length may affect the usage frequency of IGHJ. We analyzed the deletions of the 3'D end, the 5'J end and the insertion between the D-J end and found that there was a difference between the 5'J end of IGHJ4 and other IGHJ genes (Figure 3 and Supplementary Table 11). Whether it was a factor for high usage of IGHJ4 needs to be further studied. In addition, IGHD gene families may also affect the nonrandomness of IGHJ genes. VanDyk LF et al suggested that V(D)J recombination was targeted by RSSs, while the RSSs flanking D segments appeared to be equivalent. They were not randomly utilized, suggesting that the D-3' RSSs were not simply superior targets for the D-J recombinase but instead that targeting certain 12/23-bp spacer RSS combinations is more effective (60).

We found that the conserved nonamer of IGHJ4 and IGHJ6 had a higher “double-stranded complementary paired” rate than the 27 IGHD nonamer sequences (Supplementary Table 10 and Supplementary figure 1), although it did not show obvious differences. At present, no evidence to support that RAG has an effect on the “double-stranded complementary paired” of the J-heptamer to D-heptamer and J-nonamer to D-nonamer exists; the mechanism is still unknown. We hypothesize that two genes with high complementarity (7-7/9-9) may be more favorable for binding, cleaving, hairpin formation, and DSB in the recombination process (Figure 5), which is a very interesting entry point for further research in BCR gene recombination.

In summary, for the possible impacts of RSSs on IGHJ usage frequency, RSSs of human IGHJ4 genes are consistent with conserved RSSs. The length of the IGHJ6 spacer (23 bp) changed slightly, the nonamer and heptamer of IGHJ3 changed, and the length and nonamer of IGHJ5 changed. However, the nonamer, heptamer and spacer of IGHJ1 and IGHJ2 changed significantly. These may be factors that resulted in nonrandom usage of the human IGHJ gene (generally, IGHJ4>IGHJ6>IGHJ3> or ≈IGHJ5>IGHJ2≈IGHJ1) in the initial rearrangement. In the initial human BCR-H repertoires (before antigen selection), the "background" of the 6 IGHJ genes (the initial usage frequency) is different and rarely influenced by antigen selection. These results suggest that re-evaluation and further investigation are needed when analyzing the significance and mechanism of each IGHJ gene usage in self-tolerance selection and the clonal proliferative response.

DATA AVAILABILITY

Our sequencing data was deposited in ImmnunoSEQ database (https://clients.adaptivebiotech.com/login). Sequence for RSS analysis included X97051, X86356, M25625, J00256, AJ879487 from the IMGT/LIGM-DB and GenBank.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

We are grateful to all the volunteers for supporting this study, and we thank all the authors, IMGT and GenBank for using their HTS data for analysis.

FUNDING

The National Natural Science Foundation of China [81860300]; The National Natural Science Foundation of China [81660269]; The National Natural Science Foundation of China [81760300]; Guizhou Provincial High-level Innovative Talents Project [No. (2018) 5637].

Conflict of interest statement. None declared.

REFERENCES

[1] Sakano, H., Huppi, K., Heinrich, G. and Tonegawa, S. (1979) Sequences at the somatic recombination sites of immunoglobulin light-chain genes. *Nature*, **280**, 288-294.

[2] Tonegawa, S. (1983) Somatic generation of antibody diversity. *Nature,* **302**, 575-581.

[3] Hesse, J.E., M.R. Lieber, M. Gellert, and K. Mizuuchi. (1987) Extrachromosomal DNA substrates in pre-B cells undergo inversion or deletion at immunoglobulin V–(D)–J joining signals. *Cell*, **49**, 775–783.

[4] Lewis, S.M. (1994) The mechanism of V(D)J joining: lessons from molecular, immunological and comparative analyses. *Adv. Immunol.*, **56**, 27–150.

[5] Bassing, C.H., Alt, FW., Hughes, M.M., D'Auteuil, M., Wehrly, TD., Woodman, B.B., Gärtner, F., White, J.M., Davidson, L. and Sleckman, B.P. (2000) Recombination signal sequences restrict chromosomal V(D)J recombination beyond the 12/23 rule. *Nature*, **405**, 583–586.

[6] Bogue, M., and D.B. Roth. (1996) Mechanism of V(D)J recombination. *Curr. Opin. Immunol.*, **8**, 175–180.

[7] Feeney, A.J., Tang, A. and Ogwaro, K.M. (2000) B-cell repertoire formation: role of the recombination signal sequence in non-random V segment utilization. *Immunol. Rev.*, **175**, 59-69.

[8] Yancopoulos, G.D., Desiderio, S.V., Paskind, M., Kearney, J.F., Baltimore, D. and Alt, F.W. (1984) Preferential utilization of the most JH-proximal VH gene segments in pre-B-cell lines. *Nature*, **311**, 727–733.

[9] Perlmutter, R.M., Kearney, J.F., Chang, S.P. and Hood, L.E. (1985) Developmentally controlled expression of immunoglobulin VH genes. *Science*, **227**, 1597–1600.

[10] Reth, M.G., Jackson, S. and Alt, F.W. (1986) VHDJH formation and DJH replacement during pre-B differentiation: nonrandom usage of gene segments. *EMBO*, **5**, 2131–2138.

[11] Feeney, A. J., Lugo, G. and Escuro, G. (1997) Human cord blood kappa repertoire. *J. Immunol.*, **58**, 3761 -3768.

[12] Ramsden, D.A., McBlane, J.F., van Gent, D.C. and Gellert, M. (1996) Distinct DNA sequence and structure requirements for the two steps of V(D)J recombination signal Cleavage. *EMBO J.*, **15**, 3197-3206.

[13] Akamatsu, Y., Tsurushita, N., Nagawa, F., Matsuoka, M., Okazaki, K., Imai, M. and Sakano, H. (1994) Essential residues in V(D)J recombination signals. *J. Immunol.*, **153**, 4520-4529.

[14] Larijani, M., Yu, C.C., Golub, R., Lam, Q.L. and Wu, G.E. (1999) The role of components of recombination signal sequences in immunoglobulin gene segment usage: a V81x model. *Nucleic Acids Res.*, **27**, 2304-2309.

[15] Steen, SB., Gomelsky, L., Speidel, S.L. and Roth, D.B. (1997) Initiation of V(D)J recombination in vivo: role of recombination signal sequences in formation of single and paired ouble-strand breaks. *EMBO J.*, **16**, 2656–2664.

[16] Nadel, B., Tang, A., Escuro, G., Lugo, G. and Feeney, A.J. (1998) Sequence of the Spacer in the Recombination Signal Sequence Affects V(D)J Rearrangement Frequency and Correlates with Nonrandom Vk Usage In Vivo. *J. Exp. Med.*, **187**, 1495–1503.

[17] Fanning, L., Connor, A., Baetz, K., Ramsden, D. and Wu, G.E. (1996) Mouse RSS spacer sequences affect the rate of V(D)J recombination. *Immunogenetics*, **44**, 146-150.

[18] Montalbano, A., Ogwaro, K.M., Tang, A., Matthews, A.G., Larijani, M., Oettinger, M.A. and Feeney, A.J. (2003) V(D)J Recombination Frequencies Can Be Profoundly Affected by Changes in the Spacer Sequence. *J Immunol*, **171**, 5296-5304.

[19] Gu, H., Tarlinton, D., Müller, W., Rajewsky, K. and Förster, I. (1991) Most peripheral B cells in mice are ligand selected. *J. Exp. Med.*, **173,** 1357–1371.

[20] Groettrup, M., and von Boehmer, H. (1993) A role for a preT-cell receptor in T-cell development. *Immunol. Today.*, **14**, 610–614.

[21] Ten Boekel, E., Melchers, F. and Rolink, A.G. (1997) Changes in the VH gene repertoire of developing precursor B lymphocytes in mouse bone marrow mediated by the pre-B cell receptor. *Immunity.*, **7**, 357–368.

[22] Ma, L., Yang, L.W., Shi, B., He, X.Y., Peng, A., Li, Y., Zhang, T., Sun, S.H., Ma, R. and Yao, X.S. (2016) Analyzing the CDR3 Repertoire with respect to TCR－Beta Chain V-D-J and V-J Rearrangements in Peripheral T Cells using HTS. *Sci Rep.*, **6**, 29544.

[23] Shi, B., Yu, J., Ma, L., Ma, Q., Liu, C., Sun, S., Ma, R. and Yao, X.S. (2016) Short-term assessment of BCR repertoires of SLE patients after high dose glucocorticoid therapy with high-throughput sequencing. *Springerplus.*, **5**, 75.

[24] Pan, J.，Shi, B.，Ma, L. and Yao, X. S. (2015) Analysis of BCR CDR3 repertoire of peripheral blood with HBsAb titer higher than 10 000 mU/ml. *Chinese Journal of Immunology*, **3**, 300-303.

[25] Ma, L., Wang. X., Bi, X., Yang, J., Shi, B., He, X., Ma, R., Ma, Q. and Yao, X.S. (2017) Characteristics Peripheral Blood IgG and IgM Heavy Chain Complementarity Determining Region 3 Repertoire before and after Immunization with Recombinant HBV Vaccine. *PLoS One.*, **12**, e0170479**.**

[26] Carlson, C.S., Emerson, R.O., Sherwood, A.M., Desmarais, C., Chung, M.W., Parsons, J.M., Steen, M.S., LaMadrid-Herrmannsfeldt, M.A., Williamson, D.W., Livingston, R.J. et.al. (2013) Using synthetic templates to design an unbiased multiplex PCR assay. *Nat Commun*., **4**, 2680.

[27] Shi, B., Ma, L., He, X., Wang, X., Wang, P., Zhou, L. and Yao, X.S. (2014) Comparative analysis of human and mouse immunoglobulin variable heavy regions from IMGT/LIGM-DB with IMGT/HighV-QUEST. *Theor Biol Med Model.*, **11**,30.

[28] DeWitt, W.S., Lindau, P., Snyder, T.M., Sherwood, A.M., Vignali, M., Carlson, C.S., Greenberg, P.D., Duerkopp, N., Emerson, R.O. and Robins, H.S. (2016) A Public Database of Memory and Naive B Cell Receptor Sequences. *PLoS One.*, **11**, e0160853.

[29] Liu, S., Hou, X.L., Sui, W.G., Lu, Q.J., Hu, Y.L. and Dai, Y. (2017) Direct measurement of B-cell receptor repertoire's composition and variation in systemic lupus erythematosus. Genes and immunity, *Genes Immun.*, **18**, 22-27.

[30] Tan, Y.G., Wang, Y.Q., Zhang, M., Han, Y.X., Huang, C.Y., Zhang, H.P., Li, Z.M., Wu, X.L., Wang, X.F., Dong, Y. et.al. (2016) Clonal Characteristics of Circulating B Lymphocyte Repertoire in Primary Biliary Cholangitis. *J Immunol.*, **197**, 1609-1620.

[31] Martin, V.G., Wu, Y.B., Townsend, C.L., Lu, G.H., O'Hare, J.S., Mozeika, A., Coolen, A.C., Kipling, D., Fraternali, F. and Dunn-Walters, D.K. (2016) Transitional B Cells in Early Human B Cell Development - Time to Revisit the Paradigm. *Front Immunol.*, **7**, 546.

[32] Guo, C., Wang. Q., Cao, X., Yang, Y., Liu, X., An, L., Cai, R., Du, M., Wang, G., Qiu, Y., Peng, Z. et.al. (2016) High-Throughput Sequencing Reveals Immunological Characteristics of the TRB-/IgH-CDR3 Region of Umbilical Cord Blood. *J Pediatr.*, **176,** 69-78.

[33] Zhang, W., Feng, Q., Wang, C., Zeng, X., Du, Y., Lin, L., Wu, J., Fu, L., Yang, K., Xu, X. et.al. (2017) Characterization of the B Cell Receptor Repertoire in the Intestinal Mucosa and of Tumor-Infiltrating Lymphocytes in Colorectal Adenoma and Carcinoma. *J Immunol.*, **198**, 3719-3728.

[34] Roy, B., Neumann, R.S., Snir, O., Iversen, R., Sandve, G.K., Lundin, K.E.A. and Sollid, L.M. (2017) High-Throughput Single-Cell Analysis of B Cell Receptor Usage among Autoantigen-Specific Plasma Cells in Celiac Disease. *J Immunol.*, **199**, 782-791.

[35] Rother, M.B., Schreurs, M.W., Kroek, R., Bartol, S.J., Dongen, J.J. and Zelm, M.C. (2016) The Human Thymus Is Enriched for Autoreactive B Cells. *J Immunol.*, **197**, 441-448.

[36] Kerzel, S., Rogosch, T., Struecker, B., Maier, R.F., Kabesch, M. and Zemlin, M. (2016) Unlike in Children with Allergic Asthma, IgE Transcripts from Preschool Children with Atopic Dermatitis Display Signs of Superantigen-Driven Activation. *J Immunol.*, **196**, 4885-4892.

[37] Zhang, W., Du, Y., Su, Z., Wang, C., Zeng, X., Zhang, R., Hong, X., Nie., Wu, J., Cao, H., Xu, X. and Liu, X. (2015) IMonitor: A Robust Pipeline for TCR and BCR Repertoire Analysis. *Genetics.*, **201**, 459-72.

[38] Forconi, F., Potter, K.N., Wheatley, I., Darzentas, N., Sozzi, E., Stamatopoulos, K., Mockridge, C.I., Packham, G. and Stevenson, F.K. (2010) The normal IGHV1-69-derived B-cell repertoire contains stereotypic patterns characteristic of unmutated CLL. *Blood.*, **115**, 71-77.

[39] Racanelli, V., Brunetti, C., De, Re. V., Caggiari, L., Zorzi, M., Leone, P., Perosa, F., Vacca, A. and Dammacco, F. (2011) Antibody V(h) repertoire differences between resolving and chronically evolving hepatitis C virus infections. *PLoS One*., **6**, e25606.

[40] Ippolito, G.C., Hoi, K.H., Reddy, S.T., Carroll, S.M., Ge, X., Rogosch, T., Zemlin, M., Shultz, L.D., Ellington, A.D., Vandenberg, C.L. et.al. (2012) Antibody repertoires in humanized NOD-scid-IL2Rgamma(null) mice and human B cells reveals human-like diversification and tolerance checkpoints in the mouse. *PLoS One.*, **7**, e35497.

[41] Prabakaran, P., Chen, W., Singarayan, M.G., Stewart, C.C., Streaker, E., Feng, Y. and Dimitrov, D.S. (2012) Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations[J]. *Immunogenetics.*, **64**, 337-350.

[42] Briney, B.S., Willis, J.R., Finn, J.A., McKinney, B.A. and Crowe, J.E. (2014) Tissue-specific expressed antibody variable gene repertoires. *PLoS One.*, **9**, e100839.

[43] Mroczek, E.S., Ippolito, G.C., Rogosch, T., Hoi, K.H., Hwangpo, T.A., Brand, M.G., Zhuang, Y., Liu, C.R., Schneider, D.A., Zemlin, M. et.al. (2014) Differences in the composition of the human antibody repertoire by B cell subsets in the blood. *Front Immunol.*, **5**, 96.

[44] Lecerf, M., Scheel, T., Pashov, A.D., Jarossay, A., Ohayon, D., Planchais, C., Mesnage, S., Berek. C., Kaveri, S.V., Lacroix-Desmazes, S. et.al. (2015) Prevalence and gene characteristics of antibodies with cofactor-induced HIV-1 specificity. *J Biol Chem.*, **290**, 5203-5213.

[45] Martin, V., Wu, Y.C., Kipling, D. and Dunn-Walters, D.K. (2015) Age-related aspects of human IgM+ B cell heterogeneity. *Ann N Y Acad Sci.*, **1362**, 153-63.

[46] Hirokawa, M., Fujishima, N., Togashi, M., Saga, A., Omokawa, A., Saga, T., Moritoki, Y., Ueki, S., Takahashi, N., Kitaura, K. and Suzuki, R. (2019) High-throughput sequencing of IgG B-cell receptors reveals frequent usage of the rearranged IGHV4–28/ IGHJ4 gene in primary immune thrombocytopenia. *Sci Rep.*, **9**, 8645.

[47] Yin, L., Hou, W., Liu, L., Cai, Y., Wallet, M.A., Gardner, B.P., Chang, K., Lowe, A.C., Rodriguez, C.A., Sriaroon, P. et.al. (2013) IgM repertoire biodiversity is reduced in HIV-1 infection and systemic lupus erythematosus. *Front Immunol.*, **4**, 373.

[48] Kenneth Murphy，Charles A. Janeway Jr. Paul Travers. et al. Janeway's immunobiology，ISBN 978-0-8153-4243-4，Published by Garland Science, Taylor & Francis Group, LLC, an informa business. Chapter 5: The Generation of Lymphocyte Antigen Receptors.

[49] Akira, S., Okazaki, K. and Sakano. H. (1987) Two pairs of recombination signals are sufficient to cause immunoglobulin V-(D)-J joining. *Science.*, **238**, 1134–1138.

[50] Swanson, P.C., and Desiderio, S. (1998) V(D)J Recombination Signal Recognition: Distinct, Overlapping DNA-Protein Contacts in Complexes Containing RAG1 with and without RAG2. *Immunity.*, **9**, 115-125.

[51] Difilippantonio, M.J., McMahan, C.J., Eastman, Q.M., Spanopoulou, E. and Schatz, D.G. (1996) RAG1 Mediates Signal Sequence Recognition and Recruitment of RAG2 in V(D)J Recombination. *Cell.*, **87**, 253-262.

[52] Spanopoulou, E., Zaitseva, F., Wang, F.H., Santagata, S., Baltimore, D. and Panayotou, G. (1996) The homeodomain region of Rag-1 reveals the parallel mechanisms of bacterial and V(D)J recombination. *Cell.*, **87**, 263-276.

[53] Ramsden, D.A., McBlane, J.F., van, Gent, D.C. and Gellert, M. (1996) Distinct DNA sequence and structure requirements for the two steps of V(D)J recombination signal cleavage. *EMBO J.*, **15**, 3197-3206.

[54] Akamatsu, Y. and Oettinger, M.A. (1998) Distinct Roles of RAG1 and RAG2 in Binding the V(D)J Recombination Signal Sequences. *Mol Cell Biol.*, 18, 4670-4678.

[55] Kowalski, D., Natale, D.A. and Eddy, M.J. (1988) Stable DNA unwinding, not “breathing”, accounts for single-strand-specific nuclease hypersensitivity of specific A+T-rich sequences. *Proc Natl Acad Sci U S A.*, **85**, 9464-9468.

[56] Kowalski, D. and Eddy, M.J. (1989) The DNA unwinding element: a novel cis-acting component that facilitates opening of the Escherichia coli replication origin. *EMBU J.*, **8**, 4335-4344.

[57] Hesse, J.E., Lieber, M.R. Mizuuchi, K. and Gellert, M. (1989) V(D)J recombination: a functional definition of the joining signals. *Genes Dev.*, **3**, 1053-1061.

[58] Akira, S., Okazaki, K. and Sakano, H. (1987) Two pairs of recombination signals are sufficient to cause immunoglobulin V(D)J joining. *Science*., **238**, 1134-1138.

[59] Malynn, B.A., Yancopoulos, G.D., Barth, J.E., Bona, C.A., and Alt, F.W. (1990) Biased expression of JH-proximal VH genes occurs in the newly generated repertoire of neonatal and adult mice. *J. Exp. Med.* ,**171,** 843–859.

[60] VanDyk, L.F., Wise, T.W., Moore, B.B. and Meek, K. (1996) Immunoglobulin D(H) recombination signal sequence targeting: effect of D(H) coding and flanking regions and recombination partner. *J Immunol*., **157**, 4005-4015.

FIGURES LEGENDS

Figure 1. The usage frequencies of 6 IGHJ genes in the in-frame and out-of-frame BCR-H repertoire from different subjects. **(A)** The IGHJ usages of BCR-H repertoire from 6 Healthy volunteers. **(B)** The IGHJ usages of BCR-H repertoire from public data. **(C)** The IGHJ usages of BCR-H repertoire from IMGT data. **(D)** The IGHJ usages of IgM-H repertoire from volunteers before and after immunization with the HBV vaccine. **(E)** The IGHJ usages of IgG-H repertoire from volunteers before and after immunization with the HBV vaccine. **(F)** The IGHJ usages of BCR-H repertoire from SLE volunteers. **(G)** The IGHJ usages of BCR-H repertoire from breast cancer volunteers. **(H)** The IGHJ usages of BCR-H repertoire from volunteers with a high titer of HbsAb.

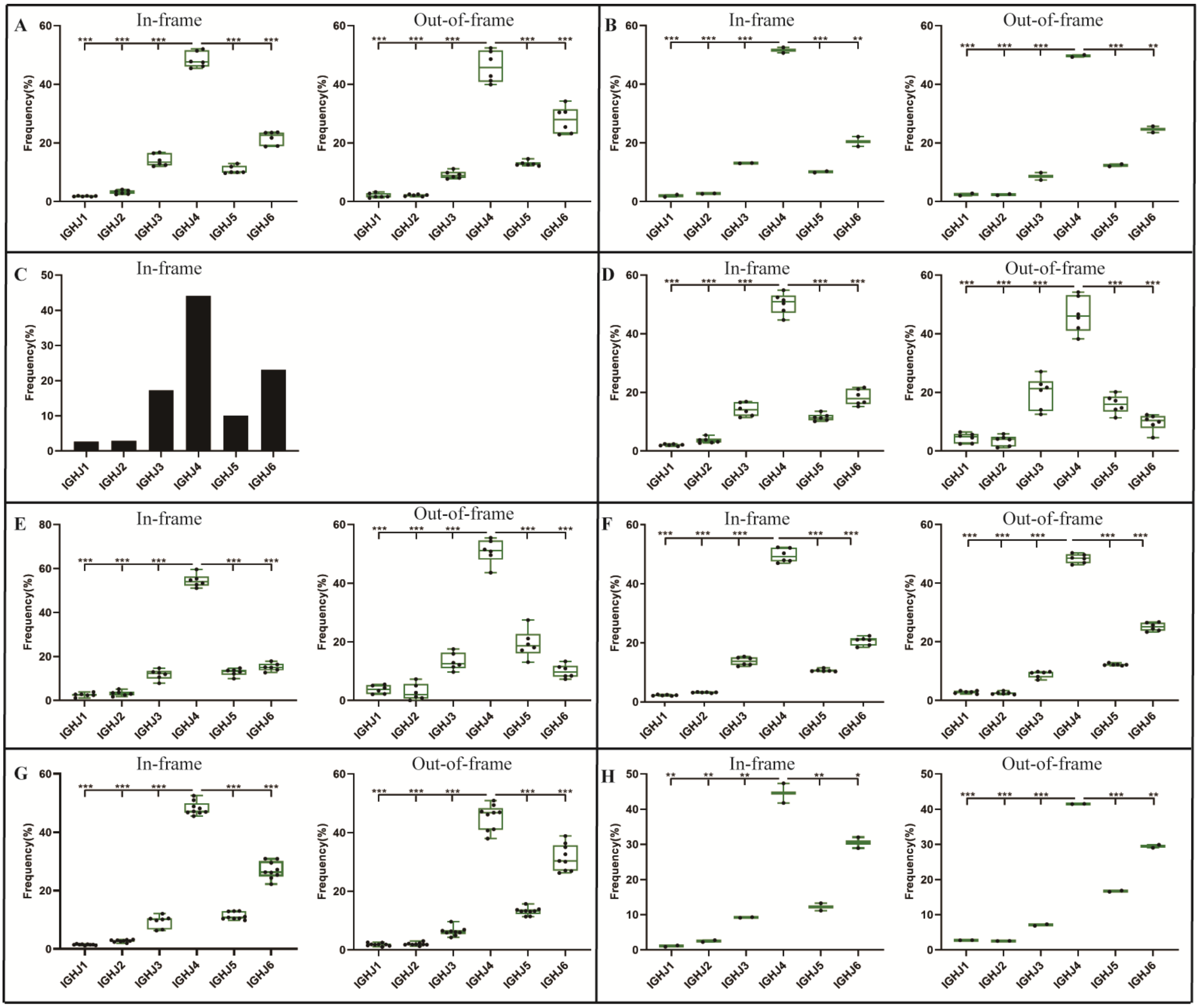


Figure 2. The ratio of unique to total sequences (U/T) of 6 IGHJ genes in the in-frame and out-of-frame BCR-H repertoires from different subjects. **(A)** The IGHJ usages of BCR-H repertoires from 6 Healthy volunteers. **(B)** The IGHJ usages of BCR-H repertoires from public data. **(C)** The IGHJ usages of IgM-H repertoires from volunteers before and after immunization with the HBV vaccine. **(D)** The IGHJ usages of IgG-H repertoires from volunteers before and after immunization with the HBV vaccine. **(E)** The IGHJ usages of BCR-H repertoires from SLE volunteers. **(F)** The IGHJ usages of BCR-H repertoires from breast cancer volunteers. **(G)** The IGHJ usages of BCR-H repertoires from volunteers with a high titer of HbsAb.

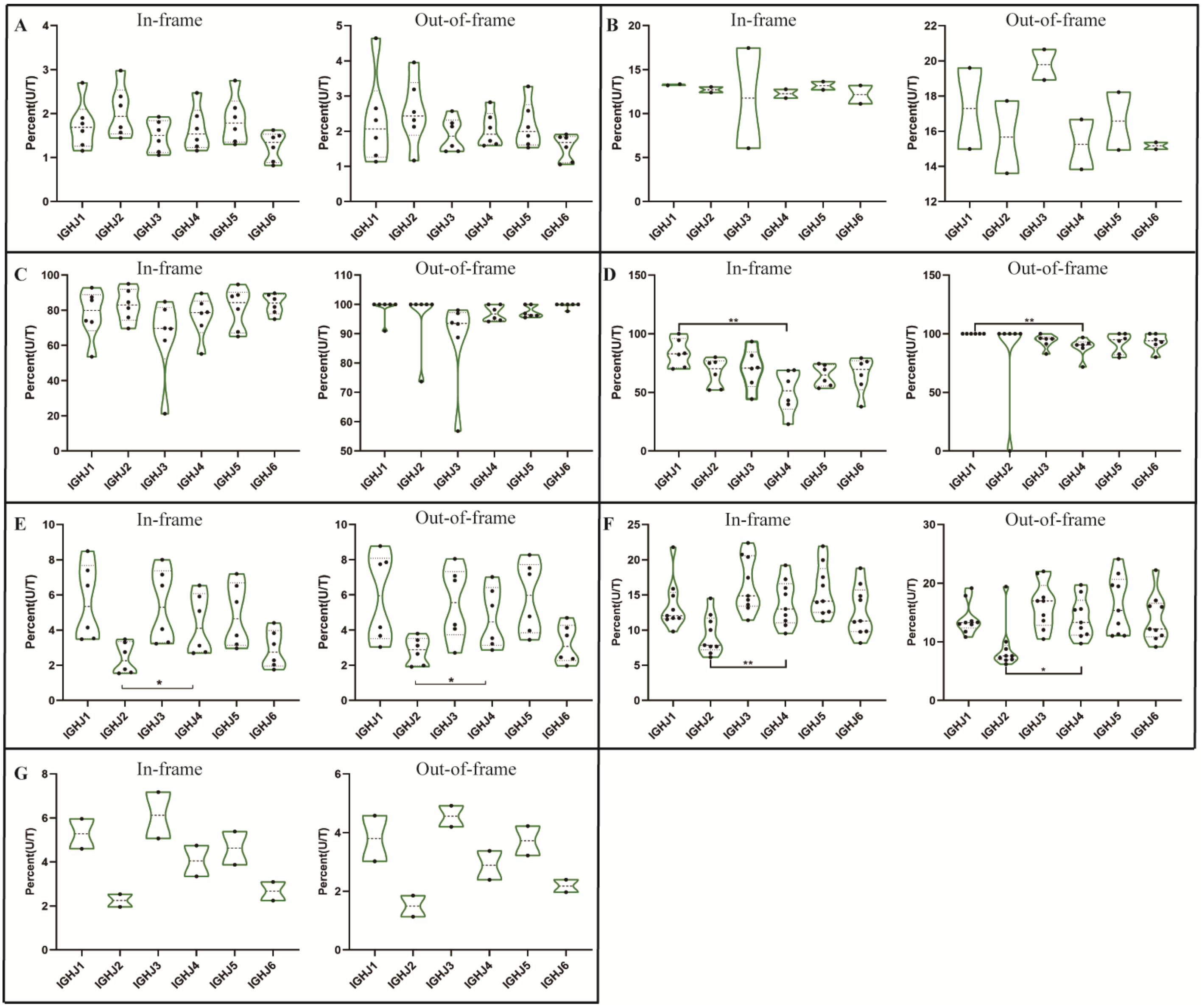


Figure 3. IGHJ-IGHD pairing in the in-frame and out-of-frame BCR-H repertoires from different subjects. **(A)** The IGHJ usages of BCR-H repertoires from 6 Healthy volunteers. **(B)** The IGHJ usages of BCR-H repertoires from public data. **(C)** The IGHJ usages of IgM-H repertoires from volunteers before and after immunization with the HBV vaccine. **(D)** The IGHJ usages of IgG-H repertoires from volunteers before and after immunization with the HBV vaccine. **(E)** The IGHJ usages of BCR-H repertoires from SLE volunteers. **(F)** The IGHJ usages of BCR-H repertoires from breast cancer volunteers. **(G)** The IGHJ usages of BCR-H repertoires from volunteers with a high titer of HbsAb.

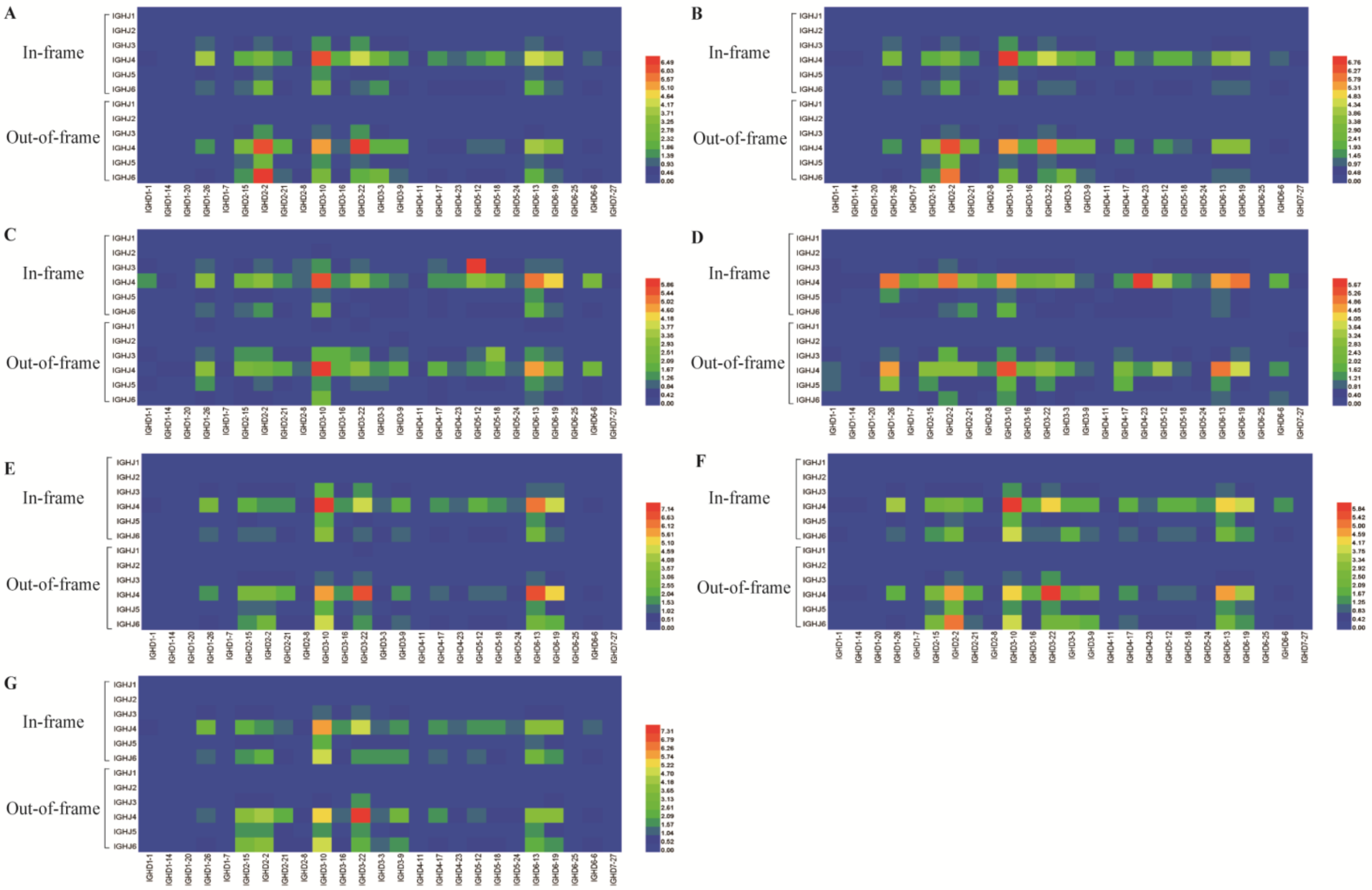


Figure 4. 3'D trimmed, 5'J trimmed and N2 insertion at IGHD-IGHJ junction in the in frame and out of frame BCR-H repertoires from different subjects. **(A)** The IGHJ usages of BCR-H repertoires from 6 Healthy volunteers. **(B)** The IGHJ usages of BCR-H repertoires from public data. **(C)** The IGHJ usages of IgM-H repertoires from volunteers before and after immunization with the HBV vaccine. **(D)** The IGHJ usages of IgG-H repertoires from volunteers before and after immunization with the HBV vaccine. **(E)** The IGHJ usages of BCR-H repertoires from SLE volunteers. **(F)** The IGHJ usages of BCR-H repertoires from breast cancer volunteers. **(G)** The IGHJ usages of BCR-H repertoires from volunteers with a high titer of HbsAb.

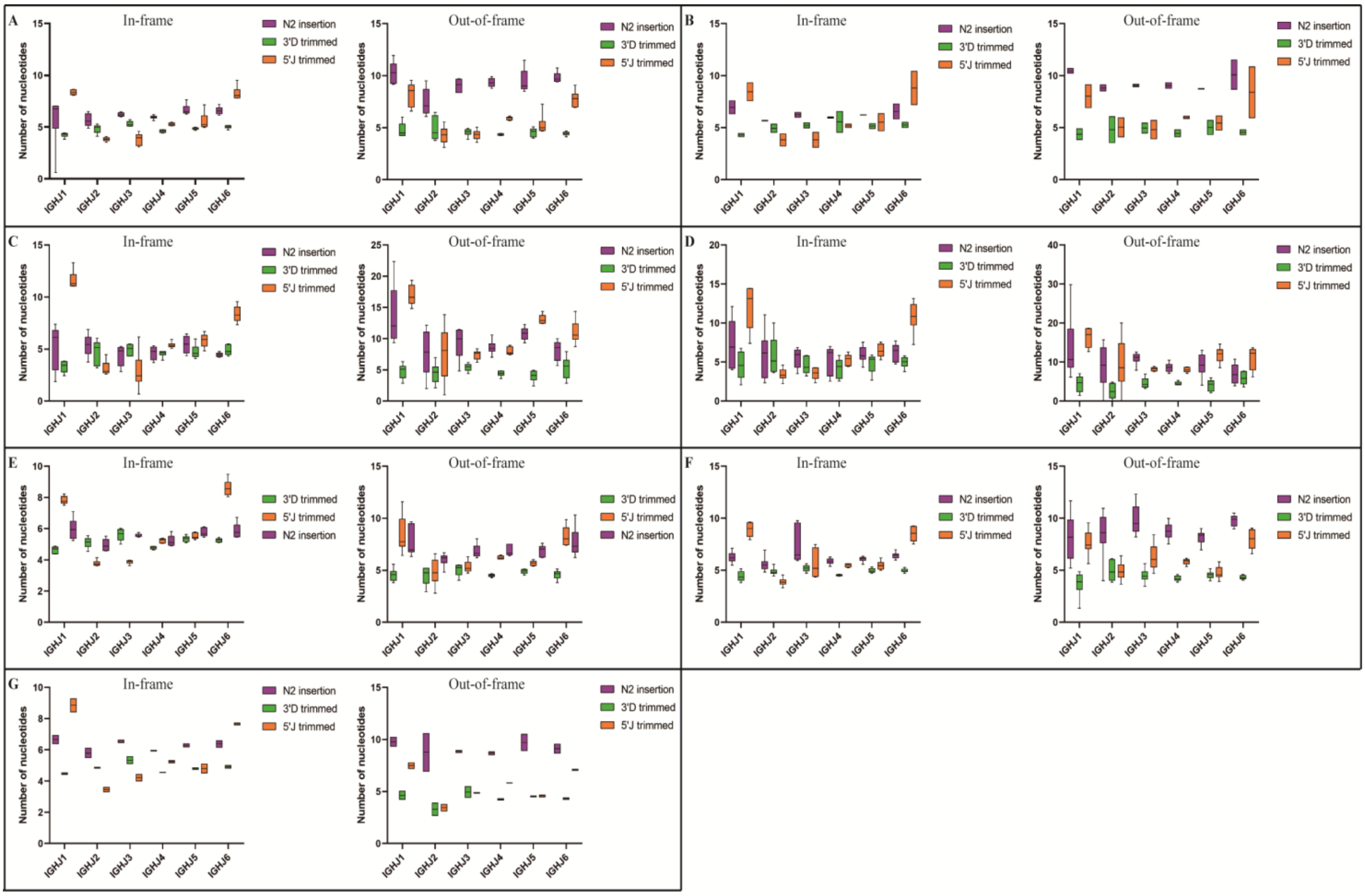
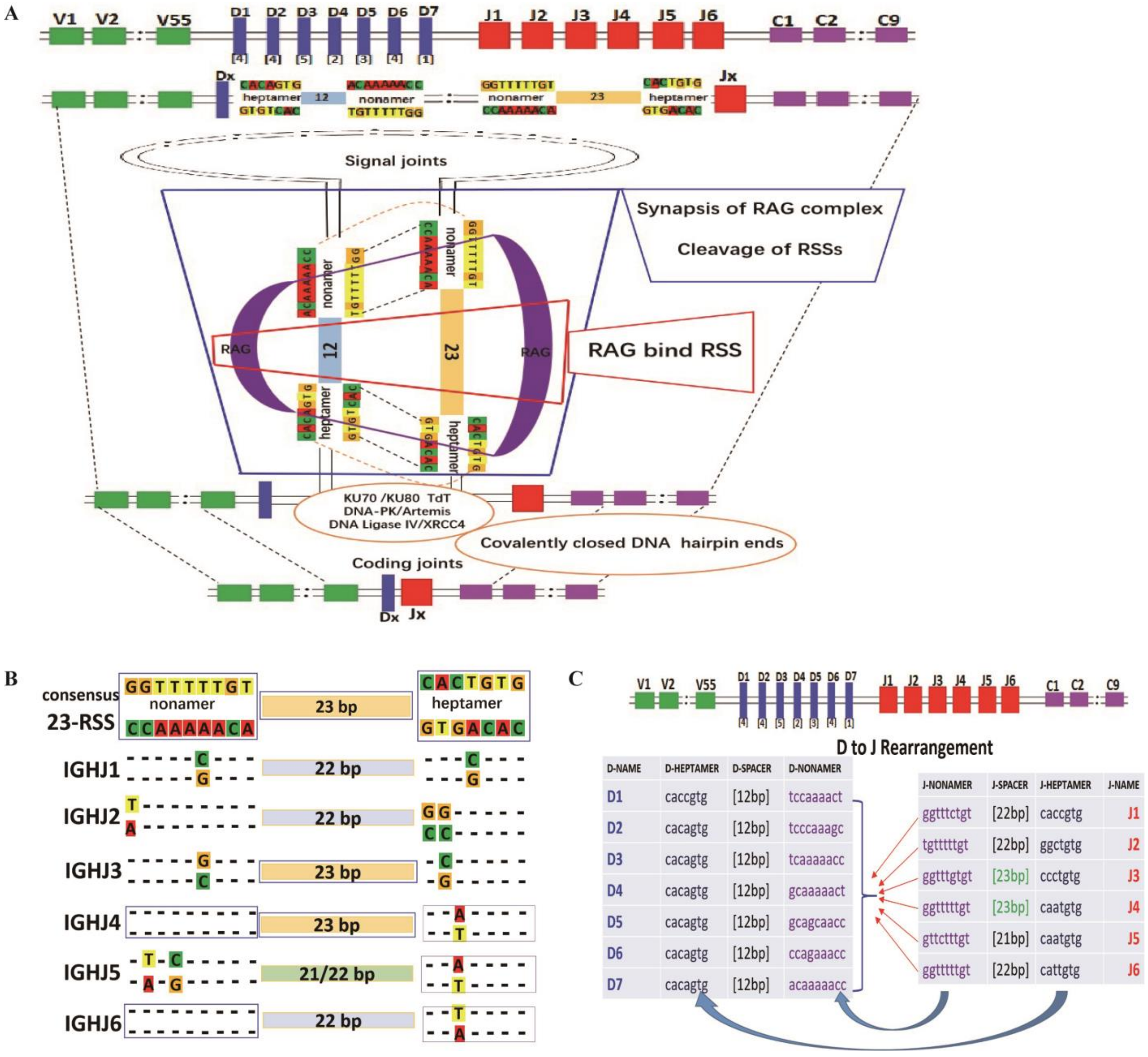


Figure 5. RSS composition characteristics during the IGHD-IGHJ recombination. **(A)** The schematic diagram of IGHJ and IGHD recombination. **(B)** The composition characteristics of human 9-23-7 RSSs (IGHJ-nonamer--IGHJ- spacer--IGHJ-heptamer). **(C)** The pairing of IGHJ (7-12-9) RSSs and IGHD (9-23-7) RSSs during the IGHD-IGHJ recombination.



D-NAME	D-HEPTAMER	D-SPACER	D-NONAMER
D1	caccgtg	[12bp]	tccaaaact
D2	cacagtg	[12bp]	tcccaaagc
D3	cacagtg	[12bp]	tcaaaaacc
D4	cacagtg	[12bp]	gcaaaaact
D5	cacagtg	[12bp]	gcagcaacc
D6	cacagtg	[12bp]	ccagaaacc
D7	cacagtg	[12bp]	acaaaaacc

J-NONAMER	J-SPACER	J-HEPTAMER	J-NAME
ggtttctgt	[22bp]	caccgtg	J1
tgtttttgt	[22bp]	ggctgtg	J2
ggtttgtgt	[23bp]	ccctgtg	J3
ggtttttgt	[23bp]	caatgtg	J4
gttctttgt	[21bp]	caatgtg	J5
ggtttttgt	[22bp]	cattgtg	J6

Supplementary Table 1. The sequence data of 6 IGHJ gene families in the in-frame and out-of-frame BCR-H repertoire from 6 Healthy volunteers

Sample	Repertoire		In-frame		IGHJ1		IGHJ2		IGHJ3		IGHJ4		IGHJ5		IGHJ6	
	total	unique	total	unique	total	unique	total	unique	total	unique	total	unique	total	unique	total	unique
H-1	550082	6351	474996	5266	6977	90	12735	184	59981	634	194010	2432	49685	681	151608	1245
H-2	556813	12353	482531	10455	7485	202	14409	429	70220	1352	219959	5438	38288	1053	132170	1981
H-3	765659	14690	658273	12285	11000	209	14764	353	112124	2028	300258	5835	55841	1190	164286	2670
H-4	1227335	20722	1052967	17223	15791	280	20565	449	177833	2894	471857	7832	89786	1725	277135	4043
H-5	499983	5777	436701	4889	8296	96	11744	184	60807	688	217167	2515	37455	487	101232	919
H-6	897745	12839	763705	10621	12348	198	15385	260	96228	1326	360141	5072	76626	1265	202977	2500
Total	4497617	72732	3869173	60739	61897	1075	89602	1859	577193	8922	1763392	29124	347681	6401	1029408	13358
U/T	72732/4497617		60739/3869173		1075/61897		1859/89602		8922/577193		29124/1763392		6401/347681		13358/1029408	
Sample	**Repertoire**		**Out-of-frame**		**IGHJ1**		**IGHJ2**		**IGHJ3**		**IGHJ4**		**IGHJ5**		**IGHJ6**	
	total	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**
H-1	550082	6351	75086	1085	718	13	1269	27	5865	84	32189	527	10311	158	24734	276
H-2	556813	12353	74282	1898	668	31	1189	47	6851	176	34443	972	7056	231	24075	441
H-3	765659	14690	107386	2405	2493	66	1254	40	12011	268	41385	991	11932	308	38311	732
H-4	1227335	20722	174368	3499	2551	59	2473	63	16117	344	66495	1397	20752	439	65980	1197
H-5	499983	5777	63282	888	1145	15	1631	19	5253	75	29222	465	6806	111	19225	203
H-6	897745	12839	134040	2218	6352	72	2116	49	11306	179	54900	949	15562	291	43804	678
Total	4497617	72732	628444	11993	13927	256	9932	245	57403	1126	258634	5301	72419	1538	216129	3527
U/T	72732/4497617		11993/628444		256/13927		245/9932		1126/57403		5301/258634		1538/72419		3527/216129	

Note: U/T represents the ratio of unique to total sequences.

Supplementary Table 2. The sequence data of 6 IGHJ gene families in the in-frame and out-of-frame BCR-H repertoire from public data (Naive and Memory B cells)

	Repertoire		**In-frame**		**IGHJ1**		**IGHJ2**		**IGHJ3**		**IGHJ4**		**IGHJ5**		**IGHJ6**	
Sample	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**
PLOS-1	354494	48167	301527	40455	4797	640	8684	1077	30191	5271	160720	20514	29304	3997	67831	8956
PLOS-2	398162	50290	363046	44349	7939	1049	9395	1224	96345	5835	198216	23315	36045	4583	75106	8343
Total	752656	98457	664573	84804	12736	1689	18079	2301	126536	11106	358936	43829	65349	8580	142937	17299
U/T	98457/752656		84804/664573		1689/12736		2301/18079		11106/126536		43829/358936		8580/65349		17299/142937	
	Repertoire		**Out-of-frame**		**IGHJ1**		**IGHJ2**		**IGHJ3**		**IGHJ4**		**IGHJ5**		**IGHJ6**	
Sample	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**
PLOS-1	354494	48167	52967	7712	1034	155	1220	166	3003	568	27883	3857	6604	986	13223	1980
PLOS-2	398162	50290	35116	5941	837	164	852	151	2833	585	17573	2929	3907	712	9114	1400
Total	752656	98457	88083	13653	1871	319	2072	317	5836	1153	45456	6786	10511	1698	22337	3380
U/T	98457/752656		13653/88083		319/1871		317/2072		1153/5836		6786/45456		1698/10511		3380/22337	

Note: U/T represents the ratio of unique to total sequences.

Supplementary Table 3. The sequence data of 6 IGHJ gene families in the in frame and out of frame BCR-H repertoire from IMGT data

Unique sequence	**IGHJ1**	**IGHJ2**	**IGHJ3**	**IGHJ4**	**IGHJ5**	**IGHJ6**
9340	245	270	1607	4118	942	2158

Note: U/T represents the ratio of unique to total sequences.

Supplementary Table 4. The sequence data of 6 IGHJ gene families in the in-frame and out-of-frame BCR-H repertoire (IgM & IgG) from volunteers before and after immunization with the HBV vaccine

IgM	**Repertoire**		**In-frame**		**IGHJ1**		**IGHJ2**		**IGHJ3**		**IGHJ4**		**IGHJ5**		**IGHJ6**		**Out-of-frame**		**IGHJ1**		**IGHJ2**		**IGHJ3**		**IGHJ4**		**IGHJ5**		**IGHJ6**	
Sample	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**
V1-QM	689	569	564	447	8	7	16	13	73	51	290	229	58	51	119	94	125	123	3	3	2	2	30	28	65	65	14	14	11	11
V1-HM	2275	1853	1904	1497	56	30	54	41	225	181	1048	821	223	180	298	244	371	356	22	20	19	14	51	50	169	166	67	64	43	42
V2-QM	994	678	834	524	15	11	22	20	113	71	478	264	109	71	97	87	160	154	7	7	7	7	33	32	74	70	32	31	7	7
V2-HM	684	619	563	499	14	12	20	19	99	84	249	223	56	53	125	108	121	120	3	3	7	7	16	15	65	65	17	17	13	13
V3-QM	1500	1140	1193	849	27	20	66	46	177	123	624	445	127	86	172	129	307	291	15	15	12	12	71	63	128	122	52	50	29	29
V3-HM	1437	869	1227	699	14	13	26	22	547	116	400	335	89	79	151	134	210	170	11	11	2	2	81	46	69	65	26	25	21	21
Total	7579	5728	6285	4515	134	93	204	161	1234	626	3089	2317	662	520	962	796	1294	1214	61	59	49	44	282	234	570	553	208	201	124	123
U/T	5728/7579		4515/6285		93/134		161/204		626/1234		2317/3089		520/662		796/962		1214/1294		59/61		44/49		234/282		553/570		201/208		123/124	
IgG	**Repertoire**		**In-frame**		**IGHJ1**		**IGHJ2**		**IGHJ3**		**IGHJ4**		**IGHJ5**		**IGHJ6**		**Out-of-frame**		**IGHJ1**		**IGHJ2**		**IGHJ3**		**IGHJ4**		**IGHJ5**		**IGHJ6**	
Sample	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**
V1-QG	1249	736	998	510	17	14	20	15	94	66	630	272	118	66	119	77	251	226	5	5	6	6	23	22	122	112	75	62	20	19
V1-HG	1239	946	1007	728	18	17	33	25	70	57	584	403	146	107	156	119	232	218	12	12	11	11	26	25	105	95	49	46	29	29
V2-QG	418	305	342	235	9	9	5	4	30	28	202	120	43	32	53	42	76	70	3	3	0	0	9	9	42	38	12	12	10	8
V2-HG	771	534	646	414	12	10	23	15	89	52	318	218	95	57	109	62	125	120	4	4	1	1	23	21	63	61	24	23	10	10
V3-QG	662	390	512	252	10	7	25	13	52	37	346	138	36	25	43	32	150	138	7	7	10	10	23	22	81	71	18	18	11	10
V3-HG	1858	681	1549	443	7	5	21	11	106	47	1150	264	101	54	164	62	309	238	5	5	3	3	35	29	184	132	54	43	28	26
Total	6197	3592	5054	2582	73	62	127	83	441	287	3230	1415	539	341	644	394	1143	1010	36	36	31	31	139	128	597	509	232	204	108	102
U/T	3592/6197		2582/5054		62/37		83/127		287/441		1415/3230		341/539		394/644		1010/1143		36/36		32/32		128/139		109/597		204/232		102/108	

Note: U/T represents the ratio of unique to total sequences.

Supplementary Table 5. The sequence data of 6 IGHJ gene families in the in-frame and out-of-frame BCR-H repertoire from SLE volunteers

	Repertoire		**In-frame**		**IGHJ1**		**IGHJ2**		**IGHJ3**		**IGHJ4**		**IGHJ5**		**IGHJ6**	
Sample	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**
S1-1	570721	16712	511377	14846	8109	336	24468	434	46695	1894	249093	7768	42177	1559	140835	2855
S1-2	293305	7458	267141	6732	4740	165	14792	228	24525	811	127812	3510	24181	773	71091	1245
S1-3	175479	4657	158713	4175	2976	105	8357	133	16310	527	78112	2102	14700	435	38258	874
S2-1	547455	32930	482878	28856	7080	601	26120	907	54027	4321	207334	13558	41951	3018	146366	6452
S2-2	614082	33199	540830	29027	7944	587	28538	941	62200	4451	234612	13873	45935	2993	161601	6182
S2-3	542654	25268	479320	22069	7361	481	26428	727	49663	3243	208396	10594	42376	2372	145096	4653
Total	2201042	94956	1960939	83636	30849	1794	102275	2643	203757	12004	896963	40811	168944	8778	558151	17608
U/T	94956/2201042		83636/1960939		1794/30849		1643/102275		12004/203757		40811/896963		8778/168944		17608/558151	
	Repertoire		**Out-of-frame**		**IGHJ1**		**IGHJ2**		**IGHJ3**		**IGHJ4**		**IGHJ5**		**IGHJ6**	
Sample	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**
S1-1	570721	16712	59344	1866	1490	62	2506	48	3218	131	25491	903	5718	227	20921	495
S1-2	293305	7458	26164	726	790	24	955	19	1370	59	10597	341	2582	89	9870	194
S1-3	175479	4657	16766	482	381	14	603	16	1738	47	7769	223	1213	58	5062	124
S2-1	547455	32930	64577	4074	1515	119	2985	93	4769	384	28238	1981	6318	523	20752	974
S2-2	614082	33199	73252	4172	1185	104	2394	91	5775	409	33766	2100	6563	494	23569	974
S2-3	542654	25268	63334	3199	878	68	1892	65	4459	304	29474	1588	5430	390	21201	784
Total	2201042	94956	240103	11320	5361	323	9443	267	16870	1030	105861	5548	22394	1391	80174	2761
U/T	94956/2201042		11320/240103		323/5361		267/9443		1030/16870		5548/105861		1391/22394		2761/80174	

Note: U/T represents the ratio of unique to total sequences.

Supplementary Table 6. The sequence data of 6 IGHJ gene families in the in-frame and out-of-frame BCR-H repertoire from breast cancer volunteers

	Repertoire		In-frame		IGHJ1		IGHJ2		IGHJ3		IGHJ4		IGHJ5		IGHJ6	
Sample	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**
B3-1	20155	2504	16665	2087	256	33	713	56	1400	208	8093	1052	1474	208	3985	445
B2-1	161983	19745	138798	16723	2001	241	6790	520	11738	1679	64510	7828	12092	1689	35977	4066
B1-1	23214	4559	20083	3902	248	54	613	89	1117	250	8867	1703	2222	487	6150	1156
B3-2	70031	7744	59316	6536	966	113	2555	196	4491	611	28080	3181	4951	615	15478	1519
B2-2	86602	8129	72850	6806	980	96	2896	178	6950	793	32509	3096	6078	684	21083	1719
B1-2	31795	5527	27230	4724	334	53	723	88	1369	284	12143	2091	2917	582	8362	1386
B3-3	145504	15503	122436	12971	1716	201	5579	375	9576	1262	56084	6004	10698	1348	33650	3279
B2-3	63404	9530	53985	8051	997	115	2309	231	4564	796	24077	3636	5249	857	14764	2107
B1-3	31184	4950	26916	4252	336	50	917	103	1349	275	11661	1863	2927	513	7925	1176
Total	633872	78191	538279	66052	7834	956	23095	1836	42554	6158	246024	30454	48608	6983	147374	16853
U/T	78191/633872		66052/538279		956/7834		1836/23095		6158/42554		3054/246024		6983/48608		16853/147374	
	Repertoire		**Out-of-frame**		**IGHJ1**		**IGHJ2**		**IGHJ3**		**IGHJ4**		**IGHJ5**		**IGHJ6**	
Sample	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**
B3-1	20155	2504	3490	417	54	7	194	12	224	38	1503	186	408	45	870	106
B2-1	161983	19745	23185	3022	405	54	658	50	1232	179	9976	1327	2432	373	7114	853
B1-1	23214	4559	3131	657	84	15	67	13	120	26	1269	250	344	83	995	221
B3-2	70031	7744	10715	1208	161	21	379	26	507	69	4919	556	1136	149	2884	305
B2-2	86602	8129	13752	1323	120	13	303	22	822	86	6037	586	1476	163	4151	379
B1-2	31795	5527	4565	803	143	19	184	14	200	44	1629	302	476	103	1524	260
B3-3	145504	15503	23068	2532	366	43	615	43	1139	137	10786	1187	2742	310	5525	612
B2-3	63404	9530	9419	1479	155	21	285	25	536	91	4136	645	817	159	2858	456
B1-3	31184	4950	4268	698	47	9	80	8	182	32	1600	247	519	102	1573	253
Total	633872	78191	95593	12139	1535	202	2765	213	4962	702	41855	5286	10350	1487	27494	3445
U/T	78191/633872		12139/95593		202/1535		213/2765		702/4962		5286/41855		1487/10350		3445/27494	

Note: U/T represents the ratio of unique to total sequences.

Supplementary Table 7. The sequence data of 6 IGHJ gene families in the in-frame and out-of-frame BCR-H repertoire from volunteers with a high titer of HbsAb

	Repertoire		**In-frame**		**IGHJ1**		**IGHJ2**		**IGHJ3**		**IGHJ4**		**IGHJ5**		**IGHJ6**	
Sample	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**
HBsAg-1	763366	20379	369481	11070	2109	97	15192	297	19605	993	133101	4456	36486	1412	152213	3415
HBsAg-2	889478	32277	409468	17329	3473	207	14856	376	21264	1527	167115	7928	34721	1869	156580	4843
Total	1652844	52656	778949	28399	5582	304	30048	673	40869	2520	300216	12384	71207	3281	308793	8258
U/T	52656/1652844		28399/778949		304/5582		673/30048		2520/40869		12384/300216		3281/71207		8258/308793	
	Repertoire		**Out-of-frame**		**IGHJ1**		**IGHJ2**		**IGHJ3**		**IGHJ4**		**IGHJ5**		**IGHJ6**	
Sample	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**	**total**	**unique**
HBsAg-1	763366	20379	393885	9309	8015	242	20156	228	15418	647	154116	3686	46467	1497	131032	2577
HBsAg-2	889478	32277	480010	14948	8564	392	19155	355	19749	971	173488	5859	55282	2335	176431	4222
Total	1652844	52656	873895	24257	16579	634	39311	583	35167	1618	327604	9545	101749	3832	307463	10006
U/T	52656/1652844		24257/873895		634/16579		583/39311		1618/35167		9545/327604		3832/101749		6799/307463	

Note: U/T represents the ratio of unique to total sequences.

Supplementary Table 8. The frequency of six IGHJ gene families in BCR-H repertoire from 19 publishing article

The Title of Paper	**Object and number**	**Sample**	**BCR-H sequences (group)**	**The usage and distribution of IGHJ1&2&3&4&5&6**	**Sources of literature**
[1]Direct measurement of B-cell receptor repertoire's composition and variation in systemic lupus erythematosus (SLE)	10 SLE patients 6 heathy controls	PBMC (DNA)	(1)SLE patients (2)heathy controls	Figure 3 (C) (2 groups) IGHJ4 >IGHJ6 > IGHJ5 > GHJ3 >IGHJ2 >IGHJ1	[29]
[2]Clonal Characteristics of Circulating B Lymphocyte Repertoire in Primary Biliary Cholangitis（PBC）	43 PBC patients 34 healthy volunteers	PBMC (RNA)	(1)PBC patients (2)healthy volunteers	Figure 3 (B) (overlapping clones) (2 groups) IGHJ4 > IGHJ3>IGHJ6 > GHJ5 >IGHJ2 >IGHJ1	[30]
[3]Transitional B Cells in Early Human B Cell Development - Time to Revisit the Paradigm?	19 healthy adult donors (aged 24–86 years)	Sorting cells from Peripheral blood and Bone marrow	(1)Pre-B; (2)Immature B (3)Transitional B ; (4)Naïve B	Figure 2 (A) (4 groups) IGHJ4 >IGHJ6 >IGHJ3 >IGHJ5 >IGHJ2≈IGHJ1	[31]
[4]High-Throughput Sequencing Reveals Immunological Characteristics of the TRB-/IgH-CDR3 Region of Umbilical Cord Blood	20 healthy adults; 56 pregnant women 40 newborns	Umbilical Cord Blood Peripheral blood	(1)Newborns; (2)Pregnant women; (3)adults	Figure 2 (F) (3 groups) IGHJ4 >IGHJ6 >IGHJ3 >IGHJ5 >IGHJ2 >IGHJ1	[32]
[5] Characterization of the B Cell Receptor Repertoire in the Intestinal Mucosa and of Tumor-Infiltrating Lymphocytes in Colorectal Adenoma and Carcinoma (CRC)	6 healthy controls 4 AD patients 6 CRC.patients	Biopsies (RNA)	(1) healthy controls;(2)AD patients and CRC patients	Figure 6 (E) (2 groups) Most of V pair to J were IGHJ4&IGHJ6	[33]
[6]High-Throughput Single-Cell Analysis of B Cell Receptor Usage among Autoantigen-Specific Plasma Cells in Celiac Disease (CD)	10 CD patients (Biopsies): 8 untreated consuming a normal diet; 2 treated consuming a gluten-free diet	Sorting Cells from Biopsies (RNA)	Celiac disease (CD) patients (1)TG2+ ; (2) TG2-	Figure 1 (E) (2 groups) IGHJ4 >IGHJ6 > IGHJ5 >GHJ3 >IGHJ2 >IGHJ1	[34]
[7]The Human Thymus Is Enriched for Autoreactive B Cells	Thymus material was obtained from three children requiring surgery for congenital heart disease (2 months and 1.5 years of age); fetal BM samples was obtained from elective abortions (three donors); Pediatric BM samples were obtained from three children (aged 2–6years) [51]	Single-cell sorting from tissue (RNA)	(1)fetal BM; (2)pediatric BM;(3)pediatric thymus	Figure 2 (C) (3 groups) Fetal BM: IGHJ4 > IGHJ2> IGHJ3 > GHJ5>IGHJ6≈IGHJ1 pediatric BM: IGHJ4 > IGHJ3 >IGHJ5 ≈IGHJ6≈IGHJ2 >IGHJ1 pediatric thymus: IGHJ4 >IGHJ6 >IGHJ3 >IGHJ5 >IGHJ2 >IGHJ1	[35]
[8]Unlike in Children with Allergic Asthma, IgE Transcripts from Preschool Children with Atopic Dermatitis Display Signs of Superantigen-Driven Activation	Five preschool children with atopic dermatitis	Peripheral blood (RNA)	IgE(IgM control) : (1)atopic dermatitis ; (2)allerjic asthma	Figure 1 (C) (2 groups) IGHJ4 >IGHJ5 >IGHJ6 >IGHJ3 >IGHJ2 >IGHJ1	[36]
[9]IMonitor: A Robust Pipeline for TCR and BCR Repertoire Analysis	Samples of peripheral blood from 2 healthy human donors (H-H-1, H-B-1)	Peripheral blood (RNA)	2 healthy donors	Figure 4 (k) (1 group) IGHJ4 >IGHJ6 >IGHJ5 >IGHJ2 >IGHJ3> IGHJ1	[37]
[10]The normal IGHV1-69-derived B-cell repertoire	3 healthy persons: D1 (69 years); D2 (69	Peripheral blood	IGHJ use in normal B cells with	Figure 1. (2 groups)	[38]

contains stereotypic patterns characteristic of unmutated CLL	years) D3 (51 years); and age-matched to that of patients with CLL [52]	(RNA)	IGHV1-69-DJ-C rearrangements:(1)previous study (n=26);(2)present study (n=72)	IGHJ4 >IGHJ6 > IGHJ3> IGHJ5 >IGHJ2>IGHJ1	
[11]Antibody V(h) repertoire differences between resolving and chronically evolving hepatitis C virus infections	7 healthy donors (HD); 6 patients (acute HCV infection, spontaneous resolvers, SR) 9 patients (chronic HCV infection, chronically evolving, CE)	Sorting cells (naive B cell clones and naive B cell clones) from Peripheral blood (DNA)	(1) healthy donors; (2) acute HCV infection, spontaneous resolvers; (3) chronic HCV infection, chronically evolving	Figure 1 (E F): (3 groups) IGHJ4 >IGHJ6 > IGHJ3 >IGHJ5>IGHJ1 （No IGHJ2）	[39]
[12]Antibody repertoires in humanized NOD- scid-IL2Rgamma(null) mice and human B cells reveals human like diversification and tolerance checkpoints in the mouse	Two healthy females humanized mouse spleens	human PBMCs; naive or total B cells from humanized mouse spleen; immature B cells pooled humanized mice immature B cells (RNA)	(1)Hu PBC-1; (2) Hu PBC-2 ; (3)HuMs-1NSpl; (4) HuMs-2NSpl; (5) HuMs-3TSpl ; (6)HuMs-ImmB	Figure 2 (C) (6 groups) IGHJ4 >IGHJ6 > GHJ3 >IGHJ5 >IGHJ2 >IGHJ1	[40]

[13]Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations	An African-American female baby
	a Caucasian male baby
		Two cord blood (RNA)	two babies IG:
			(1) CB1 (productive);
			(2) CB1 (unproductive);
			(3) CB2 (productive);
			(4) CB2 (unproductive)
				Figure 1 (C) (4 groups)
				IGHJ4 > IGHJ3 >IGHJ6 >IGHJ5 >IGHJ2 >IGHJ1
					[41]

[14]Tissue-specific expressed antibody variable gene repertoires	Healthy human subjects were obtained from a commercial source (Clontech). 39 peripheral leukocytes, 56 bone marrow, 15 small intestines, 13 lung, 7 stomach, 42 lymph node, 34 tonsil, 12 spleen and 25 thymuses	peripheral blood, bone marrow, mucosal tissues; lymph tissues (RNA)	(1) peripheral blood; (2) bone marrow; mucosal tissues (lung, small intestine, stomach); (3) lymphoid tissues (lymph node, tonsil, spleen and thymus).	Figure 1 (3 groups) IGHJ4 >IGHJ6 > GHJ3 >IGHJ1>IGHJ2 (figure 1 No IGHJ5)	[42]
[15]Differences in the composition of the human antibody repertoire by B cell subsets in the blood	One healthy female subject (age 56)	Sorting cells (RNA)	(1)immature;(2)transitional;(3)Mature;(4)Memory;IgD+IgM; (5)memory IgD−IgM;(6) plasmacytes ;IgM;(7)Memory;IgD-IgG;(8)plasmacytes IgG	Figure 7 (8 groups） IGHJ4 >IGHJ6 > IGHJ5 ≈ IGHJ3 >IGHJ2>IGHJ1	[43]
[16]Prevalence and gene characteristics of antibodies with cofactor-induced HIV-1 specificity	Human immunodeficiency virus immune -globulin (HIVIg) was obtained through the NIH	The repertoire of human antibodies (gp120)	(1)Sensitive Abs (2)Non-Sensitive Abs	Figure 6 (A [2]) (2 groups) IGHJ4 >IGHJ6 > IGHJ3 ≈ IGHJ5 >IGHJ1>IGHJ2	[44]

[17]Age-related aspects of human IgM+ B cell heterogeneity	Peripheral blood mononuclear cells were isolated from a total of 14 young (21–45 years) and 16 old (62–87 years) healthy volunteers.	Sorting cells from PBMCs（RNA）	(1)young naive;(2)old naive;(3)young IgM memory;(4)old IgM memory;(5)young IgM only CD27–;(6)old IgM only CD27–;(7)young IgM onlyCD27+;(8) old IgM onlyCD27+	Figure 3 (A) (8 groups) IGHJ4 >IGHJ6 >IGHJ3 ≈IGHJ5 >IGHJ2>IGHJ1	[45]
[18]High-throughput sequencing of IgG B-cell receptors reveals frequent usage of the rearranged IGHV4-28/IGHJ4 gene in primary immune thrombocytopenia	Eleven adult chronic Primary immune thrombocytopenia patients and nine volunteer donors	PBMCs（RNA）	IgG-BCRs (1)Primary immune thrombocytopenia; (2)volunteer donors	Figure 1 (2 groups) IGHJ4 >IGHJ6 > IGHJ5 > IGHJ3 >IGHJ2>IGHJ1	[46]
[19]IgM repertoire biodiversity is reduced in HIV-1 infection and systemic lupus erythematosus	Sixteen individuals: 4 healthy controls, 4 subjects with SLE, 4 therapy-naïve HIV, and 4 receiving combination antiretroviral therapy (cART) HIVTx	PBMCs（RNA）	IgM -BCRs (454-deep pyrosequencing)	Figure 7 (C or D) IGHJ4 >IGHJ3>IGHJ6>IGHJ5 >IGHJ2>IGHJ1	[47]

Supplementary Table 9. The composition and characteristics of human J-NONAMER--J-SPACER--J-HEPTAMER (9-23-7) recombination signal sequence (RSS) and J-REGION Subsequence & AA

Names	F	Accession number	Location	J-NONAMER Subsequence	Location	J-SPACER Subsequence	Location	J-HEPTAME Subsequence	J-REGION Location	J-REGION Subsequence & AA
IGHJ1*01	**F**	X97051	87524..875	ggtttctgt	87533..87554	agccccctggctcagggctgact [22]	87555..87561	caccgtg	87562..87613	gctgaatacttccagcactggggccagggcaccctggtcaccgtctcctcag

			32			(3a+4t+8c+7g)				AEYFQHWGQGTLVTVSS [17]
		X86356	247..255		256..277		278..284		285..336	
IGHJ2*01	**F**	X97051	87731..87739	tgttttgt	87740..87761	atggggagaagcaggagggcaga [22] (8a+1t+2c+11g)	87762..87768	ggctgtg	87769..87821	ctactggtacttcgatctctggggccgtggcaccctggtcactgtctcctcag YWYFDLWGRGTLVTVSS [17]
		X86356	454..462		463..484		485..491		492-544	
IGHJ3*01	**F**	M25625	31..39	ggtttgtgt	40..62	ctgggtctaggaacggactgtgt [23] (4a+5t+4c+9g)	63..69	ccctgtg	70..119	tgatgcttttgatgtctggggccaagggacaatggtcaccgtctcttcag DAFDVWGQGTMVTVSS [16]
IGHJ3*02		X97051	88345..88353		88354..88376	ctgggcaggaacagggactgtgt [23] (5a+4t+4c+10g)	88377..88383		88384..88433	tgatgcttttgatatctggggccaagggacaatggtcaccgtctcttcag DAFDIWGQGTMVTVSS [16]
		X86356	1068..1076		1077..1099		1100..1106		1107..1156	
IGHJ4*01	**F**	J00256	1873..1881	ggtttttgt	1882..1904	gcaccccttaatggggcctccca [23] (4a+4t+10c+5g)	1905..1911	caatgtg	1912..1959	actactttgactactggggccaaggaaccctggtcaccgtctcctcag YFDYWGQGTLVTVSS [15]
IGHJ4*02		X97051	88719..88727		88728..88750		88751..88757		88758..88805	
		X86356	1441..1449		1450..1472		1473..1479		1480..1527	
IGHJ4*03		M25625	407..415		416..438		439..445		446..493	
IGHJ5*01	**F**	M25625	855..863	gttctttgt	864..884	cggggtctggcattgttgtca [21] (2a+7t+4c+8g)	885..891	caatgtg	892..942	acaactggttcgactcctggggccaaggaaccctggtcaccgtctcctcag NWFDPWGQGTLVTVSS [16]
IGHJ5*02		X97051	89118..89126	gttcttgcc	89127..89148	tggggtcctggcattgttgtca [22] (2a+8t+4c+8g)	89149..89155		89156..89206	acaactggttcgacccctggggccagggaaccctggtcaccgtctcctcag NWFDSWGQGTLVTVSS [16]
		X86356	1840..1848		1849..1870		1871..1877		1878..1928	
IGHJ6*01	**F**	J00256	2909..2917	ggtttttgt	2918..2939	ggggtgaggatggacattctgc [22] (4a+5t+3c+10g)	2940..2946	cattgtg	2947..3009	attactactactactacggtatggacgtctgggggcaagggaccacggtcac cgtctcctcag YYYYYGMDVWGQGTTVTVSS [20]
IGHJ6*02		M25625	1442.22.1450		1451..1472	tgggtgaggatggacattctgc [22] (4a+6t+3c+9g)	1473..1479		1480..1542	
IGHJ6*03		X97051	89722..89730		89731..89752		89753..89759		89760-89760..89822	
		X86356	2444..2452		2453..2474	ggggtgaggatggacattctgc [22] (4a+5t+3c+10g)	2475..2481		2482..2543	attactactactactacggtatggacgtctggggccaagggaccacggtcacc gtctcctca YYYYYYMDVWGKGTTVTVSS [20]
IGHJ6*04		AJ879487	1..9		10..31		32..38		39..101	attactactactactacggtatggacgtctggggcaaagggaccacggtcacc gtctcctcag YYYYYGMDVWGKGTTVTVSS [20]

Supplementary Table 10. The composition of human IGHD heptamer, spacer and nonamer (7-12-9 RSSs)

Accession number	Allele or Gene	Location	D-HEPTAMER	Location	D-SPACER Subsequence	Location	D-NONAMER
X97051	IGHD1-1*01	33731..33737	caccgtg	33738..33749	agaaaaactgtg	33750..33758	tccaaaact
X97051	IGHD1-7*01	43317..43323	cactgtg	43324..43335	agaaaagcttcg	43336..43344	tccaaaacg
X97051	IGHD1-14*01	52585..52591	cactgtc	52592..52603	agaatagctacg	52604..52612	tcaaaaact
X97051	IGHD1-20*01	62032..62038	caccgtg	62039..62050	agaaaaactgtg	62051..62059	tccaaaact
X97051	IGHD1-26*01	72189..72195	cactgtg	72196..72207	agaaaagctatg	72208..72216	tccaaaact
X97051	IGHD2-2*02	36398..36404	cacagtg	36405..36416	acacagccccat	36417..36425	tccccaaagc
X97051	IGHD2-8*01	46013..46019	cacagtg	46020..46031	acacagccccat	46032..46040	tccccaaagc
X97051	IGHD2-15*01	55266..55272	cacagtg	55273..55284	acacagacccat	55285..55293	tccccaaagc
X97051	IGHD2-21*02	64672..64678	cacagtg	64679..64690	acacaaccccat	64691..64699	tcctaaagc
X97051	IGHD3-3*01	38865..38871	cacagtg	38872..38883	tcacagagtcca	38884..38892	tcaaaaacc
X97051	IGHD3-9*01	48543..48549	cacagtg	48550..48561	tcacagagtcca	48562..48570	tcaaaaacc
X97051	IGHD3-10*01	48727..48733	cacagtg	48734..48745	tcacagagtcca	48746..48754	tcaaaaacc
X97051	IGHD3-16*02	57589..57595	cacagca	57596..57607	tcacacggtcca	57608..57616	tcagaaacc
X97051	IGHD3-22*01	67192..67198	cacagtg	67199..67210	tcacagagtcca	67211..67219	tcaaaaact
X97051	IGHD4-4*01	40002..40008	cacagtg	40009..40020	atgaacccagca	40021..40029	gcaaaaact
X97051	IGHD4-11*01	49607..49613	catagtg	49614..49625	atgaacccagtg	49626..49634	gcaaaaact
X97051	IGHD4-17*01	58715..58721	cacagtg	58722..58733	atgaaactagca	58734..58742	gcaaaaact
X97051	IGHD4-23*01	68353..68359	cacagtg	68360..68371	atgaaaccagca	68372..68380	gcaaaaact
X97051	IGHD5-5*01	40967..40973	cacagtg	40974..40985	gtgctgcccata	40986..40994	gcagcaacc
X97051	IGHD5-12*01	50574..50580	cacagtg	50581..50592	gtgccgcccata	50593..50601	gcagcaacc
X97051	IGHD5-18*01	59681..59687	cacagtg	59688..59699	gtgctgcccata	59700..59708	gcagcaacc
X97051	IGHD5-24*01	69320..69326	cacagtg	69327..69338	gtgccgcccata	69339..69347	gcagcaacc
X97051	IGHD6-6*01	42813..42819	cacagtg	42820..42831	acactcgccagg	42832..42840	ccagaaacc
X97051	IGHD6-13*01	52081..52087	cacagtg	52088..52099	acactcacccag	52100..52108	ccagaaacc
X97051	IGHD6-19*01	61524..61530	cacagtg	61531..61542	acactcgccagg	61543..61551	ccagaaacc
X97051	IGHD6-25*01	71684..71690	cacaatg	71691..71702	acactgggcagg	71703..71711	acagaaacc
X97051	IGHD7-27*01	87470..87476	cacagtg	87477..87488	attggcagctct	87489..87497	acaaaaacc

Supplementary Table 11. The pairing of human IGHD heptamer-spacer-nonamer (7-12-9）RSSs and IGHJ nonamer-Spacer-heptamer (9-23-7) RSSs

Allele or Gene	D-HEPTAMER	IGHJ1(caccgtg) to D-HEPTAMER	IGHJ2 (ggctgtg)to D-HEPTAMER	IGHJ3 (ccctgtg) to D-HEPTAMER	IGHJ6(cattgtg) to D-HEPTAMER	IGHJ4&IGH5 (caatgtg) to D-HEPTAMER	D-NONAMER	IGHJ1(Ggtttctgt) to D-NONAMER	IGHJ2(tgtttttgt)) to D-NONAMER	IGHJ3(Ggtttgtgt)) to D-NONAMER	IGHJ5(gttctttgt)) to D-NONAMER	IGHJ4&IGHJ6 -(Ggtttttgt)) to D-NONAMER
IGHD1-1*01	Caccgtg	caccgtg	caccgtg	caccgtg	caccgtg	caccgtg	tccaaaact	tccaaaact	tccaaaact	tccaaaact	tccaaaact	tccaaaact
IGHD1-7*01	cactgtg	cactgtg	cactgtg	cactgtg	cactgtg	cactgtg	tccaaaacg	tccaaaacg	tccaaaacg	tccaaaacg	tccaaaacg	tccaaaacg
IGHD1-14*01	cactgtc	cactgtc	cactgtc	cactgtc	cactgtc	cactgtc	tcaaaaact	tcaaaaact	tcaaaaact	tcaaaaact	tcaaaaact	tcaaaaact
IGHD1-20*01	caccgtg	caccgtg	caccgtg	caccgtg	caccgtg	caccgtg	tccaaaact	tccaaaact	tccaaaact	tccaaaact	tccaaaact	tccaaaact
IGHD1-26*01	cactgtg	cactgtg	cactgtg	cactgtg	cactgtg	cactgtg	tccaaaact	tccaaaact	tccaaaact	tccaaaact	tccaaaact	tccaaaact
IGHD2-2*02	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	tcccaaagc	tcccaaagc	tcccaaagc	tcccaaagc	tcccaaagc	tcccaaagc
IGHD2-8*01	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	tcccaaagc	tcccaaagc	tcccaaagc	tcccaaagc	tcccaaagc	tcccaaagc
IGHD2-15*01	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	tcccaaagc	tcccaaagc	tcccaaagc	tcccaaagc	tcccaaagc	tcccaaagc
IGHD2-21*02	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	tcctaaagc	tcctaaagc	tcctaaagc	tcctaaagc	tcctaaagc	tcctaaagc
IGHD3-3*01	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	tcaaaaacc	tcaaaaacc	tcaaaaacc	tcaaaaacc	tcaaaaacc	tcaaaaacc
IGHD3-9*01	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	tcaaaaacc	tcaaaaacc	tcaaaaacc	tcaaaaacc	tcaaaaacc	tcaaaaacc
IGHD3-10*01	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	tcaaaaacc	tcaaaaacc	tcaaaaacc	tcaaaaacc	tcaaaaacc	tcaaaaacc
IGHD3-16*02	cacagca	cacagca	cacagca	cacagca	cacagca	cacagca	tcagaaacc	tcagaaacc	tcagaaacc	tcagaaacc	tcagaaacc	tcagaaacc
IGHD3-22*01	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	tcaaaaact	tcaaaaact	tcaaaaact	tcaaaaact	tcaaaaact	tcaaaaact
IGHD4-4*01	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	gcaaaaact	gcaaaaact	gcaaaaact	gcaaaaact	gcaaaaact	gcaaaaact
IGHD4-11*01	catagtg	catagtg	catagtg	catagtg	catagtg	catagtg	gcaaaaact	gcaaaaact	gcaaaaact	gcaaaaact	gcaaaaact	gcaaaaact
IGHD4-17*01	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	gcaaaaact	gcaaaaact	gcaaaaact	gcaaaaact	gcaaaaact	gcaaaaact
IGHD4-23*01	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	gcaaaaact	gcaaaaact	gcaaaaact	gcaaaaact	gcaaaaact	gcaaaaact
IGHD5-5*01	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	gcagcaacc	gcagcaacc	gcagcaacc	gcagcaacc	gcagcaacc	gcagcaacc
IGHD5-12*01	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	gcagcaacc	gcagcaacc	gcagcaacc	gcagcaacc	gcagcaacc	gcagcaacc
IGHD5-18*01	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	gcagcaacc	gcagcaacc	gcagcaacc	gcagcaacc	gcagcaacc	gcagcaacc
IGHD5-24*01	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	gcagcaacc	gcagcaacc	gcagcaacc	gcagcaacc	gcagcaacc	gcagcaacc
IGHD6-6*01	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	ccagaaacc	ccagaaacc	ccagaaacc	ccagaaacc	ccagaaacc	ccagaaacc
IGHD6-13*01	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	ccagaaacc	ccagaaacc	ccagaaacc	ccagaaacc	ccagaaacc	ccagaaacc
IGHD6-19*01	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	ccagaaacc	ccagaaacc	ccagaaacc	ccagaaacc	ccagaaacc	ccagaaacc
IGHD6-25*01	cacaatg	cacaatg	cacaatg	cacaatg	cacaatg	cacaatg	acagaaacc	acagaaacc	acagaaacc	acagaaacc	acagaaacc	acagaaacc
IGHD7-27*01	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	cacagtg	acaaaaacc	acaaaaacc	acaaaaacc	acaaaaacc	acaaaaacc	acaaaaacc
		32/189	60/189	35/189	35/189	36/189		69/243	82/243	76/243	105/243	64/243