DR. ZHUI  TU (Orcid ID : 0000-0003-3187-0132)

Article type      : Original Article

# Landscape of Variable Domain of Heavy-chain-only Antibody Repertoire from Alpaca

Zhui Tu [1,2,3,4], Xiaoqiang Huang [2], Jinheng Fu [1, 5], Na Hu [1, 4, 6], Wei Zheng [2], Yanping Li [1, 4, 5,]**Error! Bookmark not defined.**, Yang Zhang [2, 3,] **Error! Bookmark not defined.**

*[1] State Key Laboratory of Food Science and Technology, Nanchang University, Nanchang, China, [2] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA, [3] Department of Biological Chemistry, University of Michigan, Ann Arbor, MI, USA, [4] Jiangxi Province Key Laboratory of Modern Analytical Science, Nanchang University, Nanchang, China, [5] Jiangxi-OAI Joint Research Institution, Nanchang University, Nanchang, China, and [6] Maternal and Child Medical Research Institute, Shenzhen Maternity and Child Healthcare Hospital, Southern Medical University, Shenzhen, China*

**Running title:** *Tu Z et al / Analysis of Immune Repertoire in Alpaca*

Abbreviations: HCAbs,  heavy-chain-only antibodies; VHHs, the variable regions of heavy chain of HCAbs; HTS, high-throughput sequencing; CDR, complementary determining region; GSSPs, germline specific scoring profiles; AAs, amino acids; SR, substitution rate; ASR, average substitution rate; SHM, somatic hypermutation; PBMCs, peripheral blood mononuclear cells; PCR, polymerase chain reaction; MSAs, multiple sequence alignments

*Correspondence: Yanping Li, Jiangxi-OAI Joint Research Institute, Nanchang University, 235 East Nanjing Road, Nanchang, Jiangxi, 330047, China. Email: liyanping@ncu.edu.cn (Li Y); and Yang Zhang, Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109-2218, USA. Email: zhng@umich.edu (Zhang Y). Senior author: Yang Zhang

## Summary

Heavy-chain-only antibodies (HCAbs), which are devoid of light chains, have been found naturally occurring in various species including camelids and cartilaginous fish. Due to their high thermostability, refoldability, and capacity for cell permeation, the variable regions of heavy chain of HCAbs (VHHs) have been widely used in diagnosis, bio-image, food safety, and therapeutics. Most immunogenetic and functional studies of HCAbs are based on case studies or a limited number of low throughput sequencing data. A complete picture derived from more abundant high-throughput sequencing (HTS) data can help us gain deeper insights. Thus, we cloned and sequenced the full-length coding region of VHHs in Alpaca (*Vicugna pacos*) via HTS in this study. A new pipeline was developed to conduct an in-depth analysis of the HCAbs repertoires. Various critical features, including the length distribution of complementary determining region 3 (CDR3), V(D)J usage, VJ pairing, germline specific mutation rate, and germline specific scoring profiles (GSSPs), were systematically characterized. The quantitative data show that V(D)J usage and VHHs recombination are highly biased. Interestingly, we found that the average CDR3 length of classical VHHs is longer than that of non-classical ones, whereas the mutation rates are similar in both kinds of VHHs. Finally, GSSPs were built to quantitatively describe and compare sequences that originate from each VJ pairs. Overall, this study presents a comprehensive landscape of the HCAbs repertoire, which can provide useful guidance for the modeling of somatic hypermutation and the design of novel functional VHHs or VHH repertoires via evolutionary profiles.

**Keywords:** High-throughput sequencing; Nanobody; Immune repertoire; Antibody diversity; Protein design

## Introduction

The antigen binding domain of functional heavy-chain-only antibodies (HCAbs) discovered in camelids and sharks is composed of a single variable domain.[1, 2] The variable regions of heavy chain of HCAbs (VHHs), also known as nanobodies, have attracted growing interest in various applications, as they are more soluble and stable than canonical antibodies (VHs).[3-6] In camels, the ratio of HCAbs to total IgGs can reach more than 80%, which indicates that HCAbs play a significant role in immune protection.[7] However, it is obvious that the diversity of HCAbs is dramatically lower than that of canonical antibodies due to the lack of VH-VL combinational diversification. This raises a question of how HCAbs can compete with canonical antibodies. Several hypotheses and observations have been proposed over the past decades to address the problem of diversity reduction inherent to HCAbs. One hypothesis is that the complementary determining region 3 (CDR3) of VHHs contains longer loops than canonical antibody VHs (18 amino acids versus 13 amino acids), which helps compensate for the lack of diversity.[8] Evidently, longer CDR3 length increases the paratope size, as well as the three-dimensional structural diversity and contact surface area with antigens.[9] Another explanation, inferred from a structural study that compared two independently generated anti-lysozyme nanobodies, is that the *in vivo* maturation and selection systems are strong enough to compensate for the decrease in the VHHs primary repertoire.[10]

High-throughput sequencing (HTS) technology enables scientists to evaluate millions of sequences in parallel, resulting in the collection of more complete and comprehensive information for target samples. This capability makes HTS suitable for the characterization of immune repertoires that are highly plastic and diverse. Although HTS is now routinely applied in the studies of human adaptive immunity,[11] vaccine development,[12] and diagnostic research,[13] only a few studies were tried on VHHs. Fridy *et al.* developed a pipeline combining HTS and proteomics to identify specific VHHs.[14] Similarly, Turner *et al.* demonstrated that HTS can be used as a complementary tool for phage-display bio-panning to rapidly obtain additional clones from an immune VHH library.[15] For the first time, Li *et al.* compared the repertoires of classical antibodies and HCAbs of *Bactrian* camels, with analysis data including CDR3 length distribution, mutation rate, characteristic amino acids, the distribution of cysteine codons, and the non-classical VHHs.[8] Nevertheless, the features of HCAbs, such as the germline usage and mutation preferences, still remain unknown. Like classical immunoglobulin (Ig) heavy-chains, VHHs are encoded by recombined V(D)J genes that are formed from sets of Variable (V), Diversity (D), and Joining (J) genes (IGHV, IGHD, IGHJ) on the genome. An in-depth analysis of the origination and mutation profiles of VHHs would help us to better understand the diversity of the HCAb repertoire, as well as the diversity compensation. Furthermore, appropriate interpretation of the information is important to guide the design of novel functional VHHs.[16, 17]

This study is mainly focused on the HCAb repertoire. First, the coding sequences of VHHs from long-hinge HCAbs ($IgG_2$) and short-hinge HCAbs ($IgG_3$) are amplified from the non-immunized and the antigen immunized antibody repertoires of *Vicugna pacos*, where full-length coding sequences of VHHs are obtained by an Illumina MiSeq System (2×300) under the paired-end module. Next, a new pipeline combined with multiple software tools is developed to characterize the diversity and evolutionary features of the VHHs, including CDR3 length distribution, V(D)J usage, VJ pairing, DJ pairing, germline specific mutation rate, and germline specific scoring profiles (GSSPs) (Fig. 1). Considering that the diversity of antibody repertoires is position, chain, and species-dependent,[18-20] comparative studies are also made on amino acid sequences derived from different germline genes.

## Materials and methods

*RNA extraction and reverse transcription*

Peripheral blood mononuclear cells (PBMCs) were separated from peripheral blood by Ficoll-1.077 (Sangon, Shanghai, China) gradient centrifugation, separately. Three naïve blood samples were collected from three non-immunized healthy male Alpaca (*Vicugna pacos*). To collect immunized blood samples, one donor was immunized by subcutaneous, lower-back injections every two weeks. Samples of fresh blood were collected 1 week after the fifth and seventh immunization. For each blood sample, RNA was purified from approximately $2\times10^7$ PBMCs using RNAprep Kit (Tiangen, Beijing, China), following the manufacturer's instruction. First-strand complementary DNA (cDNA) was synthesized with random hexamer primers using PrimeScript$^{TM}$ RT-PCR Kit (TAKARA, Dalian, China), and then stored at -80 °C.[21]

*Library construction and Illumina sequencing*

The VHH coding region was amplified from cDNA by a nested polymerase chain reaction (PCR) as described before.[22, 23] In brief, the variable region was first amplified by primers AlpVh-LD and AlpVHH, which anneal to the conserved region of the leading sequence and CH2 region, respectively. Next, the PCR products were diluted as a template for the second round of PCR, which employed primer pairs AlpVHH-F/AlpVHH-R1 and AlpVHH-F/AlpVHH-R2 to amplify coding sequences of short and long hinge heavy chain antibodies, respectively. The PCR products that encoded VHHs (~450 bp) were purified using TAKARA gel extraction kits (Dalian, China), and then subjected to Next-Generation Sequencing by the Beijing Genomics Institute (BGI) sequencing center. Sequences were generated with a MiSeq System using a 2×300 paired-end module.

*Basic data processing*

Adapter sequences were first checked and removed from the reads. Then, the reads that bases of "N" were greater than 10% or have > 50% bases with quality values ≤ 5 were discarded, resulting in 14.13×2 million paired-end reads. The pairwise reads were joined using the fastq-join tool (version 1.3.1).[24] The main parameters were the maximum difference percentage (8%) and the minimum base overlap (6 bp). Phylogenetic trees of V germline genes were built using MEGA version X.[25]

*V(D)J assignment and numbering*

The V(D)J germline gene sequences were obtained from the international ImMunoGeneTics information system (IMGT) antibody repertoire database.[26] The out-frame bases in the 3' end of J gene sequences were manually deleted, where the elaborated germline gene sequences were used to build an IgBLAST database. The resulting 5.85 million joined sequences were subjected to IgBLAST1.8.0 with default parameters.[27] The origin of each sequence, either from long-hinge or short-hinge IgGs, was identified by BLAST 2.7.1+ according to the E-value and sequence identity of the alignments.[28] The V(D)J germline genes on the top of the resulting list of IgBLAST were assigned to the sequences. An in-house Python script was used to analyze VHHs length distribution, CDR3 length distribution, V/D/J germline gene usage, VJ pairing, DJ pairing, and amino acid substitution. The IMGT numbering system was adopted for the coding sequences of VHHs.

*Construction and comparison of GSSPs*

The sequences were translated and aligned to all alleles of the gene. Sequences with more than one stop codon or no amino acid substitution were discarded. Sequences belonging to the same VJ germline gene were parsed from IgBLAST output to build multiple sequence alignments (MSAs); redundant amino acid sequences were removed in each MSAs. To improve the accuracy of sequence alignment, the V and J segments of each amino acid sequence in an MSA were re-aligned with corresponding IMGT numbered germline sequences using an in-house NW-align program (Y. Zhang, https://zhanglab.ccmb.med.umich.edu/NW-align/) before GSSP construction. The GSSPs were built and compared as described in previous work.[20] In brief, the MSAs whose number of sequences was greater than a given threshold (e.g. 100, 500, and 1000) were used to build GSSPs. were used to build GSSPs, respectively. A divergence matrix between GSSPs was calculated; each element in the matrix was the Jensen-Shannon divergence calculated between each pair of sequences from the MSAs. The R function *cmdscale* was used for multidimensional scaling and generating coordinates for plotting. The logo plot of MSA was drawn using a stand-alone version of WebLogo 3.6. [29]

*Calculation of substitution frequencies for the 20 AAs*

The GSSPs were used to calculate the substitution frequencies. The substitution rate (SR) from each GSSP is calculated by

$$SR = \frac{\sum_{i=1}^{N} f_i}{L \times N} \times 100\% \qquad 1)$$

where $f_i$ is the mutation frequency of sequence $i$ to the corresponding germline genes. $L$ is the length of the GSSP, and $N$ is the total sequences in the MSA. The average substitution rate (ASR) for the 20 AAs of a GSSP is calculated by

$$ASR_{(a,b)} = \frac{\sum_{i=1}^{L} fi_{(a,b)}}{f_{(a)} \times N} \times 100\% \qquad 2)$$

where $ASR_{(a,\ b)}$ is the average substitution rate of amino acid $a$ in germline gene substituted by observed amino acid $b$ in MSA, $fi_{(a,\ b)}$ is the frequency of amino acid $a$ in germline gene substituted by amino acid $b$ at the position $i$ of an MSA, $f_{(a)}$ is the frequency of amino acid $a$ in germline sequence, $L$ is the length of the MSA, and $N$ is the total sequences in the MSA.

*Statistical analysis*

To investigate the likelihood of pairing preference between germline segments, we used an *in silico* simulation protocol as described in a previous study.[30] Briefly, in each simulation, an equal number of real data sequences were constructed using the same individual frequencies of V, D, and J segments observed in the real data. After 2,000 simulation steps, the DJ and VJ pairing that appeared in each simulation were counted. The relative deviation (RD) of minimum, maximum, and real frequencies of each kind of pairing were calculated by

$$RD = \frac{x - \bar{x}}{\bar{x}} \times 100\% \qquad 3)$$

where $x$ is the minimum or maximum frequencies of simulation, or frequencies of real sequence data, and $\bar{x}$ is the average frequency of each pairing in the 2,000 simulation steps.

We used the function *spearmanr* in the Python module *scipy* to calculate the Spearman's Rank Correlation Coefficient to evaluate the statistical dependence of the germline usage, VJ and DJ pairings, and the substitution preference between samples.

## Results

*Sequence data filtration and formation*

A summary of the sequencing datasets processed in this study is shown in Table 1. The MiSeq sequencing of the non-immune and antigen-experienced HCAb repertoires yielded a total of 38.25×2 million reads. Since the sample Naïve-1 generated the most sequencing reads (14.13×2 million reads), it was used to build and test the pipeline. A number of 2,550,856 unique DNA sequences were subjected to IgBLAST to identify the germline gene origination of each sequence, after the redundant DNA sequences of the joined paired-end reads were removed. Both V and J germline genes are found in more than 97% of the non-redundant DNA sequences. Following these filtrations, a total of 2,490,298 unique DNA sequences with VJ assignment hits were used to determine the coding sequence (CDS) distribution, V(D)J usage, VJ pairing, and DJ pairing. Briefly, the CDS length distribution centers around 375 bp and follows an approximately normal distribution, where the maximum CDS length is 438 bp in the dataset (Fig. S1 in Supplementary Material). A number of 1,973,186 unique amino acid sequences deduced from this dataset were used to construct multiple sequence alignments (MSAs), to analyze CDR3 length distribution, and to calculate substitution rates and construct GSSPs. VHHs from long-hinge and short-hinge HCAbs were identified and analyzed for comparison.

*Germline gene usage*

Studies of canonical antibody repertoires have demonstrated that specific V, D, and J germline genes have very different frequencies in humans and mice.[30-33] Meanwhile, HCAbs and canonical IgGs in Alpaca (*Vicugna pacos*) genome have been shown to originate from the same IgH locus, which is composed of 88 V genes (including 4 pseudogenes), 8 D genes, and 7 J genes.[34] Here, we utilized the tool IgBLAST to determine the origination of V, D, and J of each clone. The 84 functional V genes, 8 D genes, and 7 J genes were employed to create a reference database for IgBLAST. The IgBLAST results showed that the V, D, and J segment usages have strong preferences for specific germline genes (Fig. 2).

The V segments of all clones were generated from the subgroups of IgHV3. The V segments IGHV3S65*01, IGHV3S3-3*01, and IGHV3S53*01 are used by more than 10% of all the clones, while the top 11 V germline genes are used by more than 95% of all the clones (Fig. 2A). All the 17 V germline genes, which contain at least two framework region 2 (FR2) hallmark residues, F37, E44, R45, and G47 in the Kabat numbering system,[35] are in a sub-cluster of IgHV3 (Fig. S2 in Supplementary Material). Germline genes from this sub-cluster contribute more than 85% of V gene usage (Fig. 2A). These hallmark residues are considered to be important for the solubility and stability of VHHs, as well as the VH-VL association of conventional VHs. A novel promiscuous class of VHHs that do not have any FR2 imprints was reported in Sanger sequencing studies.[36, 37] It is now clear that sequences that lack FR2 imprints are generated from other V germline genes, in which IGHV3S39*01, IGHV3S41*01, IGHV3S25*01, IGHV3-1*01, IGHV3S9*01, and IGHV3S1*01 constitute the top 6 contributors. These hallmark-free V segments are responsible for about 10% of V gene usage in the dataset.

The usage of D segments was relatively evenly distributed across the germline genes, where six out of eight D germline genes have above 10% usage (Fig. 2B). Similar to the V gene usage, the J germline gene usage was also highly biased (Fig. 2C). For instance, the germline gene IGHJ4*01 was used by two-thirds of the J segments. Since only a few sequences were assigned to IGHJ5*01 and IGHJ1*01 (0.15% and 0.09%, respectively), we manually checked the DNA and corresponding amino acid sequences. The IGHJ5*01 hits of J segments were correctly assigned by IgBLAST. However, due to the defects in the 3' sequences, all the IGHJ1*01 assignments were false positives, indicating that VHHs never use IGHJ1*01. These sequences were therefore discarded in the subsequent analyses.

*V(D)J recombination preferences*

VJ pairing data showed that more than 90% of the VJ pairs are composed of genes from the top 21 most used VJ germline gene combinations (Fig. 3A), indicating that VJ pairing is biased. Theoretically, the combination of VJ pairing should be much greater than 21, even though the V and J usage are highly biased toward specific germline genes. To evaluate whether V(D)J pairing exhibits bias, simulated antibody repertoires were employed to test statistical preference. As the V(D)J recombination occurs in two steps to assemble a complete variable region *in vivo*, we firstly analyzed the DJ pairing, and then the VJ pairing. Although most relative deviation of the real data was less than 100%, DJ pairing showed a preference (Fig. 3B). The VJ pairing results indicated a stronger bias, as the relative deviations of the 6 types of VJ combination were more than 200% (Fig. 3C). Notably, all the highly biased VJ pairing were from FR2 hallmark free V germline genes, which were IGHV3-1*01, IGHV3S1*01, IGHV3S25*01, and IGHV3S39*01.

*CDR3 length and distribution*

The CDR3 length of VHHs from the HTS data mainly ranged from 4 to 34 amino acids (AAs), according to the IMGT numbering system (Fig. 4). The overall average length of HCAbs CDR3 is 18 AAs, consistent with previous studies.[8] We found that the shortest and longest CDR3 lengths were 2 and 39 AAs, respectively, although they were quite rare. Interestingly, VHHs derived from various germline genes showed different CDR3 length distributions (Table S1 in Supplementary Material), indicating a bias of insertion during the process of *in vivo* V(D)J recombination. Hence, we further compared the sequences derived from the top 11 V germline genes (Fig. 4). Notably, the results showed that the average CDR3 length of clones derived from hallmark germline genes is longer than that of hallmark free germline genes, except IGV3S9*01.

*Substitution and insertion analysis*

Since CDR3 contains random insertions and is highly diverse, only the VJ paired segments were used for substitution analysis. The substitution rate (SR), which represents the mutation strength of a VJ pair lineage, ranged from 12% to 22% (Table S2 in Supplementary Material). To analyze the substitution preference of each amino acid, we calculated the average substitution rate (ASR) of the VJ pairing that is comprised of more than 1000 lineages, and then overall ASR. The results demonstrated that partial substitutions tend to be biased and most types of mutations are rare (Fig. 5). As to the overall ASR, 79 out of 441 substitution types are higher than 1%. Insertion of glycine (20.78%) and alanine (12.18%) are preferred at the tip of CDR1 and CDR2 loop. Meanwhile, we found that each germline VJ pair showed various substitution patterns. Therefore, we further calculated the germline specific substitution profiles (GSSPs) to quantify and compare the diversity clustered by each germline VJ pair.

*Construction and comparison of VHH profiles*

A GSSP which captures the frequency at which each amino acid appears at every position in an MSA is an N by L matrix, where N is the number of residue types and L is the alignment length. The weighted average of the Jensen-Shannon divergence between GSSPs was calculated to quantitatively compare different profiles and then visualized using multidimensional scaling. In order to test the robustness of this quantification method for GSSPs, we calculated Jensen-Shannon divergence of VJ pairing types that have more than 100, 500, or 1000 lineages, respectively. The results confirmed that lineages from common V genes tend to be clustered, no matter what cutoff values were used (Fig. 6). Moreover, plotting Jensen-Shannon divergence of the top 11 V germline family showed that some classes are close to each other, indicating that mutation patterns are similar between clustered families (Fig. 6 B, D, and F).

*Comparison of long-hinge and short-hinge HACbs*

Specific primers were designed to amplify $IgG_2$ and $IgG_3$, which enabled the identification of each clone type. The IgG specific primer sequences were found in 5,674,954 sequences (97% of all unique sequences). Comparison of $IgG_2$ and $IgG_3$ showed that the ranks of J gene usage are the same, but ranks of V and D usage are different, indicating a different preference of V and D segments (Fig. 7). Notably, a bias of gene rearrangement was observed for these two types of HCAbs. The top 5 hallmark V germline genes contribute 90.91% of long-hinge ($IgG_2$) clones, but only 75.30% of short-hinge ($IgG_3$).

*Comparison of VHHs from different donors*

To test the robustness of the pipeline, the HTS sequences from four other peripheral blood samples (Naïve-2, Naïve-3, Immune-1, and Immune-2), which were collected from the non-immunized and immunized donors (Table 1), were processed following the same pipeline respectively. The V and J germline usage, VJ pairing, DJ pairing, and the substitution preference are highly correlated between the five samples (Table S3 in Supplementary Material). Interestingly, the correlations of the D germline usage are low between samples (Table S3 in Supplementary Material), especially between the naïve and the immunized samples (Spearman rank correlation coefficients: Immune-1 and Naïve-1, $\rho = 0.683$, $P = 0.042$; Immune-1 and Naïve-3, $\rho = 0.467$, $P = 0.205$).

## Discussion

HCAbs naturally occur in various species such as camelids (e.g., camels and llamas) and cartilaginous fish (e.g., sharks).[38] This remarkable evolutionary convergence implies the advantages of functional HCAbs. Thus, systematically investigation of HCAb repertoires is important to reveal the mystery of evolutionary conservation as well as to understand the compensation for the lack of diversity in HCAbs. In this study, we developed a novel pipeline to analyze the full coding sequences of variable domains. In order to automatically process data and maintain its reliability, we tried to avoid using arbitrary filters in the workflow when possible and checked the intermediate results from each step. Since MSAs are crucial for calculating substitution rates and building GSSPs, a classic NW-align algorithm was employed to re-align MSAs from IgBLAST. In order to mitigate effects from noise in the data, we set 1000 as the minimum number of lineages to calculate ASR. The pipeline can be easily extended to analyze the HTS data of antibody repertoires from other species.

V(D)J recombination is one of the mechanisms of antibody diversity. Previous study confirmed that the V germline genes of HCAbs and conventional IgGs were located in the same IgH locus on the genome.[34] The hallmark residues in FR2 regions have been known as a characteristic of VHHs. A previous study reported the presence of novel hallmark-free variable domains that can be rearranged to both camelid classical antibodies and HCAbs.[36] Here, we found that more than 10% of hallmark-free sequences are in the non-immunized HCAbs repertoire, indicating an increase in the HCAbs diversity by sharing V germline genes with tetrameric IgGs. Interestingly, the germline gene IGHV3S39*01 contributes about 60% of V segment usage among all non-hallmark V germline. The biological mechanism of how hallmark-free HCAbs are developed is still unknown. It is well accepted that Ig heavy chains (HC) are selected at the pre-B cell receptor (pre-BCR) checkpoint. Martin *et al.* found specific structural requirements (CDR3 length and amino composition) to select Ig μ heavy-chains during maturation of the pre-B stage.[39] In our dataset, the hallmark residues (F37, E44, and G47), which usually forming a contact interface with the CDR3 to stabilize the structure, show greater diversity than the others in FR2 except for the ones near CDR regions (Fig. S3 in Supplementary Material). Based on our observations, we infer that partial VH germline genes, if not all, are capable of rearranging to HCAbs, but only a small portion (~10%) pass the pre-B checkpoint. Nevertheless, our data confirm that germline gene usage shows a high preference for specific genes. Five out of 17 hallmark-containing germline V segments are responsible for 85.54% of V gene usage in the dataset (Fig.2A).

Studies of antibody repertoires from humans and mice demonstrated that germline gene usage is dynamic during vaccination or infection. Hence, we investigated non-immunized samples from three individuals and two samples from one antigen injected animal with two weeks interval. The results show that the V and J germline usage, VJ pairing, DJ pairing, and the amino acid substitution preference are highly correlated whether antigen immunized or not (Table S3 in Supplementary Material). This high similarity is in accordance with a recent work that revealed the high prevalence of shared clonotypes in human B cell repertoires.[40] Contrarily, the D germline usage shows poor correlations especially between the naïve and immunized samples, indicating the D fragment seems to be the main driving force for *in vivo* antibody maturation.

The CDR3 is the most polymorphic region of IgGs and the main contributor to antigen binding.[41, 42] The CDR3 loop in VHHs of dromedary is longer than the loop in VHs from humans or mice (average of 14 or 13 AAs, respectively).[43] Longer loop lengths increase the paratope size and consequently help compensate for the diversity loss that occurs when light chains are absent.[9] Analysis of 114 conventional camel antibodies showed that CDR3 length is dependent on the germline gene family.[44] Our HTS data demonstrated that the distribution of CDR3 length various on V germline families, indicating the length of CDR3 loops is determined by the usage of the germline gene. This finding is in accordance with studies of T cell receptor (TCR), whose repertoire distribution patterns depend on the use of the germline genes.[45, 46]

Somatic hypermutation (SHM) has long been known as a key process for increasing diversity and improving affinity of antibodies. Comparative analysis of the immune repertoire between the conventional antibodies and the HCAbs from the *Bactrian* camel showed that the nucleotide mutation rate of HCAbs is higher than that of canonical antibodies.[8] In contrast, the calculation of the amino acid mutation rate shows no significant difference in the substitution rate between hallmark and non-hallmark germline gene families (Table S2 in Supplementary Material). However, the substitution patterns of each VJ family did not converge (Fig. 6). Quantitative analysis of GSSPs also confirmed the diversity of the lineages that originated from various VJ germline genes.

A recent study on antibody maturation showed that antibodies that respond to the same antigen to a large extent share similar amino acid substitutions.[47] It appears that the conserved antibody structures that drive adaptive immune responses are highly limited and selected.[31] This is consistent with the results of dominant mutations in HCAbs, suggesting the existence of some preferred mutation patterns. For the result of overall ASR, we observed that non-polar amino residues tend to mutate to non-polar AAs; polar and charged residues are more likely to be substituted by polar and charged AAs (Fig.5, a boxed region in heat-map). Phenylalanine (F), alanine (A), serine (S), and aspartic acid (D) are the germline residues that are most preferred to be substituted (Fig.5, upper histogram), while alanine (A), serine (S), glycine (G), aspartic acid (D) are the residues to which are most likely to be mutated (Fig.5, right histogram).

The new pipeline developed in this work has revealed novel and detailed features of the HACbs repertoire, which is important for VHH engineering or design. The GSSPs built in this work can describe the mutation sequence space of variable domains of antibodies.[20] In previous studies, we have shown that coupling with appropriate evolutionary profile information, our evolution-based protein design protocol, EvoDesign, exhibits a high accuracy in designing protein folds [48, 49] and protein-protein interactions.[16, 17] The preference of germline usage and mutation of HCAbs can be very useful to reduce the effective searching space of amino acid sequences and increase the accuracy of protein design.

## Acknowledgements

This work was supported in part by the National Key R&D Program of China (Grants 2018YFC1602203), the National Natural Science Foundation of China (No.: 31860260, 31301479), the Science and Technology Innovation Platform Project of Jiangxi Province

(No.: 20192BCD40001), the Research Program of State Key Laboratory of Food Science and Technology (No.: SKLF-ZZA-201912), the National Institute of General Medical Sciences (GM083107, GM116960), the National Institute of Allergy and Infectious Diseases (AI134678), and the National Science Foundation (DBI1564756, IIS1901191). The authors gratefully acknowledge financial support from the China Scholarship Council. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation grant number ACI-1548562.[50]

## Authors' contributions

ZT and YZ conceived and designed the study. ZT developed the computer code, carried out the statistical analysis and wrote the initial draft. ZT, XH, and YZ reviewed and edited the manuscript. JF, NH, and YL conducted the experiments of sample collection and deep sequencing. WZ participated in the data visualization. All authors read and approved the final manuscript.

## Conflict of Interest

The authors have declared no competing interests.

## References

1. Hamers-Casterman C, Atarhouch T, Muyldermans S, Robinson G, Hamers C, Songa EB, et al. Naturally-occurring antibodies devoid of light-chains. *Nature* 1993; 363:446-8.
2. Greenberg AS, Avila D, Hughes M, Hughes A, McKinney EC, Flajnik MF. A new antigen receptor gene family that undergoes rearrangement and extensive somatic diversification in sharks. *Nature* 1995; 374:168-73.
3. Herce HD, Schumacher D, Schneider AFL, Ludwig AK, Mann FA, Fillies M, et al. Cell-permeable nanobodies for targeted immunolabelling and antigen manipulation in living cells. *Nat Chem* 2017; 9:762-71.
4. Bruce VJ, McNaughton BR. Evaluation of Nanobody Conjugates and Protein Fusions as Bioanalytical Reagents. *Anal Chem* 2017; 89:3819-23.
5. Bannas P, Hambach J, Koch-Nolte F. Nanobodies and Nanobody-Based Human Heavy Chain Antibodies As Antitumor Therapeutics. *Front Immunol* 2017; 8:1603.
6. Steeland S, Vandenbroucke RE, Libert C. Nanobodies as therapeutics: big opportunities for small antibodies. *Drug Discov Today* 2016; 21:1076-113.

7. Blanc MR, Anouassi A, Ahmed Abed M, Tsikis G, Canepa S, Labas V, et al. A one-step exclusion-binding procedure for the purification of functional heavy-chain and mammalian-type gamma-globulins from camelid sera. *Biotechnol Appl Biochem* 2009; 54:207-12.

8. Li X, Duan X, Yang K, Zhang W, Zhang C, Fu L, et al. Comparative Analysis of Immune Repertoires between Bactrian Camel's Conventional and Heavy-Chain Antibodies. *PLoS One* 2016; 11:e0161801.

9. Muyldermans S, Baral TN, Retamozzo VC, De Baetselier P, De Genst E, Kinne J, et al. Camelid immunoglobulins and nanobody technology. *Vet Immunol Immunopathol* 2009; 128:178-83.

10. De Genst E, Silence K, Ghahroudi MA, Decanniere K, Loris R, Kinne J, et al. Strong in vivo maturation compensates for structurally restricted H3 loops in antibody repertoires. *J Biol Chem* 2005; 280:14114-21.

11. Rubelt F, Busse CE, Bukhari SAC, Burckert JP, Mariotti-Ferrandiz E, Cowell LG, et al. Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. *Nat Immunol* 2017; 18:1274-8.

12. Galson JD, Pollard AJ, Truck J, Kelly DF. Studying the antibody repertoire after vaccination: practical applications. *Trends Immunol* 2014; 35:319-31.

13. Ye B, Smerin D, Gao Q, Kang C, Xiong X. High-throughput sequencing of the immune repertoire in oncology: Applications for clinical diagnosis, monitoring, and immunotherapies. *Cancer Lett* 2018; 416:42-56.

14. Fridy PC, Li Y, Keegan S, Thompson MK, Nudelman I, Scheid JF, et al. A robust pipeline for rapid production of versatile nanobody repertoires. *Nat Methods* 2014; 11:1253-60.

15. Turner KB, Naciri J, Liu JL, Anderson GP, Goldman ER, Zabetakis D. Next-Generation Sequencing of a Single Domain Antibody Repertoire Reveals Quality of Phage Display Selected Candidates. *PLoS One* 2016; 11:e0149393.

16. Pearce R, Huang X, Setiawan D, Zhang Y. EvoDesign: Designing Protein-Protein Binding Interactions Using Evolutionary Interface Profiles in Conjunction with an Optimized Physical Energy Function. *J Mol Biol* 2019; 431:2467-76.

17. Shultis D, Mitra P, Huang X, Johnson J, Khattak NA, Gray F, et al. Changing the Apoptosis Pathway through Evolutionary Protein Design. *J Mol Biol* 2019; 431:825-41.

18. Cohen RM, Kleinstein SH, Louzoun Y. Somatic hypermutation targeting is influenced by location within the immunoglobulin V region. *Mol Immunol* 2011; 48:1477-83.

19. Cui A, Di Niro R, Vander Heiden JA, Briggs AW, Adams K, Gilbert T, et al. A Model of Somatic Hypermutation Targeting in Mice Based on High-Throughput Ig Sequencing Data. *J Immunol* 2016; 197:3566-74.

20. Sheng Z, Schramm CA, Kong R, Program NCS, Mullikin JC, Mascola JR, et al. Gene-Specific Substitution Profiles Describe the Types and Frequencies of Amino Acid Changes during Antibody Somatic Hypermutation. *Front Immunol* 2017; 8:537.

21. Tu Z, Xu Y, He QH, Fu JH, Liu X, Tao Y. Isolation and characterisation of deoxynivalenol affinity binders from a phage display library based on single-domain camelid heavy chain antibodies (VHHs). *Food Agric Immunol* 2012; 23:123-31.

22. Tu Z, Xu Y, Liu X, He Q, Tao Y. Construction and Biopanning of Camelid NaIve Single-domain Antibody Phage Display Library. *China Biotechnol* 2011; 31:31-6.

23. Liu X, Xu Y, Xiong YH, Tu Z, Li YP, He ZY, et al. VHH phage-based competitive real-time immuno-polymerase chain reaction for ultrasensitive detection of ochratoxin A in cereal. *Anal Chem* 2014; 86:7471-7.

24. Aronesty E. Comparison of sequencing utility programs. *Open Bioinform Journal* 2013:1-8.

25. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Mol Biol Evol* 2018; 35:1547-9.

26. Lefranc MP, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, et al. IMGT (R), the international ImMunoGeneTics information system (R) 25 years on. *Nucleic Acids Res* 2015; 43:D413-D22.

27. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* 2013; 41:W34-W40.

28. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics* 2009; 10:421.

29. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator. *Genome Res* 2004; 14:1188-90.

30. Arnaout R, Lee W, Cahill P, Honan T, Sparrow T, Weiand M, et al. High-resolution description of antibody heavy-chain repertoires in humans. *PLoS One* 2011; 6:e22365.

31. Henry Dunand CJ, Wilson PC. Restricted, canonical, stereotyped and convergent immunoglobulin responses. *Philos Trans R Soc Lond B Biol Sci* 2015; 370.

32. Chen L, Kutskova YA, Hong F, Memmott JE, Zhong S, Jenkinson MD, et al. Preferential germline usage and VH/VL pairing observed in human antibodies selected by mRNA display. *Protein Eng Des Sel* 2015; 28:427-35.

33. Jayaram N, Bhowmick P, Martin AC. Germline VH/VL pairing in antibodies. *Protein Eng Des Sel* 2012; 25:523-9.

34. Achour I, Cavelier P, Tichit M, Bouchier C, Lafaye P, Rougeon F. Tetrameric and homodimeric camelid IgGs originate from the same IgH locus. *J Immunol* 2008; 181:2001-9.

35. Kabat EA, Wu TT. Identical V region amino acid sequences and segments of sequences in antibodies of different specificities. Relative contributions of VH and VL genes, minigenes, and complementarity-determining regions to binding of antibody-combining sites. *J Immunol* 1991; 147:1709-19.

36. Deschacht N, De Groeve K, Vincke C, Raes G, De Baetselier P, Muyldermans S. A novel promiscuous class of camelid single-domain antibody contributes to the antigen-binding repertoire. *J Immunol* 2010; 184:5696-704.

37. Monegal A, Olichon A, Bery N, Filleron T, Favre G, de Marco A. Single domain antibodies with VH hallmarks are positively selected during panning of llama (Lama glama) naive libraries. *Dev Comp Immunol* 2012; 36:150-6.

38. Flajnik MF, Deschacht N, Muyldermans S. A case of convergence: why did a simple alternative to canonical antibodies arise in sharks and camels? *PLoS Biol* 2011; 9:e1001120.

39. Martin DA, Bradl H, Collins TJ, Roth E, Jack HM, Wu GE. Selection of Ig mu heavy chains by complementarity-determining region 3 length and amino acid composition. *J Immunol* 2003; 171:4663-71.

40. Soto C, Bombardi RG, Branchizio A, Kose N, Matta P, Sevy AM, et al. High frequency of shared clonotypes in human B cell receptor repertoires. *Nature* 2019; 566:398-402.

41. De Genst E, Saerens D, Muyldermans S, Conrath K. Antibody repertoire development in camelids. *Dev Comp Immunol* 2006; 30:187-98.

42. Miqueu P, Guillet M, Degauque N, Dore JC, Soulillou JP, Brouard S. Statistical analysis of CDR3 length distributions for the assessment of T and B cell repertoire biases. *Mol Immunol* 2007; 44:1057-64.

43. Kaplinsky J, Li A, Sun A, Coffre M, Koralov SB, Arnaout R. Antibody repertoire deep sequencing reveals antigen-independent selection in maturing B cells. *Proc Natl Acad Sci U S A* 2014; 111:E2622-9.

44. Griffin LM, Snowden JR, Lawson AD, Wernery U, Kinne J, Baker TS. Analysis of heavy and light chain sequences of conventional camelid antibodies from Camelus dromedarius and Camelus bactrianus species. *J Immunol Methods* 2014; 405:35-46.

45. Nishio J, Suzuki M, Nanki T, Miyasaka N, Kohsaka H. Development of TCRB CDR3 length repertoire of human T lymphocytes. *Int Immunol* 2004; 16:423-31.

46. Gomez-Tourino I, Kamra Y, Baptista R, Lorenc A, Peakman M. T cell receptor beta-chains display abnormal shortening and repertoire sharing in type 1 diabetes. *Nat Commun* 2017; 8:1792.

47. Tian M, Cheng C, Chen X, Duan H, Cheng HL, Dao M, et al. Induction of HIV Neutralizing Antibody Lineages in Mice with Diverse Precursor Repertoires. *Cell* 2016; 166:1471-84.

48. Brender JR, Shultis D, Khattak NA, Zhang Y. An Evolution-Based Approach to De Novo Protein Design. *Methods in molecular biology (Clifton, N.J.)* 2017; 1529:243-64.

49. Mitra P, Shultis D, Zhang Y. EvoDesign: De novo protein design based on structural and evolutionary profiles. *Nucleic Acids Res* 2013; 41:W273-80.

50. John Towns, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, et al. XSEDE: Accelerating Scientific Discovery. *Comput Sci Eng* 2014; 16:62-74.

**Table 1. Summary of sequencing datasets in this work.**

| Sample [a] | Data [b] (counts) | Joined Data [c] (counts) | Unique DNA Data (counts) | Unique AA Data (counts) |
|---|---|---|---|---|
| Naïve-1 | 14,130,770 × 2 | 5,850,191 | 2,550,586 | 1,973,186 |
| Naïve-2 | 9,783,739 × 2 | 6,381,831 | 2,317,312 | 1,717,133 |
| Naïve-3 | 7,939,987 × 2 | 5,285,912 | 1,763,675 | 1,271,695 |
| Immune-1 | 2,935,298 × 2 | 1,841,952 | 1,270,186 | 782,529 |
| Immune-2 | 3,459,366 × 2 | 2,270,196 | 1,653,673 | 1,149,386 |

a. Naïve-1, Naïve-2, and Naïve-3 were collected from three healthy donors; Immune-1 and Immune-2 were collected from one donor after fifth and seventh immunization, respectively.

b. Data is the total sequences of paired-end reads after filter and quality control;

c. Joined Data is the number of sequences that were generated from paired-end reads.

**Figure Legends**

**Figure 1. Workflow for the analysis of HCAbs repertoire.** The coding sequences of VHHs are amplified from long and short hinge HCAbs, respectively, where the next-generation sequencing (NGS) data are processed using fastq-join to merge paired-end sequences after discarding the low-quality reads. Next, IgBLAST is used to assign V(D)J genes for each transcript. Coding sequence (CDS) length distribution, V(D)J usage, VJ pairing, and DJ pairing usage are based on DNA sequences, while other analyses are based on amino acid sequences.

**Figure 2. V, D, and J germline gene usage of VHHs repertoire is highly biased.** (A) Usage of V germline genes. (B) Usage of D germline genes. (C) Usage of J germline genes.

**Figure 3. V(D)J pairing of VHHs is biased.** (A) VJ pairing of germline genes in naïve VHHs repertoire is highly biased to pairs of certain germline genes. The top 21 VJ pair made up > 90% of the clones in the repertoire. (B) and (C) are the *in silico* simulation of DJ and VJ combination. The error bars illustrate the relative deviation of the 2,000 steps of simulation, while the columns represent the relative deviation of the VHHs repertoire.

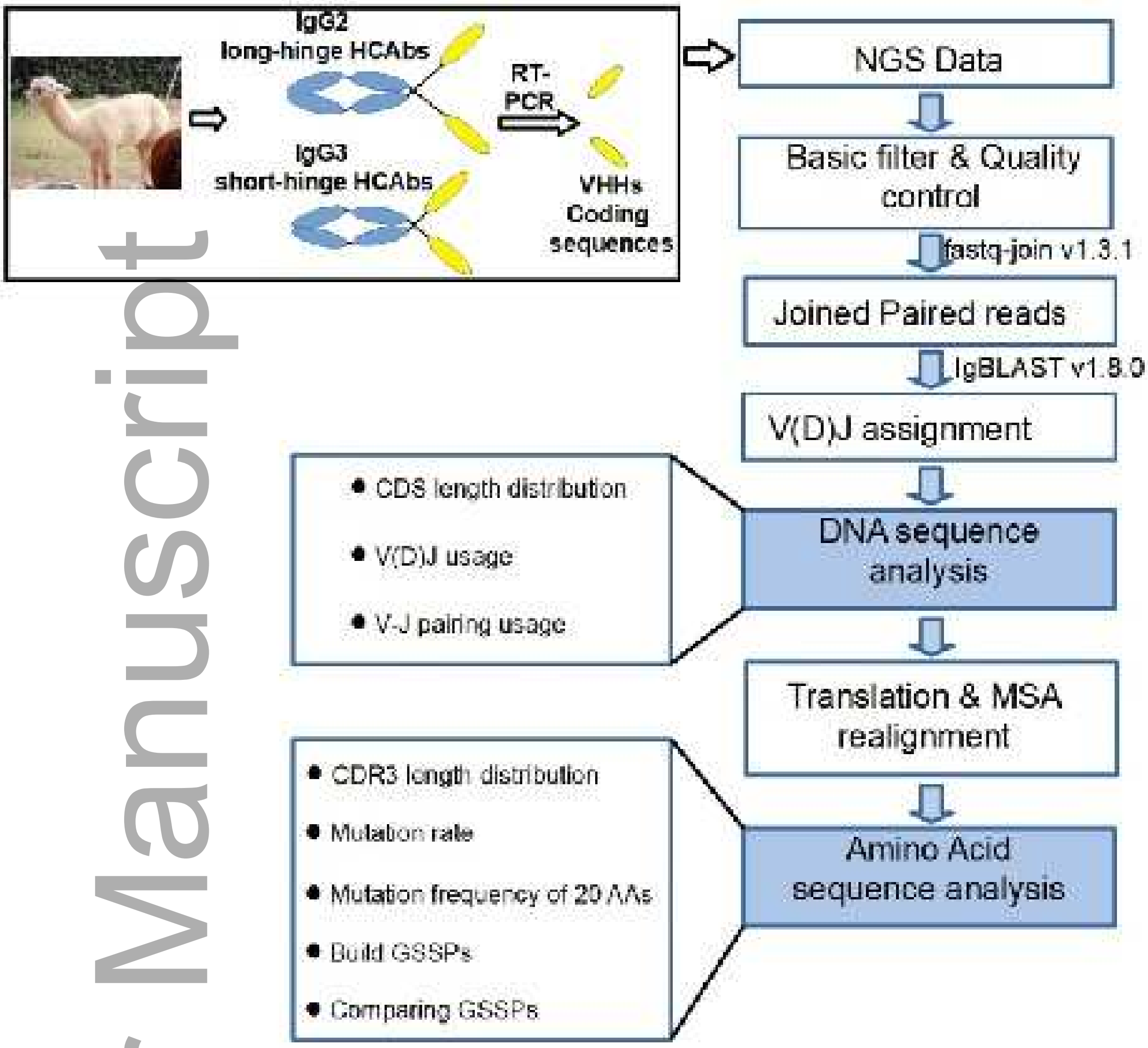
**Figure 4. Distribution of CDR3 amino acid length is depending on germline gene families.** The average CDR3 lengths of the top 11 V germline genes are presented in the upper right panel. Hallmark and non-hallmark residue V genes are shown in red or blue bars, respectively.

**Figure 5. Most substitutions are rare and partial mutations are preferred.** Due to mutation bias, most observed types of substitution occur rarely. The histogram in the upper and right are the sum of the cases in the corresponding column or row, respectively. The asterisk in mutation type means stop codon. Dashes mean gaps in both germline residues and mutation types.

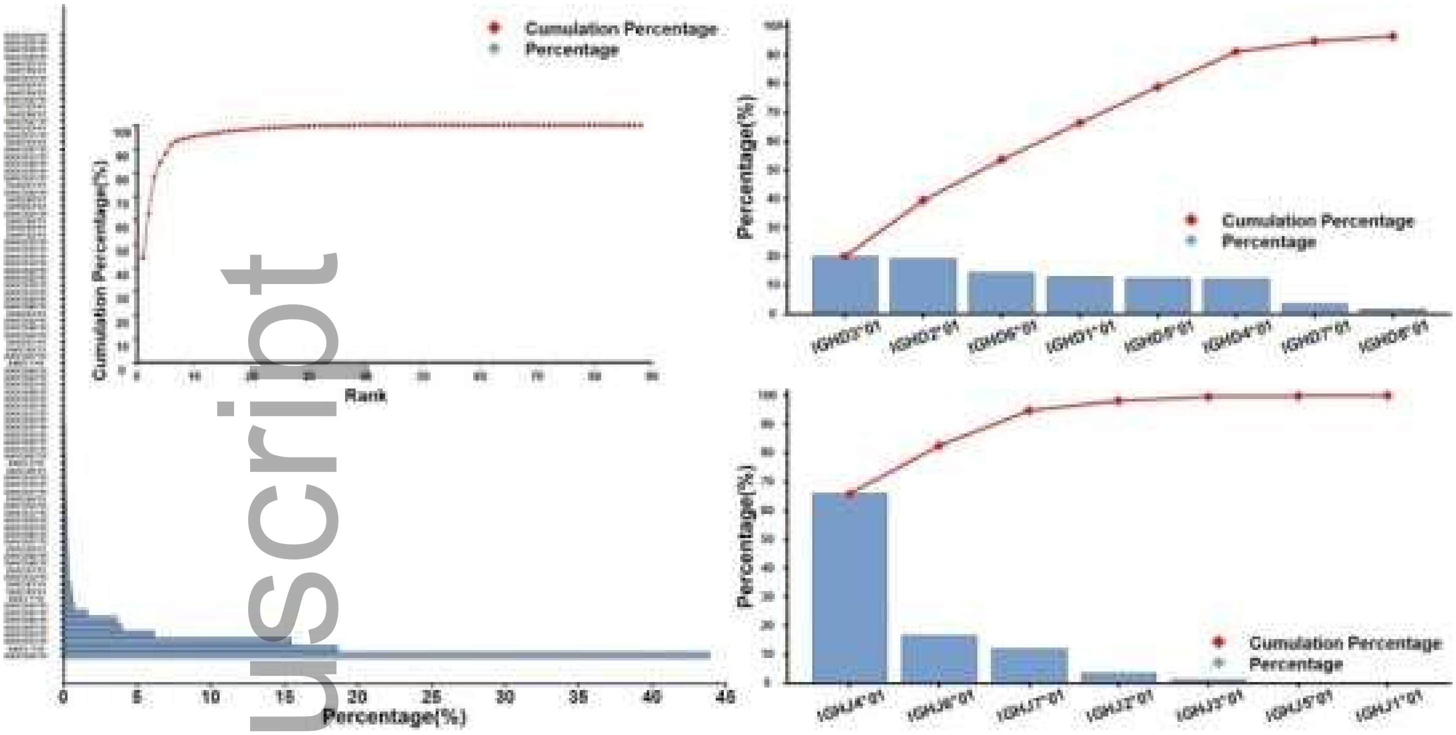
**Figure 6. The similarity of GSSPs between VJ germline gene families.** The Jensen-Shannon divergence was used to compare GSSPs, and the distance matrix was visualized

using multidimensional scaling. The VJ pair types that have more than 100 (A and B), 500 (C and D), or 1000 (E and F) lineages were calculated and plotted respectively. GSSPs of the same V gene tend to be clustered together. Meanwhile, the distance of partial VJ pair is close, which indicates sharing similar mutation patterns. VJ pairs from the top 11 V genes are shown in the right column (B, D, and F).

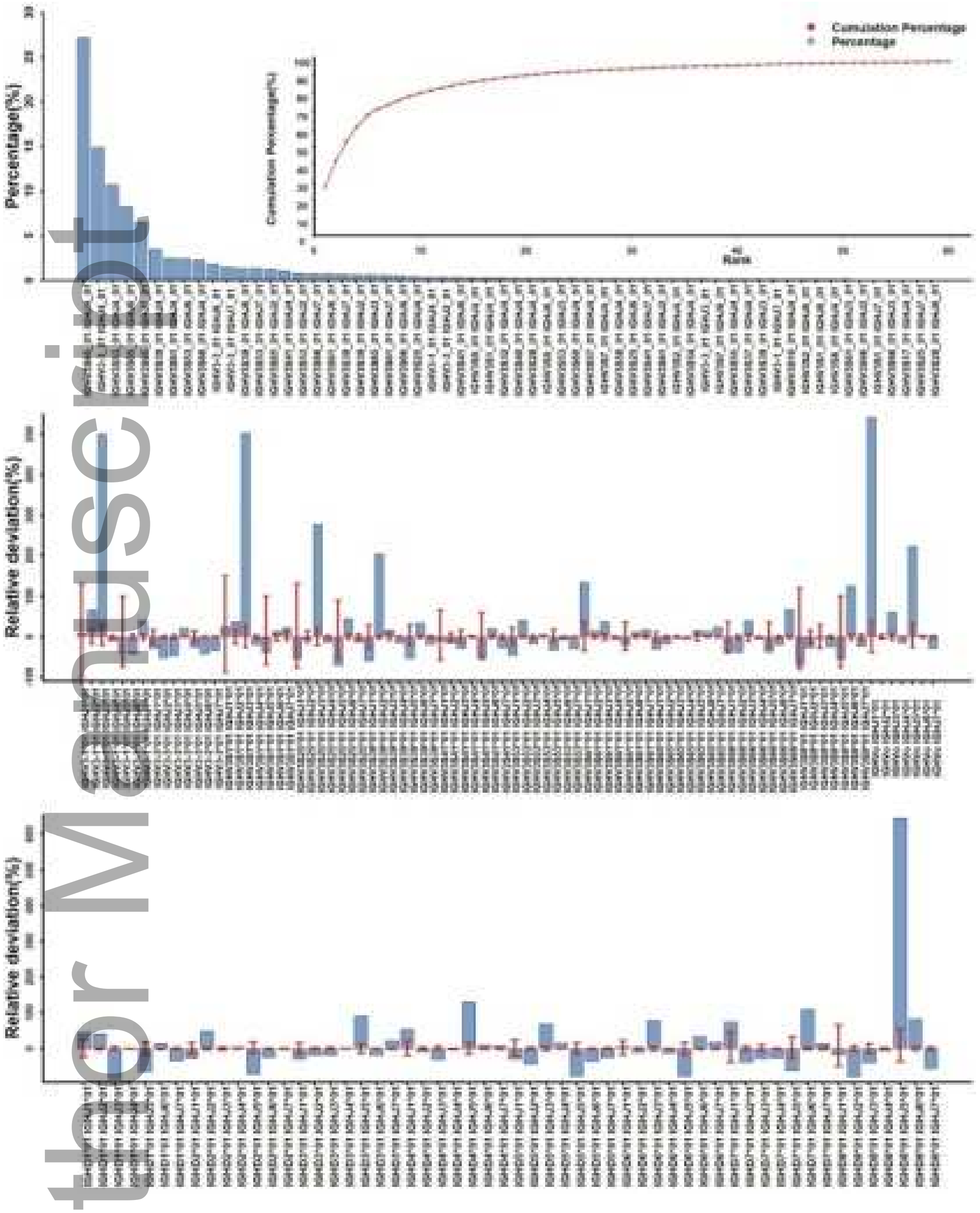
**Figure 7. Comparison of V(D)J usage of long-hinge (IgG2) and short-hinge (IgG3) HCAbs.** Sequences belonging to IgG2 or IgG3 were identified using specific primers for BLAST. (A), (B), and (C) are the V, D, and J gene usage of both types of HCAbs, respectively.



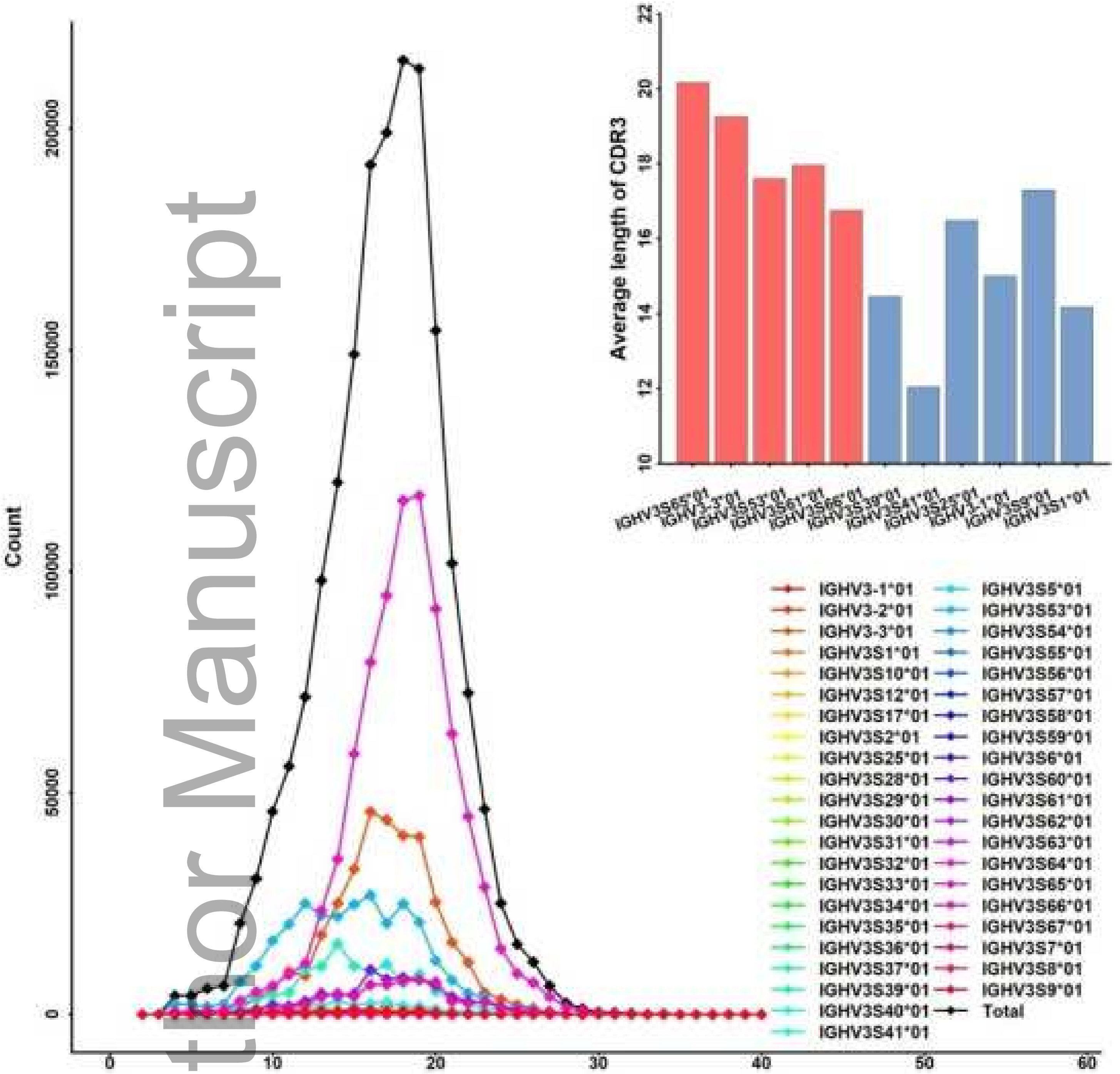
imm_13224_f1.jpg



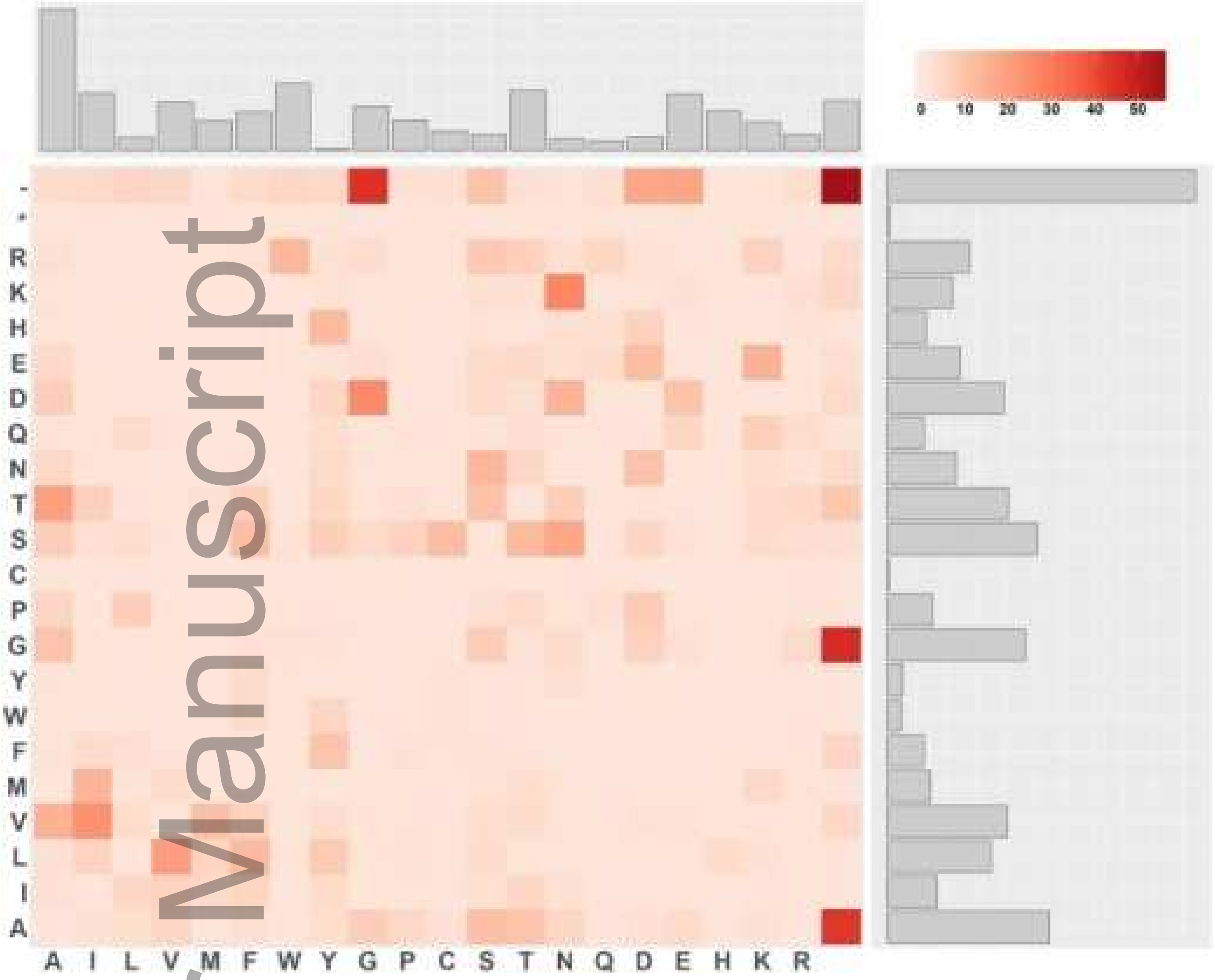
imm_13224_f2.jpg



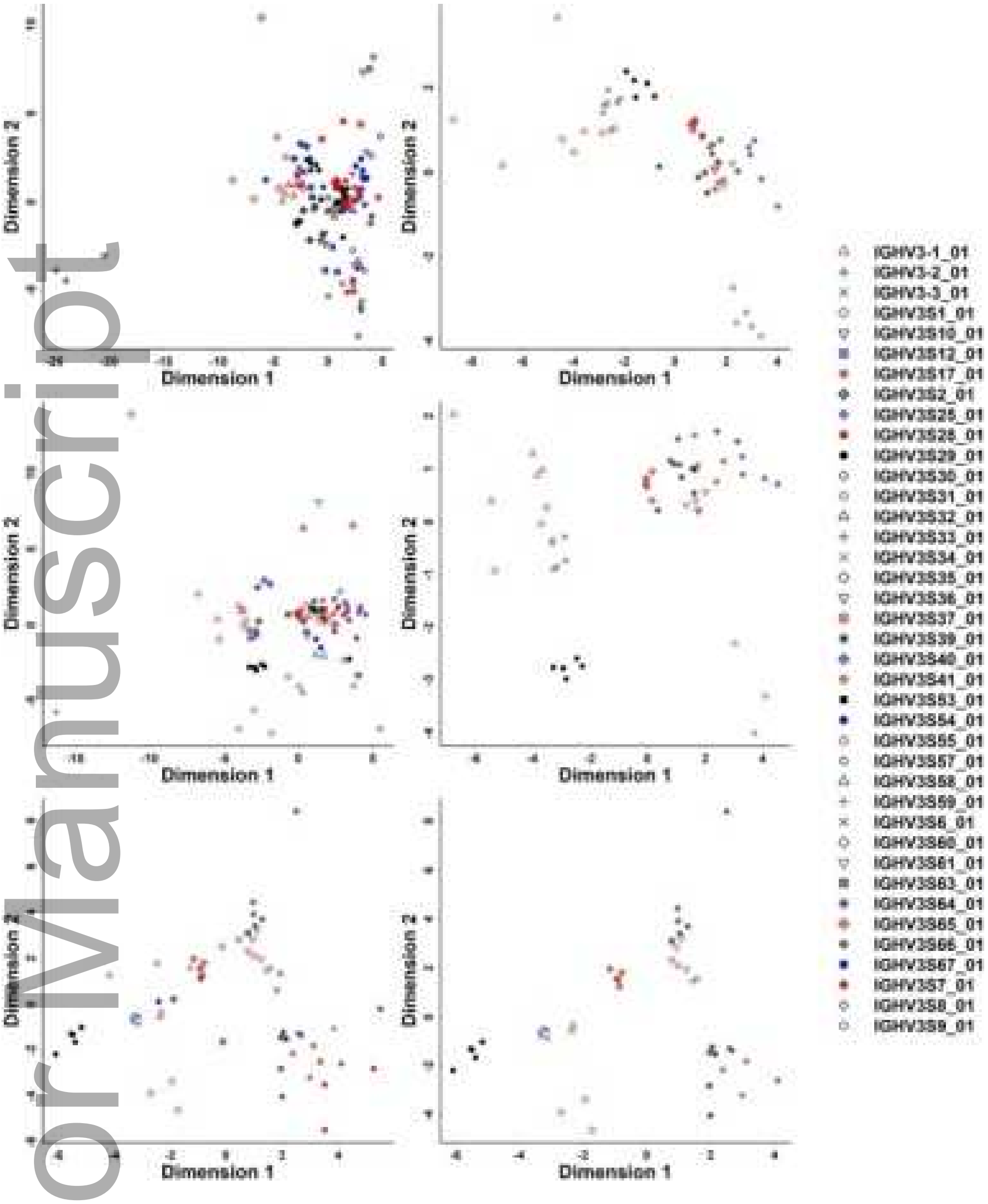
imm_13224_f3.jpg



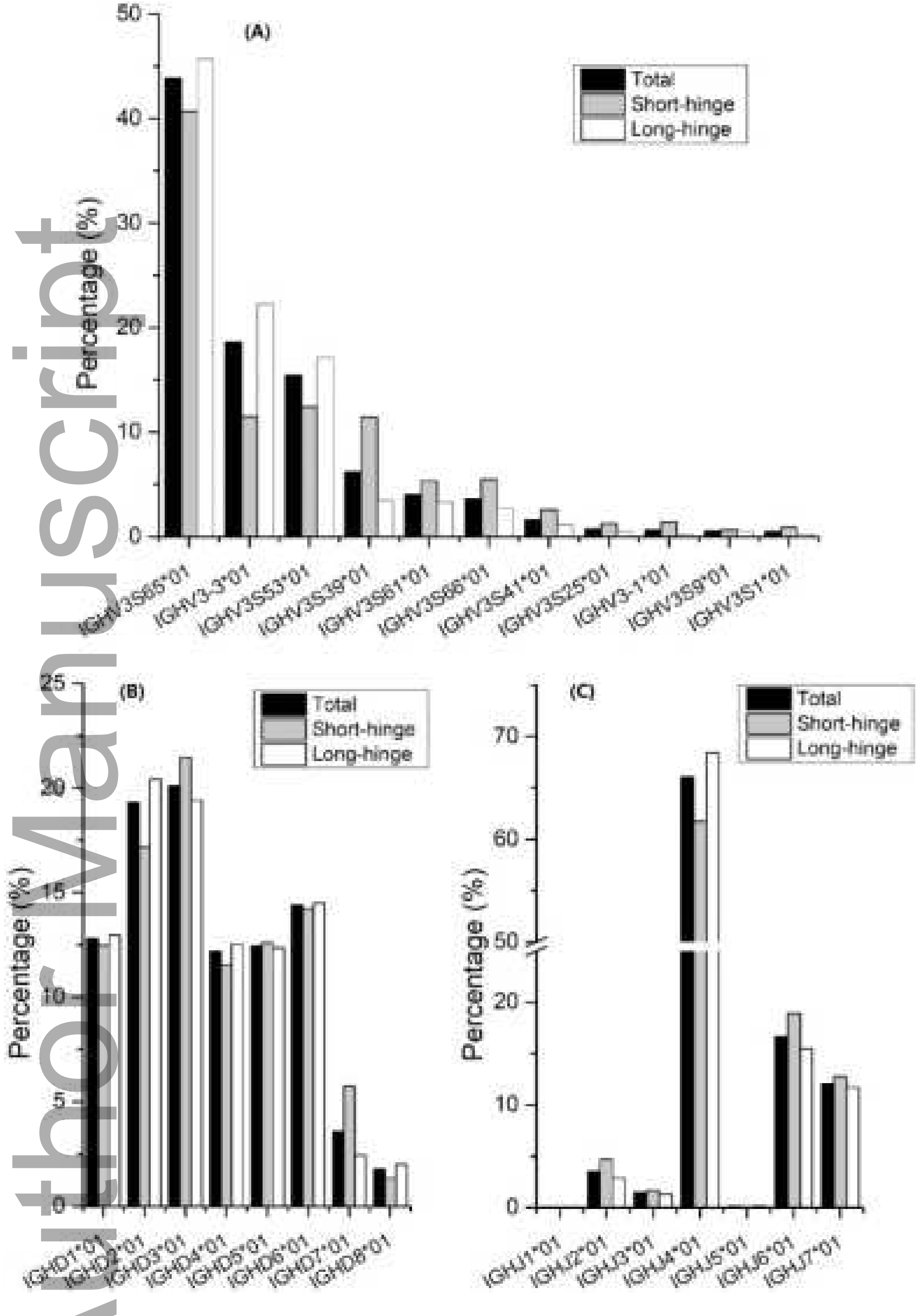
imm_13224_f4.jpg



imm_13224_f5.jpg



imm_13224_f6.jpg



imm_13224_f7.jpg