Antibody Informatics: IMGT, the International ImMunoGeneTics Information System

MARIE-PAULE LEFRANC
IMGT, the international ImMunoGeneTics information system, Laboratoire d’Immunogénétique Moléculaire LIGM, Université Montpellier 2, Institut de Généétique Humaine IGH, UPR CNRS 1142, Montpellier, 34396 cedex 5, France

ABSTRACT Antibody informatics, a part of immunoinformatics, refers to the concepts, databases, and tools developed and used to explore and to analyze the particular properties of the immunoglobulins (IG) or antibodies, compared with conventional genes and proteins. Antibody informatics is based on a unique ontology, IMGT-ONTOLOGY, created in 1989 by IMGT, the international ImMunoGeneTics information system (http://www.imgt.org). IMGT-ONTOLOGY defined, for the first time, the concept of ‘genes’ for the IG and the T cell receptors (TR), which led to their gene and allele nomenclature and allowed their entry in databases and tools. A second IMGT-ONTOLOGY revolutionizing and definitive concept was the IMGT unique numbering that bridged the gap between sequences and structures for the variable (V) and constant (C) domains of the IG and TR, and for the groove (G) domains of the major histocompatibility (MH). These breakthroughs contributed to the development of IMGT databases and tools for antibody informatics and its diverse applications, such as repertoire analysis in infectious diseases, antibody engineering and humanization, and study of antibody/antigen interactions. Nucleotide sequences of antibody V domains from deep sequencing (Next Generation Sequencing or High Throughput Sequencing) are analyzed with IMGT/HighV-QUEST, the high-throughput version of IMGT/V-QUEST and IMGT/JunctionAnalysis. Amino acid sequences of V and C domains are represented with the IMGT/Collier-de-Perles tool and analyzed with IMGT/DomainGapAlign. Three-dimensional (3D) structures (including contact analysis and paratope/epitope) are described in IMGT/3Dstructure-DB. Based on a friendly interface, IMGT/mAb-DB contains therapeutic monoclonal antibodies (INN suffix–mab) that can be queried on their specificity, for example, in infectious diseases, on bacterial or viral targets.

INTRODUCTION

The efficiency of the adaptive immune response and its capability of recognizing a large number of different antigens depend on the huge diversity of the antigen receptors, immunoglobulins (IG), or antibodies of the B lymphocytes and T cell receptors (TR) of the T lymphocytes. The genes that code the IG and TR are highly polymorphic and are organized in clusters in several loci (three loci for IG and four for TR in humans) located on different chromosomes (four in humans) in the genome (1, 2). The molecular synthesis of the IG and TR chains is particularly complex and unique. It includes several mechanisms that occur at the DNA level: combinatorial rearrangements of the variable (V), diversity (D), and joining (J) genes that code the IG and TR variable domain, exonuclease trimming at the ends of the V, D, and J genes, and the random addition of nucleotides by the terminal deoxynucleotidytransferase (TdT) that create...
the junction N diversity and, for IG, somatic hypermutations (1, 2). The IG and TR repertoires show an extraordinary diversity with a potential of $10^{12}$ IG and $10^{12}$ TR per individual, and the only limiting factor is the number of genetically programmed B and T cells for an organism. Therefore, the analysis of the IG and TR genes and of their expression represents a crucial challenge for the understanding of the immune response in normal and pathological situations.

IMGT, the international ImMunoGeneTics information system (http://www.imgt.org) (3), was created in 1989 at Montpellier, France (CNRS and Université Montpellier 2) to standardize the immunogenetics data and to manage the huge diversity of the antigen receptors, IG and TR (1, 2). New concepts, databases, and tools were developed by IMGT to explore and to analyze the particular properties of the IG and TR, compared with conventional genes and proteins. They led to the emergence of a new science, immunoinformatics, and, owing to the huge development of therapeutic antibodies, to antibody informatics, the part of immunoinformatics dealing with IG.

THE BASICS FOR ANTIBODY INFORMATICS
IMGT Gene and Allele Nomenclature
The accuracy and the consistency of the IMGT data are based on IMGT-ONTOLOGY, the first and, so far, unique ontology for immunogenetics and immunoinformatics (4, 5, 6, 7, 8, 9, 10). IMGT-ONTOLOGY defined, for the first time, the concept of ‘genes’ for the IG and TR, which led to their gene and allele nomenclature (e.g.,IGHV1-2*01) and allowed their entry in databases and tools (11). The IMGT IG and TR gene nomenclature (1, 2, 12, 13) was approved at the international level by the Human Genome Organisation (HUGO) Nomenclature Committee (HGNC) in 1999 (14, 15) and endorsed by the World Health Organization - International Union of Immunological Societies (WHO-IUIS) Nomenclature Committee (16, 17). The IMGT IG and TR gene names are the official reference for the genome projects and, as such, have been entered in IMGT/GENE-DB (18), which is the global reference for IG and TR alleles assigned by the WHO-IUIS-IMGT Nomenclature Subcommittee and, so far, the only database managing IG and TR alleles. Since 2008, IMGT gene and allele names have been used by the WHO-International Nonproprietary Names (WHO-INN) program in the definition of therapeutic monoclonal antibodies (mAb, INN suffix -mab), fusion proteins for immune applications (FPIA, INN suffix -cept), and composite proteins for clinical applications (CPCA) with Fc for increased half-life (24, 25), and the corresponding sequences and data have been entered in IMGT/2Dstructure-DB (26) and in IMGT/mAb-DB (27) (http://www.imgt.org).

IMGT Unique Numbering
A second IMGT-ONTOLOGY revolutionizing concept was the IMGT unique numbering (28, 29, 30, 31, 32, 33), which, with its two-dimensional (2D) representation, the IMGT Collier de Perles (34, 35, 36, 37, 38), bridged for the first time, and definitively, the gap between sequences and structures for the variable (V) and constant (C) domains of the immunoglobulin superfamily (IgSF) proteins (28, 29, 30, 31, 32), and for the groove (G) domains of the major histocompatibility (MH) superfamily (MhSF) (32, 33). The V domains include the V-DOMAIN of the IG and TR and the V-LIKE-DOMAIN of the IgSF other than IG and TR; the C domains include the C-DOMAIN of the IG and TR and the C-LIKE-DOMAIN of the IgSF other than IG and TR; whereas the G domains include the G-DOMAIN of the MH and the G-LIKE-DOMAIN of the MhSF other than MH (28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38).

IMGT anchors for V and C domains
One of the IMGT unique numbering major breakthroughs for antibody informatics was to define ‘anchors’ in the V and C domains, a new and basic concept for both sequences and structures, that definitively solved the heterogeneity and complexity of previous numberings, and also is valid for both V and C domains (28, 29, 30, 31). A V domain is made of nine antiparallel beta strands on two layers (designated as framework regions (FR) in a V-DOMAIN), forming a barrel structure or ‘sandwich’ with a hydrophobic inner core, and three loops (designated as complementarity determining regions (CDR) or hypervariable regions in a V-DOMAIN) (28, 29, 30) (Fig. 1A). Compared with a V domain, a C domain has seven antiparallel beta strands instead of nine, two loops instead of three, and a char-
acteristic CD transverse strand (in the absence of a C' strand, C'C" loop, and C" strand) (31) (Fig. 1B).

The IMGT anchors define the positions of the V or C domain strands that support the loops. Anchors are shown in squares in IMGT Colliers de Perles (Fig. 1). In a V domain, these anchors comprise the positions 26 and 39, 55 and 66, 104 and 118 that support the BC, C'C", and FG loops. In an IG or TR V-DOMAIN, these anchors are the positions of the FR-IMGT that support the CDR-IMGT (28, 29, 30). In a C domain, four anchors are identical to those of a V domain (positions 26 and 39, 104 and 118 that support the BC and FG loops) and two anchors (positions 45 and 77) delimit the CD strand (31).

**CDR-IMGT lengths for antibody humanization by grafting**

By allowing a correct delimitation of the FR-IMGT and CDR-IMGT, anchors led to the ‘CDR-IMGT lengths’ concept. The CDR-IMGT lengths are indicated between

---

**FIGURE 1** IMGT Collier de Perles for V-DOMAIN and C-DOMAIN. A, VH and V-KAPPA. B, CH1 and C-KAPPA. The VH, V-KAPPA, CH1, and C-KAPPA of the motavizumab antibody are shown as examples. IMGT Colliers de Perles are shown on one layer (top), on two layers (middle) (INN, from IMGT/2Dstructure-DB, http://www.imgt.org), and on two layers with hydrogen bonds (bottom) (3ixt_H and 3ixt_L, from IMGT/3Dstructure-DB, http://www.imgt.org) (26, 48). In the V-DOMAIN, anchors (positions 26 and 39, 55 and 66, 104 and 118) support the three CDR (CDR1-IMGT, CDR2-IMGT, and CDR3-IMGT that correspond to the BC, C'C", and FG loops, respectively) (29, 30, 32). The VH and V-KAPPA CDR-IMGT lengths are [10.7.12] and [5.3.9], respectively. In the C-DOMAIN, anchors (positions 26 and 39, 104 and 118) support the BC and FG loops as in a V-DOMAIN, whereas in the absence of a C' strand, C'C" loops, and C" strand, anchors 45 and 77 delimit the CD transverse strand (31, 32). A ban symbol indicates an error (K instead of CH1 R120) in the 3D structure PDB file (3ixt), detected by using the IMGT unique numbering and by comparison with the INN sequence. A similar type of error (A instead of CH1 V121) has also been detected by IMGT for the 3D structure of the anti-HIV b12 antibody (1hzh).

doi:10.1128/microbiolspec.AID-0001-2012.f1
Highly conserved amino acids

In the IMGT unique numbering, amino acids always have the same position number. Highly conserved amino acids of an IG or TR V and C domains are written in red online, in the IMGT Colliers de Perles. They comprise five positions in a V-DOMAIN: cysteine C23 (1st-CYS), tryptophan W41 (CONSERVED-TRP), hydrophobic amino acid 89, cysteine C104 (2nd-CYS), and phenylalanine F118 (J-PHE) or tryptophan W118 (J-TRP) (that belongs to the F/W-G-X-G motif encoded by the J-REGION) (Fig. 1A). In a C domain, there are four highly conserved amino acids: C23 (1st-CYS), W41 (CONSERVED-TRP), hydrophobic 89, and C104 (2nd-CYS), instead of five.

IMGT Standardized and Integrated System

In 2013, IMGT comprised seven databases, seventeen online tools, and more than 15,000 pages of Web resources and is the reference in immunogenetics and immunoinformatics. IMGT/ONTOLOGY concepts (identification, description, classification [11], numerotation [32, 37], and detailed protocols (IMGT/V-QUEST [44], IMGT/JunctionAnalysis [45], IMGT/DomainGapAlign [47], IMGT/3Dstructure-DB [48], IMGT/Collier-de-Perles [38]) have been published in Cold Spring Harbor Protocols (‘IMGT booklet’ freely downloadable in IMGT references, http://www.imgt.org).

The next sections illustrate the use of IMGT tools for the standardized IG sequences and structures analysis for the IG repertoire analysis in infectious diseases, antibody engineering and humanization, and study of antibody/antigen interactions. Nucleotide sequences from deep sequencing (Next Generation Sequencing [NGS] or High Throughput Sequencing [HTS]) of antibody V domains are analyzed with IMGT/HighV-QUEST (43), the high-throughput version of IMGT/V-QUEST and IMGT/JunctionAnalysis (44, 45, 46). Amino acid sequences of V and C domains are represented with the IMGT/Collier-de-Perles tool (38) and analyzed with IMGT/DomainGapAlign (26, 47). Three-dimensional (3D) structures (including contact analysis and paratope/epitope) are described in IMGT/3Dstructure-DB (26, 48). Based on a friendly interface, IMGT/mAb-DB (27) contains therapeutic monoclonal antibodies (and also FPIA and CPCA with Fc) that can be queried on their specificity, for example, in infectious diseases, on bacterial or viral targets.
IMGT/HighV-QUEST for Deep Sequencing
Since October 2010, the analysis of IG and TR nucleotide sequences obtained from deep sequencing (NGS or HTS) can be performed with IMGT/HighV-QUEST [43, 46] (http://www.imgt.org) that analyzes 500,000 sequences per run and performs statistical analysis on the results (version July 2012).

The IMGT/HighV-QUEST Search page is very similar to the IMGT/V-QUEST Search page [44, 46]. The results are provided in a downloadable main folder with eleven files in CSV format (results equivalent to those of the Excel file of IMGT/V-QUEST online), and one folder with the individual files (up to 500,000) of all the sequences results (results identical for each analyzed sequence to those of IMGT/V-QUEST online in Text, and corresponding to ‘Detailed view’) [43, 44, 46]. Text and CSV formats were chosen to facilitate statistical studies for further interpretation and knowledge extraction. The eleven files of the main folder provide, per analysis, a total of 513 to 576 columns of results (Table 1) and demonstrate that IMGT standards, based on IMGT-ONTOLOGY concepts, fit remarkably well with extensive and detailed NGS data analysis for antibody informatics. The results include the identification of the closest V and J genes and alleles (and D genes for VH), detailed IMGT/JunctionAnalysis results, analysis of the nucleotide (nt) sequence mutations, and analysis of the amino acid (AA) changes in the sequence translations. The AA changes are described for the hydropathy (3 classes), volume (5 classes), and IMGT physicochemical properties (11 classes that comprise aliphatic [A, V, I, L], acid [D, E], basic [H, K, R], amide [N, Q], hydroxyl [S, T], sulfur [C, M], G, P, Y, F, and W) [49] (IMGT Aide-memoire, in the ‘Amino acids’ section, http://www.imgt.org). Thus, S40→G (+ + +) means that the two AA involved in the change (S-G) at codon 40 belong to the same hydropathy (+) and volume (+) classes but to different physicochemical properties (−) classes [49]. It is the first time that such qualification of AA replacement is provided. This has led us to identify four types of AA changes: very similar (+++), similar (++−, +−+), dissimilar (−−+, −++, −−−) and very dissimilar (−−−) [44] (file 9 in Table 1).

IMGT/HighV-QUEST performs statistical analysis on the results of several runs up to 500,000 sequences. Detailed statistical analysis tables and histograms (e.g., V, D, and J usage, CDR3-IMGT [nt and AA lengths]), provided as PDF reports and separate graphical elements (figures in PNG format), are described elsewhere [43].

IMGT/HighV-QUEST achieved the same degree of resolution and high-quality results for antibody sequences analysis as IMGT/V-QUEST. Both tools use the same algorithm, and the user can evaluate the quality of his/her sequences before IMGT/HighV-QUEST analysis, by checking the results obtained with IMGT/V-QUEST on a few sequences. The eventual limitations of the IG V domain nucleotide sequence analysis using IMGT/HighV-QUEST are currently not on the antibody informatics side, but rather on the experimental and sequencing conditions that need to be improved (sequence length, avoiding PCR or sequencing errors) for a fully reliable and meaningful biological interpretation of the NGS or HTS results. By providing warnings and classification of sequences in categories (‘1 copy,’ ‘More than 1,’ ‘single allele,’ ‘several alleles’...), IMGT/HighV-QUEST helps the user in evaluating his/her own data quality and therefore represents the reference tool for antibody-standardized analysis of NGS or HTS data.

ANTIBODY V- AND C-DOMAIN 2D REPRESENTATIONS
IMGT Collier de Perles 2D Representation
IMGT Collier de Perles [34, 35, 36, 37, 38] is a graphical 2D representation of domain, based on the IMGT unique numbering [28, 29, 30, 31, 32, 33]. For the IG, the V domains include the V heavy (VH) and V light (VL) domains, the VL being V kappa (V-KAPPA) or V lambda (V-LAMBDA) in higher vertebrates, and V iota (V-IOTA) in fish. The C domains correspond either to a complete C-REGION (C-KAPPA of the IG-Light-Kappa chains or C-LAMBDA of the IG-Light-Lambda chains) or to part of it (e.g., CH1, CH2, and CH3 of the IG-Heavy-Gamma chains) [28, 29, 30, 31, 32].

In IMGT Collier de Perles (Fig. 1), amino acids are shown in the one-letter abbreviation. Anchor positions are shown in squares. Hatched circles correspond to missing positions according to the IMGT unique numbering [29, 30, 31, 32]. Strand positions at which hydrophobic amino acids (hydropathy index with positive value: I, V, L, F, C, M, A) and tryptophan (W) are found in more than 50% of sequences are shown with a blue background color. Arrows indicate the direction of the beta strands and their designations in 3D structures. The CDR-IMGT of the V-DOMAIN are colored according to the IMGT color menu which indicates the type of rearrangement, V-J or V-D-J [1, 2]. Thus, the IMGT color menu for CDR1-IMGT, CDR2-IMGT, and CDR3-IMGT is online red, orange, and purple for VH (encoded by a V-D-J-REGION resulting from a V-D-J rearrangement), and blue, green, and green-blue for
<table>
<thead>
<tr>
<th>File number</th>
<th>File name</th>
<th>Number of columns</th>
<th>Results content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>'Summary'</td>
<td>25–29</td>
<td>• Identity percentage with the closest V, D, and J genes and alleles</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• FR-IMGT and CDR-IMGT lengths</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Amino acid (AA) JUNCTION</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Description of insertions and deletions if any</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Nucleotide (nt) sequences gapped according to the IMGT unique numbering for V-D-J-REGION, V-J-REGION, V-REGION, FR1-IMGT, CDR1-IMGT, FR2-IMGT, CDR2-IMGT, FR3-IMGT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• nt sequences of CDR3-IMGT, JUNCTION, J-REGION and FR4-IMGT</td>
</tr>
<tr>
<td>2</td>
<td>'IMGT-gapped-nt-sequences'</td>
<td>18</td>
<td>• AA sequences gapped according to the IMGT unique numbering for the labels V-D-J-REGION, V-J-REGION, V-REGION, FR1-IMGT, CDR1-IMGT, FR2-IMGT, CDR2-IMGT, FR3-IMGT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• nt sequences of all labels that can be automatically annotated by IMGT/Automat.</td>
</tr>
<tr>
<td>3</td>
<td>'Nt-sequences'</td>
<td>63–78</td>
<td>• Same columns as 'IMGT-gapped-AA-sequences' (#4), but sequences of labels are without IMGT gaps.</td>
</tr>
<tr>
<td>4</td>
<td>'IMGT-gapped-AA-sequences'</td>
<td>18</td>
<td>• Results of IMGT/JunctionAnalysis (53 columns for IGL and IGK (also for TRA and TRG) sequences, 46 (if one D), 66 (if two D) or 77 (if 3 D) columns for IGH (also for TRB and TRD) sequences</td>
</tr>
<tr>
<td>5</td>
<td>'AA-sequences'</td>
<td>18</td>
<td>• List of mutations (nt mutations, AA changes, AA class identity (+) or change (−)) for V-REGION, FR1-IMGT, CDR1-IMGT, FR2-IMGT, CDR2-IMGT, FR3-IMGT, and germline CDR3-IMGT</td>
</tr>
<tr>
<td>6</td>
<td>'Junction'</td>
<td>33, 46, 66, or 77</td>
<td>• Number (nb) of nt positions including IMGT gaps, nb of nt, nb of identical nt, total nb of mutations, nb of silent mutations, nb of nonsilent mutations, nb of transitions (a→g, g→a, c→t, t→c) and nb of transversions (a→c, c→a, a→t, t→a, g→c, c→g, g→t, t→g) for V-REGION, FR1-IMGT, CDR1-IMGT, FR2-IMGT, CDR2-IMGT, FR3-IMGT, and germline CDR3-IMGT</td>
</tr>
<tr>
<td>7</td>
<td>'V-REGION-mutation-and-AA-change table'</td>
<td>11</td>
<td>• Nb of AA positions including IMGT gaps, nb of AA, nb of identical AA, total nb of AA changes, nb of AA changes according to AAclassChangeType (+ +, + + −, + − −, + − +, + − − −, − + +, − + − −, − − − +, − − − −), and nb of AA class changes according to AAclassSimilarityDegree (nb of Very similar, nb of Similar, nb of Dissimilar, nb of Very dissimilar) for V-REGION, FR1-IMGT, CDR1-IMGT, FR2-IMGT, CDR2-IMGT, FR3-IMGT, and germline CDR3-IMGT</td>
</tr>
<tr>
<td>8</td>
<td>'V-REGION-nt-mutation-statistics'</td>
<td>130</td>
<td>• Hot spots motifs (la/lta, lta/lta, la/ltaglc/lta/ltaglc, lta/ltaglc/c/ltaglc) detected in the closest germline V-REGION with positions in FR-IMGT and CDR-IMGT</td>
</tr>
<tr>
<td>9</td>
<td>'V-REGION-AA-change-statistics'</td>
<td>189</td>
<td>• Date of the analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• IMGT/V-QUEST program version, IMGT/V-QUEST reference directory release</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Parameters used for the analysis: species, receptor type or locus, IMGT reference directory set, and Advanced parameters.</td>
</tr>
</tbody>
</table>

---

**TABLE 1** List of the IMGT/HighV-QUEST\(^d\) results files with number of columns and results content

\(^d\)IMGT/HighV-QUEST\(^d\), freely available for academics on the IMGT Home page ([http://www.imgt.org](http://www.imgt.org)), analyzes up to 500,000 sequences per run. Results in the 11 files are equivalent to those of the Excel file of IMGT/V-QUEST online\(^\text{46}\). The statistical analysis (tables and histograms) is performed on the ‘Summary’ file content and on the CDR3-IMGT results (nt and AA) of results of several runs up to 500,000 sequences and provided as PDF files and PNG figures (detailed in reference\(^\text{43}\)).

\(^d\)Files 1 to 10 comprise systematically sequence identification (name, functionality, and names of the closest V, D, and J genes and alleles).
V-KAPPA or V-LAMBDA (encoded by a V-J-REGION resulting from a V-J rearrangement) (Fig. 1A).

The IMGT Colliers de Perles of the V and C domains can be displayed on two layers (Fig. 1, middle row) to get a graphical representation closer to the 3D structure, and on two layers with hydrogen bonds (Fig. 1, bottom row) if the mAb has been crystallized and data are in IMGT/3Dstructure-DB (26, 48).

The IMGT/Collier-de-Perles Tool
The IMGT/Collier-de-Perles tool (38) (IMGT Home page, http://www.imgt.org) allows the users to draw IMGT Colliers de Perles, on one or two layers, starting from their own domain amino acid sequences. Sequences have to be gapped according to the IMGT unique numbering (using, for example, IMGT/DomainGapAlign [47]). The IMGT/Collier-de-Perles tool can be customized to display the CDR-IMGT according to the IMGT color menu or the amino acids according to their hydropathy, volume, or the 11 IMGT physicochemical classes (49) (IMGT Aide-Mémoire, in the ‘Amino acids’ section, http://www.imgt.org).

Importance for Antibody Engineering and Antibody Humanization
The IMGT Colliers de Perles are used in antibody engineering and antibody humanization, and for the evaluation of the immunogenicity of therapeutic monoclonal antibodies (40, 41, 42). The information is particularly useful:

1. to precisely define the CDR1-IMGT, CDR2-IMGT, and CDR3-IMGT to be grafted in antibody humanization design based on CDR grafting. Amino acids of the CDR-IMGT loops are those involved in the interactions with the antigen (40, 41, 42),
2. to compare the physicochemical properties of AA at given positions in the user V-region sequences with those of the IMGT Collier de Perles statistical profiles for the human expressed IGHV, IGKV, and IGLV repertoires (49), and to identify AA that, given their positions, could be potentially immunogenic in chimeric or poorly humanized antibodies. The IMGT statistical profiles comprise 61 positions with conserved properties, among which 13 participate to the inner core, 10 to the inner surface of the ABED and GFCC'C' beta sheet (inside of the ‘sandwich’), 21 to the outer surface of the ABED beta sheet and of the F strand, and 10 to turns (exposed to solvent), and seven to the VH-VL interface (detailed in Table 3 of reference 49),
3. to localize unusual amino acid changes, for example, the four AA changes observed between camel IGHV1 genes expressed in conventional IgG1 and those expressed in nonconventional IgG2 and IgG3 (dimer of heavy gamma2 or gamma3 chains with deleted CH1, without light chain): V42>(F,Y), G49>(E,Q), L50>(C,R) and W52>(F,G,L,W) (IMGT Gene table in IMGT Repertoire, and IMGT Biotechnology page, http://www.imgt.org/IMGTbiotechnology/Camel_IgG.html),
4. to predict important AA interactions, even in the absence of 3D structures, by analyzing the display of conserved hydrogen bonds in crystallized V domains and C domains (Fig. 1) (IMGT Collier de Perles on two layers in IMGT/3Dstructure-DB (26, 48)). Of importance for antibody humanization are the two FR-IMGT positions, 39 and 40, that have hydrogen bonds with a CDR2-IMGT and a CDR3-IMGT position, respectively: anchor 39 with CDR2-IMGT 56 (in VH) or 57 (in VL), and position 40 with CDR3-IMGT 105 (in both VH and VL). These positions eventually need to be kept from the original antibody (26, 48), and
5. to localize the landmarks of the CH domain. These include the conserved N-glycosylation site in the CH2 of the IGHG1, IGHG2, IGHG3, and IGHG4, on the asparagine (Asn) N84.4 at the top of the DE turn. There is also a polymorphic N-glycosylation site N79 in the CH3 strand D of several alleles of the IGHG3 (replaced by a lysine [Lys] K in the other alleles) (Alignment of alleles in IMGT Repertoire, http://www.imgt.org). The positions used for the ‘knobs-into-holes’ engineering for heavy-chain dimORIZATION and the production of bispecfic IgG antibodies involve CH3 positions threonine (Thr) T22 (strand B) and tyrosine (Tyr) Y86 (strand E) with the following AA changes, T22Y on one CH3 domain and Y86>T on the other one (IMGT Biotechnology page, http://www.imgt.org).

ANALYSIS OF ANTIBODY V- AND C-DOMAIN AMINO ACID SEQUENCES
IMGT/DomainGapAlign
IMGT/DomainGapAlign (26, 47) analyzes the amino acid sequences of the IG and TR variable (V) and constant (C) domains. Several amino acid sequences can be analyzed simultaneously, provided that they belong to the same domain type (V or C). IMGT/DomainGapAlign analyzes the user amino acid domain sequences by comparison with the IMGT reference directory sets (translation of the germline V and J genes and of the C gene domains from IMGT/GENE-DB [18]). These reference amino acid sequences can be displayed by querying IMGT/DomainDisplay (IMGT Home page, http://www.imgt.org).
Antibody V domain amino acid sequence analysis

IMGT/DomainGapAlign identifies the closest germline V-REGION and J-REGION alleles. IMGT/DomainGapAlign displays the V region amino acid sequences of the user aligned with the closest V and J regions (Fig. 2A) with IMGT gaps and delimitations of the strands (FR-IMGT) and loops (CDR-IMGT), according to the IMGT unique numbering for V domain (38, 39, 40, 41). For instance, for motavizumab, the V-REGION of the VH domain is identified as having 86.9% identity with the Homo sapiens IGHV2-70*01 and the V-REGION of the VKAPPA as having 83.0% identity with the H. sapiens IGKV1-5*01. If several closest alleles are identified, the user can select the display of each corresponding alignment. The amino acid sequence is also displayed online, according to the IMGT color menu, with the delimitations of the V-REGION, J-REGION, and for VH domains, (N-D)-REGION.

The number of amino acid differences in the FR-IMGT and CDR-IMGT is one of the criteria to evaluate the potential immunogenicity (39, 40, 41, 42). The framework of a VH domain comprises 91 positions (25, 17, 38, and 11 positions for FR1-, FR2-, FR3-, and FR4-IMGT, respectively, represented as [25.17.38.11]), whereas the framework of a VL domain comprises 89 positions (26, 17, 36, 10 positions for FR1-, FR2-, FR3-, and FR4-IMGT, respectively, represented as [26.17.36.10]) (26, 47). Thus, for motavizumab, the framework of the VH has 10 AA differences (81/91 identical AA) with the framework constituted by the closest human germline IGHV2-70*01 and IGJH6*03, whereas the framework of the VKAPPA has 5 AA differences (84/89 identical AA) with the framework constituted by the closest human germline IGKV1-5*01 and IGKJ*01 (Fig. 2A) (in a humanization with a good IMGT grafting, the number of AA differences in the framework is minimal and usually between 0 and 2).

The IMGT/DomainGapAlign results include tables with the characteristics of the AA changes (49) shown in strands, turns, and loops. The IMGT Collier de Perles of the analyzed C domain is also available with highlighted AA differences (in pink circles online) with the closest allele domain sequence.

Antibody C domain amino acid sequence analysis

IMGT/DomainGapAlign displays the C-domain amino acid sequences of the user sequence aligned with the closest C-DOMAIN alleles (Fig. 2B) with IMGT gaps and delimitations of the strands, turns and loops, according to the IMGT unique numbering for C domain (31, 32). The IMGT unique numbering for C-DOMAIN contributes to antibody engineering and humanization by providing a standardized description of the Gm, Am, and Km allotypes, and by establishing, for the first time, the correlation between the G1m, G2m, G3m, A2m, and Km allotypes, and the IGHGV1, IGHGV2, IGHGV3, IGHCA, and IGKC alleles, respectively (50).

For instance, IMGT/DomainGapAlign shows that the C-REGION of motavizumab is 100% identical with the IGHG1*03 allele in its three CH domains and hinge region. The IGHG1*03 allele corresponds to the G1m3 allele (50), characterized by the G1m3 allotype (CH1 arginine [Arg] R120, associated to CH1 isoleucine [Ileu] I103) (Fig. 2B), and to the isoallotypes nG1m17 (CH1 R120) and nG1m1 (CH3 glutamate [Glu] E12 and methionine [Met] M14) (50). The C-KAPPA of motavizumab is 100% identical to the IGKGC*01 allele (Fig. 2B). IGKC*01 is one of the four IGKC alleles that correspond to a Km3 allele, characterized by an alanine (Ala) A45.1 (‘.1’ for the first position in the transverse CD strand) and a valine (Val) V101 (50).

The IMGT/DomainGapAlign results include tables with the characteristics of the AA changes (49) shown in strands, turns, and loops. The IMGT Collier de Perles of the analyzed C domain is also available with highlighted AA differences (in pink circles online) with the closest allele domain sequence.

ANTIBODY 3D STRUCTURES

IMGT/3Dstructure-DB Card

IMGT/3Dstructure-DB (26, 48) (http://www.imgt.org), the IMGT 3D structure database, is queried through a user-friendly interface and different displays for the results can be chosen from the database home page (48). Each entry in the database is detailed in an IMGT/3Dstructure-DB card that provides access to all sequence data related to that entry, and, in particular, for antibodies, the paired heavy and light chains, and for Ig/antigen (Ag) complexes, the ligands (if peptides or proteins). IMGT uses the ‘PDB code’ (4 letters and/or numbers) as ‘IMGT entry ID.’ An additional letter separated by an underscore (‘_’) identifies the different chains in a 3D structure. For example, the 3ixt entry that corresponds to the 3D structure of an Ig/antigen (Ag) complex (motavizumab Fab in complex with the antigen ‘ligand’ Fusion glycoprotein F1) comprises two crystallographic complexes (as indicated in the column CC with the colors red and blue), with for one of them (CC1, red), the following chains: 3ixt_H (VH-CH1) and...
**FIGURE 2** IMGT/DomainGapAlign alignments. A. VH and V-KAPPA. B. CH1 and C-KAPPA. The closest V-REGION and J-REGION identified at the amino acid level are aligned with the user sequence (here, motavizumab INN 8693, as example). The VH and V-KAPPA are identified as having 86.9% and 83% identity at the amino acid level with the *Homo sapiens* IGHV2-70*01 and IGKV1-5*01, respectively (% identity is shown in an upper section online). Amino acid differences are indicated below the V and J alignments. The FR-IMGT and CDR-IMGT, strands and loops are according to the IMGT unique numbering (28, 29, 30, 32). The CH1 and C-KAPPA are identified as having 100% identity at the amino acid level with the *H. sapiens* IGHG1*03 CH1 and IGKC*01, respectively. IMGT/DomainGapAlign displays the C-domain amino acid sequence of the user, with IMGT gaps and delimitations of the strands, turns, and loops, according to the IMGT unique numbering (31, 32). doi:10.1128/microbiolspec.AID-0001-2012.f2
3ixt_L (L-KAPPA) for the IG Fab, and 3ixt_P for the antigen (‘ligand’). Eight tabs are available at the top of each card: ‘Chain details,’ ‘Contact analysis,’ ‘Paratope and epitope,’ ‘3D visualization Jmol or QuickPDB,’ ‘Renumbrmed IMGT file,’ ‘IMGT numbering comparison,’ ‘References and links,’ and ‘Printable card.’

**IMGT/3Dstructure-DB Chain details**

The section ‘Chain details’ of the IMGT/3Dstructure-DB card comprises, for each chain, information first on the chain itself, then per domain.

1. The information for each chain includes:
   - ‘Chain ID’ (e.g., 3ixt_H),
   - ‘Chain length’ in amino acids (e.g., 225),
   - ‘IMGT chain description’ with the delimitations of the different domains (e.g., VH-CH1 = VH (1-120) [D1] + CH1 (121-210) [D2]),
   - ‘Chain sequence’ with delimitations of the regions and domains, highlighting of AA (in orange color) that are different from the closest genes and alleles, and links to **Sequence in FASTA format** and to **Sequence in IMGT format**.

2. The information for a V-DOMAIN, as an example, includes:
   - ‘IMGT domain description’ (e.g., VH (1-120) [D1]),
   - ‘IMGT gene and allele name’ with the percentage of identity for the V and a link to **Alignment details**
   - ‘IMGT gene and allele name’ with the percentage of identity for the J as well as other alleles giving the same percentage of identity), and a link to **Alignment details**.
   - ‘2D representation’: links to **IMGT Collier de Perles on one layer** or **IMGT Collier de Perles on two layers**,
   - ‘Contact analysis’: a link to **Domain contacts (overview)**,
   - ‘CDR-IMGT lengths’ (e.g., [10.7.12]),
   - ‘Sheet composition’ (e.g., [ABDE][A"CC"C"FG]),
   - the domain amino acid sequence with CDR-IMGT delimitations and highlighting of AA (in orange color) that are different from the closest V and J genes and alleles,
   - the link to **IMGT/DomainGapAlign results**.

**IMGT/3Dstructure-DB Contact Analysis**

The IMGT/3Dstructure-DB Contact analysis (26, 48) provides extensive information on the atom pair contacts between domains and/or chains and on the internal contacts in an IMGT/3Dstructure-DB entry. Atom pair contacts are obtained in IMGT/3Dstructure-DB by a local program in which atoms are considered to be in contact when no water molecule can take place between them and are characterized by their atom contact types (Noncovalent, Polar, Hydrogen bond, etc.) and their atom contact categories ([BB] Backbone/backbone, [SS] Side chain/side chain, etc.) (26, 48).

This information can be obtained at three different levels:

1. Domain contacts (overview),
2. Domain pair contacts (‘DomPair’) that provides information on the contacts between a pair of domains or between a domain and a ligand (e.g., between the VH domain of motavizumab (3ixt_H chain) and the ligand (3ixt_P chain) (Fig. 3A), or between the V-KAPPA domain of motavizumab (3ixt_L chain) and the ligand (3ixt_P chain) (Fig. 3B),
3. Residue pair contacts, or more precisely, IMGT Residue@Position (R@P) pair contacts.

R@P represents one of the major breakthroughs of the IMGT unique numbering for 3D structures. Indeed, in IMGT, any residue (amino acid) that belongs to an antibody domain (and more generally, any amino acid that belongs to a V or C domain of an IgSF protein, or to a G domain of a MhSF protein [32]) is characterized by its position in the domain. An R@P is defined by the IMGT position numbering, the residue name, the IMGT domain description, and the IMGT chain ID (e.g., 35 – ALA (A) – VH – 3ixt_H). Structural information and contacts for a given ‘R@P’ are provided in the IMGT Residue@Position card that includes general information (PDB file numbering, IMGT file numbering, residue full name and formula), and structural information ‘IMGT LocalStructure@Position’ (secondary structure, Phi and Psi angles [in degrees] and accessible surface area [ASA] [in square angstroms]). The IMGT Residue@Position cards can be accessed directly from the amino acid sequences of the IMGT/3Dstructure-DB card or from the IMGT Colliers de Perles, by clicking on one AA.

**IMGT Paratope and Epitope**

In an IG/Ag complex, the amino acids in contact at the interface between the IG and the Ag constitute the paratope on the IG surface, and the epitope on the Ag surface (Fig. 3C). In IMGT/3Dstructure-DB, the ‘IMGT paratope and epitope’ of IG/Ag complexes is determined automatically by combining contact analysis with an interaction scoring function and are described in a...
FIGURE 3 IMGT/3Dstructure-DB Domain pair contacts and IMGT paratope and epitope details. A, IMGT/3Dstructure-DB Domain pair contacts between the VH of motavizumab (3ixt_H) and the Fusion glycoprotein F1 (ligand) (3ixt_P). B, IMGT/3Dstructure-DB Domain pair contacts between the V-KAPPA of motavizumab (3ixt_L) and the Fusion glycoprotein F1 (ligand) (3ixt_P).

Polar, ‘Hydrogen bonds,’ and ‘Nonpolar’ were selected before display, in ‘Atom contact types.’ Amino acids belonging to the CDR1-IMGT, CDR2-IMGT, and CDR3-IMGT are colored online according to the IMGT color menu (red, orange, and purple, respectively, for VH; blue, green, and green-blue, respectively, for V-KAPPA). In this 3D structure, all but one amino acid that contact the antigen belong to the CDR-IMGT. Clicking on R@P gives access to the IMGT Residue@Position cards (26, 48).

C, IMGT paratope and epitope details of the IG/Ag complex 3ixt is shown. Each AA that belongs to the IG paratope is characterized by its position in the V domains according to the IMGT unique numbering (29, 30, 32). Thus, ‘A (35V1_A)’ means that the alanine (A) is at position 35 of the V domain 1 of 3ixt_A (VH). In the same way, ‘G(107V1_B)’ means that the glycine (G) is at position 107 of the V domain 1 of 3ixt_B (V-KAPPA). Each AA that belongs to the antigenic determinant (epitope) is characterized by its position (here, position in the chain, in the 3D structure). For example, ‘S (3_C)’ means that the serine (S) is at position 3 of the Fusion glycoprotein F1 ligand (3ixt_C), whereas ‘SN (23-24_C)’ means that the serine (S), asparagine (N) are at positions 23, 24. The ‘IMGT paratope and epitope’ analysis of the IG/Ag 3D structure (3ixt) is from IMGT/3Dstructure-DB (http://www.imgt.org).

standardized way. Thus, the IG paratope of 3ixt (motavizumab Fab) (Fig. 3C) comprises amino acids of VH (3ixt_H chain) and of V-KAPPA (3ixt_L chain). Fifteen amino acids of the antibody, 8 from VH and 7 from V-KAPPA, form the paratope (Fig. 3C). Eight of the 8 positions belong to the VH CDR-IMGT [10.7.12] (A35 to the CDR1-IMGT; W58, D59, and K64 to the CDR2-IMGT; and I109, F110, N112, and F113 to the CDR3-IMGT). Seven of the 7 positions belong to the V-KAPPA CDR-IMGT [5.3.9] (G37 and Y38 to the CDR1-IMGT; D56 to the CDR2-IMGT; and G107, S108, G109, and Y114 to the CDR3-IMGT). These results emphasize the importance of using the IMGT unique numbering for standardized antibody analysis and confirm that the CDR-IMGT structural delimitations correspond to the V-domain interactions with the antigen.

The Ag epitope of 3ixt (Fig. 3C) comprises AA of the Fusion glycoprotein F1 (3ixt_P). Eleven amino acids form the Ag epitope. Each amino acid that belongs to the epitope is defined by its position in the chain in the 3D structure (if the AA is part of a V, C, or G domain, the position is given according to the IMGT unique numbering (32).

Clicking on a residue in ‘Paratope details’ or in ‘Epitope details’ (Fig. 3C) provides the R@P card for each AA that belongs to the paratope or epitope, respectively.

IMGT/StructuralQuery and IMGT/DomainSuperimpose
IMGT/DomainSuperimpose (3) (IMGT® Home page, http://www.imgt.org) allows one to superimpose the 3D structures of two domains from IMGT/3Dstructure-DB, and demonstrate that the same IMGT numbering is found at the same positions of the framework of the heavy and light chains and for the lower part of the CDR-IMGT loops of the two domains. IMGT/StructuralQuery (3) allows the retrieval the IMGT/3Dstructure-DB entries containing a V, C, or G domain, based on specific structural characteristics of the intramolecular interactions: phi and psi angles, accessible surface area, type of atom contacts, distance in angstroms between amino acids, R@P pair contacts, and, for the V domain, CDR-IMGT length or pattern.

IMGT/2Dstructure-DB AND IMGT/mAb-DB
In a further effort to bridge the gap between sequences and 3D structures, a new extension of IMGT/3Dstructure-DB, designated as IMGT/2Dstructure-DB, was created to describe and analyze amino acid sequences of paired chains of heavy and light chains of antibodies for which no 3D structures are available. These amino acid sequences are analyzed and managed with the IMGT criteria of standardized gene and allele nomenclature (classification), standardized labels (description), and IMGT unique numbering and IMGT Colliers de Perles (numerotation) (10). IMGT/2Dstructure-DB uses the IMGT/3Dstructure-DB informatics frame and interface that allows one to analyze, manage, and query antibodies as polymeric receptors made of several chains, in contrast to the IMGT/LIGM-DB sequence database that analyzes and manages IG sequences, individually. The current IMGT/2Dstructure-DB entries include amino acid sequences of the paired heavy and light chains of antibodies from WHO-INN (24, 25) and those from Kabat (annotated by IMGT and for which there are no available nucleotide sequences, the others already being in IMGT/LIGM-DB). IMGT/2Dstructure-DB entries also include the INN amino acid sequences of FPIA (suffix -cept) and those of CPCA with a Fc (for increased half-life) (prefix ef-).

Queries can be made on an individual entry, using the Entry ID (e.g., for an INN entry, the 4-number code, e.g., 8693) or the Molecule name (e.g., motavizumab). Search can be made on Entry type (e.g., ‘INN’), and on criteria detailed in the IMGT/2Dstructure-DB Query page, http://www.imgt.org. In the results page, clicking on an IMGT entry ID gives access to the IMGT/2Dstructure-DB card (42). The IMGT/2Dstructure-DB card provides standardized IMGT information on chains and domains and IMGT Colliers de Perles on one or two layers, identical to that of an IMGT/3Dstructure-DB card; however, the information on experimental structural data (hydrogen bonds in IMGT Collier de Perles on two layers, Contact analysis) is only available in the corresponding IMGT/3Dstructure-DB card, if the antibodies have been crystallized (26, 48). For therapeutic antibodies, queries can be made from the IMGT/mAb-DB user-friendly interface (http://www.imgt.org) on search criteria, such as characteristics (conjugated, radiolabeled...), IG classes or subclasses for complete IG or format (Fab, scFv, etc.) for IG fragments, specificity (target name and species), clinical indication, development status, etc.

TOWARD TARGETED AND CUSTOMIZED THERAPEUTIC ANTIBODIES
IMGT provides antibody informatics with a standardized system for sequences and structures (6, 7), from amino acid characteristics at given positions (49), to IG
gene and allele identification, and to 3D structures analysis (antibody/antigen contact analysis, paratope/epitope description). Since its creation in 1989, and owing to a strong expertise and background in immunogenetics, IMGT has defined IG alleles and developed databases and tools and web resources for IG polymorphism, as demonstrated by IMGT/GENE-DB (18), IMGT Alignments of alleles and the correspondences between IGHG and IGKC genes and the C-domain Gm and Km allotypes (50). The extension of the IMGT unique numbering to the IgSF and to the MhSF proteins other than IG or TR has opened new perspectives for the standardized description of the polymorphism of the antigens (epitopes belonging to V, C, or G domains) and of the Fc receptors (FCGR of the IgSF, FCGRT of the MhSF) and for the characterization of their interactions (antibody/antigen, FcR/antibody). Given the importance of these interactions in the antibody specificity and affinity on the one hand and in the antibody pharmacokinetics/pharmacodynamics and half-life on the other hand, the IMGT integrated and standardized approach provides the genetic knowledge for allowing antibody informatics to answer the needs of targeted and customized therapy in the context of personalized medicine.

AVAILABILITY AND CITATION

Authors who use IMGT databases and tools are encouraged to cite this article and to quote the IMGT Home page (http://www.imgt.org). Online access to IMGT databases and tools is freely available for academics and under licenses and contracts for companies.

ACKNOWLEDGMENTS

I thank Gérard Lefranc for helpful discussions, Sophia Kossida and Souphatta Sasorith for reading the manuscript, Denis Moreno for the figures, and the IMGT team for its expertise and constant motivation. IMGT is a registered trademark of CNRS. IMGT is a member of the International Medical Informatics Association (IMIA). IMGT was funded in part by the BIOMED1 (BIOT930038), Biotechnology BIOTECH2 (BIO4CT960037), 5th PCRDT Quality of Life and Management of Living Resources (QLG2-2000-01287), and 6th PCRDT Information Science and Technology (ImmunoGrid, FP6 IST-028069) programs of the European Union (EU), and by the Agence Nationale de la Recherche ANR (BIOYS06_135457, FLAVORES). IMGT is currently supported by the Centre National de la Recherche Scientifique (CNRS), the Ministère de l’Enseignement Supérieur et de la Recherche (MESR), the Université Montpellier 2, the GIS IBiSA, the Région Languedoc-Roussillon (Grand Plateau Technique pour la Recherche [GPTR], GEPETOS), and the Labex MAbImprove (ANR-10-LABX-53). This work was granted access to the HPC resources of CINES under the allocations 2010-2014-036029 made by GENCI (Grand Equipement National de Calcul Intensif).

Conflicts of interest: The author declares no conflicts.

REFERENCES
