IMGT® Immunoinformatics Tools for Standardized V-DOMAIN Analysis

Véronique Giudicelli, Patrice Duroux, Maël Rollin, Safa Aouinti, Géraldine Folch, Joumana Jabado-Michaloud, Marie-Paule Lefranc, Sofia Kossida

Chapter 24

IMGT® Immunoinformatics Tools for Standardized V-DOMAIN Analysis

Véronique Giudicelli, Patrice Duroux, Maël Rollin, Safa Aouinti, Géraldine Folch, Joumana Jabado-Michaloud, Marie-Paule Lefranc, and Sofia Kossida

Abstract

The variable domains (V-DOMAIN) of the antigen receptors, immunoglobulins (IG) or antibodies and T cell receptors (TR), which specifically recognize the antigens show a huge diversity in their sequences. This diversity results from the complex mechanisms involved in the synthesis of these domains at the DNA level (rearrangements of the variable (V), diversity (D), and joining (J) genes; N-diversity; and, for the IG, somatic hypermutations). The recognition of V, D, and J as "genes" and their entry in databases mark the creation of IMGT by Marie-Paule Lefranc, and the origin of immunoinformatics in 1989. For 30 years, IMGT®, the international ImMunoGeneTics information system® http://www.imgt.org, has implemented databases and developed tools for IG and TR immunoinformatics, based on the IMGT Scientific chart rules and IMGT-ONTOLOGY concepts and axioms, and more particularly, the princeps ones: IMGT genes and alleles (CLASSIFICATION axiom) and the IMGT unique numbering and IMGT Collier de Perles (NUMEROTATION axiom). This chapter describes the online tools for the characterization and annotation of the expressed V-DOMAIN sequences: (a) IMGT/V-QUEST analyzes in detail IG and TR rearranged nucleotide sequences, (b) IMGT/HighV-QUEST is its high throughput version, which includes a module for the identification of IMGT clonotypes and generates immunoprofiles of expressed V, D, and J genes and alleles, (c) IMGT/StatClonotype performs the pairwise comparison of IMGT/HighV-QUEST immunoprofiles, (d) IMGT/DomainGapAlign analyzes amino acid sequences and is frequently used in antibody engineering and humanization, and (e) IMGT/Collier-de-Perles provides two-dimensional (2D) graphical representations of V-DOMAIN, bridging the gap between sequences and 3D structures. These IMGT® tools are widely used in repertoire analyses of the adaptive immune responses in normal and pathological situations and in the design of engineered IG and TR for therapeutic applications.

Key words IMGT, Immunogenetics, Immunoinformatics, Immunoglobulin, Antibody, T cell receptor, V-DOMAIN sequence analysis, Adaptive immune repertoire, IMGT Collier de Perles, IMGT-ONTOLOGY

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_24,

1 Introduction

Immunoglobulins (IG) or antibodies [1, 2] and T cell receptors (TR) [3] are antigen receptors of the adaptive immune responses in vertebrates with jaws (gnathostomata) [4]. The huge diversity of the variable domains (V-DOMAIN) of the IG and TR chains of the immune repertoires results from several mechanisms that occur during their synthesis [1–4]. In particular, the combinatorial diversity depends on the number of variable (V), diversity (D), and joining (J) genes found in the IG and TR loci, which potentially can rearrange to form V-DOMAIN encoded by V-(D)-J regions [1–4]. It is the recognition of the V, D, and J as "genes" and their entry in databases that mark the creation of IMGT in 1989 by Marie-Paule Lefranc (Université de Montpellier, CNRS) at Human Gene Mapping 10 (HGM10) and is at the origin of immunoinformatics, a new science at the interface between immunogenetics and bioinformatics [4].

Other mechanisms of diversity comprise the junctional diversity with exonuclease trimming at the ends of the V, D, and J genes and the random addition of nontemplated nucleotides, preferably "g" and "c," by the terminal deoxynucleotidyl transferase (TdT) encoded by the DNA nucleotidylexotransferase (DNTT) gene, creating the N-regions [1–3], and for IG, somatic hypermutations [1, 2]. For 30 years, IMGT®, the international ImMunoGeneTics information system® http://www.imgt.org, has implemented databases and developed tools for IG and TR immunoinformatics [5], based on the IMGT Scientific chart rules (*see* Subheading 2) and IMGT-ONTOLOGY concepts and axioms [6, 7], and more particularly, the princeps ones: IMGT genes and alleles (CLASSIFICATION axiom) [8–12] and the IMGT unique numbering [13–17] and IMGT Collier de Perles [18–21] (NUMEROTATION axiom). This chapter describes the online analysis tools for the characterization and annotation of the expressed V-DOMAIN nucleotide (nt) and amino acid (AA) sequences, available from "IMGT tools" section of the IMGT® Home page. Protocols for their use and the description of main results are presented in this chapter. These concern the following: (a) IMGT/V-QUEST [22, 23] is the IMGT® online tool for the analysis of IG and TR nucleotide rearranged sequences (*see* Subheading 3); (b) IMGT/HighV-QUEST [24–27], the high throughput version of IMGT/V-QUEST, can analyze sets of up to one million sequences. It includes a module for the identification of IMGT clonotypes (AA) and the generation of IG and TR gene profiles for the diversity and expression of IMGT clonotypes (AA) (*see* Subheading 4); (c) IMGT/StatClonotype [28, 29] is a standalone package that performs statistical pairwise comparisons of IMGT clonotype (AA) diversity or expression between two IMGT/HighV-QUEST result sets (*see* Subheading

5); (d) IMGT/DomainGapAlign [30, 31] analyses domain AA sequences and two dimensional (2D) structures, and its results are used in antibody engineering and humanization [32, 33] (*see* Subheading 6); and (e) IMGT/Collier-de-Perles tool [21] generates IMGT Colliers de Perles graphical 2D representations for AA domain sequences [18–20] (*see* Subheading 7), it is available from the IMGT Home page and is also automatically launched by IMGT/V-QUEST and IMGT/DomainGapAlign.

2 IMGT Scientific Chart Rules for the Analysis of the V-DOMAIN

2.1 IMGT Gene and Allele Nomenclature and IMGT Reference Directory Sets

The IMGT gene names of the IG and TR V, D, J, and C genes [1–4] were approved by the Human Genome Organization (HUGO) Nomenclature Committee (HGNC) in 1999 [8, 9, 12] and were endorsed by the WHO-IUIS Nomenclature Subcommittee for IG and TR [10, 11]. IMGT gene and allele names are based on the concepts of classification of IMGT-ONTOLOGY "Group," "Subgroup," "Gene," and "Allele" [1–4, 10–12]. Alleles are the polymorphic variants of a gene: they are identified by their IMGT reference sequence, which corresponds to the coding V-REGION, D-REGION, J-REGION, and C-REGION sequence at the nucleotide level of V, D, J, and C gene alleles, respectively. IMGT reference directory sets include the allele IMGT reference sequences from functional (F) genes and alleles, open reading frame (ORF), and pseudogenes (P) [5]. IMGT germline V, D, and J genes and alleles, with their characteristics, their reference sequence and other sequences from the literature are managed in IMGT/GENE-DB [34] and in IMGT Repertoire (IG and TR) Gene tables and Alignments of alleles Web resources [1–4]. The tools for V-DOMAIN analysis compare user sequences with IMGT reference directory sets for the identification of V, D, and J genes and alleles and the evaluation of mutations and AA changes.

2.2 IMGT Unique Numbering for the IG and TR V Domains

An IG or TR V-DOMAIN comprises about 100 amino acids and is made of nine antiparallel beta strands (A, B, C, C′, C″, D, E, F, and G) linked by beta turns (AB, CC′, C″D, DE, and EF) or loops (BC, C′C″, and FG) [35]. At the structural level, they form a sandwich of two sheets closely packed against each other through hydrophobic interactions and joined together by a disulfide bridge between 1st-CYS at position 23 in B-STRAND (in the first sheet) and 2nd-CYS at position 104 in F-STRAND (in the second sheet) [13].

The IMGT unique numbering for IG and TR V-DOMAIN [13] delimits (1) the four framework regions: FR1-IMGT (A and B strands, from positions 1 to 26), FR2-IMGT (C and C′ strands, from positions 39 to 55), FR3-IMGT (C″, D, E and F strands, from positions 66 to 104), FR4-IMGT (G strand, from positions 118 to 128), and (2) the three hypervariable or complementarity

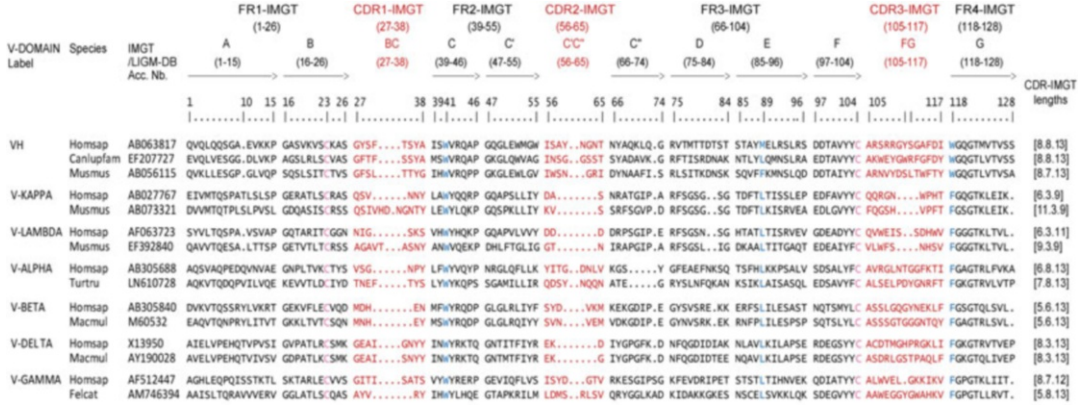


Fig. 1 Protein displays of IG and TR V-DOMAIN based on the IMGT unique numbering for V-DOMAIN [13]. The V-DOMAIN translations were obtained from the analysis by IMGT/V-QUEST [22, 23] (*see* Subheading 3) of the nucleotide sequences of shown accession numbers in IMGT/LIGM-DB [36]. The identification of FR-IMGT and CDR-IMGT and of beta strands and loops was performed by IMGT/DomainGapAlign [30, 31] (*see* Subheading 6), which provides a standardized delimitation whatever the species, the receptor type, and the chain type. CDR-IMGT lengths are indicated between brackets, separated by dots (column on the right). 1st-CYS 23 and 2nd-CYS 104 are in pink, and W 41, hydrophobic AA 89, and W or F 118 are in blue. Taxons are in the IMGT 6- or 9-letter abbreviation: Homsap for *Homo sapiens*, Canlupfam for *Canis lupus familiaris* (dog), Musmus for *Mus musculus* (mouse), Turtru for *Tursiops truncatus* (dolphin), Macmul for *Macaca mulatta* (Rhesus monkey), and Felcat for *Felis catus* (cat)

determining regions involved in the ligand recognition: CDR1-IMGT (BC loop, positions 27 to 38), CDR2-IMGT (C′C″ loop, positions 56 to 65), and CDR3-IMGT (FG loop, positions 105 to 117, with additional positions 112.1, 111.1, 112.2 etc., if longer than 13 codons (or AA)). FR-IMGT positions, which delimit the three CDR-IMGT, are designated as anchors: they are 26 and 39, 55 and 66, and 104 and 118, respectively (Fig. 1), and shown as squares in IMGT Colliers de Perles [18–21]. According to the IMGT unique numbering [13], a V-DOMAIN is characterized by five highly conserved AA: 1st-CYS 23, tryptophan 41 (CONSERVED-TRP), hydrophobic amino acid 89, 2nd-CYS 104, and J-PHE or J-TRP 118 of the J-MOTIF (F/W-G-X-G, 118–121, where F is phenylalanine, W tryptophan, G glycine, X, any AA). The three CDR-IMGT lengths characterize a V-DOMAIN. By convention, they are indicated between brackets, separated by dots (for example [8.8.13]). The CDR1-IMGT and CDR2-IMGT are encoded by the V-REGION, whereas the CDR3-IMGT results from the V-(D)-J rearrangement. The IMGT Collier de Perles [18–20] can be generated by the IMGT/Collier-de-Perles tool [21] (*see* Subheading 7).

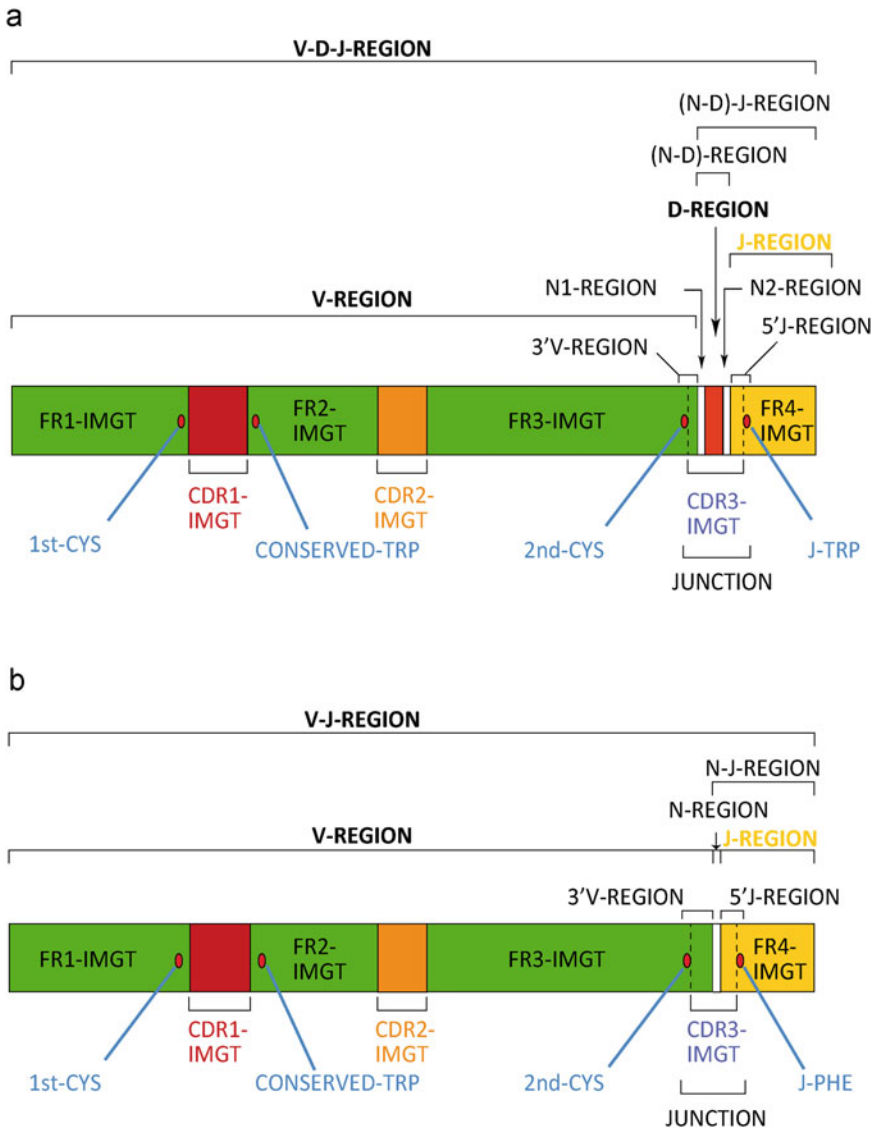


Fig. 2 Graphical representation or prototypes of IG and TR V-DOMAIN with IMGT labels at the nucleotide level. (**a**) V-D-J-REGION. (**b**) V-J-REGION [1–4]. The JUNCTION encompasses 2nd-CYS 104, CDR3-IMGT, and J-TRP or J-PHE 118, and its length is therefore two AA longer than CDR3-IMGT. Potential palindromic nucleotides ("P") identified in case of untrimmed V, D, and/or J regions during the DNA rearrangement are not shown (With permission from M-P. Lefranc and G. Lefranc, LIGM, Founders and Authors of IMGT®, the international ImMunoGeneTics information system®, http:/www.imgt.org)

2.3 IMGT Standardized Labels and Sequence Description

The IMGT tools, which perform the analysis of sequences, provide the description of the V-DOMAIN with IMGT standardized labels (written in capital letters). The V-DOMAIN corresponds either to a V-D-J-REGION (in IG heavy (IGH)), TR beta (TRB), and TR delta (TRD) chains) (Fig. 2a) or to a V-J-REGION (in IG light

lambda (IGL) and IG kappa (IGK)), TR alpha (TRA) and TR gamma (TRG) chains) (Fig. 2b), encoded by V-D-J or V-J rearrangements, respectively.

The V-DOMAIN labels according to the chain type or locus are: VH, V-KAPPA, V-LAMBDA for the IGH, IGK, and IGL, respectively, and V-ALPHA, V-BETA, V-DELTA, V-GAMMA for the TRA, TRB, TRD, and TRG, respectively [1–4].

2.4 IMGT Functionality of IG and TR Genes and Alleles and of Rearranged Sequences

The "Functionality" concept identifies the functionality based on the configuration of the IG and TR genes. The functionality of the germline (V, D and J) and undefined (C) IG and TR genes and alleles, defined on the same criteria as conventional genes and alleles, is either functional (F), open reading frame (ORF), or pseudogene (P). The functionality of the IG and TR V-(D)-J rearranged sequences is either "productive" (no stop codon and in-frame JUNCTION (2nd-CYS 104 and J-TRP/J-PHE 118 in the same reading frame)) or "unproductive" (stop codons and/or out-of-frame JUNCTION) [2].

3 IMGT/V-QUEST

IMGT/V-QUEST [22, 23] identifies the V, D, and J genes and alleles in IG and TR V domains. It characterizes the nucleotide (nt) mutations and amino acid (AA) changes resulting from somatic hypermutations in IG V-REGION. It provides a detailed characterization of the V-D-J or V-J junctions by the integrated IMGT/JunctionAnalysis tool [37, 38] and the full annotation of the V-DOMAIN with IMGT labels by IMGT/Automat [39, 40].

3.1 IMGT/V-QUEST Sequence Submission

The top of the IMGT/V-QUEST Welcome page (Fig. 3) provides two links: the first one gives access to the list of the IMGT/V-QUEST reference directory sets to which the users' own sequences can be compared (*see* **Note 1**), and the second one provides examples of human rearranged sequences to test the tool.

The page then includes five sections to configure the analysis:

3.1.1 Your Selection

1. Select the species or taxon first and then the receptor type or locus (*see* **Note 2**) in the lists.

3.1.2 Sequence Submission

IMGT/V-QUEST analyses up to 50 FASTA formatted IG or TR rearranged nucleotide sequences per run, from genomic DNA or cDNA indifferently.

1. Enter the sequences in the text area "Type (or copy/paste) your nucleotide sequence(s) in FASTA format".

Fig. 3 IMGT/V-QUEST Welcome page with the five sections: "Your selection," "Sequence submission," "Display results," "Advanced parameters," and "Advanced functionalities" [22, 23]

2. Alternatively, upload the sequences as a text file by selecting the option "Or give the path access to a local file containing your sequence(s) in FASTA format" (*see* **Note 3**).

3.1.3 Display Results

Three choices of display for the results are available [22, 23]. "A. Detailed view" and "B. Synthesis view" are displayed online in HTML (by default) or text format. Both include sequence alignments for which the user can define the number of nucleotides per line (60 by default). The third type of display, "C. Excel file," is dedicated for the download of the results.

1. Select "A. Detailed view" to get the results for each sequence individually. Results consist in a "Result summary" with the main results of the analysis and 14 detailed result sections that can be selectively checked or unchecked by the user (*see* **Note 4**). In sequence alignments, the number of IMGT reference sequences aligned with the user sequence (five by default) can be modified from 1 to 20.
2. Select "B. Synthesis view" to display the sequences that express the same V gene and allele aligned together. Results include a "Summary table" with the main results of the analysis which can be ordered by "V-GENE and allele name" (default) or by the sequence "input" order. There are eight detailed result sections that can be checked or unchecked (*see* **Note 5**).
3. Select "C. Excel file" to download the results, either in a spreadsheet (default) or as a zip archive. The results may include 11 sheets (or text files in the zip archive (*see* **Note 6**)), which can be checked or unchecked. The 12th sheet (or text file) is available if the option "Analysis of single chain Fragment variable (scFv)" is selected in "Advanced functionalities" (*see* Subheading 3.1.5) [41]. An alternative is to display the content of one given sheet in your browser ("Display 1 CSV file in your browser") or to "Download AIRR formatted results" as a zip archive (*see* **Note 7**) [42, 43].

3.1.4 Advanced Parameters

The default values of the advanced parameters are used by IMGT/V-QUEST for classical analyses [22, 23]. They may be modified for specific studies and/or unusual sequences. The user may:

1. Select the relevant set to be compared with the submitted sequences in "Selection of IMGT reference directory set": 'F+ORF', 'F+ORF+in frame P' (by default), 'F+ORF including orphons', or 'F+ORF+in frame P including orphons' (*see* **Note 8**), "With all alleles" of genes or "With allele *01 only" in order to restrict the IMGT reference directory to one representative sequence per gene only.
2. Choose to "Search for insertions and deletions in V-REGION" or not. By default, IMGT/V-QUEST does not search for

insertions and/or deletions. Selecting "Yes" allows to identify the somatic hypermutations by nucleotide insertions and deletions in the V-REGION that may occur in normal and malignant cells [44] and/or potential sequencing errors.

3. Set the values for "Parameters for IMGT/JunctionAnalysis" that include:
   (a) "Nb of accepted D-GENE" (number of D genes searched by the tool in IGH, TRB or TRD junctions).
   (b) "Nb of accepted mutations" in 3′V-REGION, D-REGION, and 5′J-REGION: by default, 2, 4 and 2 mutations are accepted in the 3′V-REGION, D-REGION, and 5′J-REGION, respectively, for IGH, 7 in the 3′V-REGION and 5′J-REGION for IGK and IGL junctions (*see* **Note 9**). By default, no mutation is accepted for the TR junctions.
4. Set "Parameters for Detailed view":
   (a) "Nb of nucleotides to exclude in 5′ of the V-REGION for the evaluation of the number of mutations" (useful in case of primer specific nucleotides).
   (b) "Nb of nucleotides to add (or exclude) in 3′ of the V-REGION for the evaluation of the alignment score" (useful in case of low (or high) exonuclease activity).

3.1.5 Advanced Functionalities

"Advanced functionalities" [22, 23] corresponds to specific analyses, with additional dedicated results, for engineered/artificial sequences, and for the search of specific sequences for clinical applications. The user may:

1. Select "Analysis of single chain Fragment variable (scFv)" if the submitted set contains engineered single chains with two V-DOMAIN connected by a linker. IMGT/V-QUEST will search for the two V-DOMAIN in the submitted sequence (*see* **Note 10**) [41]. This functionality is generic for IG and TR.
2. Select "Clinical application: search for CLL subsets #2 and #8" for sequences from patients with chronic lymphocytic leukemia (CLL). The analysis of IGH sequences includes the search of specific rearrangements and stereotyped patterns associated to the two CLL subset #2 and subset #8 (*see* **Note 11**).

3.2 IMGT/V-QUEST Results for A. Detailed View

The page "A. Detailed results for the IMGT/V-QUEST analyzed sequences" [22, 23] indicates at the top the number of analyzed sequences and the list of sequences identifiers with links allowing to browse directly the corresponding individual results. Individual results include the FASTA submitted sequence and the "Result summary" of the analysis, followed by the detailed result sections selected in the Welcome page. Importantly, the result sections allow

to explore in depth the results of the analysis regarding the identification of V, (D), J genes and alleles, the description of the V-DOMAIN with the delimitation of FR-IMGT and CDR-IMGT, and the characterization of the mutations.

3.2.1 Sequence and Result Summary

The numbers of 5′ trimmed-n and 3′ trimmed-n from the submitted sequence before the analysis if any (*see* **Note 3**), the sequence length, the sequence analysis category (*see* **Note 12**), and the IMGT reference directory set with which the sequence was compared (e.g., *Homo sapiens* (human) IG set) are indicated above the submitted sequence provided in FASTA format [22, 23] (Fig. 4). The part of the sequence corresponding to the V-DOMAIN is underlined in green. If a sequence was submitted in antisense orientation, it is complementary reversed and displayed, as well as the results, in the V gene sense orientation.

The "Result summary" provides the main characteristics of the analyzed sequence [22, 23]:

1. The evaluation of the sequence functionality: "Productive" or "Unproductive." Only productive sequences are expressed in antigen receptors.
2. The identification of the closest V, (D), and J genes and alleles: the names of the closest "V-GENE and allele" and "J-GENE and allele" are provided with their alignment score (*see* **Note 13**), the percentage of identity and the ratio of the number of identical nucleotides (nt)/number of aligned nt. The name of the closest "D-GENE and allele" determined by IMGT/JunctionAnalysis [37, 38] is indicated with the D-REGION reading frame.
3. The length of the four FR-IMGT and of the three CDR-IMGT between square brackets and separated by dots and the amino acid (AA) JUNCTION sequence.
4. The JUNCTION length (in nt) and the JUNCTION decryption [45], which describe the length (in nt) of the IMGT labels that compose the JUNCTION (*see* **Note 14**).

IMGT/V-QUEST provides warnings (not shown) that appear as notes in red to alert the user, if potential insertions or deletions are suspected in the V-REGION (*see* **Note 15**), or if other possibilities for the J-GENE and allele names are identified. Users are encouraged to check alignments in related detailed result sections.

Below the "Result summary," notes in black (not shown) may appear to indicate:

1. The number of missing nt in the 5′ part of the V-REGION and/or the number of missing nt in the 3′ part of the J-REGION in case of a partial V-DOMAIN.
2. The number of V-REGION uncertain nt number("n") within the analyzed sequence if any.

Species	Homo sapiens (human)
Receptor type or locus	IG
IMGT directory reference set	F+ORF+ in-frame P
Search for insertions and deletions	no

A. Detailed results for the IMGT/V-QUEST analysed sequences

Number of analysed sequences: 1

1. seq_1

This release of IMGT/V-QUEST uses IMGT/JunctionAnalysis for the analysis of the JUNCTION

Hyphens (-) show nucleotide identity, dots (.) represent gaps

Sequence: 1 seq_1

Analysed sequence length: 417.

Sequence analysis category: 1 (no indel search).

Sequence compared with the *Homo sapiens* (human) IG set from the IMGT reference directory (set: F+ORF+ in-frame P)

```
>seq_1
atggagtttgggctgagctggctttttcttgtggctattttaaaaggtgtccagtgtgaa
gtgcagctggtggagtctgagggaggcttggtacagcctggcaggtcccagagactctcc
tgtgcagcctctggattcacctttgatgattatgccatgcactgggtccggcaagctcca
gggaagggcctggagtgggtctcaggtattagttggaatagtggtagcataggctatgca
gactctgtgaagggccgattcaccatctccagagacaacgccaagaactccctgtatctg
caaatgaacagtctgagagctgaggacacggccttgtattactgtgcaaaggggatttt
ggagtggttaacccccttgactactggggccagggaaccctggtcaccgtctcctca
```

Result summary: seq_1	**Productive IGH rearranged sequence** (no stop codon and in-frame junction)		
V-GENE and allele	Homsap IGHV3-9*01 F	score = 1413	identity = **98.96%** (285/288 nt)
J-GENE and allele	Homsap IGHJ4*02 F	score = 208	identity = 93.62% (44/47 nt)
D-GENE and allele by IMGT/JunctionAnalysis	Homsap IGHD3-3*01 F	D-REGION is in reading frame 3	
FR-IMGT lengths, CDR-IMGT lengths and AA JUNCTION	[25.17.38.11]	[8.8.13]	CAKGIFGVVNPLDYW
JUNCTION length (in nt) and decryption	45 nt = (8)-5{3}-7(17)-7{6}-6(11)	(3'V)3'{N1}5'(D)3'{N2}5'(5'J)	

Fig. 4 IMGT/V-QUEST "Detailed results" [22, 23]. The parameters of the analysis are recalled on the top of the page. The first part the "Detailed results" for "seq_1" (IMGT/LIGM-DB [36] accession number X81732) includes the sequence in FASTA format (the first 57 nt not underlined in green are not part of the V-DOMAIN) and the "Result summary." Seq_1 functionality is "productive." This human IGHV sequence expresses the IGHV3-9*01, IGHD3-3*01, and IGHJ4*02 genes and alleles. The lengths of the four FR-IMGT are 25, 17, 38, and 11. The lengths of the three CDR-IMGT are 8, 8, and 13. The JUNCTION length is 45 nt, and the decryption [45] shows that it is composed of 8 nt for the 3′V-REGION (5 nt were trimmed from the germline V during DNA rearrangement), 3 nt for N1-REGION, and 17 nt for the D-REGION (7 nt in 5′ and 7 in 3′ were trimmed from the germline D, 6 nt for the N2-REGION, and 11 for the 5′J-REGION (6 nt were trimmed from the germline J))

3.2.2 Detailed Result Sections

If selected in the Welcome page, the 14 detailed result sections are displayed [22, 23]. They allow to verify, detail, and complete the "Result summary."

1. Detailed result sections for V, D, and J genes and alleles identification: in sections 1–3, the alignments for V, D, and J genes and alleles display the alignments of the user sequence with the five (default value in option "Nb of aligned reference sequences") closest germline V, D, and J gene alleles, respectively, with their alignment score and their identity percentage. All V or J genes and alleles with an identical highest identity percentage in alignments are solutions and are provided in the "Result summary" table (*see* **Note 16**). The alignment for D-GENE and allele should be considered with caution since it may show discrepancies with the results obtained by IMGT/JunctionAnalysis [37, 38] (*see* **Note 17**).
2. Detailed analysis of the JUNCTION: the section 4 provides the Results of IMGT/JunctionAnalysis [37, 38], which include:
   (a) The "Analysis of the JUNCTION" (Fig. 5) [22, 23] shows the details of the junction at the nucleotide level with delimitation of the IMGT labels (Fig. 2 in Subheading 2.3). Dots indicate the number of nucleotides trimmed at the germline V, D, and J gene ends. Vmut, Dmut, and Jmut indicate the number of mutations in the 3′V-REGION, D-REGION, and 5′J-REGION, respectively, and the corresponding mutated nucleotides are underlined in the sequence. "Ngc" corresponds to the ratio of the number of g+c nucleotides to the total number of N nucleotides. The JUNCTION decryption is also provided [45] (*see* **Note 14**). If selected "Eligible D genes" (not shown), all D genes, which match the junction with their corresponding score, are displayed below.
   (b) The "Translation of the JUNCTION" displays the AA JUNCTION with AA colored according to the eleven IMGT physicochemical classes [46] (*see* **Note 18**), the JUNCTION frame ('+' for in-frame, and '-' for out-of-frame), the CDR3-IMGT length, the molecular mass, the isoelectric point (pI), and a link to detailed physicochemical descriptor (not shown). Gaps (represented by dots) are inserted in "out-of-frame" JUNCTION to maintain the J-REGION frame, and the corresponding codon, which cannot be translated, is represented by "#" in AA translation (not shown).
3. The section "5. Sequence of the JUNCTION ("nt" and "AA")" provides the JUNCTION in nt and AA with IMGT unique numbering for in-frame JUNCTION, and in the

4.Results of IMGT/JunctionAnalysis

Maximum number of accepted mutations in: 3'V-REGION = 2, D-REGION = 4, 5'J-REGION = 2
Maximum number of accepted D-GENE: 1

Analysis of the JUNCTION

D-REGION is in reading frame 2

Click on mutated (underlined) nucleotide to see the original one:

Input	V name	3'V-REGION	N1	D-REGION	N2	5'J-REGION
seq_2	Homsap IGHV3-11*05 F	tgtgtgaga..	ataatc	.ggaatagcagcagctgg...	ccc	....ctttgaccagtgg

J name	D name	Vmut	Dmut	Jmut	Ngc	JUNCTION decryption
Homsap IGHJ4*02 F	Homsap IGHD6-13*01 F	1	1	2	4/9	(9)-2{6}-1(17)-3{3}-4(13)

Translation of the JUNCTION

Click on mutated (underlined) amino acid to see the original one:

	104	105	106	107	108	109	110	111	112.1	112	113	114	115	116	117	118	Frame	CDR3-IMGT length	Molecular mass	pI
	C	V	R	I	I	G	I	A	A	A	G	P	F	D	Q	W				
seq_2	tgt	gtg	aga	ata	atc	gga	ata	gca	gca	gct	ggc	ccc	ttt	gac	cag	tgg	+	14	1,717.03	6.15

Fig. 5 IMGT/V-QUEST "Detailed results" [22, 23]. Results of IMGT/JunctionAnalysis [37, 38] for AB063867 IMGT/LIGM-DB accession number. This human IGH sequence results from the rearrangement of IGHV3-11*05 F, IGHD6-13*01 F, and IGHJ4*02 F. The JUNCTION is in-frame. The length of the CDR3-IMGT is 14 AA (42 nt). The JUNCTION length is of 48 nt, and the decryption [45] shows that it is composed of 9 nt for the 3′V-REGION (2 nt were trimmed from the germline V), 6 nt for N1-REGION, and 17 nt for the D-REGION (1 nt in 5′ and 3 in 3′ were trimmed from the germline D, 3 nt for the N2-REGION, and 13 for the 5′J-REGION (4 nt was trimmed from the germline J)

FASTA format with the formatted header required as input by IMGT/JunctionAnalysis online [37, 38]. These results are provided even if IMGT/JunctionAnalysis gives no results.

4. Delimitation of the FR-IMGT and CDR-IMGT in V-REGION: the sections 6, 7 and 8 provide three displays of the V-REGION [22, 23]:
   (a) "6. V-REGION alignment according to the IMGT unique numbering" for the nt sequences with the FR-IMGT and CDR-IMGT delimitations according to the IMGT unique numbering [13].
   (b) "7. V-REGION translation" for the nt sequence and its AA translation, aligned with the closest germline V-REGION.

(c) "8. V-REGION protein display" for the AA translation of the input sequence, aligned with the V-REGION translation of the closest germline V-GENE, and with, on the third line of the alignment and shown in bold, the AA of the input sequence which are different from the closest germline V-REGION.

5. Analysis of the mutations: the sections 9, 10 and 11 are dedicated to the analysis of the nt mutations and AA changes observed in the V-REGION by comparison with the closest germline V gene and allele [22, 23]:

   (a) "9. V-REGION mutation and AA change table" lists the nt mutations and, if nonsilent, the corresponding AA changes. They are described for each FR-IMGT and CDR-IMGT with their nt and codon positions according to the IMGT unique numbering [13]. In parentheses, the "AA class Change Type" indicates if, between germline AA and replaced AA, the hydropathy, volume, and physicochemical properties have been conserved (+) or not (−) according to the IMGT physicochemical classes [46].

   (b) "10. V-REGION mutation and AA change statistics" comprises two tables for the detailed and complete characterization of nt mutations and AA changes: "Nucleotide (nt) mutations" table quantifies nt positions with or without gaps, the identical nt, the total number of mutations, and the silent and nonsilent ones for the V-REGION and per FR-IMGT and CDR-IMGT. It then details the same evaluation for the four types of transitions and of the eight types of transversions. "Amino acid (AA) changes" table quantifies the codons or amino acid positions, with or without gaps, the unchanged AA, and AA changes for the V-REGION and per FR-IMGT and CDR-IMGT (*see* **Note 19**). It then evaluates the number of changes in 4 "AA class Similarity Degree": "Very similar" (the three properties hydropathy, volume, and physicochemical properties are conserved), "Similar" (one of the three properties is changed), "Dissimilar" (two of the three properties are changed), and "Very dissimilar" (the three properties are changed).

   (c) "11. V-REGION mutation hot spots" shows the localization of the hot spot patterns (a/t)a (or wa) and (a/g) g(c/t)(a/t) (or rgyw) and their complementary reverse motifs t(a/t) (or tw) and (a/t)(a/g)c(c/t) (or wrcy) in the closest germline V gene and allele. Finally, this section includes a table for the "Correlation between V-REGION mutations, AA changes, codons changes, and hotspot motifs." It provides a synthesis for each mutation: the position in nt, the AA change and its position according

to the AA numbering [13, 16, 17], the AA class Change Type, the germline and mutated codon, and the corresponding hotspot if any. An illustration is provided in Fig. 6.

6. Sequence annotation with IMGT labels:
   (a) "12. V-REGION and V-(D)-J-REGION" provides nt and AA FASTA sequences with gaps according to the IMGT unique numbering [13, 16, 17] of the V-REGION (nt sequence with access to the IMGT/PhyloGene tool [47]) and of V-J or V-D-J-REGION. In case of out-of-frame junctions V-J or V-D-J-REGION, a note is added, and the V-J or V-D-J-REGION is shown in red.
   (b) "13. Annotation by IMGT/Automat" provides a full automatic annotation for the V-J-REGION or V-D-J-REGION by IMGT/Automat [39, 40] with IMGT labels (*see* Subheading 2.3).
7. Graphical representation of the V-DOMAIN [22, 23]: "14. IMGT Collier de Perles" allows to display the IMGT Collier de Perles for analyzed V-DOMAIN either through a "link to IMGT/Collier-de-Perles tool" [21] (*see* Subheading 7 IMGT/Collier-de-Perles) or as a direct representation integrated in IMGT/V-QUEST results depending on the user selection.

3.2.3 Sequence and Result Summary with the Search for Insertions and Deletions in V-REGION

The insertions and/or deletions that are detected by using the "Advanced parameters" and "Search for insertions" are described in the "Result summary" row [22, 23] (Fig. 7) with their localization in FR-IMGT or CDR-IMGT, the number of inserted or deleted nt, and, for insertions, the inserted nucleotides, the presence or absence of frameshift, the V-REGION codon from which the insertion or deletion starts, and the nt position in the user sequence.

3.2.4 Top of Detailed Results for the Analysis of single chain Fragment variable (scFv)

The Advanced functionality "Analysis of single chain Fragment variable (scFv)" [41] allows the analysis of scFv sequences from phage-display combinatorial libraries [48, 49]. IMGT/V-QUEST [22, 23] identifies, localizes, and characterizes the two V-DOMAIN of a scFv (Fig. 8). At the top of the result page, the number of analyzed sequences and the number of identified V-DOMAIN are indicated. V-DOMAIN identifiers are automatically generated by adding to the sequence identifier a suffix composed of an underscore plus a letter for the locus (H, K, L for IGH, IGK, IGL or A, B, D, G for TRA, TRB, TRD, TRG, respectively). Below the list of V-DOMAIN identifiers is a table that indicates the positions of each V-DOMAIN and of the linker in the identified scFv. The detailed analysis of each individual V-DOMAIN is then provided classically.

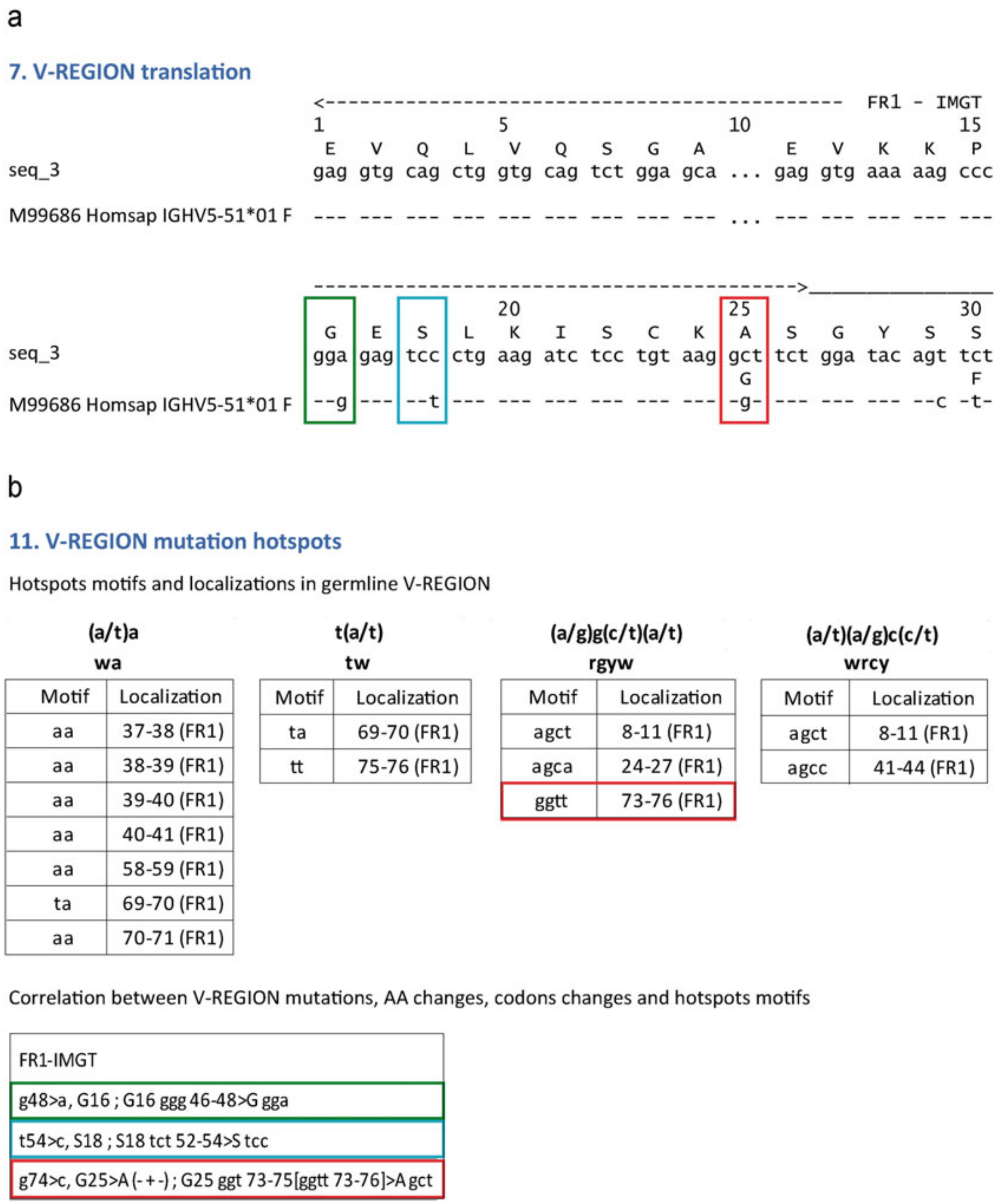


Fig. 6 IMGT/V-QUEST "Detailed results" [22, 23]. Correlation between V-REGION mutations, AA changes, codons changes, and hotspots motifs in FR1-IMGT of seq_3 (accession number AJ006165 of IMGT/LIGM DB [36]). (**a**) "7. V-REGION translation" and (**b**) "11. V-REGION mutation hotspots." Only the FR1-IMGT parts of the results are displayed: the two silent mutations g48>a (G16) and t54>c (S18) are shown in green and light blue rectangles, respectively. The nonsilent mutation g74>c (shown in the red rectangle) leads to the AA dissimilar change G25>A (hydropathy and physicochemical properties are not conserved). The codon "ggt" in position 73-75 is changed in "gct." The nt mutation occurs in the hotspot "ggtt" in position 73–76

Species	Homo sapiens (human)
Receptor type or locus	IG
IMGT directory reference set	F+ORF+ in-frame P
Search for insertions and deletions	yes

A. Detailed results for the IMGT/V-QUEST analysed sequences

Number of analysed sequences: 1

1. seq_4

This release of IMGT/V-QUEST uses IMGT/JunctionAnalysis for the analysis of the JUNCTION

Hyphens (-) show nucleotide identity, dots (.) represent gaps

Sequence: 1 seq_4

Analysed sequence length: 379.
Sequence analysis category: 2 (indel search & correction).
Sequence compared with the *Homo sapiens* (human) IG set from the IMGT reference directory (set: F+ORF+ in-frame P)

>seq_4
caggtgcagctacagcagtggggcgcaggactgttgaagccttcggagaccctgtccctc
acctgcgctgtctatggtgggtccttcagtggttactactggagctggatccgccagtcc
ccagagacggggctggagtggcttggcgaaTTCGATCTTGGTGGAAGCatcactcatagt
agaggcaccaactacaacccgtcgctcaagagtcgagtcaccatctcaggagacacgtcc
aagaaccagttctccctgaaactgacctctgtgaccgccgcggacacggctgtctattac
tgtgcgagaggtttagcaatgggtggaactaaggagtttgactcctggggccagggaacc
ctggtcaccgtctcctcag

Result summary: seq_4

Nucleotide insertions have been detected and automatically removed for this analysis: they are displayed as capital letters in the user submitted sequence above.

localization in V-REGION	nb of inserted nt	inserted nt	causing frameshift	from V-REGION codon	from nt position in user submitted sequence
CDR2-IMGT	18	TTCGATCTTGGTGGAAGC	no	56	151

IMGT/V-QUEST results after removal of the insertion(s)
Potentially productive IGH rearranged sequence (no stop codon and in-frame junction)
(Check also your sequence with BLAST against IMGT/GENE-DB reference sequences to eventually identify out-of-frame pseudogenes)

V-GENE and allele	Homsap IGHV4-34*01 F, or Homsap IGHV4-34*12 F	score = 1299	identity = 95.09% (271/285 nt) [94.74% (270/285 nt)]
J-GENE and allele	Homsap IGHJ4*02 F (a)	score = 204	identity = 91.67% (44/48 nt)
D-GENE and allele by IMGT/JunctionAnalysis	Homsap IGHD6-19*01 F	D-REGION is in reading frame 2	
FR-IMGT lengths, CDR-IMGT lengths and AA JUNCTION	[25.17.38.11]	[8.7.14]	CARGLAMGGTKEFDSW
JUNCTION length (in nt) and decryption	48 nt = (11)0{2}-5(16)0{7}-5(12)	(3'V)3'{N1}5'(D)3'{N2}5'(5'J)	

(a) Other possibilities: Homsap_IGHJ5*02 (highest number of consecutive identical nucleotides)

Fig. 7 IMGT/V-QUEST "Detailed results" [22, 23]. Sequence and Result summary with "Search for insertions and deletions." An insertion of 18 nt is identified in seq_4 (IMGT/LIGM-DB [36] accession number MG950400) from CDR2-IMGT position 56 (from nt 151 in the submitted sequence). The insertion is shown in capital letters

3.3 IMGT/V-QUEST Results for B. Synthesis View

At the top of the page, the parameters used for the analysis are recalled, and the number of analyzed sequences is indicated. The results include a summary table and potentially eight detailed result sections if selected by the user.

3.3.1 Summary Table

The "Summary table" (Fig. 9) displays one row for each input sequence with the corresponding results, including 22 columns [22, 23]: (1) the sequence order in the submission; (2) the sequence identifier (Sequence ID); (3) the name of the closest V-GENE and allele; (4) the functionality of the sequence (when found, the presence of stop codons is indicated); (5) the V-REGION score; (6) the V-REGION percentage of identity with, between parentheses, the ratio of number of identical nucleotides (nt)/number of aligned nt; (7) the name of the closest J-GENE and allele; (8) the J-REGION score; (9) the J-REGION percentage of identity and the ratio of number of identical nucleotides (nt)/number of aligned nt; and provided according to the IMGT/JunctionAnalysis results [37, 38] (10) the D-GENE and allele name; (11) the D reading frame; (12) the CDR-IMGT lengths; (13) the AA JUNCTION; and (14) the JUNCTION frame (in the absence of results of IMGT/JunctionAnalysis, only the AA JUNCTION defined by IMGT/V-QUEST [22, 23] is displayed); (15) the JUNCTION nt length and decryption [45]; (16) the number of missing in 5′ partial V-REGION; (17) the number of uncertain nt; (18) the number of missing nt in 3′ partial J-REGION; (19) and (20) the numbers of 5′ and 3′ trimmed 'n' nucleotides; (21) the length of the sequence; and (22) the sequence analysis category (*see* **Note 12**). Clicking on the sequence ID provides the corresponding Detailed View in a separate tab (depending on your browser). Warnings in red may be indicated to highlight specific features of the sequence (Fig. 9).

3.3.2 Detailed Analysis of the JUNCTION

A link to access the IMGT/JunctionAnalysis [37, 38] results is provided for sequences of the same locus. AA translations are aligned on the longest CDR3 length according to the IMGT unique numbering [13] (Fig. 10).

3.3.3 Detailed Result Sections for Alignment of Sequences Expressing the Same V Gene and Allele

In "Alignment with the closest alleles" below the summary table, the V genes and alleles are listed with the number of assigned sequences in parentheses [22, 23]. Click on the associated link to reach the corresponding detailed result sections. They provide six different displays (if all were selected) of alignment of sequences

Fig. 7 (continued) in the FASTA sequence. IMGT/V-QUEST then performs a classical analysis (for gene and allele identification, analysis of the JUNCTION, evaluation of nt mutation, and AA changes) after removal of the insertion(s) and addition of gaps to replace the deletions. The evaluation of the identity percentage added in square brackets includes each insertion or deletion as an additional mutation

Species	Homo sapiens (human)
Receptor type or locus	IG
IMGT directory reference set	F+ORF+ in-frame P
Search for insertions and deletions	no
Analysis of scFv	yes

A. Detailed results for the IMGT/V-QUEST analysed sequences

Number of analysed sequences: 2
Number of analysed V-DOMAIN: 4

1. scFv_1_H, 2. scFv_1_K, 3. scFv_2_H, 4. scFv_2_K

Identified scFv:

Sequence ID	5'V-DOMAIN ID	5'V-DOMAIN positions	5'V-DOMAIN length	linker positions	linker length	3'V-DOMAIN ID	3'V-DOMAIN positions	3'V-DOMAIN length
scFv_1	1_scFv_1_H	1..349	349	350..390	41	2_scFv_1_K	391..714	324
scFv_2	3_scFv_2_H	1..361	361	362..405	44	4_scFv_2_K	406..727	322

This release of IMGT/V-QUEST uses IMGT/JunctionAnalysis for the analysis of the JUNCTION
Hyphens (-) show nucleotide identity, dots (.) represent gaps

V-DOMAIN: 1 scFv_1_H (associated V-DOMAIN: 2 scFv_1_K)

Analysed sequence length: 714
Sequence analysis category: 1 (no indel search).
Sequence compared with the *Homo sapiens* (human) IG set from the IMGT reference directory (set: F+ORF+ in-frame P)

>scFv_1_H
gaggtgcagctgttggagtctgggggaggcttggtacagcctgggggtccctgagactc
tcctgtgcagcctctggattcacctttagcagctatgccatgagctgggtccgccaggct
ccagggaaggggctggagtgggtctcagctattagtggtagtggtggtagcacatactac
gcagactccgtgaagggccggttcaccatctccagagacaattccaagaacacgctgtat
ctgcaaatgaacagcctgagagccgaggacacggccgtatattactgtgcgaaatctctt
cttctttttgactactggggccagggaaccctggtcaccgtctcgagtggcgatgggtcc
agtggcggtagcgggggcgcgtcgactggcgaaattgtgttgacgcagtctccaggcacc
ctgtctttgtctccaggggaaagagccaccctctcctgcagggccagtcagagtgttagc
agcagctacttagcctggtaccagcagaaacctggccaggctcccaggctcctcatctat
ggtgcatccagcagggccactggcatcccagacaggttcagtggcagtgggtctgggaca
gacttcactctcaccatcagcagactggagcctgaagattttgcagtgtattactgtcag
cagtggggtgagaagcccttgacgttcggccaagggaccaaggtggaaatcaaa

Result summary: scFv_1_H	**Productive IGH rearranged sequence** (no stop codon and in-frame junction)		
V-GENE and allele	Homsap IGHV3-23*01 F, or Homsap IGHV3-23D*01 F	score = 1440	identity = **100.00%** (288/288 nt)
J-GENE and allele	Homsap IGHJ4*02 F	score = 159	identity = 81.25% (39/48 nt)
D-GENE and allele by IMGT/JunctionAnalysis	Homsap IGHD2-15*01 F	D-REGION is in reading frame 1	
FR-IMGT lengths, CDR-IMGT lengths and AA JUNCTION	[25.17.38.11]	[8.8.9]	CAKSLLLFDYW
JUNCTION length (in nt) and decryption	33 nt = (9)-2{3}-21(8)-2{1}-5(12)	(3'V)3'{N1}5'(D)3'{N2}5'(5'J)	

Fig. 8 IMGT/V-QUEST "Detailed results" [22, 23]. Top of Detailed results for the "Advanced functionality" "Analysis of single chain fragment variable (scFv)". scFv_1 and scFv_2 sequences correspond to the accession numbers AJ006120 and AF117956 in the IMGT/LIGM-DB database [36]. The 5′V-DOMAIN are VH and 3′ V-DOMAIN are V-KAPPA for both scFv. The detailed results are then provided for each domain

B. Synthesis for the IMGT/V-QUEST analysed sequences

Number of analysed sequences: 4

Sequence compared with the *Homo sapiens* (human) IG set from the IMGT reference directory (set: F+ORF+ in-frame P)

Summary table:

a

Sequence Number	Sequence ID	V-GENE and allele	V-DOMAIN Functionality	V-REGION score	V-REGION identity % (nt)	J-GENE and allele	J-REGION score	J-REGION identity % (nt)	D-GENE and allele	D-REGION reading frame	CDR-IMGT lengths
1	AB021529	Homsap IGHV3-9*01 F	productive	1359	96.88% (279/288 nt)	Homsap IGHJ4*02 F	129	78.57% (33/42 nt)	Homsap IGHD3-16*01 F	2	[8.8.14]
2	X81732	Homsap IGHV3-9*01 F	productive	1413	98.96% (285/288 nt)	Homsap IGHJ4*02 F	208	93.62% (44/47 nt)	Homsap IGHD3-3*01 F	3	[8.8.13]
3	AB245095	Homsap IGHV5-51*01 F	productive	1251	92.71% (267/288 nt)	Homsap IGHJ4*02 F, or Homsap IGHJ4*03 F (a)	141	77.08% (37/48 nt)	Homsap IGHD5-12*01 F	2	[8.8.13]
4	AJ006165	Homsap IGHV5-51*01 F	productive	1287	94.10% (271/288 nt)	Homsap IGHJ6*02 F	238	87.10% (54/62 nt)	Homsap IGHD3-10*01 F	2	[8.8.15]

b

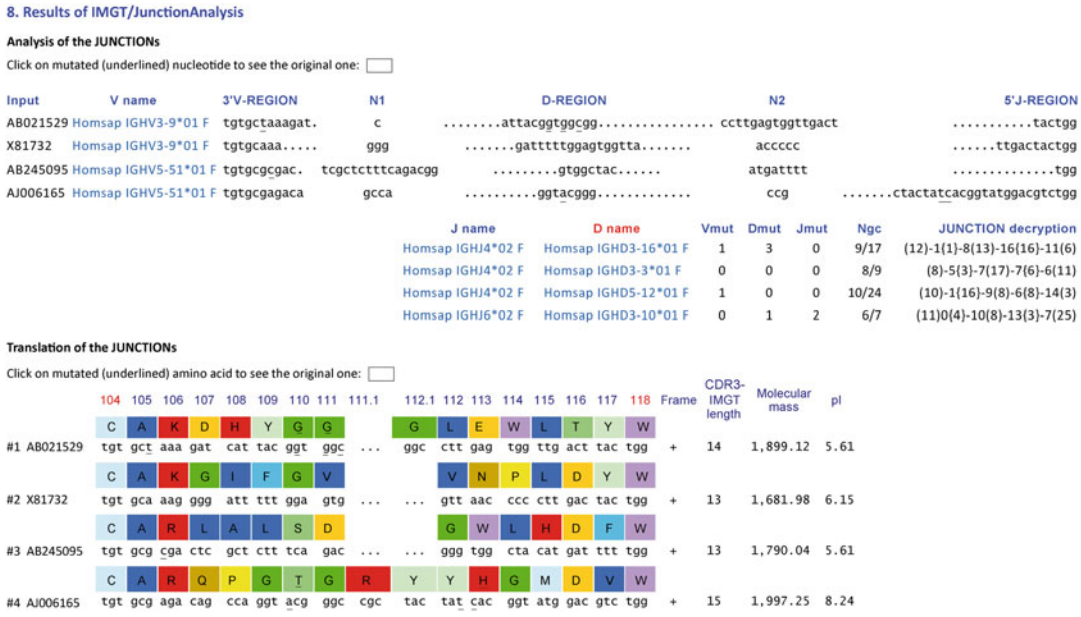
	AA JUNCTION	JUNCTION frame	JUNCTION length (in nt) and decryption	V-REGION partial 5prime missing nt nb	V-REGION uncertain nt nb	J-REGION partial 3prime missing nt nb	5prime trimmed-n nb	3prime trimmed-n nb	Analysed sequence length	Sequence analysis category
1	CAKDHYGGGLEWLTYW	in-frame	48 nt = (12)-1(1)-8(13)-16(16)-11(6)	0	0	5	-	-	358	1 (noindelsearch)
2	CAKGIFGVVNPLDYW	in-frame	45 nt = (8)-5(3)-7(17)-7(6)-6(11)	0	0	0	-	-	417	1 (noindelsearch)
3	CARLALSDGWLHDFW	in-frame	45 nt = (10)-1(16)-9(8)-6(8)-14(3)	0	0	0	-	-	684	1 (noindelsearch)
4	CARQPGTGRYYHGMDVW	in-frame	51 nt = (11)0(4)-10(8)-13(3)-7(25)	0	0	0	-	-	412	1 (noindelsearch)

Fig. 9 IMGT/V-QUEST Synthesis view [22, 23]. (**a**) The 12 first columns of the "Summary table": the four analyzed sequences are shown in the "V-GENE and allele name" order in the Summary table. The sequences ID are accession numbers of IMGT/LIGM-DB [36]. The hyperlinks allow to get the corresponding results in "A Detailed view." The two sequences assigned to IGHV3–9 and the two sequences assigned to IGHV5-51 will be, respectively, aligned together in the detailed result sections 1 to 6. In the column "J-GENE and allele," a warning "(**a**)" in red indicates that other IGHJ genes and alleles may be solutions for the sequence 3 (not shown). (**b**) The last 12 columns of the "Summary table": the V-DOMAIN of sequence 1 is partial: 5 nt are missing in the 3′ part of the J-REGION

that express the same V gene and alleles: "1. Alignment for V-GENE," "2. V-REGION alignment according to the IMGT unique numbering" [13], "3. V-REGION translation," and three different formats for the "V-REGION protein display." Section "7. V-REGION most frequently occurring AA per position and per FR-IMGT and CDR-IMGT" shows the most frequent AA in sequences expressing the same V genes and alleles per FR-IMGT and CDR-IMGT and per position according to the IMGT unique numbering [13].

3.4 IMGT/V-QUEST Output for Excel File

"Excel file" allows the users to open and save a spreadsheet including the results of the IMGT/V-QUEST analysis [22, 23]. The file contains 11 sheets or 12 for the Advanced Functionality "Analysis of single chain Fragment variable (scFv)" [41] (*see* Subheading 4.3 for the detail of their content in IMGT/HighV-QUEST sequence analysis results).



8. Results of IMGT/JunctionAnalysis

Analysis of the JUNCTIONs

Click on mutated (underlined) nucleotide to see the original one:

Input	V name	3'V-REGION	N1	D-REGION	N2	5'J-REGION
AB021529	Homsap IGHV3-9*01 F	tgtgctaaagat.	c	.....attacggtggcgg.....	ccttgagtggttgact	..........tactgg
X81732	Homsap IGHV3-9*01 F	tgtgcaaa.....	ggg	.....gatttttggagtggtta.....	accccc	.....ttgactactgg
AB245095	Homsap IGHV5-51*01 F	tgtgcgcgac.	tcgctctttcagacgg	.....gtggctac.....	atgatttt	.....tgg
AJ006165	Homsap IGHV5-51*01 F	tgtgcgagaca	gcca	.....ggtacggg.....	ccg	.....ctactatcacggtatggacgtctgg

J name	D name	Vmut	Dmut	Jmut	Ngc	JUNCTION decryption
Homsap IGHJ4*02 F	Homsap IGHD3-16*01 F	1	3	0	9/17	(12)-1(1)-8(13)-16(16)-11(6)
Homsap IGHJ4*02 F	Homsap IGHD3-3*01 F	0	0	0	8/9	(8)-5(3)-7(17)-7(6)-6(11)
Homsap IGHJ4*02 F	Homsap IGHD5-12*01 F	1	0	0	10/24	(10)-1(16)-9(8)-6(8)-14(3)
Homsap IGHJ6*02 F	Homsap IGHD3-10*01 F	0	1	2	6/7	(11)0(4)-10(8)-13(3)-7(25)

Translation of the JUNCTIONs

Click on mutated (underlined) amino acid to see the original one:

	104	105	106	107	108	109	110	111	111.1		112.1	112	113	114	115	116	117	118	Frame	CDR3-IMGT length	Molecular mass	pI
	C	A	K	D	H	Y	G	G			G	L	E	W	L	T	Y	W				
#1 AB021529	tgt	gct	aaa	gat	cat	tac	ggt	ggc	...		ggc	ctt	gag	tgg	ttg	act	tac	tgg	+	14	1,899.12	5.61
	C	A	K	G	I	F	G	V				V	N	P	L	D	Y	W				
#2 X81732	tgt	gca	aag	ggg	att	ttt	gga	gtg	...		...	gtt	aac	ccc	ctt	gac	tac	tgg	+	13	1,681.98	6.15
	C	A	R	L	A	L	S	D				G	W	L	H	D	F	W				
#3 AB245095	tgt	gcg	cga	ctc	gct	ctt	tca	gac	...		...	ggg	tgg	cta	cat	gat	ttt	tgg	+	13	1,790.04	5.61
	C	A	R	Q	P	G	T	G	R		Y	Y	H	G	M	D	V	W				
#4 AJ006165	tgt	gcg	aga	cag	cca	ggt	acg	ggc	cgc		tac	tat	cac	ggt	atg	gac	gtc	tgg	+	15	1,997.25	8.24

Fig. 10 IMGT/V-QUEST Synthesis view [22, 23]. Results of IMGT/JunctionAnalysis for four IGH junctions [37, 38]: "Analysis of the JUNCTIONs" displays for the four junctions, the sequences of the IMGT labels 3′V-REGION, N1, D-REGION, N2, and 5′J-REGION. The mutated nt are underlined. "Translation of the JUNCTIONs" displays the four junctions aligned per position according to the IMGT unique numbering [13]

4 IMGT/HighV-QUEST

IMGT/HighV-QUEST [24–27] is the high throughput version of IMGT/V-QUEST [22, 23]. It is freely available for academics, but it requires the user's registration. This allows the tool to automatically notify the users on the availability of the results. A link to the "New user" form is provided in the IMGT/HighV-QUEST welcome page. When the user logs in, the tool uses reCAPTCHA (https://developers.google.com/recaptcha) to protect the site from spam and abuse.

IMGT/HighV-QUEST provides two main functionalities [24–27]:

1. The high throughput analysis of IG and TR rearranged sequences based on the IMGT/V-QUEST algorithm (use the "IMGT/HighV-QUEST Search page" for sequence submission (*see* Subheading 4.1) and use the "Analysis history" page for the download of sequence analysis results (*see* Subheading 4.2)).
2. The evaluation of the diversity and of the expression of the IMGT clonotypes (AA) in analyzed sequence sets (use the "Launch statistics" page for IMGT clonotypes evaluation (*see*

Subheading 4.4), and use "Statistics history" page for the download of IMGT clonotypes results (*see* Subheading 4.5)).

Links to the four pages are displayed on the top of the IMGT/HighV-QUEST web interface.

4.1 IMGT/HighV-QUEST Sequence Set Submission

IMGT/HighV-QUEST Search page (Fig. 11) is provided when the user logs in. It includes the four following sections [24–27]:

4.1.1 The Sequence Submission Form

1. Provide an analysis title, select the species (*see* **Note 1**), and the receptor type or locus (*see* **Note 2**) as for IMGT/V-QUEST.
2. Upload a simple text-formatted file containing your FASTA sequences (up to 1,000,000 of IG or TR rearranged sequences can be submitted in a single run).
3. When an analysis is launched ("Start" button), it is firstly dispatched and queued on the IMGT servers and is then performed depending on the available resources. Choose to be notified by e-mail "when analysis is queued" and/or "when analysis is completed" (selected by default).

4.1.2 Display Results

1. Select Result format: the default "CSV" result format includes 11 (or 12 with the Advanced Functionality "Analysis of single chain Fragment variable (scFv)") CSV files equivalent to those provided by the "Excel file" of IMGT/V-QUEST. Result format "AIRR" [42, 43] (*see* **Note 7**) or "Both formats" can be selected.
2. The individual result files (equivalent to IMGT/V-QUEST "Detailed view" in text format) can be included in the results for submissions of maximally 200,000 sequences only.

4.1.3 Advanced Parameters

The analysis can be customized with exactly the same advanced parameters as proposed by IMGT/V-QUEST (*see* Subheading 3.1.4).

4.1.4 Advanced Functionalities

"Analysis of single-chain Fragment variable (scFv)" [41] can be included in the analysis (default is "no") (*see* Subheading 3.1.5).

4.2 IMGT/HighV-QUEST Analysis History Page: Follow-Up and Download of Results

The "Analysis history" page allows the user to check the status of the submitted analyses [24–27]. A table displays for each of them its title, its status (queued, running, or completed), the submission date, the number of submitted sequences, the species and the receptor type or locus (as selected by the user), and the actions that can be performed. When the analysis is completed, the user can download the results as a single archive file in TXZ format (commonly supported by archive tools for windows and other operating systems). The availability of the results is guaranteed for two weeks

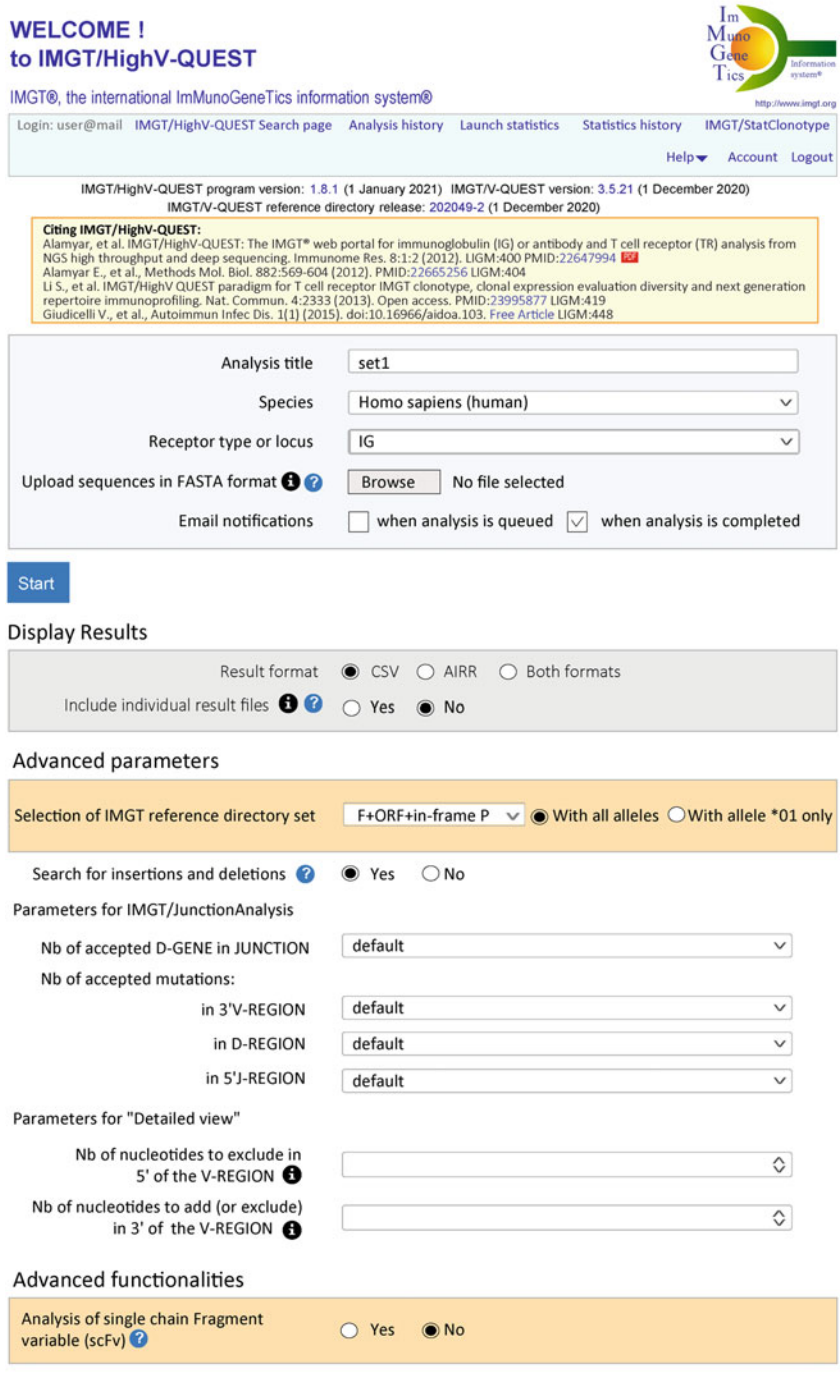


Fig. 11 The IMGT/HighV-QUEST Search page [24–27]

after the analysis is completed. After that, the files can be removed by the system. In that case, "File removed" is indicated in red instead of the archive logo.

A user may delete an analysis at any time except if it is used by the second module "Statistics" of IMGT/HighV-QUEST. In such cases, "Used by Statistics" is indicated in place of the "delete" button.

4.3 IMGT/HighV-QUEST Sequence Analysis Results

The content of the TXZ file depends on the selected "Result format" ("CSV," "AIRR," or "Both formats") [24–27]:

1. "CSV" format contains a tar folder (which needs to be extracted by an archive tool) with 11 (or 12) files (equivalent to the results of the excel file provided by the classical IMGT/V-QUEST) in CSV format, and, if selected in the IMGT/HighV-QUEST Search page, one subfolder with individual result files, in text format for each sequence (equivalent to the classical IMGT/V-QUEST "Detailed view" results (*see* Subheading 3.2)).

   The content of each CSV file is indicated in Table 1 [27].
2. "AIRR" format contains a "vquest_airr.tsv" file, generated by the tool in AIRR format [42, 43] (described in the IMGT/V-QUEST [22, 23] Documentation http://www.imgt.org/IMGT_vquest/vquest_airr) and the "11_Parameters.txt" for the parameters used for the analysis.
3. "Both formats" selection includes "CSV" and "AIRR" format results.

4.4 IMGT/HighV-QUEST Launch Statistics Page for the Evaluation of IMGT Clonotypes

An IMGT clonotype (AA) is defined by a unique V-(D)-J-rearrangement (V and J genes and alleles), with a unique CDR3-IMGT amino acid sequence and the presence of the conserved anchors C 104 and W/F 118 [26]. An IMGT clonotype (AA) is linked to one or more IMGT clonotype (nt): they are defined by a unique V-(D)-J-rearrangement with a unique CDR3-IMGT nucleotide sequence, whose translation corresponds to the CDR3-IMGT of the IMGT clonotype (AA) [26]. When IMGT/HighV-QUEST "Statistics" is launched, the tool evaluates IMGT clonotypes in batches of analyzed sequence sets per locus and provides immunoprofiles for IMGT clonotypes (AA) diversity (number of different IMGT clonotypes per V, D, and J gene and allele) and expression (number of sequences assigned to IMGT clonotypes (AA) per V, D, and J gene and allele) [26]. Moreover, IMGT/HighV-QUEST can perform the comparison of multiple batches and provide the list of the IMGT clonotype (AA), which are common to two or more batches [26].

Table 1
List of the IMGT/HighV-QUEST CSV files with the number of columns and result content [27]

File number	Result type	File name	Number of columns	Results content (*see* Note 20)
#1	Gene and allele identification, summary of the analysis, characterization of insertions and deletions	"Summary"	33 (or 29)	Result overview: · Sequence order and sequence ID · V-DOMAIN Functionality · Identification of V, D, and J genes and alleles · Identity percentage with the closest V and J genes and alleles and alignment scores · FR-IMGT and CDR-IMGT lengths · Amino acid (AA) JUNCTION · Description of insertions and deletions (indels) in V-REGION if any · User sequence in the direct orientation · Sequence orientation at the submission, the number of trimmed "n" before analysis if any, sequence length, sequence analysis category This file may include notes regarding the evaluation of the functionality, the V, D, and J gene and allele identification, potential indels in V-REGION in three dedicated columns (V-DOMAIN Functionality comment, V-REGION potential ins/del, J-GENE, and allele comment)
#2	Sequence description and annotation	"IMGT-gapped-nt-sequences"	18	Sequences for main labels of the gapped nt V-(D)-J-REGION: · 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 · nt sequences of CDR3-IMGT, JUNCTION, J-REGION and FR4-IMGT

(continued)

Table 1
(continued)

File number	Result type	File name	Number of columns	Results content (*see* Note 20)
#3		"Nt-sequences"	118	Full annotation of V-(D)-J-REGION nucleotide sequence with IMGT labels: · nt sequences of all labels that can be automatically annotated by IMGT/Automat · Start and end positions of annotated labels The four last columns evaluate the V-REGION reading frame, the number of missing 5′ and 3′ nt for partial V-(D)-J-REGION, and the number of uncertain nt in V-REGION
#4		"IMGT-gapped-AA-sequences"	18	Sequences for main IMGT labels of the gapped AA V-(D)-J-REGION: · 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 · AA sequences of CDR3-IMGT, JUNCTION, J-REGION, and FR4-IMGT
#5		"AA-sequences"	18	Sequences for main IMGT labels of the non-gapped AA V-(D)-J-REGION: · Same columns as "IMGT-gapped-AA-sequences" (#4), but sequences of IMGT labels are without IMGT gaps
#6		"Junction"	85	Results of IMGT/JunctionAnalysis [37, 38]: 37 columns for IGL, IGK, TRA and TRG sequences, 51 (if one D), 63 (if two D) or 78 (if 3 D) columns for IGH, TRB, and TRD sequences

#7	Analysis of mutations	"V-REGION-mutation-and-AA-change table"	11	Correlation between V-REGION mutations, AA changes [46], codons changes and hotspots motifs Description of the mutations for V-REGION, FR1-IMGT, CDR1-IMGT, FR2-IMGT, CDR2-IMGT, FR3-IMGT and germline CDR3-IMGT, each of them characterized by: the nt mutation, the AA changes and the 3 AA class identity (+) or change (−), the codon change, and the corresponding hotspot with their localization
#8		"V-REGION-nt-mutation-statistics"	130	Characteristics and number of nt mutations: · Number of nt positions including IMGT gaps, number of nt, number of identical nt, total number of mutations, number of silent mutations, and number of nonsilent mutations · Number of transitions (a>g, g>a, c>t, t>c) and number 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
#9		"V-REGION-AA-change-statistics"	109	Number of AA positions including IMGT gaps, number of AA, number of identical AA, total number of AA changes, number of AA changes according to the AA class Change Type (+++, ++−, +−+, +−−, −+−, −−+, −−), and number of AA class changes according to AA class Similarity Degree (Very similar, Similar, Dissimilar, and Very dissimilar) for V-REGION, FR1-IMGT, CDR1-IMGT, FR2-IMGT, CDR2-IMGT, FR3-IMGT, and germline CDR3-IMGT [46]

(continued)

Table 1
(continued)

File number	Result type	File name	Number of columns	Results content (*see* Note 20)
#10		"V-REGION-mutation-hotspots"	8	Hotspot motifs (a/t)a, t(a/t), (a/g)g(c/t)(a/t), and (a/t)(a/g)c(c/t) detected in the closest germline V-REGION with their localization in FR-IMGT and CDR-IMGT
#11	"Parameters"	"Parameters"		· Date of the analysis · IMGT/V-QUEST program version, IMGT/V-QUEST reference directory release · Parameters used for the analysis: species, receptor type or locus, IMGT reference directory set, advanced parameters, advanced functionalities
#12	Sequence description	scFv	40	Available only for the advanced functionality "Analysis of single chain Fragment variable (scFv)," one line per scFv: Positions and length of the V-(D)-J-REGION, CDR_lengths, JUNCTION for the 2 V-DOMAIN of the scFv, positions and length of the linker [41]

For launching statistics, the following steps should be followed:

1. Provide a statistical analysis title.
2. Indicate if an email notification should be sent when the statistical analysis is completed.
3. Provide comments on the analysis (optional).
4. Choose if the "Multiple batch comparison" will be performed (yes is selected by default).
5. Define a list of batches: in order to define a batch, click on the "show" button in the form "Define a batch" of the page. It allows to list the available sequences analyses already performed by IMGT/HighV-QUEST. They are displayed in a table with their title, the user name, the status of the analysis, the number of submitted sequences, the species and receptor type of locus, and the main information for analysis parameters (IMGT reference directory set and Search for insertions/deletions) (*see* **Note 21**). Provide a short title for the batch (six characters or less) before adding it to the list. Up to 15 batches can be defined.
6. Click on the start button to launch the run.

4.5 IMGT/HighV-QUEST Statistics History Page: Follow-Up and Download of Statistics

It allows to follow the status of the submitted statistics analysis and to download the results once completed [24–27]. The IMGT/HighV-QUEST statistical output is provided as a zip file.

1. Extract the archive.
2. Open the file "open_to_start.html" localized in the main folder with a web browser.

4.5.1 Results Sections to be Displayed in the User Web Browser

The IMGT/HighV-QUEST statistics output [26] is organized in the sections listed in Table 2 (see also http://www.imgt.org/HighV-QUEST/doc.action#statistical-outputs-results).

The illustration of the content of file 4.2.1 is shown in Fig. 12: it shows the first seven most expressed IMGT clonotypes (AA) of a list of 27,080.

4.5.2 "Data" Directory

Importantly, the archive includes a "data" directory: it contains text files named 'stats_xxx' where 'xxx' is composed of 'batch name'_'-locus'. They include the list of all the IMGT clonotypes (AA) (that are displayed through html sections of Table 2) and their characteristics separated by tabulations. These files include the fields needed by the external IMGT/StatClonotype [28, 29] tool (*see* Subheading 5). Their content is described in the IMGT/HighV-QUEST Documentation at http://www.imgt.org/HighV-QUEST/doc.action#datastatsxxx.

Table 2
Documents included in IMGT/HighV-QUEST statistics output [26]

Documents	File type	Content description
1. "Selected parameters" and "batch list table"	html	Species, Receptor type or locus, IMGT reference directory set, Search for insertions/deletions (yes or no), the total number of sequences, Batch IMGT clonotype comparison (yes or no), and then the list of batches with the titles of the sequence analyses, the number of sequences, the species, Receptor type (or locus), the program version of IMGT/HighV-QUEST and IMGT/V-QUEST, and the release of the IMGT reference directory
2. Result summary for batches	html	List of batches including their ID, nb of sequences, of "1 copy," "1 copy with indels," "More than 1," "More than 1 with indels" (*see* **Note 22**), the number of sequences with no J-GENE, No junction, Warnings, Unknown functionality, and with No results
3. Result summary for IMGT clonotypes (AA)	html	Number of IMGT clonotypes (AA), of assigned sequences, number of in-frame sequences not assigned to IMGT clonotypes (AA), number of productive sequences, in-frame unproductive sequences, out-of-frame sequences, sequences "1 copy" + "More than 1," "single gene," "several genes," and of submitted sequences per batch and per locus
4. Detailed results per batch	html	
4.1 "Results categories" and V, D, and J genes and alleles for genotype analysis ("1 copy" "single gene" for V and J)	zip	Includes five pdf reports including the list of filtered out sequences, the number of "1 copy single gene" and of "1 copy several genes," and a folder of graphics
4.2 Detailed IMGT clonotype (AA and nt) results per locus	html	
4.2.1 IMGT clonotypes (AA) per Nb	html	List of IMGT clonotypes (AA) ordered by decreasing number of assigned sequences with the IMGT clonotype (AA) definition, the IMGT clonotype (AA) representative sequence, and access of corresponding FASTA "1 copy" sequences
4.2.2 IMGT clonotypes (AA) per number with detailed clonotypes (nt)	html	Same as 4.2.1 with associated IMGT clonotype (nt)
4.2.3 IMGT clonotypes (AA) per V gene	html	Identical to 4.2.1 ordered by V gene

(continued)

Table 2
(continued)

Documents	File type	Content description
4.2.4 IMGT clonotypes (AA) per V gene with detailed clonotypes (nt)	html	Identical to 4.2.2 ordered by V gene
4.2.5 IMGT clonotypes (AA) per CDR3-IMGT length (AA)	html	Identical to 4.2.1 ordered by CDR3 length
4.2.6 IMGT clonotypes (AA) per CDR3-IMGT length (AA) with detailed clonotypes (nt)	html	Identical to 4.2.2 ordered by CDR3 length
4.2.7 IMGT clonotypes (AA) with identical CDR3-IMGT (AA) with detailed clonotypes (nt) per CDR3-IMGT length (AA)	html	IMGT clonotypes (AA) grouped by CDR3-IMGT (AA) with detailed clonotypes (nt) per CDR3-IMGT length (AA)
4.2.8 IMGT clonotype (AA) diversity and expression histograms: per V, (D), J-GENE and per CDR3-IMGT length	html	· IMGT clonotype (AA) expression histograms: number of sequences assigned to an IMGT clonotype (AA) per V-GENE (green color), D-GENE (for IGH, TRB, TRD) (red color) and J-GENE (yellow color) and per CDR3-IMGT length · IMGT clonotype (AA) diversity histograms: number of different IMGT clonotypes (AA) per V-GENE, D-GENE (for IGH, TRB, TRD) and J-GENE (pink color) and per CDR3-IMGT length
4.2.9 IMGT clonotype (AA) diversity and expression tables: per V, (D), J-GENE and per CDR3-IMGT length	html	Tables for the number of sequences assigned to an IMGT clonotype (AA) and the number of IMGT clonotypes (AA) per V-GENE, D-GENE (for IGH, TRB, TRD) and J-GENE, and per CDR3-IMGT length
4.2.10 V gene and allele table: Rearrangements, number of sequences and number IMGT clonotypes (AA) per V-GENE and allele	html	Number of sequences assigned to an IMGT clonotype (AA), number of different IMGT clonotypes (AA), number of out-of-frame sequences, and number of sequences of other categories per V-GENE and allele
5. IMGT clonotype (AA) results comparison	html	
5.1 IMGT clonotype (AA) comparison: Full results	html	Lists of IMGT clonotypes (AA) unique for each batch and lists for common IMGT clonotypes (AA) in two or more batches
5.2 IMGT clonotype (AA) comparison: Synthesis table	html	Number of IMGT clonotypes (AA) (diversity) and the number of sequences assigned to IMGT clonotypes (AA) (expression) only present ('exclusive') in a single batch or common to two or more batches
	html	"Number of IMGT clonotypes (AA)," "Number of in-frame sequences assigned to

(continued)

Table 2
(continued)

Documents	File type	Content description
5.3 IMGT clonotype (AA) comparison: Result summary table per V-GENE, D-GENE (for IGH, TRB, TRD), J-GENE		IMGT clonotypes (AA)," per gene, and for each batch

a

ID		Nb			IMGT clonotype (AA) definition							IMGT clonotype (AA) representative sequence			IMGT clonotypes (nt)
#	Exp. ID	Total nb of '1 copy'	Total nb of 'More than 1'	Total	V gene and allele	D gene and allele	J gene and allele	CDR3-IMGT length (AA)	CDR3-IMGT sequence (AA)	Anchors 104, 118	V %	Sequence length	Functio -nality	Sequence ID	Sequences file ('1 copy')
1	21129-S3	224	11	235	Homsap IGHV4-4*07 F	Homsap IGHD1-1*01 F	Homsap IGHJ6*03 F	13 AA	ARGTTFYYYYMDV	C,W	100	497	productive	SRR1168790.43 G9YJURD01DTO length=497_NA	Sequences file
2	15294-S3	148	0	148	Homsap IGHV4-4*07 F	Homsap IGHD3-16*01 F	Homsap IGHJ4*02 F	15 AA	ARDPLGGNSALTFDY	C,W	98.6	557	productive	SRR1168790.29 G9YJURD01AKHJ length=557_NA	Sequences file
3	13555-S3	125	0	125	Homsap IGHV4-39*01 F	Homsap IGHD5-12*01 F	Homsap IGHJ4*02 F	16 AA	ARLAPQSKSHVSAPDY	C,W	99.66	515	productive	SRR1168790.63 G9YJURD01AZDA length=515_NA	Sequences file
4	5172-S3	121	1	122	Homsap IGHV4-59*01 F	Homsap IGHD6-6*01 F	Homsap IGHJ6*03 F	20 AA	ARTPIGHYSSSSKRYYYMDV	C,W	100	521	productive	SRR1168790.60 G9YJURD01B57Z length=521_NA	Sequences file
5	23185-S3	106	1	107	Homsap IGHV4-61*02 F	Homsap IGHD6-25*01 F	Homsap IGHJ3*01 F	12 AA	ARGSGIAPVMDV	C,W	90.34	499	productive	SRR1168790.349 G9YJURD01B3TE length=499_NA	Sequences file
6	5153-S3	105	0	105	Homsap IGHV4-34*01 F	Homsap IGHD2-21*01 F	Homsap IGHJ4*02 F	20 AA	ARSWGYCGSIDCQTPVGLGY	C,W	97.19	509	productive	SRR1168790.59 G9YJURD01ARW length=509_NA	Sequences file
7	17162-S3	96	2	98	Homsap IGHV4-31*03 F	Homsap IGHD2-15*01 F, or Homsap IGHD2-21*01 F	Homsap IGHJ4*02 F	14 AA	ACDVQTSQYVAFDY	C,W	87.29	517	productive	SRR1168790.42 G9YJURD01BH3 length=517_NA	Sequences file

b

#	CDR3-IMGT length (nt)	Nb diff CDR3-IMGT (nt)	CDR3-IMGT sequence (nt)	Nb diff nt	V gene and allele	D gene and allele	J gene and allele	Anchors 104, 118	V % mean	V-REGION length mean	J % mean	J-REGION length mean	Sequence length mean	Total nb of '1 copy'	Total nb of 'More than 1'	Total
7	42	3	gcgtgcgacgtccagacgtcacaatatgtagcttttgactac	1	Homsap IGHV4-31*03 F	Homsap IGHD2-15*01 F	Homsap IGHJ4*02 F	C,W	86.55	298	60.42	43	510	1	0	1
			gcgtgtgacgtccagacgtcacaatatgtagccttgactac	1	Homsap IGHV4-31*03 F	Homsap IGHD2-21*01 F	Homsap IGHJ4*02 F	C,W	87.24	298	81.25	44	506	1	0	1
			gcgtgtgacgtccagacgtcacaatatgtagcttttgactac	0	Homsap IGHV4-31*03 F	Homsap IGHD2-15*01 F	Homsap IGHJ4*02 F	C,W	86.4	298	80.59	43	514	94	2	96

Fig. 12 Top of the file 4.2.1 'IMGT clonotypes (AA) per Nb' [26]. (**a**) List of IMGT clonotypes (AA) ordered by decreasing number of assigned sequences. The first seven of IMGT clonotypes (AA) are shown. The table provides the Exp. ID (IMGT clonotype (AA) identifier in the set), the numbers of "1 copy," "More than one," and the total. The IMGT clonotype (AA) definition includes the names of the V, D, and J genes and alleles, the CDR3-IMGT length, the AA CDR3-IMGT, and the anchors. The IMGT clonotype (AA) representative sequence is characterized by the identity percentage with the closest V gene and allele, the sequence length, the sequence functionality, and a link to the FASTA sequence. An additional link allows to display all "1 copy" assigned to the IMGT clonotype (AA). The batch S3 results from the analysis of the run SRR1168790 available on Sequence Read Archive (SRA) (https:/www.ncbi.nlm.nih.gov/sra). (**b**) Example of IMGT clonotypes (nt) linked to the IMGT clonotypes (AA) #7 (extracted from file 4.2.2) to which 96 "1 copy" sequences were assigned. Ninety-four of them are assigned to the same IMGT clonotype (nt) with a CDR3-IMGT of 42 nucleotide "gcgtgtgacgtccagacgtcacaatatgtagcttttgactac". Two other IMGT clonotypes (nt) (with one sequence each) are also linked to #7. One shows a mutation (t>c) on the nt 6 of the CDR3 and the second a mutation (t>c) on the nt 33 of the CDR3 (shown in red in the figure)

5 IMGT/StatClonotype

IMGT/StatClonotype [28, 29] provides statistical pairwise comparison of the diversity and of the expression of the IMGT clonotypes (AA) between two IMGT/HighV-QUEST statistics output results [24, 26]. The tool evaluates the statistical significance of the differences in proportions per variable (V), diversity (D), and joining (J) gene and allele of a given IG or TR locus according to seven multiple testing procedures for the adjustment of the *p*-value: this allows the user to choose the stringency, which is the most relevant for the aim of a given study [28, 29]. IMGT/StatClonotype includes the characterization of the CDR-IMGT with the analysis of the distribution of CDR-IMGT length (for IMGT clonotype (AA) diversity or expression) and, for a given length, the distribution per position of the amino acids according to IMGT AA physicochemical classes [46] and variability indexes. Results for the evaluation of V-(D)-J associations are provided through heatmaps.

5.1 IMGT/StatClonotype Installation and Launch

IMGT/StatClonotype [28, 29] is a standalone IMGT® tool that needs to be installed locally on the user's computer. Running IMGT/StatClonotype requires the prior installation of the R program (*see* **Note 23**):

1. Install R program and IMGTStatClonotype R package following the steps described in "IMGTStatClonotype R package installation" (http://www.imgt.org/StatClonotype/IMGTStatClonotypeDoc.html#pack).
2. In the console of R program, following the prompt ">," enter the two command lines:

   library(IMGTStatClonotype)

   launch()

   The IMGT/StatClonotype web interface is launched on your default web browser.

5.2 IMGT/StatClonotype Uploading of Input Sets of IMGT Clonotypes (AA)

1. In the left panel of IMGT/StatClonotype welcome page, choose the IMGT/HighV-QUEST set 1 and IMGT/HighV-QUEST set 2 to be compared (Fig. 13). The IMGT/StatClonotype input sets must be selected from IMGT/HighV-QUEST statistical analysis output folders, which are already stored on your computer (*see* Subheading 4.5, IMGT/HighV-QUEST "Statistics history" page: follow-up and download of statistics, Subheading 4.5.2 "data" directory).
2. Select CDR3-IMGT length range of IMGT clonotypes (AA) in order to eliminate outliers from the statistical procedures (default range for CDR3-IMGT lengths is ≥4 and ≤45).

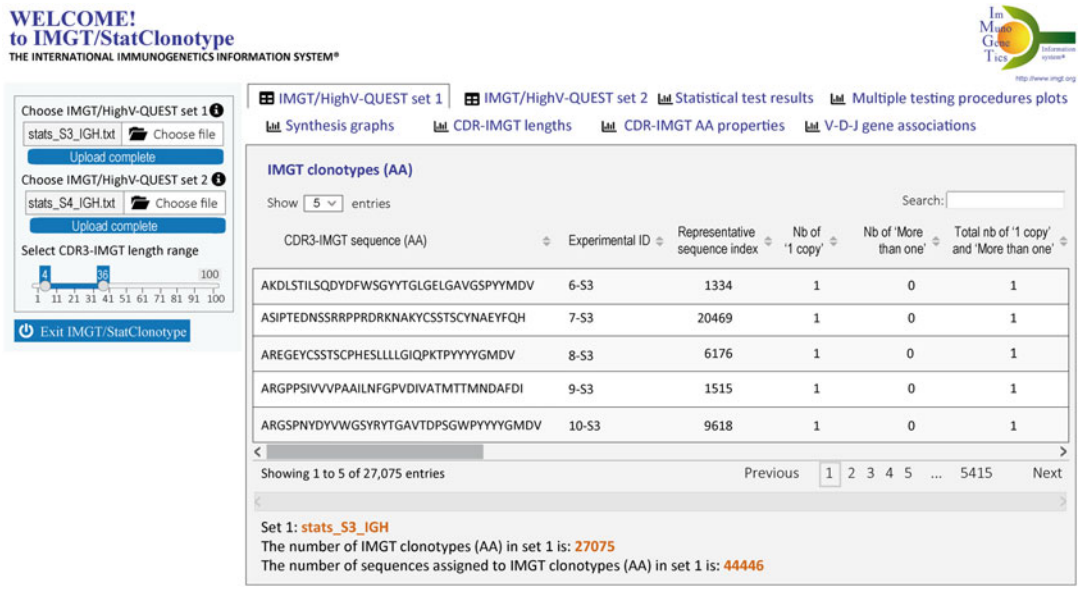


Fig. 13 IMGT/StatClonotype Welcome page [28, 29]. The files stats_S3_IGH.txt (IMGT/HighV-QUEST set 1) and stats_S4_IGH.txt (IMGT/HighV-QUEST set 2) were uploaded from the "data" directory of IMGT/HighV-QUEST statistical output (obtained from the runs SRR1168790 and SRR1168789, respectively, available on Sequence Read Archive (SRA) (https:/www.ncbi.nlm.nih.gov/sra)). The range for CDR3-IMGT lengths is $\geq$4 and $\leq$36. IMGT/HighV-QUEST set 1 includes 27,075 IMGT clonotypes (AA) to which 44,446 sequences were assigned

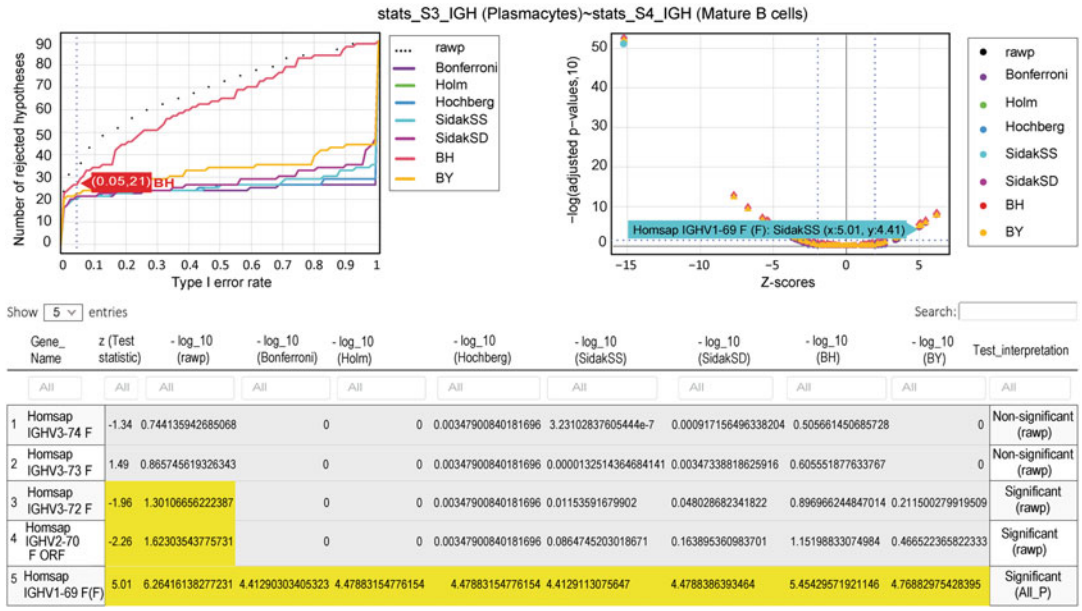
5.3 IMGT/StatClonotype Results

IMGT/StatClonotype [28, 29] results are displayed in the major right panel in eight distinct tabs:

1. The "IMGT/HighV-QUEST set 1" and "IMGT/HighV-QUEST set 2" tabs display, for set 1 and set 2, respectively, a table of the IMGT clonotypes (AA) with a CDR3-IMGT length in the range selected in the left panel (Fig. 13). The table includes [28, 29]: CDR3-IMGT sequence (AA), experimental ID, representative sequence index, Nb of "1 copy," Nb of "More than one" (*see* **Note 22**), total number of "1 copy" and "More than one," "1 copy" indexes, V gene, V allele, D gene, D allele, J gene, J allele, CDR1-IMGT, CDR2-IMGT, CDR1-IMGT-gapped sequence (AA), CDR2-IMGT-gapped sequence (AA), V-REGION %identity, sequence length, C104, F/W118, anchors (true or false), sequence ID, functionality, sequence file number, and sequence clonotype number.

   The total number of the selected IMGT clonotypes (AA) and the number of sequences assigned are indicated below the table. At the bottom of the page, a second table lists the IMGT clonotypes (AA) corresponding to CDR3-IMGT length outliers that are not taken into account for statistical procedures (not shown).

2. The "statistical test results" tab [28, 29] displays the statistical test results of differences in proportions for the IMGT clonotypes (AA) in both sets, for genes (top of the page) and for alleles (bottom of the page; *see* **Note 24**), without adjusted *p*-values or with adjusted *p*-values according to the seven multiple testing procedures (Bonferroni, Holm, Hochberg, ŠidákSS, ŠidákSD, Benjamini & Hochberg, Benjamini & Yekutieli) [28]. The results are displayed in a table of 21 columns (*see* **Note 25**). Columns 1–12 provide gene (or allele) name, gene (or allele) type, "Nb of IMGT clonotypes (AA)," "Proportion" and "normalized proportion" for set 1 and set 2 (*see* **Note 26**), "Difference in proportions," *z*-scores values, and lower and upper bound confidence interval (CI) of the difference in proportions. Unadjusted *p*-values (rawp) and adjusted *p*-values from multiple testing are given from column 13 to column 20 of the table. The column 21 provides a test interpretation for the significance of the difference in proportion. The "Download" button allows to save the tables as CSV files. Below the main table is the "Show/Hide Table" button that displays (or not) the list of genes (or alleles) with null or small occurrences. Use the left panel to customize the display: (1) select the results for IMGT clonotype (AA) diversity or for IMGT clonotype (AA) expression, (2) include results for several genes (or alleles) or for single genes (or alleles) only, (3) display or not the null or smallest gene (or allele) occurrences, (4) select or unselect one or more columns of the "Statistical test results for genes" and "Statistical test results for alleles" tables to be shown.
3. "Multiple testing procedures plots" tab [28, 29] displays, in the major right panel, interactive line graphs, and scatter plots for genes (on the top) and for alleles (at the bottom), for the comparison of the differences in proportions for IMGT clonotypes (AA) between sets 1 and 2 (Fig. 14). On the left, the line graphs display the number of rejected null hypotheses (therefore the number of significant differences in proportions) for a chosen Type I error for the seven procedures. On the right, the scatter plots show negative decimal logarithms (−log10) of unadjusted p-values (black symbols) and adjusted *p*-values obtained by each multiple testing procedure (colored symbols): it highlights the V, D, and J genes of a locus with the most significant differences (positive or negative) in proportions. Numerical values and z-scores are reported in a table below, the plots with a yellow background for significant positive or negative differences in proportions.
4. "Synthesis Graphs" tab [28, 29] displays a synthesis graph that combines a normalized bar graph of proportions (*see* **Note 26**) and the differences in proportions with significance and



	Gene_ Name	z (Test statistic)	- log_10 (rawp)	- log_10 (Bonferroni)	- log_10 (Holm)	- log_10 (Hochberg)	- log_10 (SidakSS)	- log_10 (SidakSD)	- log_10 (BH)	- log_10 (BY)	Test_interpretation
1	Homsap IGHV3-74 F	-1.34	0.744135942685068	0	0	0.00347900840181696	3.23102837605444e-7	0.000917156496338204	0.505661450685728	0	Non-significant (rawp)
2	Homsap IGHV3-73 F	1.49	0.865745619326343	0	0	0.00347900840181696	0.000013251436468414 1	0.00347338818625916	0.605551877633767	0	Non-significant (rawp)
3	Homsap IGHV3-72 F	-1.96	1.30106656222387	0	0	0.00347900840181696	0.01153591679902	0.048028682341822	0.896966244847014	0.211500279919509	Significant (rawp)
4	Homsap IGHV2-70 F ORF	-2.26	1.62303543775731	0	0	0.00347900840181696	0.0864745203018671	0.163895360983701	1.15198833074984	0.466522365822333	Significant (rawp)
5	Homsap IGHV1-69 F(F)	5.01	6.26416138277231	4.41290303405323	4.47883154776154	4.47883154776154	4.4129113075647	4.4788386393464	5.45429571921146	4.76882975428395	Significant (All_P)

Fig. 14 IMGT/StatClonotype Multiple testing procedures plots for genes [28, 29]. In the left panel, "IMGT clonotype (AA) diversity," "Single gene," and "Hide null or smallest gene occurrences" were selected (not shown). "Multiple testing procedures plots" displays an interactive line graph on the left and a scatter plot on the right for genes. Hovering the mouse on the interactive the line graph on the left allows the display of the exact number of significant differences in proportions, that is 21 for a Type I error $\alpha = 0.05$ with the multiple testing procedure BH. On the right, the scatter plot shows the coordinates of z-score and −log_10 (SidakSS) for IGHV1-69, for which the difference in proportion is significant whatever the multitesting procedure as indicated in the table. The graphs can be saved in PNG, JPG, or PDF and the tables in CSV format

confidence intervals (CI), for genes (on the top) (Fig. 15) and for alleles (at the bottom) (*see* **Note 27**). In synthesis graphs for genes, IMGT gene names are ordered by their positions in the locus with their known functionalities. Below the normalized bar graph are listed the not ordered genes (not shown). They are grouped and shown at the bottom of the gene list in the synthesis graph. The values for the normalized proportions of genes (or alleles) in set 1 and set 2, the differences in proportions, the lower and upper bound of the confidence indices for differences in proportions, and the Test interpretation are recorded in "Statistical test results" tab Tables.

5. "CDR-IMGT lengths" tab [28, 29] displays, in the right panel, interactive bar graphs for set 1 and set 2 showing the distribution of the number of IMGT clonotypes (AA) (for IMGT clonotype (AA) diversity) or of the number of sequences assigned to IMGT clonotypes (AA) (for IMGT clonotype (AA) expression), per CDR-IMGT length (*see* **Note 28**). The left panel allows to choose the CDR-IMGT (CDR1-IMGT, CDR2-IMGT, or CDR3-IMGT) and to select the length of

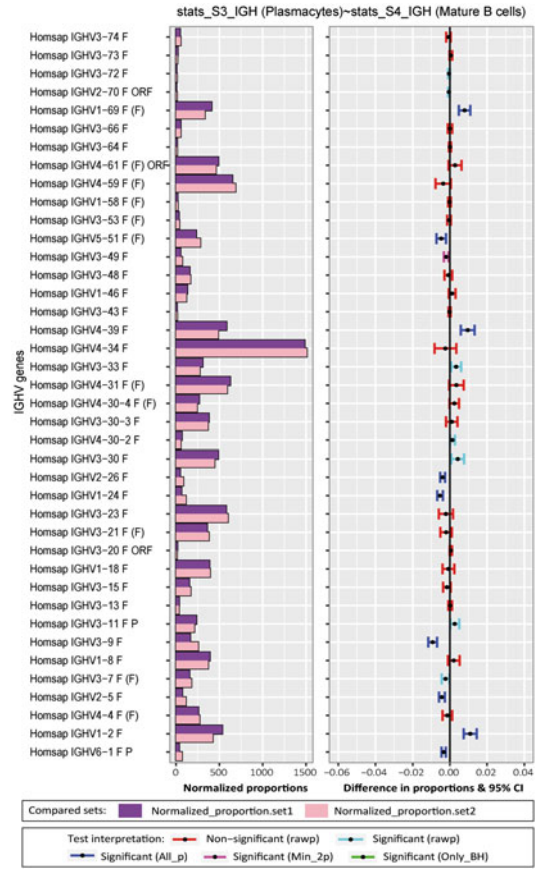


Fig. 15 IMGT/StatClonotype synthesis graph for IMGT clonotype (AA) diversity per V gene [28, 29]. It displays visual comparison of the normalized proportions of IMGT clonotype (AA) diversity of the IGHV genes between sets 1 and 2. For example, the diversity of IMGT Clonotypes (AA) expressing the IGHV1-2, IGHV4-39, and IGHV1-69 genes is significantly higher in set 1 than in set 2 whatever the multiple testing procedure. In the left panel, "Single gene" and "Hide null or smallest gene occurrences" were selected. Synthesis graphs are downloadable in PNG, JPG, or PDF format

the CDR-IMGT for the "List of IMGT clonotypes (AA) with selected CDR3-IMGT length" displayed below the bar graphs for sets 1 and 2.

6. "CDR-IMGT AA Properties" tab displays the distribution of the IMGT classes [46] of the 20 amino acids at CDR-IMGT (CDR1-IMGT, CDR2-IMGT, or CDR3-IMGT) positions in sets 1 and 2 for a given CDR-IMGT length. The left panel allows (1) to select the IMGT classes to be displayed for the amino acids: "20 amino acids," "Physicochemical" (Fig. 16), "Hydropathy," "Volume," "Chemical," "Charge," "Hydrogen donor or acceptor atoms," and "Polarity"; (2) to show

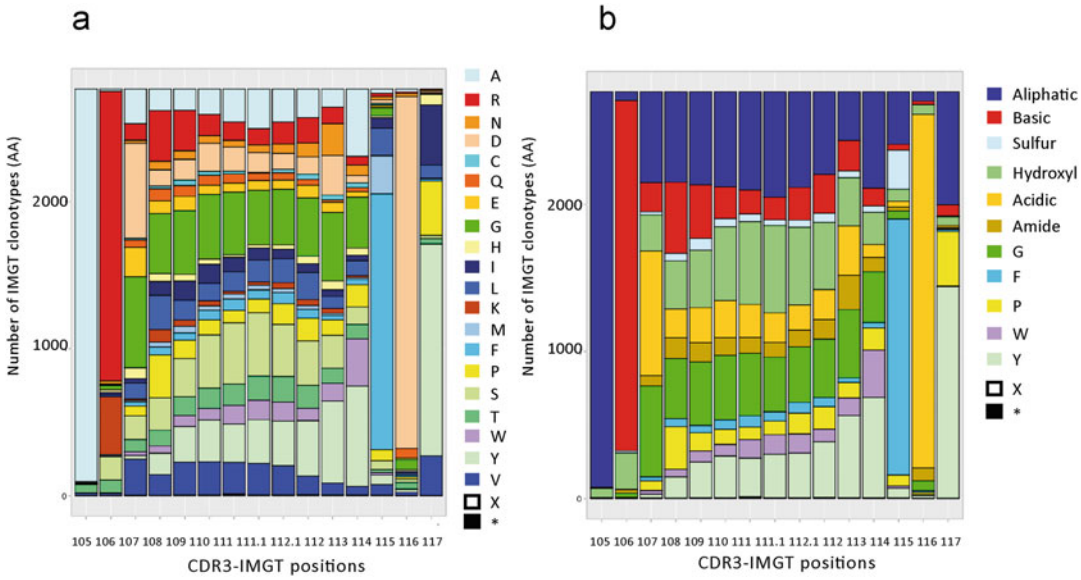


Fig. 16 IMGT/StatClonotype CDR-IMGT AA properties distribution [28, 29]. Examples for the CDR3-IMGT of length 15 in set 1: (**a**) "20 amino acids" and (**b**) "Physicochemical"

results by absolutes values (number of occurrences of an amino acid (or IMGT amino acid class) at a given position, for a given CDR length) or percentages; and (3) to modify the length and width of the graphs. The major right panel includes, for each set, a table with numbers (or percentages) of the amino acids (or IMGT amino acid classes) (in rows), at a given position (in columns). The table includes a row for undefined amino acids ("X") and for stop codons. The tables can be downloaded as CSV files. The corresponding graphical representation is shown as an interactive bar graph to visualize the amino acid distribution per position. At the bottom of the page, the variability plots based on the indexes according to "Shannon entropy," 'Wu-Kabat variability," or "Simpson index" with tables for numerical values are displayed. Comparisons of two sets are useful in detecting the characteristics of amino acids at positions important for the V domain antibody diversity or, by contrast, for maintaining its structure.

7. "V-D-J gene associations" tab [28, 29] displays interactive heat maps to represent V-J, V-D, or D-J gene associations in set 1 and set 2. The left panel allows (1) to display the Dendrogram for V-J, V-D, or D-J gene association, (2) to get the results with clustering or not, (3) to get the results in normalized values, and (4) to select the color palettes. The major right panel includes interactive heat maps to represent V-J, V-D, or D-J gene associations in set 1 and set 2. If the "Results with clustering" is selected, a double Ward hierarchical clustering with Euclidean distance is performed (this classification

operates simultaneously on the lines and columns of a matrix intersecting two different types of genes), otherwise heat maps are shown without dendrograms and ordering. Such an analysis permits to detect genes with similar diversity or expression profiles, which can be further explored for given and/or related specificities in immune repertoire comparative analysis. Under heat maps, tables crossing the V-J, V-D, or D-J gene occurrences in set 1 and set 2 are given.

6 IMGT/DomainGapAlign

IMGT/DomainGapAlign [30, 31] analyzes the amino acid sequences of the IG and TR V-DOMAIN (*see* **Note 29**). IMGT/DomainGapAlign identifies the closest V and genes and alleles of the user's amino acid domain sequences by comparison with the IMGT reference directory sets composed of the translations of the germline V and J regions of the genes managed in IMGT/GENE-DB [34]. The reference amino acid sequences are available by querying IMGT/DomainDisplay (IMGT® Home page, http://www.imgt.org). Importantly, IMGT/DomainGapAlign can analyze V-DOMAIN from different species and different locus in a single run. The tool gaps the sequences, numbers the AA of each V-DOMAIN, and provides the delimitations of the FR-IMGT and CDR-IMGT and those of the beta strands and loops by applying the IMGT unique numbering [13]. It also characterizes the amino acid changes (*see* **Note 30**).

6.1 IMGT/DomainGapAlign Query and Customization of the Analysis

6.1.1 Standard Parameters and Sequence (s)

1. Paste your FASTA amino acid sequences in the text area or upload them from a text file.
2. By default, the analysis is performed on the V-DOMAIN (Domain type "V′) [30, 31] (Fig. 17).
3. Select the species (by Latin name or English name) of the reference directory sets with which the sequences will be compared or let the tool (with default "any") detecting the best alignments whatever the species.
4. Set, with the option "Smith-Waterman score above," the threshold of the Smith-Waterman score above, which the alignments will be displayed in the results (*see* **Note 31**) (default is 0).

   Select the number of alignments displayed for each V-DOMAIN in the results (default is 5).
5. Check "IMGT Colliers de Perles" [21] to include the IMGT Collier de Perles [18–20] in the results (*see* Subheading 7)

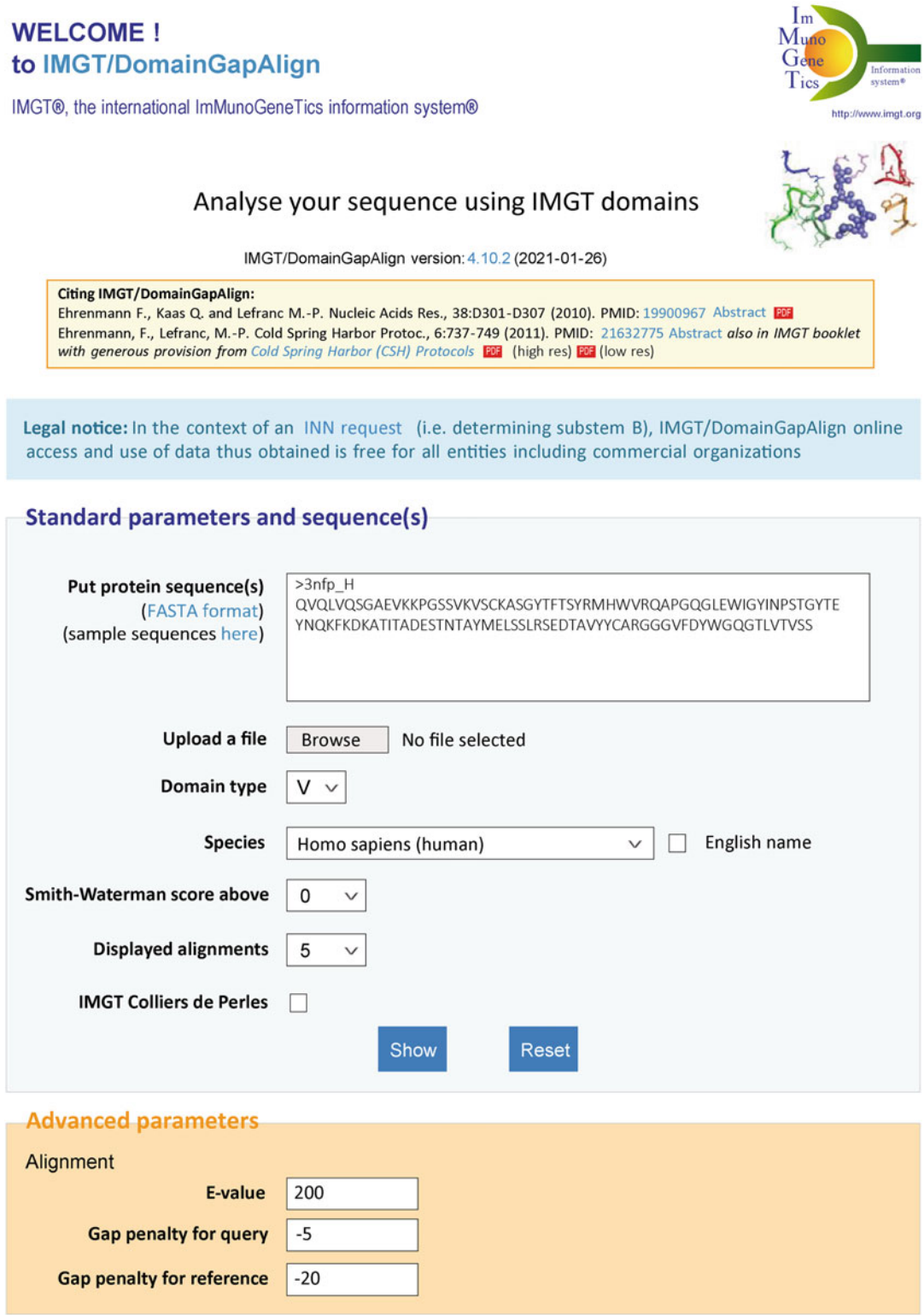


Fig. 17 IMGT/DomainGapAlign Welcome page [30, 31]

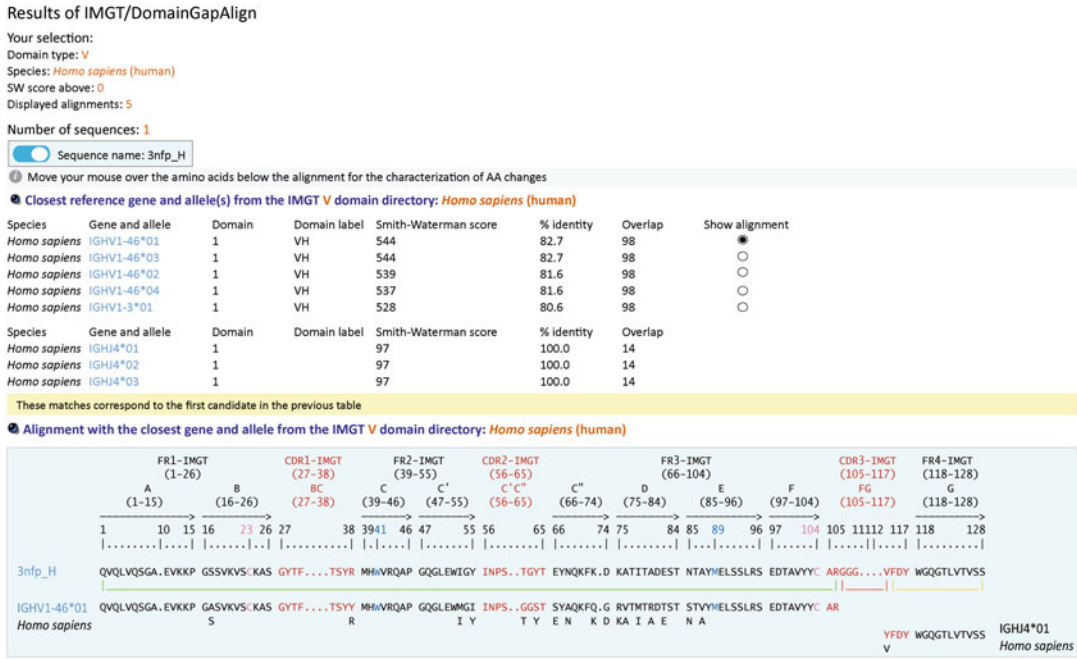


Fig. 18 IMGT/DomainGapAlign Results [30, 31]. Top of the result page: the VH domain of daclizumab 3nfp_H chain (PDB code 3nfp of IMGT/3Dstructure-DB [30, 50, 51]) is compared with the *Homo sapiens* reference directory. It is aligned with the human IGHV1-46*01 and IGHJ4*01 genes alleles

6.1.2 Advanced Parameters

Modify if necessary the "E-value," the "Gap penalty," and "Gap penalty for reference" used for Smith-Waterman alignments.

6.2 IMGT/DomainGapAlign Results

"Your selection," on the top of the IMGT/DomainGapAlign results page [30, 31] (Fig. 18), recalls the parameters and values selected for the AA sequence submission. Following the "Number of sequences", a switch button allows to display (or not) the results of the corresponding sequence.

1. "Closest reference gene and allele(s) from the IMGT V domain directory" (Fig. 18) shows the name of the species of which the AA references sequences are compared with the user sequence ("All species" is indicated if "any" was selected).

   A table summarizes the five (selected by default) best aligned V genes and alleles including the species, the IMGT V gene and allele name, the number of the domain, the Domain label, and the Smith-Waterman alignment score, with the identity percentage and the overlap (number of aligned amino acids assigned to the V-REGION) between the user and the IMGT AA reference sequence, and a radio button for the alignment to display (if the Smith-Waterman score is equal or higher than the threshold selected for the submission).

A second table is displayed for J genes and alleles with the species, the IMGT J gene and allele name, the number of the domain, the Smith-Waterman alignment score, the identity percentage, and the overlap.

2. "Alignment with the closest gene and allele from the IMGT V domain directory" [30, 31] (Fig. 18): the header of the alignment indicates the length and delimitation of the 4 FR-IMGT and 3 CDR-IMGT and of the 9 beta strands (A, B, C, C′, C″, D, E, F, and G) and 3 loops (BC, C′C″, and FG) according to the IMGT unique numbering for V domain [13]. The submitted AA-gapped sequence is aligned with the V region of the germline gene and allele. Between both sequences, a green line delimits the V-REGION, a red line delimits the (N-D)-REGION, and a yellow one delimits the J-REGION. Below the alignment, the AA changes compared with the germline are shown. The AA J-REGION is aligned with the closest J gene and allele.
3. "Region(s) and domain(s) identified in your sequence (by comparison with the closest genes and alleles" (Fig. 19) allows to download the V-DOMAIN amino acid sequence with or without IMGT gaps.
4. "Results summary (by comparison with the closest genes and alleles)" (Fig. 19) provides the first table, which includes the percentage identity with the V-REGION, the CDR-IMGT lengths, the total number of different AA in CDR1-IMGT and CDR2-IMGT, the FR-IMGT lengths, the number of different AA in FR-IMGT, and the total number of amino acid changes.

   Below are displayed two additional parallel tables: on the left the "AA changes in strands and loops" and on the right the "AA changes in FR-IMGT and CDR-IMGT" with the number of different AA, the description of the AA change with the "AA class Change Type'" (+) or not (−) (for hydropathy, volume and physicochemical characteristics [46] according to the AA IMGT classes), and "AA class Similarity Degree" (very similar, similar, dissimilar, and very dissimilar).
5. IMGT Colliers de Perles [18–20] (*See* Subheading 7) are shown, if selected, on one or two layers, without or with AA change positions shown in pink circles (or squares for CDR-IMGT anchors).

7 IMGT/Collier-de-Perles

The IMGT/Collier-de-Perles tool [21] generates "'IMGT Colliers de Perles" [18–20]. For V-DOMAIN, IMGT Colliers de Perles are obtained on one or two layers, provided that the V-DOMAIN

Region(s) and domain(s) identified in your sequence (by comparison with the closest genes and alleles: *Homo sapiens* IGHV1-46*01 and IGHJ4*01)

QVQLVQSGAEVKKPGSSVKVSCKASGYTFTSYRMHWVRQAPGQGLEWIGY
INPSTGYTEYNQKFKDKATITADESTNTAYMELSSLRSEDTAVYYCARGG
GVFDYWGQGTLVTVSS

Sequence without gaps Sequence with gaps

Results summary (by comparison with the closest genes and alleles *Homo sapiens* IGHV1-46*01 and IGHJ4*01)

Sequence name	V-REGION identity percentage	CDR-IMGT lengths	Number of different AA in CDR1- and CDR2-IMGT	FR-IMGT lengths	Number of different AA in FR-IMGT	Total number of AA changes in V-DOMAIN
3nfp_H	82.7%	[8.8.9]	3	[25.17.38.11] = 91 AA	14	17

AA changes in strands and loops

Strands	Number of different AA	AA changes
A (1-15)	0	-
B (16-26)	1	A17>S (- + -) dissimilar
C (39-46)	0	-
C' (47-55)	2	M53>I (+ + -) similar I55>Y (- - -) very dissimilar
C'' (66-74)	4	S66>E (- - -) very dissimilar A68>N (- - -) very dissimilar Q72>K (+ - -) dissimilar G74>D (- - -) very dissimilar
D (75-84)	5	R75>K (+ + +) very similar V76>A (+ - +) similar M78>I (+ + -) similar R80>A (- - -) very dissimilar T82>E (- - -) very dissimilar
E (85-96)	2	S85>N (- - -) very dissimilar V87>A (+ - +) similar
F (97-104)	0	-
G (118-128)	0	-

Loops	Number of different AA	AA changes
BC (27-38)	1	Y38>R (- - -) very dissimilar
C'C'' (56-65)	2	G62>T (+ - -) dissimilar S64>Y (+ - -) dissimilar
FG (105-117)	0	-

AA changes in FR-IMGT and CDR-IMGT

FR-IMGT	Number of different AA	AA changes
FR1-IMGT (1-26)	1	A17>S (- + -) dissimilar
FR2-IMGT (39-55)	2	M53>I (+ + -) similar I55>Y (- - -) very dissimilar
FR3-IMGT (66-104)	11	S66>E (- - -) very dissimilar A68>N (- - -) very dissimilar Q72>K (+ - -) dissimilar G74>D (- - -) very dissimilar R75>K (+ + +) very similar V76>A (+ - +) similar M78>I (+ + -) similar R80>A (- - -) very dissimilar T82>E (- - -) very dissimilar S85>N (- - -) very dissimilar V87>A (+ - +) similar
FR4-IMGT (118-129)	0	-

CDR-IMGT	Number of different AA	AA changes
CDR1-IMGT (27-38)	1	Y38>R (- - -) very dissimilar
CDR2-IMGT (56-65)	2	G62>T (+ - -) dissimilar S64>Y (+ - -) dissimilar
CDR3-IMGT (105-117)	0	-

Fig. 19 IMGT/DomainGapAlign Results [30, 31]. Bottom of the result page for VH domain of daclizumab 3nfp_H chain (PDB code 3nfp of IMGT/3Dstructure-DB [30, 50, 51]): the CDR-IMGT lengths are [8.8.9] with a total of three AA changes. The FR-IMGT lengths are [25.17.38.11] with a total of 14 AA changes

(AA) sequence is gapped according to the IMGT unique numbering [13] (*see* **Note 32**). Resulting IMGT Colliers de Perles show the standardized delimitation of FR-IMGT and CDR-IMGT, and of beta strands with their orientation in the IG and TR V-DOMAIN, allowing the visualization of the amino acids, which are important for a 3D structural configuration and bridging the gap between sequences and structures.

7.1 IMGT/Collier-de-Perles Launched from IMGT Sequence Analysis Tools

1. In order to generate the IMGT Colliers de Perles from a V-DOMAIN nucleotide sequence, use IMGT/V-QUEST [22, 23] (*see* Subheading 3) and select A. Detailed view and result section 14.
2. Starting from a V-DOMAIN amino acid sequence, use IMGT/DomainGapAlign [30, 31] (*see* Subheading 6) to generate the IMGT Colliers de Perles and select "IMGT Colliers de Perles" in the submission form (*see* **Note 33**).

7.2 IMGT/Collier-de-Perles Submission Interface

Alternatively, using the IMGT/Collier-de-Perles interface [21] (Fig. 20) offers complete display options. The submitted V domain (AA) sequence must be gapped according to the IMGT unique numbering for V-DOMAIN [13] and the CDR3-IMGT length must be 13 or longer. The user may:

1. Select the "Domain type" ("Variable (V)"), the number of layers for the IMGT Collier de Perles representation (1 or 2) (*see* **Note 34**).
2. Select the "CDR-IMGT color type" [46] according to the locus of the sequence (1 for IGH, TRB, or TRD sequences and 2 for IGK, IGL, TRA or TRG sequences) and the "Background color," which will be applied to the FR-IMGT positions (*see* **Note 35**).
3. Enter the CDR3-IMGT length.
4. Enter the gapped AA sequence without any header.
5. In case of detected amino acid insertions compared with the IMGT unique numbering for V domain [13], provide in "Amino acid insertions" the position that precedes the insertion, its length in AA, and the numbering label for each inserted position.
6. A title for the resulting IMGT Collier de Perles can be optionally provided.
7. Click on "Show" to launch the tool.

7.3 IMGT/Collier-de-Perles Results

The IMGT Collier de Perles for a V-DOMAIN [18–21] displays the graphical representation of a V-DOMAIN with one position (1 AA) per bead (circle or square). Numbers allow an easy delimitation of the FR-IMGT, of the CDR-IMGT, and of the beta strands

Fig. 20 The IMGT/Collier-de-Perles Welcome page [21]

and the localization of the conserved amino acids. The anchor positions of CDR-IMGT are in square (*see* Subheading 2.2). The hatched positions represent gaps according to the IMGT unique numbering for V domain [13]. AA written in red letters indicate the five conserved positions in V-DOMAIN (1st-CYS 23, CONSERVED-TRP 41, hydrophobic 89, 2nd-CYS 104 and J-TRP or J-PHE 118). CDR-IMGT are colored according to "IMGT CDR-IMGT color type" [46] of the corresponding locus and

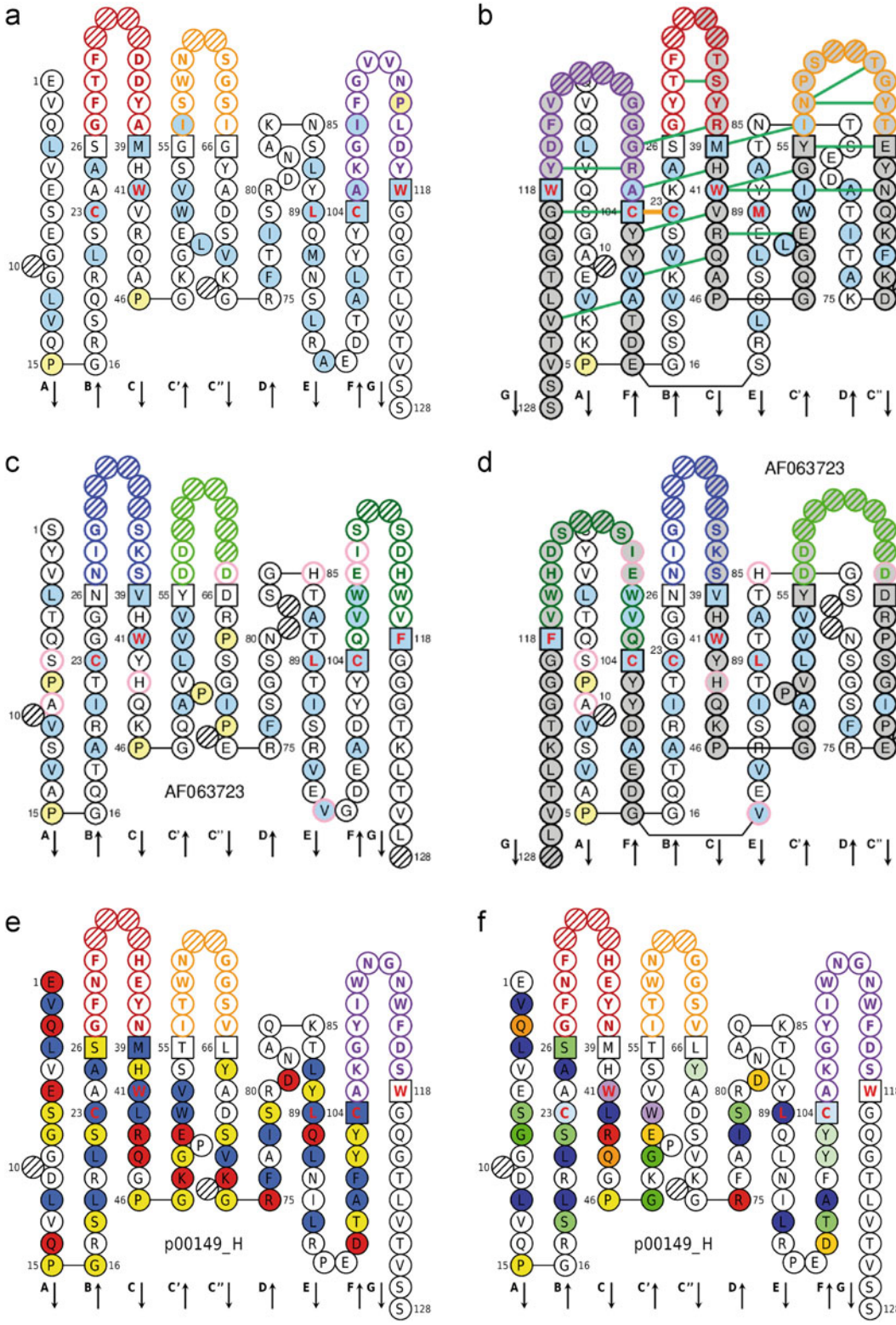


Fig. 21 IMGT Colliers de Perles for V-DOMAIN [18–21]. (**a**–**d**) Background color is "50% Hydrophobic positions," and Proline (P) is in yellow [46]. (**a**) IMGT Collier de Perles on one layer generated from the

FR-IMGT according to the "Background color" (*see* **Note 35**) selected by the user. The orientations of the nine beta-strands are indicated at the bottom of the IMGT/Collier-de-Perles. Illustrations of IMGT/Collier-de-Perles output are shown in Fig. 21.

8 Notes

1. The IMGT/V-QUEST reference directory sets [22, 23] are defined for species, which have been extensively studied, such as human, mouse, and dog, as well as for the species of which germline IG or TR repertoires are not fully available. The IMGT/V-QUEST results should be interpreted according to the available IMGT reference directories for a given species or taxon and a given locus. IMGT/V-QUEST reference directories have been also set for groups of taxons (e.g., Teleostei or Chondrichthyes), which contain pooled reference sequences from several species: these latter are not available in IMGT/HighV-QUEST.
2. Selecting "IG" or "TR" allows to submit sequences of different locus ("IGH," "IGK," and "IGL" for "IG" and "TRA," "TRB," "TRD," and "TRG" for "TR") in a same run (the locus will be automatically determined for each sequence), while the selection of a given locus forces the tool to compare the user sequences to the IMGT/V-QUEST reference directory set of the selected locus only.
3. The nucleotides "n" at 5′ and/or 3′ end of the submitted sequences are automatically trimmed before the analysis. The numbers of 5′ trimmed-n and 3′ trimmed-n are indicated in the results if any (*see* "Trimming of nucleotides "n" at 5' and/or 3' end of the submitted sequences" in the IMGT/V-QUEST [22, 23] Documentation).
4. The 14 sections are linked to the corresponding IMGT/V-QUEST [22, 23] Documentation in order to help the user in choosing of checking or unchecking them (see also Subheading 3.2.2). When "Uncheck all" is selected, only the "Result summary" is displayed.

Fig. 21 (continued) IMGT/V-QUEST [22, 23] analysis of a VH (nt) (accession number X81732 of IMGT/LIGM-DB [36]). (**b**) IMGT Collier de Perles on two layers of a VH with hydrogen bonds between the amino acids of the C, C′, C″, F, and G strands and those of the CDR-IMGT (daclizumab 3nfp_H, PDB code 3nfp of IMGT/3Dstructure-DB [30, 50, 51]). (**c**, **d**) IMGT Colliers de Perles with AA changes of a V-LAMBDA domain generated from the IMGT/DomainGapAlign [30, 31] analysis (translation of the AF063723 IMGT/LIGM-DB accession number), (**c**) on one layer, and (**d**) on two layers. (**e**, **f**) IMGT/Collier-de-Perles [21] results on one layer for the entry code p00149 from IMGT/2Dstructure-DB (**e**) with background color "IGH 80% hydrophopathy classes [46] and (**f**) with background color "IGH 80% physicochemical classes" [46]

5. The eight sections are linked to the corresponding IMGT/V-QUEST [22, 23] Documentation in order to help the user in choosing of checking or unchecking them. When "Uncheck all" is selected, only the "Summary table" is displayed.
6. The contents of the 11 or 12 text files are identical to those of the results provided by IMGT/HighV-QUEST [24–27], the high throughput version of IMGT/V-QUEST [22, 23].
7. The "AIRR formatted results" archive includes two text files: vquest_airr.tsv and Parameters.txt (IMGT/V-QUEST [22, 23] parameters used for the analysis). The vquest_airr.tsv contains the fields of the "Rearrangement Schema" provided by Adaptive Immune Receptor Repertoire (AIRR) Community [42, 43] plus additional IMGT fields (see http://www.imgt.org/IMGT_vquest/vquest_airr).
8. Including orphons [1–3] in the IMGT reference sets is relevant for genomic studies only.
9. In case of unmutated IG V-REGION (no mutations in FR1-IMGT, CDR1-IMGT, FR2-IMGT, CDR2-IMGT and FR3-IMGT), the number of accepted mutations is adjusted to 0 in 3'V-REGION and 5'J-REGION, and 2 in D-REGION for IGH, and 2 in 3'V-REGION and 5'J-REGION of IGK and IGL, in order to reflect the low probability of somatic hypermutations.
10. Both V-DOMAIN of a scFv must be in the same orientation [41]. In addition to the results for each V-DOMAIN individually, "Detailed view" includes a table for the identified scFv that links and localizes the two V-DOMAIN. In "Synthesis view," the two V-DOMAIN of a scFv are always displayed in consecutive rows of the "Summary table." In "Excel file," an additional 12th sheet provides one row per scFv with main characteristics and positions of the two V-DOMAIN.
11. Stereotyped sequences of Chronic Lymphocytic Leukemia (CLL) of subset #2 are characterized by a IGHV3-21/IGHJ6 rearrangement, a CDR3-IMGT of 9 AA with pattern "XX[D/E]XXXMDV" (X is for any AA, [D/E] means D or E). Sequences of subset #8 are characterized by an IGHV4-39/IGHJ5 rearrangement, an IGHV identity % is >98% and a CDR3-IMGT of 19 AA with a pattern "AXXXXXSSXWXXXXXWFDP". CLL patients whose malignant B clone carries a B-cell receptor with a heavy chain of subset #2 or subset #8 a re clinically associated with a poor prognosis [52, 53].
12. Four categories for sequence analysis are defined: (1) analysis without "Search for insertions and deletions in V-REGION," (2) analysis with "Search for insertions and deletions in V-REGION" and corrections if any, (3) analysis on complementary reverse sequence without "Search for insertions and

deletions in V-REGION," and (4) analysis on complementary reverse sequence with "Search for insertions and deletions in V-REGION" and corrections if any.

13. The score of the alignment for two sequences is calculated by counting +5 for each identical nt at a given position (match) and −4 for position with different nt (mismatch) [22, 23].
14. The JUNCTION decryption [45] for sequences with 1 D gene and allele provides lengths (in nt) of 3′V-REGION (3′V), D-REGION (D), and 5′J-REGION (5′J) (numbers between parentheses) of N1-REGION {N1} and N2-REGION {N2} (numbers between braces), and numbers between these regions indicate at the 3′ of the end of V, at 5′ or 3′ of D, and at the 5′ of the end of J, either trimmed nt (negative (−) values) or palindromic P nucleotides (positive (+) values) (trimmed or P nt are mutually exclusive) [45]. See IMGT/JunctionAnalysis [37, 38] Documentation (http://www.imgt.org/IMGT_jcta/decryption) [45] for sequences with 2 or 3 D genes and alleles.
15. Potential insertions or deletions are suspected by IMGT/V-QUEST [22, 23] when the V-REGION score is very low (less than 200), and/or the percentage of identity is less than 85%, and/or when the input sequence has different CDR1-IMGT and/or CDR2-IMGT lengths, compared with those of the closest germline V.
16. Several V or J genes and alleles with same highest identity percentage can be found generally: (1) if the sequence is partial in 5′ (for V) and/or in 3′ (for J), (2) if the numbers of mutations are identical (whatever their positions), (3) if reference sequences are identical (in case of duplicated genes or alleles), and (4) in case of polymorphism between different alleles in the germline CDR3-IMGT.
17. The algorithms for D gene and allele identification differ between IMGT/V-QUEST [22, 23] and IMGT/JunctionAnalysis [37, 38] and may provide different solutions. The results of IMGT/JunctionAnalysis are the most precise and are those reported in the "Summary of results." IMGT/V-QUEST results may be helpful to solve ambiguous cases and when IMGT/JunctionAnalysis does not provide results.
18. The 20 amino acids have been classified in 11 "IMGT physicochemical classes," which are based on "Hydrophathy," "Volume," and "Chemical" characteristics of the AA (http://www.imgt.org/, section 'Amino acids' in IMGT Education > Aide-mémoire) [46].
19. In case of differences due to the 5′ primer in V-REGION, it is possible to exclude a given number of nucleotides (IMGT/V-QUEST [22, 23] Search page, "Advanced parameters," Parameters for "Detailed view," and "Nb of nucleotides to exclude

in 5' of the V-REGION for the evaluation of the number of mutations") before launching the analysis.

20. Files from #2 to #6 also include six additional columns: the order of the sequence in file, the Sequence identifier, the V-DOMAIN Functionality, and the names of the V, D, and J genes and alleles. Files from #7 to #10 include four additional columns: the order of the sequence in file, the sequence identifier, the V-DOMAIN functionality, and the name of the V gene and allele.
21. The selection of several "completed" sequence analyses in the same time will combine them as a given batch. Only pertinent combinations or comparisons are allowed. For example, the selection of the sequence analyses from different species or receptor types is forbidden. The selected analyses must include the result format "CSV."
22. "1 copy" are unique sequences from which is built the list of IMGT clonotypes (AA) or (nt). "More than 1" are sequences which are fully identical to one of the "1 copy" set: they are taken into account for the evaluation of the number of sequences assigned to a given IMGT clonotype (AA) or (nt).
23. R is a language and environment for statistical computing and graphics available as free software and downloadable at the CRAN (Comprehensive R Archive Network) website (http://cran.r-project.org/) for Windows, Linux, or Macintosh operating systems. If R is already installed on your computer, please check that the R version is equal or higher to the one indicated in the IMGT/StatClonotype [28, 29] Documentation (http://www.imgt.org/StatClonotype/IMGTStatClonotypeDoc.html#pack).
24. Only alleles of genes having significant differences in proportions validated by all multiple testing procedures are analyzed [28, 29]. By displaying statistical test results per allele, in the case of individuals heterozygous for a given gene, it becomes possible to detect if significant differences in gene proportions, validated by all multiple testing procedures, depend on one allele or not.
25. Above the table, on the left, the number of displayed rows, five by default, can be modified [28, 29]. On the right "Search" allows to enter value in order to filter rows with one or several fields containing it. Clicking on the column title allows to sort the values (alphabetical or number order depending on the column type). Below the name of each column, a filter allows to select values for text fields (e.g., a gene name in column "Gene_Name") or a range for numerical values.

26. Normalized proportions for set 1 and for set 2 represent the numbers of IMGT clonotypes (AA) for a given gene obtained from the IMGT/HighV-QUEST [24–27] outputs normalized for 10,000 IMGT clonotypes (AA) (for clonotype diversity) or for 10,000 sequences assigned to IMGT clonotypes (AA) (for clonotype expression).
27. In addition to other parameters, the left panel allows: (1) the selection of the gene type (V, D, or J); (2) the addition of the locus type to graph axis title: IGH, IGK, IGL, TRA, TRB, TRG, or TRD; (3) the change of the bar colors for Normalized_proportion.set1 and Normalized_proportion.set2; (4) the addition of a title to the graphs for genes and alleles; and (5) the selection of the height and width of the graphs for genes and alleles [28, 29].
28. The CDR-IMGT lengths in the x-axis are not necessarily consecutive values: only CDR-IMGT lengths found in one or both of compared sets are displayed in the graphs.
29. IMGT/DomainGapAlign [30, 31] can analyze also amino acid sequences of C-DOMAIN of IG and TR [14], of V-LIKE-DOMAIN and C-LIKE-DOMAIN of the IgSF other than IG or TR [13, 14, 16, 17], of G-DOMAIN of major histocompatibility (MH) [15], and of G-LIKE-DOMAIN of MhSF other than MH [15–17].
30. In the context of humanization, IMGT/DomainGapAlign [30, 31] allows to precisely define the CDR1-IMGT, CDR2-IMGT, and CDR3-IMGT to be grafted and to select the most appropriate human FR-IMGT by alignment of V-DOMAIN amino acid sequence of the original species (mouse or other species) with the *Homo sapiens* V-REGION and J-REGION reference sets [32, 33].
31. The Smith-Waterman algorithm is used for local sequence alignments of the user AA sequences with the AA IMGT reference directories for V-REGION and J-REGION. The highest alignment scores correspond to the highest sequence similarities.
32. IMGT/Collier-de-Perles [21] provides also 2D graphical representations of C-DOMAIN of IG and TR [14], of V-LIKE-DOMAIN and C-LIKE-DOMAIN of the IgSF other than IG or TR [13, 14, 16, 17], of G-DOMAIN of major histocompatibility (MH), and of G-LIKE-DOMAIN of MhSF other than MH [15–17].
33. "IMGT Colliers de Perles" [18–21] are also provided in IMGT/3Dstructure-DB and IMGT/2Dstructure-DB database entries [30, 50, 51]: the hydrogen bonds within a V-DOMAIN, determined from experimental structural data, are shown as green lines in generated IMGT Collier de Perles on two layers.

34. The number of layers "2" allows to display the two sheets of beta strands of a V-DOMAIN.
35. The background color by default "50% Hydrophobic positions" displays in blue, the positions that have an hydrophobic amino acid (hydropathy index with positive value) or a tryptophan (W) in 50% or more of analyzed V domains [46]. Other background colors have been set for each IGH, IGK, and IGL AA sequences showing the positions, which belong to the same hydropathy classes, volume classes, or physicochemical classes in 80% or more of the analyzed V-DOMAIN.

Acknowledgements

We are very grateful to Gérard Lefranc, founder of the Laboratoire d'ImmunoGénétique Moléculaire LIGM (Université de Montpellier and CNRS), for his unique contribution in the creation of IMGT® in 1989 and his unwavering support for these 30 years. We thank all members of the IMGT® team for their expertise and constant motivation. IMGT® was funded in part by the BIOMED1 (BIOCT930038), 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). IMGT® received financial support from the GIS IBiSA, the Agence Nationale de la Recherche (ANR) Labex MabImprove (ANR-10-LABX-53-01), the Région Occitanie Languedoc-Roussillon (Grand Plateau Technique pour la Recherche (GPTR), and BioCampus Montpellier. IMGT® is currently supported by the Centre National de la Recherche Scientifique (CNRS), the Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI), the University of Montpellier, and the French Infrastructure Institut Français de Bioinformatique (IFB) ANR-11-INBS-0013. IMGT® is a registered trademark of CNRS. IMGT® is member of the International Medical Informatics Association (IMIA) and a member of the Global Alliance for Genomics and Health (GA4GH). This work was granted access to the High Performance Computing (HPC) resources of Meso@LR and of Centre Informatique National de l'Enseignement Supérieur (CINES), to Très Grand Centre de Calcul (TGCC) of the Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) and to Institut du développement et des ressources en informatique scientifique (IDRIS) under the allocation 036029 (2010-2022) made by GENCI (Grand Equipement National de Calcul Intensif).

References

1. Lefranc M-P, Lefranc G (2001) The Immunoglobulin FactsBook. Academic Press, London, UK
2. Lefranc M-P, Lefranc G (2020) Immunoglobulins or antibodies: IMGT® bridging genes, structures and functions. Biomedicines 8(9): 319 https://doi.org/10.3390/biomedicines8090319
3. Lefranc M-P, Lefranc G (2001) The T cell receptor FactsBook. Academic Press, London, UK
4. Lefranc M-P (2014) Immunoglobulin and T cell receptor genes: IMGT(®) and the birth and rise of immunoinformatics. Front Immunol 5:22. https://doi.org/10.3389/fimmu.2014.00022
5. Lefranc M-P, Giudicelli V, Duroux P et al (2015) IMGT®, the international ImMunoGeneTics information system® 25 years on. Nucleic Acids Res 43:D413–D422. https://doi.org/10.1093/nar/gku1056
6. Giudicelli V, Lefranc M-P (2012) IMGT-ONTOLOGY 2012. Front Genet 3:79. https://doi.org/10.3389/fgene.2012.00079
7. Duroux P, Kaas Q, Brochet X et al (2008) IMGT-Kaleidoscope, the formal IMGT-ONTOLOGY paradigm. Biochimie 90: 570–583. https://doi.org/10.1016/j.biochi.2007.09.003
8. Lefranc M-P (2000) Nomenclature of the human immunoglobulin genes. In: Coligan JE, Bierer BE, Margulies DE, Shevach EM, Strober W (eds) Current Protocols in Immunology. John Wiley and Sons, Hoboken N.J, pp A.1P.1–A.1P.37
9. Lefranc M-P (2000) Nomenclature of the human T cell receptor genes. In: Coligan JE, Bierer BE, Margulies DE, Shevach EM, Strober W (eds) Current Protocols in Immunology. John Wiley and Sons, Hoboken N.J, pp A.1O.1–A.1O.23
10. Lefranc M-P (2007) WHO-IUIS Nomenclature Subcommittee for immunoglobulins and T cell receptors report. Immunogenetics 59: 899–902. https://doi.org/10.1007/s00251-007-0260-4
11. Lefranc M-P (2008) WHO-IUIS Nomenclature Subcommittee for immunoglobulins and T cell receptors report August 2007, 13th International Congress of Immunology, Rio de Janeiro, Brazil. Dev Comp Immunol 32: 461–463. https://doi.org/10.1016/j.dci.2007.09.008
12. Lefranc M-P (2011) From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT standardized gene and allele nomenclature: for immunoglobulins (IG) and T cell receptors (TR). Cold Spring Harb Protoc 2011:627–632. https://doi.org/10.1101/pdb.ip84
13. Lefranc M-P, Pommié C, Ruiz M et al (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev Comp Immunol 27:55–77
14. Lefranc M-P, Pommié C, Kaas Q et al (2005) IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. Dev Comp Immunol 29:185–203. https://doi.org/10.1016/j.dci.2004.07.003
15. Lefranc M-P, Duprat E, Kaas Q, Tranne M, Thiriot A, Lefranc G (2005) IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. Dev Comp Immunol 29: 917–938. https://doi.org/10.1016/j.dci.2005.03.003
16. Lefranc M-P (2011) IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. Cold Spring Harb Protoc 2011:633–642. https://doi.org/10.1101/pdb.ip85
17. Lefranc M-P (2014) Immunoinformatics of the V, C, and G domains: IMGT® definitive system for IG, TR and IgSF, MH, and MhSF. Methods Mol Biol 1184:59–107. https://doi.org/10.1007/978-1-4939-1115-8_4
18. Ruiz M, Lefranc M-P (2002) IMGT gene identification and colliers de Perles of human immunoglobulins with known 3D structures. Immunogenetics 53:857–883. https://doi.org/10.1007/s00251-001-0408-6
19. Kaas Q, Lefranc M-P (2007) IMGT Colliers de Perles: standardized sequence-structure representations of the IgSF and MhcSF superfamily domains. Curr Bioinforma 2:21–30
20. Kaas Q, Ehrenmann F, Lefranc M-P (2007) IG, TR and IgSF, MHC and MhcSF: what do we learn from the IMGT Colliers de Perles? Brief Funct Genomic Proteomic 6:253–264. https://doi.org/10.1093/bfgp/elm032
21. Ehrenmann F, Giudicelli V, Duroux P, Lefranc M-P (2011) IMGT/collier de Perles: IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains). Cold Spring

Harb Protoc 2011:726–736. https://doi.org/10.1101/pdb.prot5635

22. Brochet X, Lefranc M-P, Giudicelli V (2008) IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. Nucleic Acids Res 36:W503–W508. https://doi.org/10.1093/nar/gkn316
23. Giudicelli V, Brochet X, Lefranc M-P (2011) IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. Cold Spring Harb Protoc 2011:695–715. https://doi.org/10.1101/pdb.prot5633
24. Alamyar E, Giudicelli V, Shuo L et al (2012) IMGT/HighV-QUEST: the IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. Immun Res 8(1):3
25. Alamyar E, Duroux P, Lefranc M-P, Giudicelli V (2012) 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 882:569–604. https://doi.org/10.1007/978-1-61779-842-9_32
26. Li S, Lefranc M-P, Miles JJ et al (2013) IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. Nat Commun 4:2333. https://doi.org/10.1038/ncomms3333
27. Giudicelli V, Duroux P, Lavoie A, Aouinti S, Lefranc M-P, Kossida S (2015) From IMGT-ONTOLOGY to IMGT/HighVQUEST for NGS immunoglobulin (IG) and T cell receptor (TR) repertoires in autoimmune and infectious diseases. Autoimmune Infect Dis 1(1). https://doi.org/10.16966/2470-1025.103
28. Aouinti S, Malouche D, Giudicelli V, Kossida S, Lefranc MP (2015) 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 10(11): e0142353. https://doi.org/10.1371/journal.pone.0142353
29. Aouinti S, Giudicelli V, Duroux P, Malouche D, Kossida S, Lefranc MP (2016) IMGT/StatClonotype for pairwise evaluation and visualization of NGS IG and TR IMGT clonotype (AA) diversity or expression from IMGT/HighV-QUEST. Front Immunol 7: 339. https://doi.org/10.3389/fimmu.2016.00339
30. Ehrenmann F, Kaas Q, Lefranc M-P (2010) IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. Nucleic Acids Res 38: D301–D307. https://doi.org/10.1093/nar/gkp946
31. Ehrenmann F, Lefranc M-P (2011) IMGT/DomainGapAlign: IMGT standardized analysis of amino acid sequences of variable, constant, and groove domains (IG, TR, MH, IgSF, MhSF). Cold Spring Harb Protoc 2011: 737–749. https://doi.org/10.1101/pdb.prot5636
32. Lefranc M-P, Ehrenmann F, Ginestoux C, Giudicelli V, Duroux P (2012) Use of IMGT (®) databases and tools for antibody engineering and humanization. Methods Mol Biol 907: 3–37. https://doi.org/10.1007/978-1-61779-974-7_1
33. Lefranc M-P (2014) IMGT® immunoglobulin repertoire analysis and antibody humanization. In: Alt F, Honjo T, Radbruch A, Reth M (eds) Molecular Biology of B Cells, vol 27, 2nd edn. Elsevier Ltd., London, UK, pp 481–514
34. Giudicelli V, Chaume D, Lefranc M-P (2005) IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res 33: D256–D261. https://doi.org/10.1093/nar/gki010
35. Lefranc M-P, Lefranc G (2019) IMGT® and 30 years of immunoinformatics insight in antibody V and C domain structure and function. Antibodies 8:29. https://doi.org/10.3390/antib8020029
36. Giudicelli V, Duroux P, Ginestoux C et al (2006) IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. Nucleic Acids Res 34:D781–D784. https://doi.org/10.1093/nar/gkj088
37. Yousfi Monod M, Giudicelli V, Chaume D, Lefranc M-P (2004) IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. Bioinformatics 20(Suppl 1):i379–i385. https://doi.org/10.1093/bioinformatics/bth945
38. Giudicelli V, Lefranc M-P (2011) IMGT/JunctionAnalysis: IMGT standardized analysis of the V-J and V-D-J junctions of the rearranged immunoglobulins (IG) and T cell receptors (TR). Cold Spring Harb Protoc 2011:716–725. https://doi.org/10.1101/pdb.prot5634

39. Giudicelli V, Protat C, Lefranc M-P (2003) The IMGT strategy for the automatic annotation of IG and TR cDNA sequences: IMGT/Automat. In: Proceedings of the European Conference on Computational Biology (ECCB 2003). INRIA (DISC/Spid), Paris, DKB-31, pp 103–104
40. Giudicelli V, Chaume D, Jabado-Michaloud J, Lefranc M-P (2005) Immunogenetics sequence annotation: the strategy of IMGT based on IMGT-ONTOLOGY. Stud Health Technol Inform 116:3–8
41. Giudicelli V, Duroux P, Kossida S, Lefranc M-P (2017) IG and TR single chain fragment variable (scFv) sequence analysis: a new advanced functionality of IMGT/V-QUEST and IMGT/HighV-QUEST. BMC Immunol 18:35. https://doi.org/10.1186/s12865-017-0218-8
42. Rubelt F, Busse CE, Bukhari SAC et al (2017) Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. Nat Immunol 18:1274–1278. https://doi.org/10.1038/ni.3873
43. Vander Heiden JA, Marquez S, Marthandan N et al (2018) AIRR Community Standardized Representations for Annotated Immune Repertoires. Front Immunol 9:2206. https://doi.org/10.3389/fimmu.2018.02206
44. Belessi CJ, Davi FB, Stamatopoulos KE et al (2006) IGHV gene insertions and deletions in chronic lymphocytic leukemia: "CLL-biased" deletions in a subset of cases with stereotyped receptors. Eur J Immunol 36:1963–1974. https://doi.org/10.1002/eji.200535751
45. Rollin M, Giudicelli V, Lefranc M-P IMGT/JunctionAnalysis: IMGT JUNCTION decryption values for (3′V)3′{N}[5′(D)3′{N}]5′(5′J). http://www.imgt.org/IMGT_jcta/decryption. Accessed 14 Jan 2022
46. Pommié C, Levadoux S, Sabatier R, Lefranc M-P (2004) IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. J Mol Recognit 17:17–32. https://doi.org/10.1002/jmr.647
47. Elemento O, Lefranc M-P (2003) IMGT/PhyloGene: an on-line tool for comparative analysis of immunoglobulin and T cell receptor genes. Dev Comp Immunol 27:763–779
48. Hemadou A, Giudicelli V, Smith ML et al (2017) Pacific Biosciences Sequencing and IMGT/HighV-QUEST analysis of full-length single chain Fragment variable from an in vivo selected phage-display combinatorial library. Front Immunol 8:1796. https://doi.org/10.3389/fimmu.2017.01796
49. Han SY, Antoine A, Howard D et al (2018) Coupling of single molecule, long read sequencing with IMGT/HighV-QUEST analysis expedites identification of SIV gp140-specific antibodies from scFv phage display libraries. Front Immunol 9:329. https://doi.org/10.3389/fimmu.2018.00329
50. Kaas Q, Ruiz M, Lefranc M-P (2004) IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. Nucleic Acids Res 32:D208–D210. https://doi.org/10.1093/nar/gkh042
51. Ehrenmann F, Lefranc M-P (2011) IMGT/3Dstructure-DB: querying the IMGT database for 3D structures in immunology and immunoinformatics (IG or antibodies, TR, MH, RPI, and FPIA). Cold Spring Harb Protoc 2011:750–761. https://doi.org/10.1101/pdb.prot5637
52. Agathangelidis A, Darzentas N, Hadzidimitriou A et al (2012) Stereotyped B-cell receptors in one-third of chronic lymphocytic leukemia: a molecular classification with implications for targeted therapies. Blood 119:4467–4475. https://doi.org/10.1182/blood-2011-11-393694
53. Agathangelidis A, Chatzidimitriou A, Gemenetzi K et al (2020) Higher-order connections between stereotyped subsets: implications for improved patient classification in CLL. Blood 137(10):1365–1376. https://doi.org/10.1182/blood.2020007039