

REPORT



A window into the human immune system: comprehensive characterization of the complexity of antibody complementary-determining regions in functional antibodies

Oscar Mejias-Gomez ^a, Andreas V. Madsen ^a, Kerstin Skovgaard ^a, Lasse E. Pedersen ^a, J. Preben Morth ^a, Timothy P. Jenkins ^a, Peter Kristensen ^b, and Steffen Goletz ^a

^aDepartment of Biotechnology and Biomedicine, Technical University of Denmark, Kgs. Lyngby, Denmark; ^bDepartment of Chemistry and Bioscience, Aalborg University, Aalborg, Denmark

ABSTRACT

The human immune system uses antibodies to neutralize foreign antigens. They are composed of heavy and light chains, both with constant and variable regions. The variable region has six hypervariable loops, also known as complementary-determining regions (CDRs) that determine antibody diversity and antigen specificity. Knowledge of their significance, and certain residues present in these areas, is vital for antibody therapeutics development. This study includes an analysis of more than 11,000 human antibody sequences from the International Immunogenetics information system (IMGT). The analysis included parameters such as length distribution, overall amino acid diversity, amino acid frequency per CDR and residue position within antibody chains. Overall, our findings confirm existing knowledge, such as CDRH3's high length diversity and amino acid variability, increased aromatic residue usage, particularly tyrosine, charged and polar residues like aspartic acid, serine, and the flexible residue glycine. Specific residue positions within each CDR influence these occurrences, implying a unique amino acid type distribution pattern. We compared amino acid type usage in CDRs and non-CDR regions, both in globular and transmembrane proteins, which revealed distinguishing features, such as increased frequency of tyrosine, serine, aspartic acid, and arginine. These findings should prove useful for future optimization, improvement of affinity, synthetic antibody library design, or the creation of antibodies *de-novo in silico*.

ARTICLE HISTORY

Received 22 June 2023
Revised 3 October 2023
Accepted 4 October 2023

KEYWORDS



Amino acid diversity; antibodies/immunoglobulins; antibody sequencing; antibody therapeutics development; antibody-antigen complexes; Complementary-Determining Regions (CDRs); *de-novo in silico* antibody design; IMGT (International Immunogenetics Information System); immunology; synthetic antibody libraries


Introduction

Antibodies, also known as immunoglobulins, are the main components of the humoral immune response. In humans, antibodies are generated by activated B cells as either membrane-anchored molecules (B-cell receptors, BCR) or secreted proteins. Each chain is composed of a constant chain that includes an identical pair of heavy and light chains (C_H and C_L) extending into a variable region also divided into a heavy and light chains (V_H and V_L). The constant and variable and constant regions are fundamental for the effector functions of immunoglobulins and their ability to recognize different antigens, respectively. The human immune system can generate repertoires of 5×10^{13} different antibody specificities, each capable of recognizing a distinct antigen.¹ The diversity of immunoglobulins is highly complex and influenced by several processes. One key process is combinatorial diversity in V(D)J gene rearrangements, which is responsible for a considerable portion of the V-region diversity.^{2,3} Additionally, junctional diversity contributes via insertion and deletion of nucleotides at the joints between the different segments, and the different possible combinations of heavy and light chains play a significant role in the potential of the antibody repertoire. Somatic

hypermutations, which occur in mature B cells, add a final variation process that aims to increase the affinity of antibodies to their targets. This extensive diversity is located in the variable region, which contains six hypervariable loops where most antigen recognition occurs, the complementary-determining regions (CDRs).^{4,5} The CDRs constitute the paratope of the immunoglobulins forming highly specific interactions surface, interacting with the epitope on the antigen through non-covalent forces.⁶

Previous studies have suggested general notions regarding the amino acid diversity, length variability, the unique importance of specific CDRs, and the enrichment and absence of certain residues in these regions.^{7–10} For example, aromatic residues were found to be overrepresented in antibody-antigen recognition sites, while others like alanine (A), valine (V), cysteine (C), and methionine (M) were underrepresented.¹¹ This knowledge has been essential for the development of the antibody therapeutics field, which aims to understand and mimic the specificity and diversity of the immune system to discover and engineer powerful agents against different types of diseases.¹² While the field has advanced rapidly during the past decade, there remains a need for additional data that can

CONTACT Steffen Goletz  sgoletz@dtu.dk  Biotherapeutic Glycoengineering and Immunology, Department of Biotechnology and Biomedicine, Technical University of Denmark, Søtofts Plads, Building 224, Kgs. Lyngby 2800, Denmark

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/19420862.2023.2268255>

© 2023 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

be used to design, engineer, and improve the tools used in antibody discovery, including prediction and design of antibody-antigen binding *in silico*.^{13,14} Previous studies using a smaller subset of antibody sequences compared the amino acid composition of human variable regions with mice,¹⁵ and chicken¹⁶ or focused solely on mice antibodies.¹⁷ Although these studies added valuable insights, such as indicating a higher frequency of somatic hypermutations and junctional diversity in human antibodies compared to mice, as well as higher frequencies of C residues in chicken antibodies, their scope was limited to the investigation of the CDR3 region of the heavy chain. Moreover, specific details regarding the amino acid composition for each CDR length were not explored, and the studies did not delve into the potential roles that individual amino acids might play in binding interactions.

Our study aims to describe in detail the characteristics of human antibody CDRs of both heavy and light chains by analyzing next-generation sequencing data from IMGT¹⁸ comprising over 11,000 productive human antibody sequences. Aspects such as length distribution, overall amino acid diversity, and amino acid occurrence on average per CDR and per position within each region are discussed for both heavy and light chains.

Results

Human heavy chain analysis

Specific peptide lengths in the CDRH1 and CDRH2 regions contrast with broad length diversity in CDRH3

The diversity of CDRH1 and CDRH2 is constrained to essentially one length, an eight residue short peptide stretch (Figure 1 (a–1), Table S2), that comprises 80.1% and 62.9% of all CDRH1 and CDRH2 regions, respectively. It is worth mentioning, however, that both CDRH1 and CDRH2 have one additional length that constitutes a considerable number of sequences, i.e., 10 amino acids for CDRH1 (14.7% of all sequences) and 7 amino acids for CDRH2 (28.9% of all sequences). On the contrary, CDRH3 showed length diversity ranging from 4 to 36 amino acids. However, when only considering lengths with more than 50 sequences in the database, the length range is reduced to 8 to 25 amino acids. The two most common CDRH3 lengths for human heavy chains are 14 and 15 amino acids each, occurring in nearly 12% of human CDRH3. The median CDRH3 lengths is 14 amino acids (mean 14.54 with variance of 10.92) (Table S2). The frequencies of CDRH3 lengths gradually decrease for 8 amino acids and 25 amino acids, respectively, while those lengths occurring at least in 5% of antibodies range from 11 to 20 amino acids. CDRH3 lengths follow a normal distribution with a D'Agostino and Person test p-value of 0.456 (Figure 1h and Table S3).

Diverse amino acid usage across CDRHs indicates unique structural and functional roles in antibody–antigen interactions

We assessed the average usage of each residue in CDRH1, CDRH2, and CDRH3 when all possible lengths are considered (Figure 2a, Table S4). For comparison, the amino acid

distribution of framework regions (FRs) within the variable domains and constant regions (Con) of heavy chain immunoglobulins, as well as for globular (Glob) and transmembrane proteins (TMB)¹⁹ are shown. CDRH1 shows increased usage of tyrosine (Y), phenylalanine (F), serine (S), threonine (T), and glycine (G) compared to FRs, Con, Glob, and TMB. Likewise, CDRH2 of both lengths show increased occurrence of G, S, T, Y, and isoleucine (I). CDRH3s also present interesting dissimilarities with FRs, Con, Glob, and TMB. Amino acids like Y, arginine (R), and aspartic acid (D) can be found with higher occurrences in CDRH3s. On the contrary, hydroxyl amino acids S and T, aliphatic amino acid valine (V), leucine (L), and amide amino acid glutamine (Q) appear to be higher in FR and Con. Phenylalanine and G have higher abundances in CDRH3s compared to FRs and Con but not to Glob and TMB. When comparing the different CDRHs to each other, it becomes clear that CDRH1 and CDRH2 display a higher preference for the hydroxyl residues S and T than CDRH3. CDRH1 shows a much higher occurrence of F (17.3%) than CDRH2 (1.2% and 1.5%) and CDRH3 (between 3.7% and 6.7%, latter for 12 amino acids long CDRH3). Glycine occurrence also seems to be higher for CDRH1 than CDRH2 and CDRH3, but to a lesser extent. CDRH2 has a characteristically higher occurrence of I (13.2% and 15.2%) than CDRH1 (1.6%) and CDRH3 (between 2.3% and 3.7%). Tyrosine seems to be highly represented in CDRH1 (16.6%) and CDRH2 with 7 amino acids (12.6%) in contrast to CDRH2 with 8 amino acids (4%), which constitutes the majority of the possible CDRH2 lengths (62.9%). For CDRH3, the frequency of Y follows an increasing trend from short to long CDRH3 (9.8% for 8 amino acids to 14.3% for 25 amino acids long CDRH3). Moreover, CDRH3 displays a higher occurrence of A, R, and D when compared to CDRH1 and CDRH2. The amount of A decreases the longer the CDRH3 becomes, ranging from 12.7% at 8 amino acids-long CDRH3 to 7.2% at 25-amino acids long CDRH3. Arginine and D follow a similar trend, but to a lesser extent (9.5% to 7.4%, and 12.2% to 9.0%, respectively).

Specific amino acid distributions and physicochemical properties across CDRH1, CDRH2, and CDRH3 highlight functional adaptations and the diversity in antibodies

The amino acid distribution in each position within each CDRH with more than 50 sequences in the database was also investigated. Figure 3 shows the relative abundance of each of the 20 natural amino acids per position, while Figure S1 displays their physicochemical properties.²⁰ In Figure 3, the most abundant CDRH1 length, the two most abundant CDRH2 lengths, the shortest CDRH3 with more than 50 sequences, the most abundant CDRH3, CDRH3 with 20 amino acids, and the longest CDRH3 with more than 50 sequences are shown. Figure S1 shows the physicochemical properties of the most common CDRHs lengths for H1, H2, and H3, respectively. The amino acid distribution for all CDRH lengths can be found in the supplementary materials, together with the physicochemical properties of CDRH1-8aa, CDRH2-7aa, CDRH2-8aa, CDRH3-8aa, CDRH3-14aa, CDRH3-20aa, and CDRH3-25aa.

In CDRH1 (Figure 3a), position 38 presents a high diversity where the most abundant amino acid is Y (25%). This position also shows high versatility with a preference for both very large uncharged aromatic (I, Y, F, M) and very small aliphatic

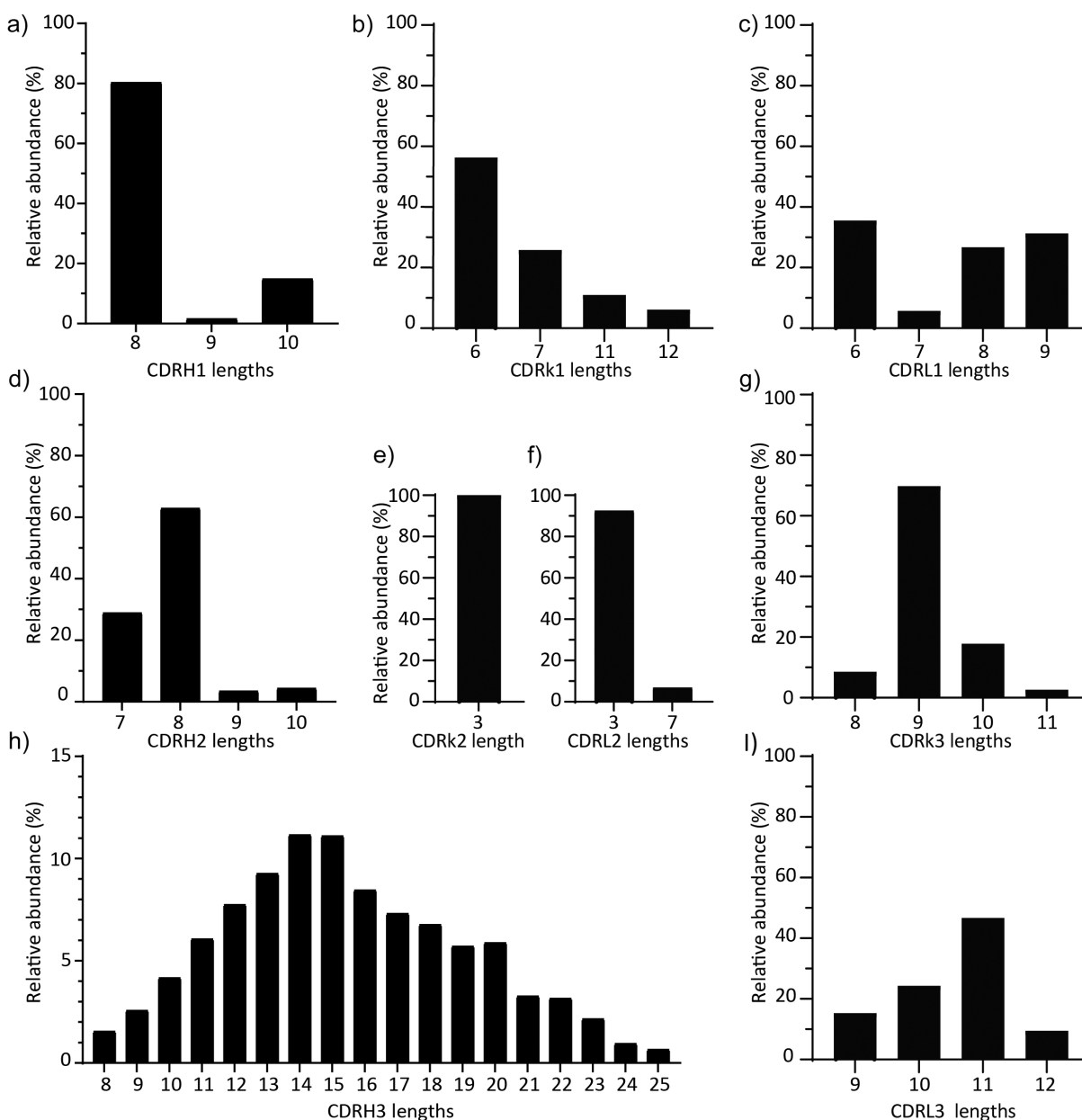


Figure 1. Length distribution of human immunoglobulin heavy and light chain CDRs. a) CDRH1. 80,1% of the sequences fall under a CDRH1 length of 8 amino acids (7779 out of 9684 sequences). 14,7% of the sequences have 10 amino acid long CDRH1 (1442 out of 9684 sequences). b) CDRk1. CDR1 lengths for κ are either 6, 7, 11, or 12 amino acids, interestingly with no lengths of 8 to 10 amino acids, while the frequency 6 (56,4%) and 7 amino acids (25,9%) is highest c) CDRL1. CDR1 lengths for λ are either 6, 7, 8, or 9 amino acids long, while all are comparably frequent (6 amino acids highest with 35,6%), 7 amino acids is rarely used (5,8%). d) CDRH2. 62,9% of the sequences have an 8 amino acids long CDRH2, constituting the majority of the CDRH2 diversity (6088 out of 9684 sequences). A considerable number of sequences have 7 amino acids long CDRH2 (28,9%, 2795 out of 9684 sequences). e) CDRk2. CDR2 for κ is exclusively 3 amino acids long (99,8%) f) CDRL2. CDR2 for λ seems to be limited to 3 amino acids (92,4%) with a rare occurrence of 7 amino acids. g) CDRk3. CDR3 for κ shows lower diversity than its counterpart in heavy chain with only four lengths and where 9 amino acids comprises the majority of sequences (69,9%) h) CDRH3. The length diversity of CDRH3 is much greater than those of CDRH1 and CDRH2. There are 18 lengths with more than 50 sequences, where 14 and 15 amino acids are the two most abundant lengths. Human heavy chain CDRH3 can be remarkably long up to 36 (only 1 sequence entry in database). i) CDRL3. CDR3 for λ also shows low diversity compared to CDRH3 with four lengths spanning the CDR3 length diversity and where 11 amino acids comprises the majority of sequences (46,8%).

residues with low polarity (Figure S1). Positions 27 to 37 (27, 28, 29, 30, 35, 36, 37) show a rather low diversity and are largely dominated by either 1 or 2 amino acids at each position. The residues S and T dominate at positions 29, 35, and 36, while aromatic residues dominate at positions, 28 (F and Y), 30 (F), and 37 (Y), and prevalence of G at position 27, resulting in a certain alternating pattern of hydroxyl and aromatic groups. Positions 29, 35, 36, 37, and 38 are also abundant in uncharged, hydrogen-donor and -acceptor, and hydrophilic and neutral residues. Position 28 shows higher abundance of

neutral and nonpolar amino acids (Figure S1). The remaining two positions, 27 and 30, appear to be very limited in diversity with two main amino acids observed with high abundance in each position (Gly and Tyr, respectively). The amino acid composition of CDRH1 positions stands out by the considerable deficiency in M, Q, and C, and a low representation of L, V, histidine (H), lysine (K), glutamic acid (E), and proline (P).

CDRH2 displays higher overall diversity than CDRH1, which is more pronounced for CDRH2 with 8 amino acids than those with 7 amino acids (Figure 3a–3). The flanking

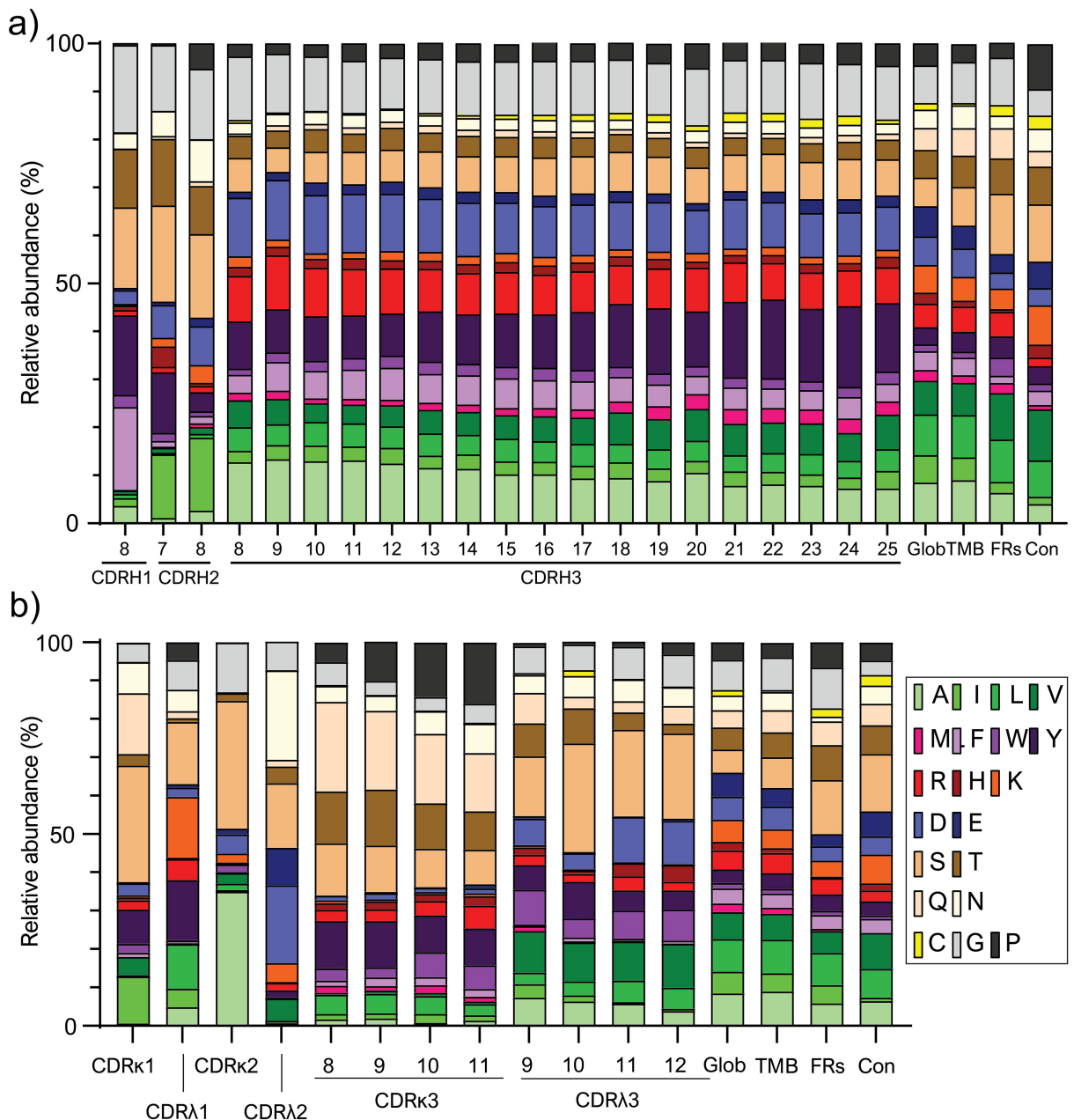


Figure 2. Average of relative abundance of each amino acid per CDR of heavy (a) and light chain (b), globular (Glob) and transmembrane proteins (TMB). The frequencies of globular proteins, transmembrane proteins, framework regions within human antibody sequences, and constant regions taking Trastuzumab as model IgG are shown for comparison a) the average relative abundances of the 20 natural amino acids are shown for CDRH1 of 8 amino acids long, CDRH2 of 7 and 8 amino acids long, all CDRH3 lengths with more than 50 sequences in the database. B) average distribution of amino acids in κ and λ CDRs. The average relative abundances of the 20 natural amino acids are shown for the most abundant κ and λ light chain lengths for CDR1, CDR2, and CDR3.

positions in both CDRH2 lengths, 56 and 65, show a similar bias of amino acid occurrence, i.e., almost exclusively (>87% for both CDRH2 lengths) I and high abundance (77% and 50% for 7 and 8 amino acids CDRH2, respectively) of T, respectively. Interestingly, CDRH2 with 8 amino acids versus those with 7 amino acids show some clearly distinct features. Positions 57 and 58 are quite different in both lengths. In position 57, Y occurs predominantly with a relative abundance of 54.3% in CDRH2-7aa, while it only reaches 9.31% in the longer CDRH2, where instead S occurs in 39.3%, making this position versatile in hydrogen bonding and highly polar residues. In position 58, H and Y appear with 26.2% and 32.8% in CDRH2-7aa, while in CDRH2-8aa they are 1.32% and 10.36%, respectively. Position 58 in CDRH2-8aa shows a remarkably

high abundance of P (39.3%), not observed in any other CDR. Position 59 in CDRH2-7aa shows low diversity, with S dominating the position with an abundance of 63%. On the contrary, the same position in CDRH2-8aa displays high diversity where only D and S surpass 20% in abundance. This position is also highly diverse in terms of hydrophobicity, hydrogen bonding, size, and chemical behavior (Figure S1).

The same diversity pattern can be observed in positions 63 and 64; for CDRH2-7aa both positions are mainly occupied by one amino acid, G (82.7%) and S (58.9%), respectively. For CDRH2-8aa S and G are almost equally distributed in position 63 (37.5% and 43%, respectively). Position 64 in CDRH2-8aa has the highest diversity of any position in both CDRH2s, where D, E, S, T, and Q show relatively high abundances

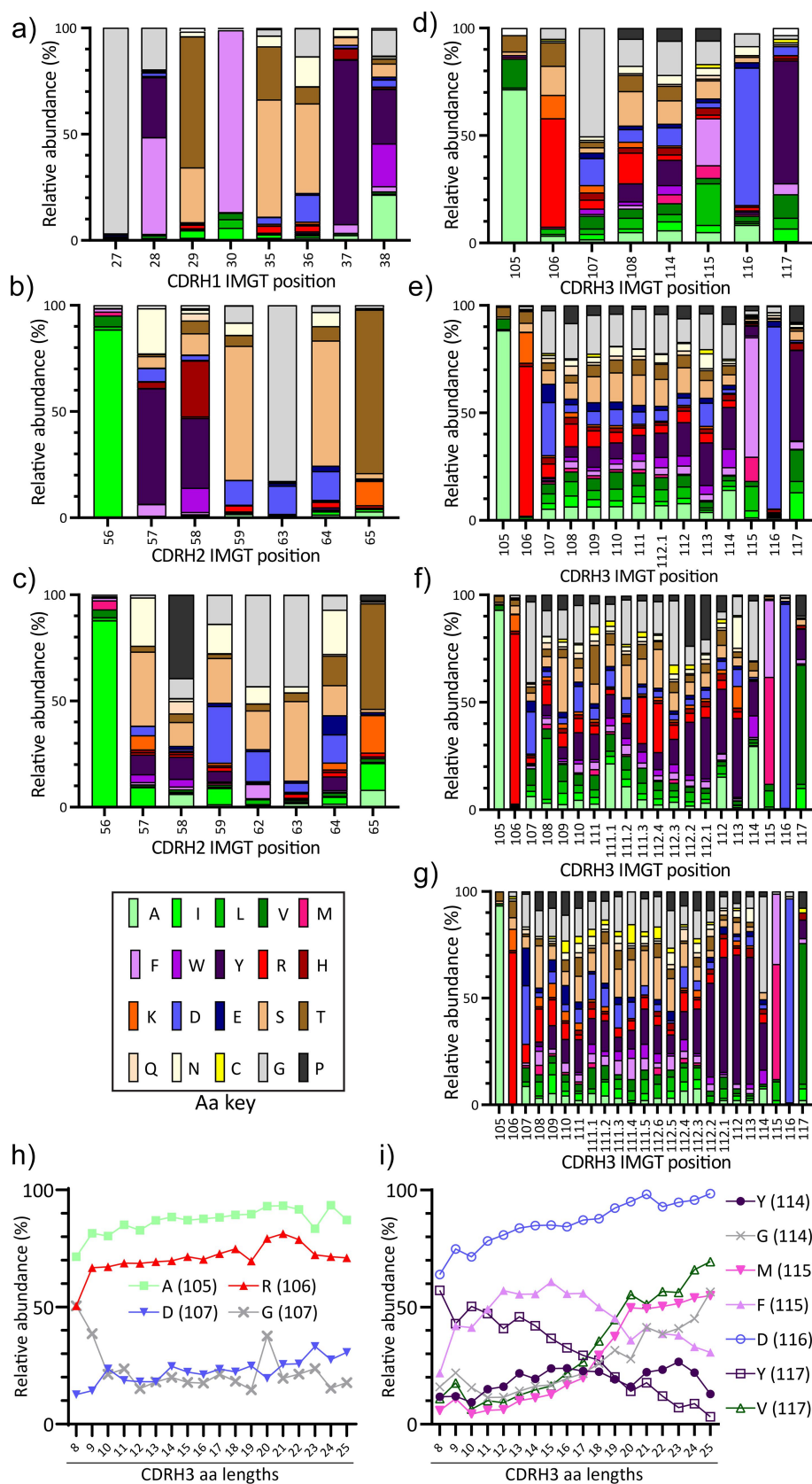


Figure 3. Amino acid distributions per position in CDRH1, CDRH2, and CDRH3. The y-axis shows relative abundance in percentage (%) and the x-axis shows the specific position within each CDRH for plots a to G. The x-axis in plot H and I shows the different CDRH3 lengths studied. a) CDRH1 b) CDRH2-7aa c) CDRH2-8aa d) CDRH3-8aa e) CDRH3-14aa f) CDRH3-20aa g) CDRH3-25aa h) amino acid changes over different CDRH3 length at flanking CDRH3 IMGT positions i) amino acid changes over different CDRH3 length at inner CDRH3 IMGT positions.

(13.3%, 8.9%, 14.2%, 13.9%, and 20.8%, respectively). Position 62, which is unique for CDRH2-8aa, displays a similar pattern to that of position 63, with a small difference in lower usage of S and higher occurrence of D. Overall, aromatic residues are in general rare in CDRH2, in contrast to the aromatic residue-dominated positions in CDRH1, while this is true for all positions in CDRH2-8aa, there are considerable occurrences in CDRH2-7aa at positions 57 and 58. Other somewhat outstanding features differentiating 7 and 8 amino acids long CDRH2 are the G at position 63 for 7 amino acids long CDRH2 and the high occurrence of P at position 58 of CDRH2-8aa.

The diversity in all lengths of CDRH3 (Figure 3d–3) is higher than for CDRH1 and CDRH2. Nevertheless, some positions within CDRH3 such as 105, 106, 115, 116, and 117 have low diversity and are highly conserved. Interesting patterns can be observed in these positions (Figure 3h–3); A, R, and D in positions 105, 106, and 116, respectively, show approximately the same abundance in all lengths. Interestingly, positions 106 and 116 are dominated by residues with opposite charges, hydrogen bonding potential, and size (Figure S1). On the contrary, F and M in position 115, and Y and V in position 117, display an opposite occurrence pattern throughout the different CDRH3 lengths. In position 115, F's abundance decreases the longer the CDRH3 gets while M frequency increases in parallel and the same occurs in position 116 for Y and V. Moreover, the diversity of positions 107 and 114 decreases the longer the CDRH3 becomes, with one amino acid gradually becoming preferred in each position: D and G for 107 and 114, respectively.

The remaining positions in CDRH3 (108 to 113) show the highest diversity in amino acid distribution and physicochemical properties of all CDRH positions. However, amino acids such as G, S, R, D, and Y are preferred within each amino acid chemical class (Figure S1). We noted that unfavored amino acids in globular and transmembrane proteins (Table S4), such as C, appear to have higher occurrences in the middle positions of long CDRH3s (Figure 3f and supplementary materials). Interestingly, besides the flanking positions of CDRH3 being present in the afore described featured frequencies independent of the lengths of the CDRH3, additional positions are added with increasing lengths with a diversity resembling positions 110 and 111 in CDRH3-14aa. The higher numbered positions seem to acquire an increasingly higher frequency for Y in positions closer to 114 as the CDRH3 becomes longer. For long CDRH3, there is a patch of increasing high neighboring Y frequencies toward 114. Interestingly, the flanking positions of CDRH3 have consistent amino acid frequencies regardless of CDRH3 length, but, as the length of CDRH3 increases, additional positions are added that have a diversity of amino acids similar to those found in positions 110 and 111 in CDRH3-14aa. These positions tend to have an increasing frequency of the amino acid Y as they approach position 114. Furthermore, in long CDRH3s, there is a patch of positions with high neighboring Y frequencies toward 114.

Human light chain analysis

Diverse lengths and frequencies observed in κ and λ light chain CDRs highlight structural variability

We found that kappa (κ) and lambda (λ) light chain CDR1s appear to have at least four different lengths (Figure 1(b, 1);

Table S5, S6). CDR λ 1 lengths for κ are either 6, 7, 11, or 12 amino acids, while lengths of 8 to 10 amino acids are not observed, the frequency 6 (56.4%) and 7 amino acids (25.9%) are highest. In contrast, CDR λ 1 lengths for λ are either 6, 7, 8, or 9 amino acids long, while all are comparably frequent (6 amino acids highest with 35.6%), 7 amino acids is rarely used (5.8%). CDR2 (Figure 1e, 1) for the two types of light chains is almost exclusively restricted to a length of 3 amino acids (99.8% for κ and 92.4% for λ), with an occasional occurrence of a length of 7 amino acids for λ . CDR3 (Figure 1g, 1) shows much less diversity than the correspondent region in the heavy chain, possibly due to the lack of a D gene segment, with essentially four lengths in both types, lengths from 9 to 12 amino acids for λ and lengths from 8 to 11 amino acids for κ and with one particular length comprising the majority of sequences, 9 amino acids for κ (69.9%) and 11 amino acids for λ (46.8%).

Diverse amino acid distributions in κ and λ light chain CDRs signify unique structural characteristics

We also evaluated the average amino acid distribution per CDR for both κ and λ light chains (Figure 2b; Table S7). For comparison, the amino acid distribution for FRs and Con of light chain immunoglobulins, as well as for globular and transmembrane proteins¹⁹ are shown. In general, the occurrence of hydroxyl residues is higher than in FRs, Con, Glob, and TMB, being specially marked for CDR κ 1, CDR κ 2, and CDR λ 3-10aa. The usage of Q is also increased for CDR κ 1 and CDR κ 3 of all lengths, which seems to decrease in longer CDR κ 3. CDR1 of both types display a similar distribution pattern but with some key differences; the usage of S is almost twice as high in CDR κ 1 than CDR λ 1 and the occurrence of K is unusually high in CDR λ 1 (15.9%), while it is less than 1% in CDR κ 1. CDR2 of both types are remarkably different; A and S are the two most abundant amino acids in CDR κ 2, comprising more than two-thirds of the total, however; on CDR λ 2, S only reaches 16.9% and the two most abundant amino acids are Q (23.2%) and D (20.3%). The CDR3s of κ and λ light chains exhibit similar distributions of amino acids with minor differences; the occurrence of Q, Y, and P is higher for CDR κ 3s than CDR λ 3s, while S is more abundant in CDR λ 3s. In CDR λ 3, the usage of P increases the longer the region becomes while Q decreases. Moreover, there is a negligible occurrence of the negatively charged residues aspartic and E. In CDR κ 3, aliphatic amino acids constitute more than 20%, where V is found at the highest frequency.

Distinct amino acid distribution patterns in κ and λ light chains highlight positional preferences and differing diversity levels across CDRs

We analyzed the amino acid distribution per position for κ and λ light chains (Figure 4). For CDR1 and CDR2, the most abundant length is shown. For CDR3, the amino acid occurrence per position of those lengths with more than 50 sequences in the dataset are shown. The values for all lengths are available in supplementary materials. In general, the diversity of CDR1 in κ and λ light chain is low with one amino acid dominating at most positions (Figure 4a–4). CDR κ 1 and CDR λ 1 show a distinct characteristic distribution of amino acids where only position 38 has similar proportions of Y (51.78% for κ and 67.01%

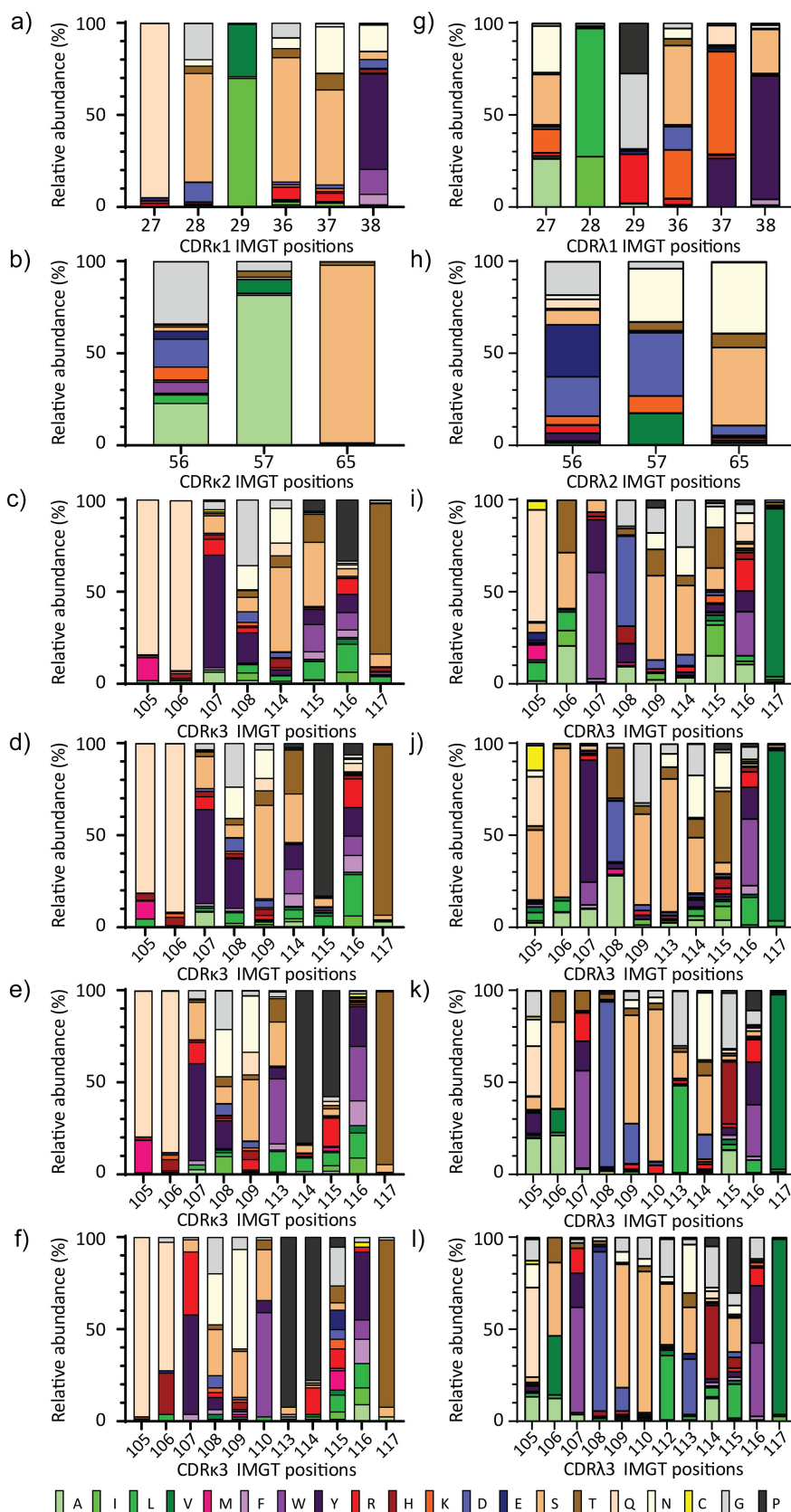


Figure 4. Position specific amino acid distribution in κ and λ light chain CDRs. The y-axis shows the relative abundance of each amino acid while the x-axis shows the positions for CDR1, CDR2, and CDR3 according to IMGT numbering scheme. a) CDRk1-6aa b) CDRk2 c) CDRk3-8aa d) CDRk3-9aa e) CDRk3-10aa f) CDRk3-11aa g) CDRλ1-6aa h) CDRλ2 i) CDRλ2-9aa j) CDRλ2-10aa k) CDRλ2-11aa l) CDRλ2-12aa.

in λ). Positions 29 in κ and 28 in λ have similar physicochemical properties in terms of chemical class and hydrophobicity; however, the most abundant amino acid in each position is I for κ and L for λ (69.7% and 69.6%, respectively). Position 27 is governed by Q for κ but for λ , with almost equal representation of four different amino acids; A (26.3%), K (12.9%), S (27.32%), and asparagine (N) (25.26%). Positions 28, 36, and 37 in κ light chains are dominated by S with an occurrence of more than 50% for all cases. Leucine, G, and K are the dominant amino acids in these λ positions, with L and K exceeding 50% abundance. Position 29 in λ light chain appears to be either positively charged or displaying short slightly hydrophobic amino acids such as P and G. Position 36 in λ light chains is the second highest in diversity below position 27 with S, D, and K as the biggest contributors.

The CDR2 has low diversity except for positions 56 and 57, which show moderate diversity. Position 56 in CDR κ 2 has the most moderate diversity, while position 65 is always S and position 57 is mostly A (over 80%). Positions 56 and 57 are unique to κ and λ CDR2 and show differences in the types of amino acids. In the κ light chain, position 56 has a high presence of G, A, and D, while in the λ light chain, only D is highly present. This position prefers negatively charged residues such as D and E (49%). Position 57 is dominated by A in κ (81.6%) and in λ , it has a mix of V, D, and Q (17%, 34.3%, 28.9%, respectively). Both chains have S as the main amino acid at position 65, but λ also has a presence of Q (38.5%).

The diversity of CDR3s in light chains is higher than for CDR1 and CDR2. Nevertheless, there are strong preferences for specific amino acids in the different positions and lengths. K light chain shows a high predilection for Q in positions 105 and 106, and T in position 117. The λ light chain, on the contrary, shows higher diversity in 105 and 106, and an almost exclusive usage of V in position 117. The occurrence of P is prominent for κ in the seventh position of each length, with over 80% in CDR2 lengths from 9 to 11 amino acids and for 10 and 11 amino acids with an additional neighboring position with another 60% to 80% frequencies of P. The remaining inner positions seem to follow a similar pattern throughout the different CDR κ 3 lengths, with an overall preference for Y, S, G, and Q. Positions 107 and 108 display a similar pattern throughout the different lengths of CDR λ 3, where aromatic in the former and negatively charged residues in the latter are largely predominant and gradually increase in abundance. The remaining inner positions in all lengths are dominated by S, L, Q, H, and Y. For the κ and λ light chains, two groups can be formed based on amino acid distribution patterns, which divide the shorter from the longer lengths, i.e., lengths 8 and 9 for κ , and 9 and 10 for λ resemble much more to each other than if they were to be compared with the longer sequences.

Discussion

The study of the amino acid composition of antibodies can be traced back to the pioneers of immunogenetics and antibody structure, Wu and Kabat,²¹ who in 1977 discussed the importance of specific amino acids in antigen

recognition, such as H and tryptophan (W). Understanding how antibodies vary in their amino acid makeup is key to developing new, targeted treatments for diseases. By studying how the immune system creates specific antibodies for different threats, we can improve therapies that rely on knowledge of these antibodies. This study aims to give a detailed overview of the diversity and amino acid composition of CDRs in human immunoglobulins.

The complexity of immunoglobulin diversity is the result of various genetic processes such as variable, diversity, and junction (V(D)J) gene rearrangements, junctional diversity, the combination of heavy and light chains, and somatic hypermutation. These processes contribute to the wide-ranging antibody repertoire, capable of recognizing billions of different targets. The CDR length diversity observed in this study (Figure 1) is explained by a combination of the rearrangement and junction diversification processes. The number of possible rearrangements for heavy chains is 20 times higher than for light chains, due to the lack of D gene segments for the latter. This partly explains why heavy chains show higher diversity in general than light chains. A bias toward certain gene segments, and therefore specific V(D)J associations, such as VH3–23/J4, VL3–19/J3, and VK3–20/J1, has been described before^{22,23} and it was confirmed in this study (Supplementary materials). A preference for specific gene segments could potentially bias the amino acid distribution shown in this study to the most common rearrangements. We observed a narrow length diversity in CDR1 and CDR2 of both chains and in the CDR3 of the light chain, while a large length diversity was observed for CDR3 of the heavy chain, ranging from 4 to 36 amino acids where 14 and 15 amino acids are the most common CDRH3 length, in accordance with previous studies.^{24,25} The differences in length variations between CDR1, CDR2, and CDR3 find explanation in several aspects. CDR1 and CDR2 for both chain types are encoded by only V-segments while CDR3 is encoded by the joining of V-genes with J- or D- and J-segments for light and heavy chains, respectively. Additionally, CDR3 diversity is further increased by the addition and deletion of P- and N-nucleotides at the junctions of V-J, V-D, and D-J. The latter plays an essential role in the larger length diversity of CDRH3, as it occurs in early B-cell differentiation where the enzyme deoxynucleotidyl transferase (TdT) is expressed during heavy chain rearrangements, causing the addition of non-templated nucleotides to DNA ends within the different segments. TdT expression decreases significantly before light chain rearrangements; therefore, N-diversity is mainly seen in heavy chain.²⁶ Moreover, it has been observed that V(DD)J rearrangements exist at low rates in human immunoglobulin heavy chains, which might be accountable for the presence of very long CDR3s in the heavy chain.²⁷ In our study, it seems that N-diversity is the primary process responsible for the very large length variation in the heavy chain CDR3, as the same region in the two types of light chain show a much more limited diversity.

CDRs of human immunoglobulins seem to overall favor certain amino acids compared to the general amino acid

usage in globular and transmembrane proteins, as well as when compared to non-CDR regions within antibodies such as S, D, Y, R, A, and G (Figure 2). This is in accordance with a previous similar study by Chen et al.,²⁸ focusing on CDRH3 of VH₃₋₂₃ × 01 antibody sequences, where >42% of the total residues in CDRH3 corresponded to Y, G, and D. Nevertheless, our study shows that this difference is especially noticeable for CDR1 and CDR2 of heavy and light chains and CDR3 of light chains. In contrast, CDRH3 shows an overall amino acid bias closer to globular proteins than any other region. These preferences can be understood in terms of the functional roles these amino acids play and explain the evolutionary mechanisms that enable antibodies to bind diverse antigens.

Aromatic residues like Y can be found at a higher frequency in the CDRs compared to other proteins.²⁹ These amino acids contribute to binding by several non-covalent interactions such as pi-stacking, cation and anion pi-interaction, hydrogen bonding, and sulfur-arene interactions, which might explain their higher preference compared to other amino acids as they provide binding versatility to antibodies targeting different antigens. In this study, we found that the proportion of aromatic residues, particularly Y, in the heavy chain, is in fact larger than for globular or transmembrane proteins, but the difference is less apparent for light chains, where aromatic residues appear at a lower frequency than other residues. In these cases, the amino acid that stands out is in most cases is the hydrophilic and polar S, which is often regarded as providing space and conformational flexibility and, in some cases, hydrogen-bonding capacities,³⁰ enhancing the binding capabilities of antibodies.

Tryptophan is frequently described to be the second most frequent aromatic residue found in CDRs of antibodies following Y, while F is less common.^{31,32} However, our results indicate that F is more frequent than W, particularly in CDRH1 in positions 28 and 30, where it is almost as prevalent as Y. The fact that light chain is rich in S while heavy chain is abundant in Y and F, might point toward light chain acting as a conformational adjuvant when both are paired, which together with its own binding capabilities, might be responsible for enhancing affinity and specificity of antibodies. Other hydrophilic and polar residues are also over-represented in CDRs, such as D and T in heavy and light chains, and Q in light chains. These amino acids are often regarded as contributing with short electrostatic interactions, as well as hydrogen bond formations, in the paratope/epitope interface.³³

Traditionally, R is indicated as the second most frequent amino acid below Y found in binding hotspots, defined as specific regions on a protein's surface that contributes significantly to the binding energy during the interaction with another protein.³⁴ However, contradictory effects on binding specificity in antibodies have been linked to Y and R, with increased usage of Y associated with improved specificity and increased usage of R linked to reduced specificity.³⁵⁻³⁷ This might explain the opposite distribution trends found in our analysis for CDRH3, with R gradually decreasing and Y increasing the longer the CDRH3 gets, and further elucidates the complex interplay of amino acid preferences in CDRs. We observed an overall large difference in amino acid usage

between κ and λ CDRs, which was also observed in a previous study by DeKosky et al.,³⁸ comparing human naïve and antigen-exposed antibody repertoires from three healthy human donors.

The differences in amino acid occurrence observed between heavy and light chains can be explained by examining the frequency of amino acids in specific positions in the CDRs of both chain types. In CDR1, only position 38 is similar, with high occurrences of Y in both heavy and light chains. Only position 65 in CDR2 shows some resemblance, with a preference for hydroxyl amino acids (S, T) and the aromatic Y, but with a clear predilection for T in heavy chains and S in light chains. CDR3 shows the highest dissimilarity in terms of the specific amino acids used in each position. However, in both heavy and light chains, the flanking positions have less diversity than the inner positions, though the specific amino acids used in these flanking positions are different between heavy and light chains. The position specific analysis of heavy chain reveals striking patterns of amino acid occurrence in distinct locations throughout all CDRs (Figure 3). Tyrosine seems to be used mainly in inner positions in all CDRH3, except for position 117, where it accounts for 57.1% in CDRH3-8aa and it gradually decreases to only 3.2% in CDRH3-25aa. Even though Y's occurrence in 117 diminishes with longer CDRH3s, its abundance increases in inner positions that are close to 117 (112.3 to 113). In 117, as Y decreases, another amino acid appears to increase, V, which is often disregarded as having a strong contribution in binding events.³⁹ Therefore, V, which is a largely hydrophobic and nonpolar residue, in this case, might have more structural importance for the correct conformation of these long CDRH3s. A similar pattern can be observed for F and M in position 115. In all other positions, M's occurrence is negligible, but in this position for CDRH3s longer than 20 amino acids, its abundance reaches 50%.

Methionine is normally underrepresented in proteins.⁴⁰ Nevertheless, the sulfur-aromatic interaction that M is capable of has been linked to increased stability in protein-protein interactions.⁴¹ Therefore, the high M occurrence in this specific position within CDRH3 might be key for the binding stability in long CDRHs. Moreover, in long CDRH3s, C seems to appear in higher abundance in inner positions than for short CDRH3s. Cysteines are often found in the paratopes of camelid and shark antibodies where they confer stability by creating intra- and inter-disulfide bridges that increase protein stability, in the extended CDRH3s commonly observed in these species.⁴² The presence of C residues in long human CDRH3s might indicate a similar function.⁴³ An interesting pattern that can be observed in CDRH3s of all lengths, is the high abundance of two amino acids with opposite charges, R, and D in positions 106 and 116, respectively. This might indicate a structural role of these positions within the CDR by stabilizing the loop with an intra ionic bridge. While the flanking positions in CDRH3 seem very fixed in particular amino acids, the increased diversity observed in inner positions (107 to 114) confers wide chemical versatility (Figure 4). These positions are rich in neutral amino acids of almost all possible sizes and chemical classes with a close to equal distribution of polar, nonpolar and hydrophobic, neutral, and hydrophilic residues. Amino acids with dual hydrogen bond

capabilities (donor and acceptor) are also abundant in these positions. CDRH1 displays positions with a very high abundance of aromatic residues, such as 28, 30, 37, and 38, which in many cases are flanked by structurally important residues that confer flexibility to the loop (S, T, and G). This pattern is also seen in CDRH2-7aa, but not in CDRH2-8aa, where aromatic residues are greatly underrepresented. Instead, amino acids that confer flexibility and structural stability, such as I, G, and P, are observed with high abundance in specific positions. The inner positions of CDRH2-8aa show the highest diversity in this region, with S, D and its amide derivative, N, playing an important role. The latter has been observed to be favored over Q in binding hotspots.³⁴

In κ and λ CDRs position-specific analysis (Figure 4), it can be observed that Y is constrained to position 38 in CDR1 and the inner positions of CDR3. Surprisingly, CDR2 is devoid of any aromatic residues and has a rather strong preference for negatively charged residues in positions 56 and 57, especially for λ light chains. CDR1, on the contrary, displays a tendency toward polar uncharged amino acids in κ and positively charged in λ . In both cases, one position with the exclusive occurrence of the isomers L (λ) and I (κ) can be observed, which might play a role in the structural stability of the loops. In CDR3 of both types, Y and W are found at high frequencies in position 107 often with S, D, N, and G in flanking positions. This might indicate that position 107 is of high importance in binding. The flanking positions of CDR3 (105 and 117) are very limited in diversity with mainly one amino acid dominating the position: Q in 105 of κ and λ , and T or V in 117 of κ and λ , respectively. In the work of Hu et al.,⁴⁴ Q was found to have a medium enrichment in binding hotspots, while T and V were in the top four residues with lowest enrichment. This may suggest that these positions are not involved in direct antigen binding, but rather aid in the overall stability of the light chain CDR3. In fact, N, Q, D, and especially P are more frequent in loops than in other secondary structures.⁴⁵ Furthermore, the usage of P is remarkable in positions 114 and 114 of CDR κ 3-10aa and 113 and 114 in CDR κ 3-11aa.

In conclusion, a thorough depiction of the defining features of an antibody was conducted by analyzing the CDRs using more than 11,000 heavy-chain sequences and 5,000 light-chain sequences for amino acid composition and length distribution. The results were consistent with current knowledge about immunoglobulins, including high diversity in length and amino acid variation in CDRH3, increased use of aromatic amino acids, such as Y, and charged and polar amino acids like D and S, as well as flexible amino acids like G. Additionally, these patterns could be linked to specific positions within each CDR. Further, we hope, it will improve and streamline applications in enhancing the binding strength of existing antibodies, designing synthetic antibody libraries, or creating new antibodies through computational methods.

Materials and methods

The data set was retrieved, extracted, and analyzed using The International Immunogenetics Information System (IMGT) databases and tools: IMGT/LIGM-DB⁴⁶ for data acquisition IMGT/V-QUEST (Version 3.5.30) IMGT/HighV-QUEST

(Version 1.8.5)^{47–49} for sequence analysis based on IMGT ontology concepts and IMGT scientific rules, and IMGT/StatClonotype⁵⁰ for the statistical analysis of IMGT/HighV-QUEST output. IMGT numbering scheme is used here to refer to framework positions (FR) and CDR positions in the variable domains. Statistical analysis and graphics were generated with GraphPad Prism version 9.4.1 for Windows, GraphPad Software, San Diego, California, USA, www.graphpad.com. Additional data for calculating average amino acid distributions in the framework of heavy and light chains variable regions were obtained using the database abYsis.⁵¹ Trastuzumab⁵² was used as an IgG model for constant region average amino acid occurrence calculation.

The following ontology concepts were used for retrieving the data in IMGT/LIGM-DB; Species = "Homo Sapiens", Molecule type = "cDNA", Configuration type = "rearranged", Gene type = "Variable", Functionality = "productive", Chain type = "IG-Heavy" or "IG-Light". The query rendered 11,469 and 5505 sequences for heavy and light chains, respectively. The sequences were submitted for analysis to IMGT/HighV-QUEST with the following parameters: Species = "Homo Sapiens (Human)", Receptor type or locus = IGH, IGK or IGL for the heavy chain, kappa (κ) light chain, and lambda (λ) light chain, respectively, and IMGT/V-QUEST reference directory set = "F+ORF+ in-frame P". The results were analyzed statistically using IMGT/HighV-QUEST and IMGT/StatClonotype. The statistical analysis allowed for the removal of redundant, out-of-frame and unproductive antibody sequences (Table S1).

Of the 11,469 submitted sequences for the heavy chain, 99.83% were in-frame productive sequences and 91.92% could be assigned to an IMGT Clonotype. This initial analysis filtered out 15.6% of sequences and the rest were analyzed for their length distribution and amino acid composition in the CDRs of the heavy chain (CDRHs). For the light chain, 5505 sequences were submitted for analysis at IMGT/HighV-QUEST selecting κ and λ locus separately. A total of 2907 of the sequences could be assigned to κ light chain, where 98.91% corresponded to in-frame productive sequences and 96.33% of these could be assigned to a κ light chain IMGT clonotype. For λ light chain, 1803 sequences were assigned to λ genes, with 99.15% corresponding to in-frame productive sequences. However, only 57.96% of these could be assigned to λ light chain IMGT clonotypes.

Abbreviation

Ab	Antibodies
BCR	B-cell Receptors
cDNA	Complementary Deoxyribonucleic Acid
CDR	Complementarity Determining Regions
CDRH	Complementarity Determining Regions in Heavy Chains
CDRH1	Complementarity Determining Regions Heavy Chain 1
CDRH2	Complementarity Determining Regions Heavy Chain 2
CDRH3	Complementarity Determining Regions Heavy Chain 3

CDRHs	Complementary Determining Regions of the Heavy Chain
CDRk1	Complementary Determining Region Kappa 1
CDRk2	Complementary Determining Region Kappa 2
CH	Constant Heavy Chain
CL	Constant Light Chain
Con	Constant Regions
FR	Framework Positions
Ig	Immunoglobulins
IgG	Immunoglobulin G
IGH	Immunoglobulin Heavy
IGK	Immunoglobulin Kappa
IGL	Immunoglobulin Lambda
IMGT	International Immunogenetics Information System
TdT	Terminal Deoxynucleotidyl Transferase
TMB	Transmembrane Proteins
V(D)J	Variable (Diversity) Joining
V(DD)J	Variable (Diversity Diversity) Joining
VH	Variable Heavy Chain
VH3–23/J4	Variable heavy 3–23/Junction 4
VK3–20/J1	Variable kappa 3–20/Junction 1
VL	Variable Light Chain
VL3–19/J3	Variable light 3–19/Junction 3









Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by The Novo Nordisk Foundation under Grant NNF19SA0056783, NNF19SA0057794, and NNF20SA0066621.

ORCID

Oscar Mejias-Gomez  <http://orcid.org/0000-0002-1026-6692>
 Andreas V. Madsen  <http://orcid.org/0000-0002-8449-9691>
 Kerstin Skovgaard  <http://orcid.org/0000-0001-5663-4879>
 Lasse E. Pedersen  <http://orcid.org/0000-0002-6064-919X>
 J. Preben Morth  <http://orcid.org/0000-0003-4077-0192>
 Timothy P. Jenkins  <http://orcid.org/0000-0003-2979-5663>
 Peter Kristensen  <http://orcid.org/0000-0001-7205-6853>
 Steffen Goletz  <http://orcid.org/0000-0003-1463-5448>

References

- Kienzler A-K, Eibel H. Human B cell development and tolerance. In: Ratcliffe MJH, editor. *Encyclopedia of immunobiology*. United Kingdom: Elsevier; 2016. p. 105–21.
- Chi X, Li Y, Qiu X. V(D)J recombination, somatic hypermutation and class switch recombination of immunoglobulins: mechanism and regulation. *Immunology*. 2020;160(3):233–47. doi:10.1111/imm.13176.
- Schroeder HW. The evolution and development of the antibody repertoire. *Front Immunol*. 2015;6:6. doi:10.3389/fimmu.2015.00033.
- Tonegawa S. Somatic generation of antibody diversity. *Nature*. 1983;302(5909):575–81. doi:10.1038/302575a0.
- Kunik V, Peters B, Ofran Y, Baker B. Structural consensus among antibodies defines the antigen binding site. *PLoS Comput Biol* [Internet]. 2012;8(2):e1002388. [accessed 2022 Nov 16]. doi:10.1371/journal.pcbi.1002388.
- Davies DR, Cohen GH. Interactions of protein antigens with antibodies. *Proc Natl Acad Sci USA*. 1996;93(1):7–12.
- Davies DR, Padlan EA, Sheriff S. Antibody-antigen complexes. *Annu Rev Biochem*. 1990;59(1):439–73. doi:10.1146/annurev.bi.59.070190.002255.
- Fellouse FA, Wiesmann C, Sidhu SS. Synthetic antibodies from a four-amino-acid code: a dominant role for tyrosine in antigen recognition. *Proc Natl Acad Sci U S A*. 2004;101(34):12467–72. doi:10.1073/pnas.0401786101.
- Fellouse FA, Esaki K, Birtalan S, Raptis D, Cancasci VJ, Koide A, Jhurani P, Vasser M, Wiesmann C, Kossiakoff AA, et al. High-throughput generation of synthetic antibodies from highly functional minimalist phage-displayed libraries. *J Mol Biol*. 2007;373(4):924–40. doi:10.1016/j.jmb.2007.08.005.
- Nguyen MN, Pradhan MR, Verma C, Zhong P, Valencia A. The interfacial character of antibody paratopes: analysis of antibody-antigen structures. *Bioinformatics* [Internet]. 2017;33:2971–76. [accessed 2022 Nov 16]. doi:10.1093/bioinformatics/btx389. <https://pubmed.ncbi.nlm.nih.gov/28633399/>.
- Wang M, Zhu D, Zhu J, Nussinov R, Ma B. Local and global anatomy of antibody-protein antigen recognition. *J Mol Recogn* [Internet]. 2018; 31:e2693. accessed 2022 Nov 16. doi:10.1002/jmr.2693.
- Winter G, Milstein C. Man-made antibodies. *Nature*. 1991;349(6307):293–99. doi:10.1038/349293a0.
- Lu R-M, Hwang Y-C, Liu I-J, Lee C-C, Tsai H-Z, Li H-J, Wu H-C. Development of therapeutic antibodies for the treatment of diseases. *J Biomed Sci*. 2020;27(1):1. doi:10.1186/s12929-019-0592-z.
- Akbar R, Robert PA, Pavlović M, Jeliazkov JR, Snapkov I, Slabodkin A, Weber CR, Scheffer L, Miho E, Haff IH, et al. A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *Cell Rep*. 2021;34(11):108856. doi:10.1016/j.celrep.2021.108856.
- Shi B, Ma L, He X, Wang X, Wang P, Zhou L, Yao X. Comparative analysis of human and mouse immunoglobulin variable heavy regions from IMGT/LIGM-DB with IMGT/HighV-QUEST. *Theor Biol Med Model*. 2014;11(1):30. doi:10.1186/1742-4682-11-30.
- Wu L, Oficjalska K, Lambert M, Fennell BJ, Darmanin-Sheehan A, Ni Shuilleabháin D, Autin B, Cummins E, Tchistiakova L, Bloom L, et al. Fundamental characteristics of the immunoglobulin VH repertoire of Chickens in comparison with those of Humans, mice, and camelids. *J Immunol*. 2012;188(1):322–33. doi:10.4049/jimmunol.1102466.
- Ivanov II, Schelonka RL, Zhuang Y, Gartland GL, Zemlin M, Schroeder HW. Development of the Expressed Ig CDR-H3 repertoire is Marked by focusing of constraints in length, amino acid use, and charge that are first established in early B cell progenitors. *J Immunol*. 2005;174:7773–80. doi:10.4049/jimmunol.174.12.7773.
- Lefranc M-P, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, Carillon E, Duvergey H, Houles A, Paysan-Lafosse T, et al. IMGT®, the international ImmunoGeneTics information system® 25 years on. *Nucleic Acids Res*. 2015;43(D1):D413–22. doi:10.1093/nar/gku1056.
- Gromiha MM, Suwa M. A simple statistical method for discriminating outer membrane proteins with better accuracy. *Bioinformatics* [Internet]. 2005;21(7):961–68. [accessed 2022 Nov 10]. doi:10.1093/bioinformatics/bti126.
- Pommié C, Levadoux S, Sabatier R, Lefranc G, Lefranc MP. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J Mol Recognit* [Internet]. 2004;17:17–32. [accessed 2022 Nov 11]. doi:10.1002/jmr.647. <https://pubmed.ncbi.nlm.nih.gov/14872534/>.
- Kabat EA, Wu TT, Bilofsky H. Unusual distributions of amino acids in complementarity-determining (hypervariable) segments of heavy and light chains of immunoglobulins and their possible roles in specificity of antibody-combining sites. *J Biol Chem*. 1977;252(19):6609–16. doi:10.1016/S0021-9258(17)39891-5.

22. Lee A, Desravines S, Hsu E. IgH diversity in an individual with only one million B lymphocytes. *Dev Immunol*. 1993;3(3):211–22. doi:10.1155/1993/17249.
23. Volpe JM, Kepler TB. Large-scale analysis of human heavy chain V (D)J recombination patterns. *Immunome Res [Internet]*. 2008;4:3. [accessed 2022 Nov 14]. doi:10.1186/1745-7580-4-3. /pmc/articles/PMC2275228/.
24. Zemlin M, Klinger M, Link J, Zemlin C, Bauer K, Engler JA, Schroeder HW, Kirkham PM. Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. *J Mol Biol*. 2003;334(4):733–49. doi:10.1016/j.jmb.2003.10.007.
25. Wu T, Te Johnson G, Kabat EA. Length distribution of CDRH3 in antibodies. *Proteins: Struct Funct Genet*. 1993;16:1–7. doi:10.1002/prot.340160102.
26. Hong B, Wu Y, Li W, Wang X, Wen Y, Jiang S, Dimitrov DS, Ying T. In-depth analysis of human neonatal and adult IgM antibody repertoires. *Front Immunol [Internet]*. 2018; 9:128. [accessed 2022 Nov 14]. doi:10.3389/fimmu.2018.00128. /pmc/articles/PMC5807330/.
27. Briney BS, Willis JR, Hicar MD, Thomas JW, Crowe JE. Frequency and genetic characterization of V(DD)J recombinants in the human peripheral blood antibody repertoire. *Immunology [Internet]*. 2012;137:56. [accessed 2022 Nov 14]. doi:10.1111/j.1365-2567.2012.03605.x. /pmc/articles/PMC3449247/.
28. Chen L, Duan Y, Benatuil L, Stine WB. Analysis of 5518 unique, productively rearranged human VH3-23*01 gene sequences reveals CDR-H3 length-dependent usage of the IGHD2 gene family. *Protein Eng Des Sel*. 2017;30(9):603–09. doi:10.1093/protein/gzx027.
29. Peng HP, Lee KH, Jian JW, Yang AS. Origins of specificity and affinity in antibody-protein interactions. *Proc Natl Acad Sci U S A*. 2014;111:E2656–65. accessed 2022 Nov 14. doi:10.1073/pnas.1401131111.
30. Bischoff R, Schlüter H. Amino acids: chemistry, functionality and selected non-enzymatic post-translational modifications. *J Proteomics*. 2012;75(8):2275–96. doi:10.1016/j.jprot.2012.01.041.
31. Ofra Y, Schlessinger A, Rost B. Automated identification of complementarity Determining regions (CDRs) reveals peculiar characteristics of CDRs and B cell epitopes. *J Immunol [Internet]*. 2008;181(9):6230–35. [accessed 2022 Nov 14]. doi:10.4049/jimmunol.181.9.6230.
32. Crowley PB, Golovin A. Cation- π interactions in protein-protein interfaces. *Proteins Struct Funct Bioinf*. 2005;59(2):231–39. doi:10.1002/prot.20417.
33. Yu C-M, Peng H-P, Chen I-C, Lee Y-C, Chen J-B, Tsai K-C, Chen C-T, Chang J-Y, Yang E-W, Hsu P-C, et al. Rationalization and design of the complementarity Determining region sequences in an antibody-antigen recognition interface. *PLoS One*. 2012;7(3):e33340. doi:10.1371/journal.pone.0033340.
34. Moreira IS, Fernandes PA, Ramos MJ. Hot spots—A review of the protein-protein interface determinant amino-acid residues. *Proteins: Struct, Funct, And Bioinf [Internet]*. 2007;68:803–12. [accessed 2022 Nov 14]. doi:10.1002/prot.21396.
35. Birtalan S, Zhang Y, Fellouse FA, Shao L, Schaefer G, Sidhu SS. The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies. *J Mol Biol*. 2008;377(5):1518–28. doi:10.1016/j.jmb.2008.01.093.
36. Koide S, Sidhu SS. The importance of being tyrosine: lessons in molecular recognition from minimalist synthetic binding proteins. *ACS Chem Biol [Internet]*. 2009;4:325–34. [accessed 2022 Nov 14]. doi:10.1021/cb800314v.
37. Tiller KE, Li L, Kumar S, Julian MC, Garde S, Tessier PM. Arginine mutations in antibody complementarity-determining regions display context-dependent affinity/specificity trade-offs. *J Biol Chem*. 2017;292(40):16638–52. doi:10.1074/jbc.M117.783837.
38. DeKosky BJ, Lungu OI, Park D, Johnson EL, Charab W, Chrysostomou C, Kuroda D, Ellington AD, Ippolito GC, Gray JJ, et al. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc Natl Acad Sci USA* 2016;113(19):113.
39. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. *J Mol Biol*. 1998;280(1):1–9. doi:10.1006/jmbi.1998.1843.
40. Aledo JC, Cantón FR, Veredas FJ. Sulphur atoms from Methionines Interacting with aromatic residues are less prone to oxidation. *Sci Rep [Internet]*. 2015;5:1–14. [accessed 2022 Nov 14]. doi:10.1038/srep16955.
41. Valley CC, Cembran A, Perlmutter JD, Lewis AK, Labello NP, Gao J, Sachs JN. The methionine-aromatic motif plays a unique role in stabilizing protein structure. *J Biol Chem [Internet]*. 2012;287(42):34979–91. [accessed 2022 Nov 14]. doi:10.1074/jbc.M112.374504.
42. Arbabi-Ghahroudi M. Camelid Single-Domain antibodies: historical perspective and future outlook. *Front Immunol*. 2017;8:1589. doi:10.3389/fimmu.2017.01589.
43. Almagro JC, Raghunathan G, Beil E, Janecki DJ, Chen Q, Dinh T, Lacombe A, Connor J, Ware M, Kim PH, et al. Characterization of a high-affinity human antibody with a disulfide bridge in the third complementarity-determining region of the heavy chain. *J Mol Recog [Internet]*. 2012;25:125–35. [accessed 2022 Nov 14]. doi:10.1002/jmr.1168.
44. Hu Z, Ma B, Wolfson H, Nussinov R. Conservation of polar residues as hot spots at protein interfaces. *Proteins Struct Funct Bioinf [Internet]*. 2000;39:331–42. doi:10.1002/(SICI)1097-0134(20000601)39:4<331:AID-PROT60>3.0.CO.
45. Otaki JM, Tsutsumi M, Gotoh T, Yamamoto H. Secondary structure characterization based on amino acid composition and availability in proteins. *J Chem Inf Model [Internet]*. 2010;50:690–700. [accessed 2022 Nov 15]. doi:10.1021/ci900452z.
46. Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, Lefranc MP. IMGT/LIGM-DB, the IMGT® comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Res [Internet]*. 2006;34:D781–4. [accessed 2022 Jun 17]. doi:10.1093/nar/gkj088. https://academic.oup.com/nar/article/34/suppl_1/D781/1133248.
47. Alamyar E, Duroux P, Lefranc MP, Giudicelli V. IMGT(*) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods Mol Biol [Internet]*. 2012;882:569–604. [accessed 2022 Jun 17]. https://pubmed.ncbi.nlm.nih.gov/22665256/.
48. Li S, Lefranc MP, Miles JJ, Alamyar E, Giudicelli V, Duroux P, Freeman JD, Corbin VDA, Scheerlinck JP, Frohman MA, et al. Imgt/highv QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nat Commun [Internet]*. 2013;4(1): [accessed 2022 Jun 17]. doi:10.1038/ncomms3333.
49. G V. From IMGT-ONTOLOGY to IMGT/HighVQUEST for NGS immunoglobulin (IG) and T cell receptor (TR) repertoires in autoimmune and infectious diseases. *Autoimmun And Infect Dis: Open Access (ISSN 2470-1025)*. 2015;1:1. doi:10.16966/2470-1025.103.
50. Aouinti S, Malouche D, Giudicelli V, Kossida S, Lefranc MP, Allen RL. IMGT/HighV-QUEST statistical significance of IMGT clonotype (AA) diversity per gene for standardized comparisons of next generation sequencing immunoprofiles of immunoglobulins and T cell receptors. *PLoS One*. 2015;10(11):10. doi:10.1371/journal.pone.0142353.
51. Swindells MB, Porter CT, Couch M, Hurst J, Abhinandan KR, Nielsen JH, Macindoe G, Hetherington J, Martin ACR. abYsis: Integrated antibody sequence and structure—management, analysis, and prediction. *J Mol Biol*. 2017;429:356–64. doi:10.1016/j.jmb.2016.08.019.
52. Pegram MD, Lipton A, Hayes DF, Weber BL, Baselga JM, Tripathy D, Baly D, Baughman SA, Twaddell T, Glaspy JA, et al. Phase II study of receptor-enhanced chemosensitivity using recombinant humanized anti-p185HER2/neu monoclonal antibody plus cisplatin in patients with HER2/neu-overexpressing metastatic breast cancer refractory to chemotherapy treatment. *J Clin Oncol*. 1998;16:2659–71. doi:10.1200/JCO20161682659.