Autoimmune and Infectious Diseases: Open Access

From IMGT-ONTOLOGY to IMGT/HighV-QUEST for NGS Immunoglobulin (IG) and T cell Receptor (TR) Repertoires in Autoimmune and Infectious Diseases

Véronique Giudicelli, Patrice Duroux, Arthur Lavoie, Safa Aouinti, Marie-Paule Lefranc*, and Sofia Kossida*

IMGT®, the international ImMunoGenetics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Institut de Génétique Humaine IGH, UPR CNRS 1142, Montpellier University, Montpellier, France

***Corresponding authors:** Marie-Paule Lefranc and Sofia Kossida, IMGT®, the international ImMunoGenetics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Institut de Génétique Humaine IGH, UPR CNRS 1142, Montpellier University, 141 rue de la Cardonille, 34396 Montpellier cedex 5; Tel.: 33-434-359965; Fax: 33-434-359901; E-mail: Marie-Paule.Lefranc@igh.cnrs.fr, Sofia.Kossida@igh.cnrs.fr

Abstract

The analysis of adaptive immune repertoires in autoimmune and infectious diseases represents a fascinating challenge because of the natural huge diversity of the antigen receptors that are the immunoglobulins or antibodies and T cell receptors, expressed by B cells and T cells, respectively. The high throughput analysis resulting from the next generation sequencing technologies offers an invaluable large scale characterization of the expressed repertoires and, consequently, a better understanding of the protective and pathogenic immune responses. IMGT®, the international ImMunoGeneTics information system® (Centre National de la Recherche Scientifique and Montpellier University) is the global reference in immunogenetics and immunoinformatics. By its creation in 1989, IMGT® marked the advent of immunoinformatics, which emerged at the interface between immunogenetics and bioinformatics. IMGT® is specialized in the immunoglobulins or antibodies, T cell receptors, major histocompatibility proteins of humans and other vertebrate species, and in the proteins of the immunoglobulin superfamily and major histocompatibility superfamily of any species. IMGT® has been built on the IMGT-ONTOLOGY axioms and concepts, which bridged the gap between genes, sequences and three-dimensional structures. The concepts include the IMGT® standardized keywords (identification), IMGT® standardized labels (description), IMGT® standardized nomenclature (classification), IMGT unique numbering and IMGT Colliers de Perles (numerotation). IMGT® comprises seven databases, 15,000 pages of web resources and 17 tools, and includes data from fishes to humans, for basic to veterinary, medical and translational research. IMGT/HighV-QUEST, the high throughput version of IMGT/V-QUEST and IMGT/Junction Analysis, created in 2010, is the first and, so far, the only online portal for analysis of the immunoglobulin and T cell receptor repertoires obtained from next generation sequencing. High-quality results contribute to the large scale vision of antigen receptor repertoires and understanding of the adaptive immune responses (protective in vaccination, cancer and infections and pathogenic in autoimmunity) for diagnosis, prognosis and therapy.

Keywords: IMGT; Immunogenetics; Immunoinformatics; IMGT-ONTOLOGY; Immunoglobulin (IG); T cell receptor (TR); Immune repertoire; IMGT unique numbering; Next generation sequencing

IMGT®: the Birth of Immunoinformatics

IMGT®, the international ImMunoGeneTics information system® [1,2], was created in 1989 by Marie-Paule Lefranc at Montpellier, France (CNRS and Montpellier University). The founding of IMGT® marked the advent of immunoinformatics, a new science, which emerged at the interface between immunogenetics and bioinformatics. For the first time, immunoglobulin (IG) or antibody and T cell receptor (TR) variable (V), diversity (D), joining (J) and constant (C) genes were officially recognized as "genes" as well as the conventional genes [3-6]. This major breakthrough allowed genes and data of the complex and highly diversified adaptive immune responses to be managed in genomic databases and tools.

The adaptive immune response was acquired by jawed vertebrates (or *gnathostomata*) more than 450 million years ago and is found in all extant jawed vertebrate species from fishes to humans. It is characterized by a remarkable immune specificity and memory, which are properties of the B and T cells owing to an extreme diversity of their antigen receptors. The specific antigen receptors comprise the IG or antibodies of the B cells and plasmacytes [3], and the TR [4]. The IG recognizes antigens in their native (unprocessed) form, whereas the TR recognizes processed protein antigens, which are presented as peptides by the highly polymorphic major histocompatibility (MH, in humans HLA for human leucocyte antigens) proteins.

The potential antigen receptor repertoire of each individual is estimated to comprise about 2×10^{12} different IG and TR, and the limiting factor is only the number of B and T cells that an organism is genetically programmed to produce [3,4]. This huge diversity results from the complex molecular synthesis of the IG and TR chains and, more particularly, of their variable domains (V-DOMAIN) which, at their N-terminal end, recognize and bind the antigens [3,4]. The IG and TR synthesis includes several unique mechanisms that occur at the DNA level: combinatorial rearrangements of the V, D and J genes that code the V-DOMAIN (the V-(D)-J being spliced to the C gene that encodes the C-REGION in the transcript); exonuclease trimming at the ends of the V, D and J genes and random addition of nucleotides by the DNA nucleotidylexotransferase (DNTT, terminal deoxynucleotidyl transferase, TdT) that creates the junctional N-diversity regions, and later during B cell differentiation, for the IG, somatic hypermutations, gene conversion (e.g., in birds), and class or subclass switch in higher vertebrates [3,4].

IMGT® manages the diversity and complexity of the IG and TR and the polymorphism of the MH of humans and other vertebrates. IMGT® is also specialized in the other proteins of the immunoglobulin superfamily (IgSF) and MH superfamily (MhSF) and related proteins of the immune system (RPI) of vertebrates and invertebrates [2]. IMGT® provides a common access to standardized data from genome, proteome, genetics, and two-dimensional (2D) and three-dimensional (3D) structures. IMGT® is the acknowledged high-quality integrated knowledge resource in immunogenetics for exploring immune functional genomics. IMGT® comprises seven databases (for sequences, genes, and 3D structures) [7-12] and 17 online tools [13-28], as well as more than 15,000 pages of web resources (e.g., IMGT Scientific chart, IMGT Repertoire, IMGT Education > Aide-mémoire [29], the IMGT Medical page, the IMGT Veterinary page, the IMGT Biotechnology page, the IMGT Immunoinformatics page) [2]. IMGT® is the global reference in immunogenetics and immunoinformatics [30-45]. Its standards have been endorsed by the World Health Organization-International Union of Immunological Societies (WHO-IUIS) Nomenclature Committee since 1995 (first IMGT® online access at the 9th International Congress of Immunology, San Francisco, CA, USA) [46,47] and the WHO International Nonproprietary Names (INN) Programme [48,49].

The accuracy and the consistency of the IMGT® data are based on IMGT-ONTOLOGY [50-52], the first, and so far unique, ontology for immunogenetics and immunoinformatics [50-69]. IMGT-ONTOLOGY manages the immunogenetics knowledge through diverse facets that rely on seven axioms: IDENTIFICATION, DESCRIPTION, CLASSIFICATION, NUMEROTATION, LOCALIZATION, ORIENTATION, and OBTENTION [51,52,56]. The concepts generated from these axioms led to the elaboration of the IMGT® standards that constitute the IMGT Scientific chart: e.g., IMGT® standardized keywords (IDENTIFICATION) [57], IMGT® standardized labels (DESCRIPTION) [58], IMGT® standardized gene and allele nomenclature (CLASSIFICATION) [59], IMGT unique numbering [60-65] and its standardized graphical 2D representation or IMGT Colliers de Perles [66-69] (NUMEROTATION).

With a focus on IG and TR immune repertoires, we first review the fundamental information generated from these IMGT-ONTOLOGY concepts which led to the IMGT Scientific chart rules, and then IMGT/HighV-QUEST [18,23,24], the online web portal for the analysis of rearranged IG and TR repertoire nucleotide sequences from next generation sequencing (NGS). IMGT/HighV-QUEST is the high throughput version of IMGT/V-QUEST [13-18] with the results of the integrated IMGT/JunctionAnalysis [19,20] and of IMGT/Automat [21,22]. IMGT/HighV-QUEST runs against IMGT reference directories built from sequences annotated in IMGT/LIGM-DB [7], the IMGT® nucleotide database (176,948 sequences from 347 species in July 2015), and from IMGT/GENE-DB [8], the IMGT® gene database (3567 genes and 5258 alleles from 22 species, of which there were 715 genes and 1469 alleles for *Homo sapiens* and 869 genes and 1319 alleles for *Mus musculus* in July 2015).

The unified IMGT® approach is of major interest for bridging knowledge from IG and TR repertoires in physiological and pathological situations [70-75], for guiding antibody discovery, humanization and engineering [76-86], and providing the same standardized results for NGS repertoires in autoimmune and infectious diseases.

Fundamental IMGT-ONTOLOGY Concepts for NGS

IDENTIFICATION: IMGT® Standardized Keywords

More than 325 IMGT® standardized keywords (189 for sequences and 137 for 3D structures) were precisely defined [57]. They represent the controlled vocabulary assigned during the annotation process and allow standardized search criteria for querying the IMGT® databases and for the extraction of sequences and 3D structures. They have been entered in BioPortal [87] at the National Center for Biomedical Ontology (NCBO) in 2010.

Standardized keywords are assigned at each step of the molecular synthesis of an IG. Those assigned to a nucleotide sequence are found in the "DE" (definition) and "KW" (keyword) lines of the IMGT/LIGM-DB files [7]. They characterize, for instance, the gene type, the configuration type and the functionality type [57]. There are six gene types: variable (V), diversity (D), joining (J), constant (C), conventional-with-leader, and conventional-without-leader. Four of them (V, D, J, and C) identify the IG and TR genes and are specific to immunogenetics. There are four configuration types: germline (for the V, D, and J genes before DNA rearrangement), rearranged (for the V, D, and J genes after DNA rearrangement), partially-rearranged (for D gene after only one DNA rearrangement) and undefined (for the C gene and for the conventional genes, which do not rearrange). The functionality type depends on the gene configuration. The functionality type of genes in germline or undefined configuration is functional (F), open reading frame (ORF) or pseudogene (P). The functionality type of genes in rearranged or partially-rearranged configuration is either productive (no stop codon in the V-(D)-J region and in-frame junction) or unproductive (stop codon (s) in the V-(D)-J region, and/or out-of-frame junction).

The 20 usual amino acids (AA) have been classified in 11 IMGT physicochemical classes (IMGT®[1], IMGT Education > Aide-mémoire > Amino acids). The amino acid changes are described according to the hydropathy (three classes), volume (five classes) and IMGT physicochemical classes (11 classes) [29]. For example Q1 > E (+ + −) means that in the amino acid change (Q > E), the two amino acids at codon 1 belong to the same hydropathy (+) and volume (+) classes but to different IMGT physicochemical properties (−) classes [29]. Four types of AA changes are identified in IMGT®: very similar (+ + +), similar (+ + −, + − +), dissimilar (− − +, − + −, + − −), and very dissimilar (− − −).

DESCRIPTION: IMGT® Standardized Labels

More than 560 IMGT® standardized labels (277 for sequences and 285 for 3D structures) were precisely defined [58]. They are written in capital letters (no plural) to be recognizable without creating new terms. Standardized labels assigned to the description of sequences are found in the "FT" (feature) lines of the IMGT/LIGM-DB files [7]. Querying these labels represent a big plus compared to the generalist databases (GenBank/European Nucleotide Archive (ENA)/DNA Data Bank of Japan (DDBJ)). Thus it is possible to query for the "CDR3-IMGT" of the human rearranged productive sequences of IG-Heavy-Gamma (e.g., 1828 CDR3-IMGT obtained, with their sequences at the nucleotide or amino acid level). The core labels include V-REGION, D-REGION, J-REGION, and C-REGION which correspond to the coding region of the V, D, J and C genes. IMGT structure labels for IG chains and domains allow describing the chain and domains of 3D structures (see for example http://www.imgt.org/3Dstructure-DB/cgi/details.cgi?pdbcode=1HZH). Correspondence between human IG structure labels and sequence labels, links and correspondence between standardized keywords for IDENTIFICATION and standardized labels for DESCRIPTION in IMGT/PROTEIN-DB and IMGT/3Dstructure-DB are detailed in IMGT Scientific chart (see http://www.imgt.org/IMGTScientificChart/SequenceDescription/IMGT3Dkeywords.html). These labels are necessary for a standardized description of the IG sequences and structures in databases and tools [58].

Highly conserved amino acids at a given position in a domain have IMGT labels [58]. Thus three amino acid labels are common to the V and C domains: 1st-CYS (cysteine C at position 23), CONSERVED-TRP (tryptophan W at position 41), and 2nd-CYS (C at position 104) [60-63,65]. Two other labels are characteristic of the V-DOMAIN and correspond to the first amino acid of the canonical F/W-G-X-G motif

(where F is phenylalanine, W tryptophan, G glycine, and X any amino acid) encoded by the J-REGION: J-PHE or J-TRP (F or W at position 118) [60-62,65].

CLASSIFICATION: IMGT® Standardized Genes and Alleles

The IMGT-ONTOLOGY CLASSIFICATION axiom was the trigger of immunoinformatics birth [45]. Indeed the IMGT® concepts of classification allowed us, for the first time, to classify the antigen receptor genes (IG and TR) for any locus (e.g., immunoglobulin heavy (IGH), T cell receptor alpha (TRA)), for any gene configuration (germline, undefined, or rearranged) and for any species (from fishes to humans). In higher vertebrates, there are three IG major loci (other loci correspond to chromosomal orphon sets, genes of which are orphons, not used in the IG chain synthesis). The IG major loci include the immunoglobulin heavy (IGH), and for the light chains, the immunoglobulin kappa (IGK) and the immunoglobulin lambda (IGL) in higher vertebrates and the immunoglobulin iota (IGI) in fishes (IMGT® [1], IMGT Repertoire).

Since the creation of IMGT® in 1989, at New Haven during the 10th Human Genome Mapping Workshop (HGM10), the standardized classification and nomenclature of the IG and TR of humans and other vertebrate species have been under the responsibility of the IMGT Nomenclature Committee (IMGT-NC).

IMGT® gene and allele names are based on the concepts of classification of "Group", "Subgroup", "Gene" and "Allele" [59]. "Group" allows classification of a set of genes that belong to the same multigene family, within the same species or between different species. For example, there are 10 groups for the IG of higher vertebrates: IGHV, IGHD, IGHJ, IGHC, IGKV, IGKJ, IGKC, IGLV, IGLJ, and IGLC. "Subgroup" allows classification of a subset of genes that belong to the same group and that, in a given species, share at least 75% identity at the nucleotide level, e.g., *Homo sapiens* IGHV1 subgroup. Subgroups, genes, and alleles are always associated to a species name. An allele is a polymorphic variant of a gene, which is characterized by the mutations of its sequence at the nucleotide level, identified in its core sequence, and compared to the gene allele reference sequence, designated as allele *01. For example, *Homo sapiens* IGHV1-2*01 is the allele *01 of the *Homo sapiens* IGHV1-2 gene that belongs to the *Homo sapiens* IGHV1 subgroup which itself belongs to the IGHV group. For the IGH locus, the constant genes are designated by the letter (and eventually number) corresponding to the encoded isotypes (IGHM, IGHD, IGHG3…), instead of using the letter C. IG and TR genes and alleles are in capital letters whatever the species and are not italicized in publications. IMGT-ONTOLOGY concepts of classification have been entered in the NCBO BioPortal [87].

The IMGT® IG and TR gene names [3-6] were approved by the Human Genome Organization (HUGO) Nomenclature Committee (HGNC) in 1999 [88,89] and were endorsed by the WHO-IUIS Nomenclature Subcommittee for IG and TR [46,47]. The IMGT® IG and TR gene names are the official international reference and, as such, have been entered in IMGT/GENE-DB [8], in the Genome Database (GDB) [90], in LocusLink at the National Center for Biotechnology Information (NCBI) USA [91], in Entrez Gene (NCBI) when this database (now designated as "Gene") superseded LocusLink [92], in NCBI MapViewer, in Ensembl at the European Bioinformatics Institute (EBI) [93], and in the Vertebrate Genome Annotation (Vega) Browser [94] at the Wellcome Trust Sanger Institute (UK). HGNC, Gene NCBI, Ensembl, and Vega have direct links to IMGT/GENE-DB [8]. IMGT® human IG and TR genes were also integrated in IMGT-ONTOLOGY on the NCBO BioPortal and, on the same site, in the HUGO ontology and in the National Cancer Institute (NCI) Metathesaurus. The *Mus musculus* IG and TR genes were provided in 2005 to the Mouse Genome Database (MGD) [95], following the IMGT® presentation of the 7 mouse IG and TR loci at the 19th International

Mouse Genome Conference (IMGC) in Strasbourg, France. Amino acid sequences of human IG and TR constant genes (e.g., *Homo sapiens* IGHM, IGHG1, IGHG2) were provided to UniProt [96] in 2008. Since 2007, IMGT® IG gene and allele names have been used for the description of the therapeutic monoclonal antibodies (-mab suffix) and fusion proteins for immune applications (FPIA) (-cept suffix) of the WHO-INN programme [48,49].

NUMEROTATION: IMGT unique numbering and IMGT Collier de Perles

The IMGT-ONTOLOGY NUMEROTATION axiom is acknowledged as the "IMGT® Rosetta stone" that has bridged the biological and computational spheres in bioinformatics [38]. The IMGT® concepts of numerotation comprise the IMGT unique numbering [60-65] and its graphical 2D representation, the IMGT Collier de Perles [66-69]. Developed for and by the "domain", these concepts integrate sequences, structures, and interactions into a standardized domain-centric knowledge for functional genomics. The IMGT unique numbering has been defined for the variable V domain (V-DOMAIN of the IG and TR, and V-LIKE-DOMAIN of IgSF other than IG and TR) [60-62], the constant C domain (C-DOMAIN of the IG and TR, and C-LIKE-DOMAIN of IgSF other than IG and TR) [63] and the groove G domain (G-DOMAIN of the MH, and G-LIKE-DOMAIN of MhSF other than MH) [64]. Thus the IMGT unique numbering and IMGT Collier de Perles provide a definitive and universal system across species, including invertebrates, for the sequences and structures of the V, C and G domains of IG, TR, MH, IgSF and MhSF [65,69,84].

IMGT/HighV-QUEST reference Web portal for NGS

IMGT/HighV-QUEST overview

IMGT/HighV-QUEST [23], created in October 2010, is the high throughput version of IMGT/V-QUEST. It is so far the only online tool available on the Web for the direct analysis of complete IG and TR V-DOMAIN nucleotide sequences from NGS. IMGT/HighV-QUEST analyzes up to 500,000 sequences per run and performs statistical analysis on the results [23,24], with the same degree of resolution and high-quality results as IMGT/V-QUEST [13-18,70] and IMGT/JunctionAnalysis [19,20]. Indeed IMGT/HighV-QUEST uses the same algorithm and runs against the same IMGT reference directory. In July 2015, more than 6.62 billions of sequences were analyzed by IMGT/HighV-QUEST [18,23,24], by 1157 users from 41 countries (44% users from USA, 37% from EU, 19% from the remaining world).

IMGT/HighV-QUEST functionalities and results

IMGT/HighV-QUEST functionalities are basically the same as IMGT/V-QUEST [13-18], the IMGT® online tool for the analysis of nucleotide sequences of the IG and TR V-DOMAIN (Table 1). IMGT/HighV-QUEST numbers the user sequences according to the IMGT unique numbering [62] and introduces gaps accordingly. It identifies the variable (V), diversity (D) and junction (J) genes in rearranged IG and TR sequences and, for the IG, characterizes the nucleotide (nt) mutations and amino acid (AA) changes resulting from somatic hypermutations [22,24] by comparison with the IMGT/V-QUEST reference directory. The tool integrates IMGT/JunctionAnalysis [19,20] for the detailed characterization of the V-D-J or V-J junctions, IMGT/Automat [21,22] for a complete sequence annotation with the delimitation of the IMGT labels of description. By default IMGT/HighV-QUEST identifies the insertions/deletions (indels) which are NGS errors resulting from homopolymer hybridization and corrects them [24]. Statistical analysis and characterization of the IMGT clonotypes (AA) [24] are functionalities specific to IMGT/HighV-QUEST (Table 1).

IMGT/HighV-QUEST reference directory

The IMGT/HighV-QUEST uses the same reference directory as IMGT/V-QUEST [2]. It is organized per species and per locus and comprises the IMGT reference sequences. It includes the germline V, D and J nucleotide sequences of the core regions (V-REGION, D-REGION, and J-REGION) from all functional (F) genes and alleles, all open reading frame (ORF) and all in-frame pseudogenes (P) alleles from IMGT/GENE-DB [8]. It corresponds to the "F+ORF+ in-frame P" set, selected by default in "Advanced parameters". The other options allow to exclude pseudogenes from the reference directory ("F+ORF") or, for genomic analyses for example, to include orphon genes localized outside the main locus ("F+ORF including orphons" or "F+ORF+ in-frame P including orphons").

By definition, the IMGT/V-QUEST reference directory contains one reference sequence for each allele. By default, the user sequences are compared with all alleles of the genes of the selected set. However, the "Advanced parameters" option "With allele *01 only" is useful if the user sequences need to be compared with different genes; and/or if the user sequences that use the same gene need to be aligned together (independently of the allelic polymorphism).

The IMGT/V-QUEST reference directories have been set up for species which are extensively studied, such as human and mouse for which a definitive gene nomenclature has been defined. This also holds true for the other species or taxons with incomplete IMGT reference directory sets with provisional nomenclature.

IMGT/HighV-QUEST updates

Each IMGT/V-QUEST programme version used by IMGT/HighV-QUEST is identified by a number with a date (list of updates and modifications in the IMGT/V-QUEST programme versions page at http://www.imgt.org/IMGT_vquest/share/textes/programversions.html). Each IMGT/V-QUEST reference directory used by IMGT/HighV-QUEST is similarly identified by a number with a date (list of changes in the IMGT/V-QUEST reference directory releases page at http://www.imgt.org/IMGT_vquest/share/textes/datareleases.html).

The IMGT/HighV-QUEST own version number reflects the updates of the functionalities that are specific to the system.

IMGT/HighV-QUEST Welcome and Search pages

As for the other IMGT® databases and tools, IMGT/HighV-QUEST is freely available for academics. However, the IMGT/HighV-QUEST Welcome page requires user identification and provides, for new users, a link to register. User identification has been set to avoid non-relevant use and overload of the server, and to contact the user if needed. The user identification gives access to the IMGT/HighV-QUEST Search page (Figure 1).

The IMGT/HighV-QUEST Search page allows launching the analysis of sequences batches. The menu bar at the top of the page recalls the Login and provides a link to "Analysis history" to follow the process of the batches and to download the results (the other two menus, "Launch statistics" to perform statistics on sequences batch results, and "Statistics history" for the downloading of the corresponding statistics results, are not detailed here).

In the top part of the IMGT/HighV-QUEST Search page, the users have to provide a title for the analysis, to select the species (e.g., *Homo sapiens*) and the receptor type (IG or TR) or locus (e.g., IGL) from the drop down list. The file to be uploaded should contain IG or TR rearranged nucleotide sequences in FASTA and must be formatted in text. Users can choose to receive an e-mail notification when the analysis is queued in the local analysis queue and/or when the analysis is submitted for computer processing and/or when the analysis is completed and/or when the results can be downloaded and/or 5 days before the analysis is expired.

In Display results, the users can select the results files for the outputs: A. "Detailed View" (individual results file for each sequence) which is optional and only available for batches containing fewer than 150,000 sequences. B. "Files in CSV" which comprises up to eleven files if all are selected (the options are equivalent to those of the IMGT/V-QUEST "Excel files").

In Advanced parameters, the users can choose parameters (identical to those of IMGT/V-QUEST). "Search for insertions and deletions" (an option in IMGT/V-QUEST) is selected by default in IMGT/HighV-QUEST.

IMGT/HighV-QUEST outputs

The results obtained from IMGT/HighV-QUEST are fully identical to those from IMGT/V-QUEST provided that the sequences have been processed with the same IMGT/V-QUEST programme version, the same IMGT/V-QUEST reference directory release, with an identical selection of species and receptor type or locus, and with the same options chosen in advanced parameters. The results are provided in a .txz archive containing a main folder with the eleven CSV files (if all were selected), and may include one subfolder with individual files, in text, for each submitted sequence (if the option "Include individual result files" of "Detailed View" was selected in the IMGT/HighV-QUEST Search Page). Text and CSV formats have been chosen in order to facilitate statistical studies for further interpretation and knowledge extraction. The format of the IMGT/HighV-QUEST results is the same as that of the option "Download in a zip archive" of the "Excel file" display of IMGT/V-QUEST.

The content of the eleven CSV files is summarized in Table 2. The CSV files contain one line per analysed sequence, and together may comprise up to 539 columns for a complete results report.

1. The "Summary" file #1 (29 columns with "Search for insertions and deletions" by default, or 25 without it) provides the synthesis of the analysis, i.e. the sequence names, the sequence functionality, the names of the closest V, D and J genes and alleles, the alignment scores and identity percentage for V and J, the D-REGION reading frame, the FR-IMGT and CDR-IMGT lengths, the amino acid (AA) JUNCTION (with frameshift(s) shown by # if any) provided by IMGT/JunctionAnalysis, the description

1.	Application of the IMGT unique numbering [62] and introduction of IMGT gaps*
2.	Identification of indels and their correction [17] (by default)**
3.	Identification of the closest V, D and J genes and alleles [22-24]*
4.	IMGT/JunctionAnalysis results [19,20]*
5.	Description of mutations and amino acid changes*
6.	IMGT/Automat annotation [21,22]*
7.	Statistical analysis [23]***
8.	Characterization of the IMGT clonotypes (AA) [24]***

Table 1: IMGT/HighV-QUEST functionalities and results
*common to IMGT/HighV-QUEST and IMGT/V-QUEST (same algorithm and same IMGT reference directories).
**option in IMGT/V-QUEST.
***specific to IMGT/HighV-QUEST.

Figure 1: IMGT/HighV-QUEST Search page.

of insertions and deletions if any, the warnings related to identification of the V and J gene and alleles and the sequence functionality, the orientation ("-" for opposite or "+" for direct) of the sequence at the submission, and the nucleotide sequence in the direct orientation with insertions in capital letters if any. The results of the IMGT/HighV-QUEST "Summary" file #1 are equivalent to those of the IMGT/V-QUEST "Results summary" (Figure 2).

2. The "IMGT-gapped-nt-sequences" file #2 (18 columns) includes the nucleotide (nt) 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, and the nucleotide sequences of CDR3-IMGT, JUNCTION, J-REGION and FR4-IMGT. The "Nt-sequences" file #3 (66, 81, 102 or 114 columns, depending

on the number of D (0, 1, 2 or 3) identified) includes all labels (ungapped nt sequences) delimited automatically by IMGT/Automat. The "IMGT-gapped-AA-sequences" file #4 (18 columns) and the "AA-sequences" file #5 (18 columns) includes the AA sequences for the same labels as file #2, gapped (according to the IMGT unique numbering) and ungapped, respectively. The results of the files #2 to #5 are equivalent to those illustrated in IMGT/V-QUEST "7. V-REGION translation" (Figure 3A) and "13. Annotation by IMGT/Automat" (Figure 3B).

3. The "Junction" file #6 includes all results of IMGT/JunctionAnalysis (40 columns for IGL, IGK, TRA and TRG sequences (no D), 53 (if one D), 73 (if two D) or 84 (if 3 D) columns for IGH, TRB and TRD sequences). All labels described in the junction are included with their nucleotide sequence and length (number of nt), as well as the number of trimmed nt for the 3'V-REGION, 5' and 3' D-REGION and 5' J-REGION, the translation of the CDR3-IMGT and the translation of the JUNCTION (which includes 2nd-CYS (C104) and J-PHE (F118) or J-TRP (W118)),

with frameshifts indicated by a # if any and without frameshift. The results of the "Junction" file #6 are equivalent to those of IMGT/V-QUEST "4. Results of IMGT/JunctionAnalysis" (Figure 4).

4. The "V-REGION-mutation-and-AA-change-table" file #7 (11 columns) includes the list of nt mutations and AA changes for the V-REGION, FR1-IMGT, CDR1-IMGT, FR2-IMGT, CDR2-IMGT, FR3-IMGT and CDR3-IMGT. The results of file #7 are equivalent to those of IMGT/V-QUEST "9. V-REGION mutation and AA change table" (Figure 5). The nt mutations and the corresponding AA changes for nonsilent mutations, are described for each FR-IMGT and CDR-IMGT, with their nt and codon position according to the IMGT unique numbering [62] and for the AA changes according to the IMGT AA classes [29] (IMGT Education > Aide-mémoire).

5. The "V-REGION-nt-mutation-statistics" file #8 (130 columns) and the "V-REGION-AA-change-statistics" file #9 (109 columns) report the statistics of the nt mutations and AA changes, respectively. File #8

File number	File name	Number of columns filled	Results content *
#1	"Summary"	29 (or 25)	- Alignment score and identity percentage with the closest V and J genes and alleles, - D-REGION reading frame, - FR-IMGT and CDR-IMGT lengths, - Amino acid (AA) JUNCTION, - Description of insertions and deletions if any, - User sequence in the direct orientation, - Sequence orientation at the submission.
#2	"IMGT-gapped-nt-sequences"	18	- Nucleotide (nt) 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, - nt sequences of CDR3-IMGT, JUNCTION, J-REGION and FR4-IMGT.
#3	"Nt-sequences"	66, 81, 102 or 114	- nt sequences of all labels that can be automatically described and delimitated by IMGT/Automat (66 columns for IGL, IGK, TRA and TRG sequences, 81 (if one D), 102 (if two D) or 114 (if 3 D) columns for IGH, TRB and TRD sequences).
#4	"IMGT-gapped-AA-sequences"	18	- 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	Same columns as "IMGT-gapped-AA-sequences" (#4), but sequences of labels are without IMGT gaps.
#6	"Junction"	40, 53, 73, or 84	- Results of IMGT/JunctionAnalysis (40 columns for IGL, IGK, TRA and TRG sequences, 53 (if one D), 73 (if two D) or 84 (if 3 D) columns for IGH, TRB and TRD sequences).
#7	"V-REGION-mutation-and-AA-change-table"	11	- List of mutations (nt mutations, AA changes, AA class identity (+) or change (-)) for V-REGION, FR1-IMGT, CDR1-IMGT, FR2-IMGT, CDR2-IMGT, FR3-IMGT and germline CDR3-IMGT.
#8	"V-REGION-nt-mutation-statistics"	130	- Number (nb) of nt positions including IMGT gaps, nb of nt, nb of identical nt, total nb of mutations, nb of silent mutations, nb of nonsilent mutations, nb of transitions (a>g, g>a, c>t, t>c) and nb of transversions (a>c, c>a, a>t, t>a, g>c, c>g, g>t, t>g) for V-REGION, FR1-IMGT, CDR1-IMGT, FR2-IMGT, CDR2-IMGT, FR3-IMGT and germline CDR3-IMGT.
#9	"V-REGION-AA-change-statistics"	109	- nb of AA positions including IMGT gaps, nb of AA, nb of identical AA, total nb of AA changes, nb of AA changes according to AAclassChangeType (+++, ++-, +-+, +--, -+-, --+, ---), and nb of AA class changes according to AAclassSimilarityDegree (nb of Very similar, nb of Similar, nb of Dissimilar, nb of Very dissimilar) for V-REGION, FR1-IMGT, CDR1-IMGT, FR2-IMGT, CDR2-IMGT, FR3-IMGT and germline CDR3-IMGT.
#10	"V-REGION-mutation-hotspots"	8	- Hot spots motifs ((a/t)a, t(a/t), (a/g)g(c/t)(a/t), (a/t)(a/g)c(c/t)) detected in the closest germline V-REGION with positions in FR-IMGT and CDR-IMGT.
#11	"Parameters"		- Date of the analysis, - IMGT/V-QUEST programme version, IMGT/V-QUEST reference directory release, - Parameters used for the analysis: species, receptor type or locus, IMGT reference directory set and Advanced parameters.

Table 2: Content of the eleven IMGT/HighV-QUEST results files in CSV format (results equivalent to those of the "Excel file" of IMGT/V-QUEST online). ***:** Files #1 to #10 comprise systematically sequence identification, i.e. the sequence name, the functionality, the names of the closest V gene and allele, and files #1 to #6 also include the D and J genes and alleles. The files #7 to #10 that report the analysis of mutations are used mostly for immunoglobulins (IG). Files #1 to #10 include one line per submitted sequence.

Citation: Giudicelli V, Duroux P, Lavoie A, Aouinti S, Lefranc M-P, et al. (2015) From IMGT-ONTOLOGY to IMGT/HighV-QUEST for NGS Immunoglobulin (IG) and T cell Receptor (TR) Repertoires in Autoimmune and Infectious Diseases. Autoimmun Infec Dis 1(1): doi http://dx.doi.org/10.16966/2470-1025.103

Sequence number 1: seq1

Sequence compared with the human IG set from the IMGT reference directory

```
>seq1
cgataagcttgatatgaaacacctgtggttcttcctcctgctggtggcagctcccagatg
ggtcctgtcccgggtgccgctgcaggagtcggggcccaggactggtgaagccctcacagac
cctgtccctcacctgcactgtcTCTGGTGACTCCTTCACCtctggtgactccttcagcag
tggtggttacttctggagctggctccgccagctcccagggaagggcctggagtggattgg
gcacatcgattacagtgggggcaccgactacaaccgtccctcaagagtcgacttaccat
gacgattgacacctctaagaacctgttctccctgaagctggcctctctgactgccgcgga
cacggccgtgtactactgtgcgagatcccctcattggcggggttattatgcgaaagtaca
tgatgcttttgatatctggggccaagggagaatggtcaccgtctcctcagcctccaccaa
gggcccatcggtcttcccctggcacctcctccaagagcacctctggggcacagcggc
cctgggctgcctggtcaaggactacttccccgaaccggtgacggtgtcttggaaactcag
gcgccctgacc
```

Result summary:	📧☞ 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
	FR1-IMGT	18	TCTGGTGACTCCTTCACC	no	26	143

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 IGHV4-31*03 F, or Homsap IGHV4-31*04 F	score = 1234	identity = **91.75%** (267/291 nt) [**91.41%** (266/291 nt)]
J-GENE and allele	Homsap IGHJ3*02 F	score = 232	identity = 96.00% (48/50 nt)
D-GENE and allele by IMGT/JunctionAnalysis	Homsap IGHD3-3*01 F	D-REGION is in reading frame 2	
FR-IMGT lengths, CDR-IMGT lengths and AA JUNCTION	[25.17.38.11] || [10.7.19]	CARSPHWRGYYAKVHDAFDIW

Figure 2: The results of the IMGT/HighV-QUEST "Summary" file #1 (29 columns) are equivalent to those of the IMGT/V-QUEST "Result summary table" with the option "Search for insertions and deletions" shown here (seq1: accession number AF013616 of IMGT/LIGM-DB).

A

7. V-REGION translation

B

13. Annotation by IMGT/Automat

```
        Label                 Location/Qualifiers
V-D-J-REGION                  1..355
                              /CDR_length="[8.8.11]"
                              /FR_length="[25.17.38.11]"
                              /nucleotide sequence
                              gaggtgcagctgttggagtctggggggaggcgtggtccagcctgggaggtccctgagactc
                              tcctgtatagcctctggattcaccttcagtagctatcctatgacctgggtccgccaggct
                              ccaggcaaggggctggagtgggtggcaagtatatcatatgacggaagttataaatataag
                              gtagactccatgaagggccgactcaccatctccagagacaattccaagaacacgctgtat
                              ttggaaatgaacagcctgacagctgaggacacggctgtgtattactgtgcgaggacagct
                              ttctttaacgcctatgacttctgggggccagggaaccctggtcaccgtctcctcag
                              /translation
                              EVQLLESGGGVVQPGRSLRLSCIASGFTFSSYPMTWVRQAPGKGLEWVASISYDGSYKYK
                              VDSMKGRLTISRDNSKNTLYLEMNSLTAEDTAVYYCARTAFFNAYDFWGQGTLVTVSS
V-REGION                      1..293
                              /allele="Homsap IGHV3-30*04 F or Homsap IGHV3-30-3*03 F"
                              /gene="Homsap IGHV3-30 or Homsap IGHV3-30-3"
                              /nucleotide sequence
                              gaggtgcagctgttggagtctggggggaggcgtggtccagcctgggaggtccctgagactc
                              tcctgtatagcctctggattcaccttcagtagctatcctatgacctgggtccgccaggct
                              ccaggcaaggggctggagtgggtggcaagtatatcatatgacggaagttataaatataag
                              gtagactccatgaagggccgactcaccatctccagagacaattccaagaacacgctgtat
                              ttggaaatgaacagcctgacagctgaggacacggctgtgtattactgtgcgag
                              /translation
                              EVQLLESGGGVVQPGRSLRLSCIASGFTFSSYPMTWVRQAPGKGLEWVASISYDGSYKYK
                              VDSMKGRLTISRDNSKNTLYLEMNSLTAEDTAVYYCA
FR1-IMGT                      1..75
                              /AA_IMGT="AA 1 to 26, AA 10 is missing"
                              /nucleotide sequence
                              gaggtgcagctgttggagtctggggggaggcgtggtccagcctgggaggtccctgagactc
                              tcctgtatagcctct
                              /translation
                              EVQLLESGGGVVQPGRSLRLSCIAS
1st-CYS                       64..66
                              /nucleotide sequence
                              tgt
                              /translation
                              C
CDR1-IMGT                     76..99
                              /AA_IMGT="AA 27 to 38, AA 31, 32, 33, 34 are  missing"
                              /nucleotide sequence
                              ggattcaccttcagtagctatcct
                              /translation
                              GFTFSSYP
FR2-IMGT                      100..150
                              /AA_IMGT="AA 39 to 55"
                              /nucleotide sequence
                              atgacctgggtccgccaggctccaggcaaggggctggagtgggtggcaagt
                              /translation
                              MTWVRQAPGKGLEWVAS
CONSERVED-TRP                 106..108
                              /nucleotide sequence
                              tgg
                              /translation
                              W
CDR2-IMGT                     151..174
                              /AA_IMGT="AA 56 to 65, AA 60, 61 are  missing"
                              /nucleotide sequence
                              atatcatatgacggaagttataaa
                              /translation
                              ISYDGSYK
FR3-IMGT                      175..288
                              /AA_IMGT="AA 66 to 104, AA 73 is missing"
                              /nucleotide sequence
                              tataaggtagactccatgaagggccgactcaccatctccagagacaattccaagaacacg
                              ctgtatttggaaatgaacagcctgacagctgaggacacggctgtgtattactgt
                              /translation
                              YKVDSMKGRLTISRDNSKNTLYLEMNSLTAEDTAVYYC
2nd-CYS                       286..288
                              /nucleotide sequence
                              tgt
                              /translation
                              C
CDR3-IMGT                     289..321
                              /AA_IMGT="AA 105 to 117, AA 111, 112 are missing"
                              /nucleotide sequence
                              gcgaggacagctttctttaacgcctatgacttc
                              /translation
                              ARTAFFNAYDF
```

Figure 3: The results of the IMGT/HighV-QUEST files #2, #4 and #5 (18 columns each) and file #3 (66, 81, 102 or 114 columns, depending on the number of D (0, 1, 2 or 3) identified) are equivalent to the results of IMGT/V-QUEST "7. V-REGION translation" (A) and "13. Annotation by IMGT/Automat" (B) shown here (seq 2: accession number AB012909 of IMGT/LIGM-DB).

Citation: Giudicelli V, Duroux P, Lavoie A, Aouinti S, Lefranc M-P, et al. (2015) From IMGT-ONTOLOGY to IMGT/HighV-QUEST for NGS Immunoglobulin (IG) and T cell Receptor (TR) Repertoires in Autoimmune and Infectious Diseases. Autoimmun Infec Dis 1(1): doi http://dx.doi.org/10.16966/2470-1025.103

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 1.

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

☐

Input	V name	3'V-REGION	N1	D-REGION	N2	5'J-REGION	J name	D name	Vmut	Dmut	Jmut	Ngc
seq2	Homsap IGHV3-30*04	tgtgcgag...	g	......acagctttcttta.	acgc	....ctatgacttctgg	Homsap IGHJ4*02	Homsap IGHD5-18*01	0	3	2	4/5

Translation of the JUNCTION

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

☐

	104	105	106	107	108	109	110	113	114	115	116	117	118	Frame	CDR3-IMGT length	Molecular mass	pI	PhysicoChemical Descriptor (by BRFAA)
	C	A	R	T	A	F	F	N	A	Y	D	F	W					
seq2	tgt	gcg	agg	aca	gct	ttc	ttt	aac	gcc	tat	gac	ttc	tgg	+	11	1,611.8	6.15	CARTAFFNAYDFW

Figure 4: The results of the IMGT/HighV-QUEST "Junction" file #6 (40, 53, 73 or 84 columns, depending on the number of D (0, 1, 2 or 3) identified) are equivalent to those of IMGT/V-QUEST "4. Results of IMGT/JunctionAnalysis" shown here. Numbers of trimmed nt (dots in "Analysis of the JUNCTION", here) are indicated in dedicated columns of the CSV file, for example in "3'V-REGION trimmed-nt nb", "5'D-REGION trimmed-nt nb", "3'D-REGION trimmed-nt nb", and "5'J-REGION trimmed-nt nb" for a sequence with one D gene and allele. Numbers of mutated nt compared to the germline (underlined in "Analysis of the JUNCTION", here) are indicated in the dedicated columns of the CSV file: "3'V-REGION mut-nt nb", "D-REGION mut-nt nb", "D1-REGION mut-nt nb", "D2-REGION mut-nt nb", "D3-REGION mut-nt nb", "5'J-REGION mut-nt nb" (seq 2: accession number AB012909 of IMGT/LIGM-DB).

9. V-REGION mutation and AA change table

FR1-IMGT	CDR1-IMGT	FR2-IMGT	CDR2-IMGT	FR3-IMGT	CDR3-IMGT
c1>g, Q1>E (+ + -)	g112>c, A38>P (- - -)	c118>a, H40>T (+ - -)	t177>c	c198>t	a318>g
g13>t, V5>L (+ - +)		a119>c, H40>T (+ - -)	a190>t, N64>Y (- - -)	t199>a, Y67>K (- - -)	g319>a
g70>a, A24>I (+ - +)		g163>a, V55>S (- - -)		c201>g, Y67>K (- - -)	a320>c
c71>t, A24>I (+ - +)		t164>g, V55>S (- - -)		c203>t, A68>V (+ - +)	
				g211>a, V71>M (+ - -)	
				t226>c, F76>L (+ - -)	
				c265>t	
				c268>g, Q90>E (+ + -)	
				g284>c, R95>T (- - -)	

Figure 5: The results of the IMGT/HighV-QUEST "V-REGION-mutation-and-AA-change-table" file #7 (11 columns) are equivalent to those of IMGT/V-QUEST "9. V-REGION mutation and AA change table" shown here. As an example, in the FR1-IMGT column, c1>g, Q1>E (++-) means that the nt mutation (c>g) at nt 1 leads to an AA change (Q>E) at codon 1, with same hydropathy (+) and volume (+) classes but different IMGT physicochemical (-) classes [29] (accession number AB012909 of IMGT/LIGM-DB)

includes the total number of mutations, the number of silent and nonsilent mutations, the number of transitions and transversions. File #9 includes the number of AA changes according to AA class Change Type (+++, ++-, +-+, +--, -+-, --+, ---), and the number of AA class changes according to AA class Similarity Degree (Nb of Very similar, Nb of Similar, Nb of Dissimilar, Nb of Very dissimilar). In both files the results are given for V-REGION, FR1-IMGT, CDR1-IMGT, FR2-IMGT, CDR2-IMGT, FR3-IMGT and CDR3-IMGT. The results of files #8 and #9 are equivalent respectively to those of "Nucleotide (nt) mutations" and "Amino acids (AA) changes" of IMGT/V-QUEST "10. V-REGION mutation and AA change statistics" (Figure 6).

6. The "V-REGION-mutation-hotspots" file #10 (8 columns) includes the hot spots motifs ((a/t)a, t(a/t), (a/g)g(c/t)(a/t), (a/t)(a/g)c(c/t)) identified in the closest germline V-REGION and with CDR-IMGT and FR-IMGT localizations. The results of file #10 are equivalent to those of IMGT/V-QUEST "11. V-REGION mutation hot spots" (Figure 7).

7. The "Parameters" file #11 includes the date of the analysis, the IMGT/V-QUEST programme version, the IMGT/V-QUEST reference directory release, and the parameters used for the analysis: the species, the receptor type or locus, the IMGT reference directory set, "with allele *01" (if selected), "Search for insertions and deletions", and the number of nucleotides to add (or exclude) in 3' of the V-REGION for the evaluation of the alignment score, and the number of nucleotides to exclude in 5' of the V-REGION for the evaluation of the number of mutations, if these 2 numbers are not 0 (default value).

Conclusions

IMGT-ONTOLOGY and IMGT® data and information system, which

10. V-REGION mutation and AA change statistics

Nucleotide (nt) mutations

IMGT labels		V-REGION	FR1-IMGT	CDR1-IMGT	FR2-IMGT	CDR2-IMGT	FR3-IMGT	CDR3-IMGT
Nb of positions including IMGT gaps (nt)		**317** (320)	78	36	51	30	117	5 (8)
Nb of nucleotides		**293** (296)	75	24	51	24	114	5 (8)
Nb of identical nucleotides		**273**	71	23	47	22	105	5
Nb of mutations		**20** (23)	4	1	4	2	9	0 (3)
Mutations	Silent	**3** (6)	0	0	0	1	2	0 (3)
	Nonsilent	**17**	4	1	4	1	7	0
Transitions	a>g	**0** (1)	0	0	0	0	0	0 (1)
	g>a	**3** (4)	1	0	1	0	1	0 (1)
	c>t	**4**	1	0	0	0	3	0
	t>c	**2**	0	0	0	1	1	0
Transversions	a>c	**1** (2)	0	0	1	0	0	0 (1)
	c>a	**1**	0	0	1	0	0	0
	a>t	**1**	0	0	0	1	0	0
	t>a	**1**	0	0	0	0	1	0
	g>c	**2**	0	1	0	0	1	0
	c>g	**3**	1	0	0	0	2	0
	g>t	**1**	1	0	0	0	0	0
	t>g	**1**	0	0	1	0	0	0

Amino acid (AA) changes

IMGT labels			V-REGION	FR1-IMGT	CDR1-IMGT	FR2-IMGT	CDR2-IMGT	FR3-IMGT	CDR3-IMGT
Nb of positions including IMGT gaps (AA)			**105** (106)	26	12	17	10	39	1 (2)
Nb of AA			**97** (98)	25	8	17	8	38	1 (2)
Nb of identical AA			**84** (85)	22	7	15	7	32	1 (2)
Nb of AA changes			**13**	3	1	2	1	6	0
AA changes	Very similar	(+ + +)	**0**	0	0	0	0	0	0
	Similar	(+ + -)	**2**	1	0	0	0	1	0
		(+ - +)	**3**	2	0	0	0	1	0
	Dissimilar	(+ - -)	**3**	0	0	1	0	2	0
		(- + -)	**0**	0	0	0	0	0	0
		(- - +)	**0**	0	0	0	0	0	0
	Very dissimilar	(- - -)	**5**	0	1	1	1	2	0

Figure 6: The results of the IMGT/HighV-QUEST file #8 (130 columns) and file #9 (109 columns) are equivalent of those of IMGT/V-QUEST "10. V-REGION mutation and AA change statistics" two tables shown here. The results are given for the V-REGION and for FR-IMGT and CDR-IMGT. Statistics are calculated up to the 3' end of the V-REGION identified in the input sequence (this includes the 3' last two identical nucleotides with the closest germline V-REGION). The numbers in parentheses, in the V-REGION and CDR3-IMGT columns, correspond to the statistics calculated up to the 3' end of the closest germline V-REGION and therefore may include nt and AA differences due to the junction diversity (accession number AB012909 of IMGT/LIGM-DB).

Citation: Giudicelli V, Duroux P, Lavoie A, Aouinti S, Lefranc M-P, et al. (2015) From IMGT-ONTOLOGY to IMGT/HighV-QUEST for NGS Immunoglobulin (IG) and T cell Receptor (TR) Repertoires in Autoimmune and Infectious Diseases. Autoimmun Infec Dis 1(1): doi http://dx.doi.org/10.16966/2470-1025.103

11. V-REGION mutation hot spots (motifs and localizations in germline V-REGION)

(a/t)a — wa

Motif	Localization
ta	105-106 (CDR1)
ta	109-110 (CDR1)
ta	114-115 (CDR1)
aa	142-143 (FR2)
ta	165-166 (FR2-CDR2)
ta	167-168 (CDR2)
ta	172-173 (CDR2)
aa	186-187 (CDR2)
ta	189-190 (CDR2)
ta	192-193 (CDR2)
aa	194-195 (CDR2)
ta	196-197 (FR3)
ta	199-200 (FR3)
aa	214-215 (FR3)
aa	244-245 (FR3)
aa	250-251 (FR3)
aa	253-254 (FR3)
ta	262-263 (FR3)
aa	269-270 (FR3)
aa	274-275 (FR3)
ta	304-305 (FR3)
ta	307-308 (FR3)

t(a/t) — tw

Motif	Localization
tt	82-83 (CDR1)
tt	88-89 (CDR1)
ta	105-106 (CDR1)
ta	109-110 (CDR1)
ta	114-115 (CDR1)
tt	164-165 (FR2)
ta	167-168 (CDR2)
ta	172-173 (CDR2)
ta	189-190 (CDR2)
ta	192-193 (CDR2)
ta	196-197 (FR3)
ta	199-200 (FR3)
tt	226-227 (FR3)
tt	246-247 (FR3)
ta	262-263 (FR3)
ta	304-305 (FR3)
tt	306-307 (FR3)

(a/g)g(c/t)(a/t) — rgyw

Motif	Localization
agct	8-11 (FR1)
agta	103-106 (CDR1)
ggct	132-135 (FR2)
ggca	139-142 (FR2)
ggct	146-149 (FR2)
ggca	159-162 (FR2)
agta	187-190 (CDR2)
agct	285-288 (FR3)
ggct	297-300 (FR3)

(a/t)(a/g)c(c/t) — wrcy

Motif	Localization
agct	8-11 (FR1)
agcc	41-44 (FR1)
agcc	72-75 (FR1)
agct	106-109 (CDR1)
tgct	111-114 (CDR1)
tact	196-199 (FR3)
agcc	277-280 (FR3)
agct	285-288 (FR3)
tact	307-310 (FR3)

Figure 7: The results of the IMGT/HighV-QUEST "V-REGION-mutation-hotspots" file #10 (8 columns) are equivalent to those of IMGT/V-QUEST "11. V-REGION mutation hot spots" shown here.

are at the origin of immunoinformatics [45], provide the concepts, knowledge and informatics frame for a standardized analysis of the IG and TR repertoire analysis. IMGT-ONTOLOGY allowed human IG and TR genes and alleles to be identified, described and classified in IMGT® [3-6], before the human genome was published [97,98]. The IMGT Repertoire and the IMGT reference directories are constantly updated with newly sequenced IG and TR of vertebrate species [99-105] and were recently enriched with new human IG genes and alleles owing to the analysis of different haplotypes and CNV in humans [106,107]. These biocurated data and the standardized IMGT® analysis tools are key for maintaining exploration of the immune repertoires with high-quality data at the forefront of basic and clinical research. Indeed, the IMGT® standards for IG and TR are used in clinical applications. IMGT/V-QUEST is frequently used by clinicians in order to identify the repertoires against pathogens and for the analysis of IG somatic hypermutations in leukemia, lymphoma and myeloma, and more particularly in chronic lymphocytic leukemia (CLL) [16,72–75] in which the percentage of mutations of the rearranged IGHV gene in the VH of the leukemic clone has a prognostic value for the patients. For this evaluation, IMGT/V-QUEST is the standard recommended by the European Research Initiative on CLL (ERIC) for comparative analysis between laboratories [72]. The sequences of the V-(D)-J junctions determined by IMGT/Junction Analysis [19,20] are also used in the characterization of stereotypic patterns in CLL [73,74] and for the synthesis of probes specific of the junction for the detection and follow-up of minimal residual diseases (MRD) in leukemias and lymphomas. A new era is opening in hemato-oncology with the use of NGS for analysis of the clonality and MRD identification, and making IMGT® standards use needed more than ever. These standards are also used in the development and description of therapeutical antibodies [40-42,82-86] and the closest human V, J and C genes and alleles (results obtained with the IMGT/DomainGapAlign tool [25-26], approved by the WHO/IUIS nomenclature subcommittee for IG and TR amino acid sequences) are required for the INN submission

of humanized and human antibodies [48]. These are the same standards as those utilized by IMGT/HighV-QUEST. The NGS analysis of IG and TR repertoires in physiological conditions [108,109], vaccination [24], autoimmune diseases [110-111] and infectious diseases [112-115] opens new insight for understanding the protective and pathogenic adaptive immune responses. The aims are to characterize the IG and TR repertoires from B and T subsets in many individuals, to identify the skewing of the B and T cell repertoires in immunodeficiency or in autoimmune diseases, to explore the elicitation of broadly neutralizing antibodies (bnAbs) against HIV-1, to understand the potential of the human immune system to develop protective and pathogenic antibodies. The ultimate goal is to find specific antibodies and T cell receptors which could be of great value for diagnostic, prognostic, vaccine development and novel immunotherapy.

Acknowledgments

Availability and Citation

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

Conflicts of Interest

The authors declare no conflict of interest.

References

1. IMGT® The International ImMunoGeneTics information system®. http://www.imgt.org

2. Lefranc MP, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, et al. (2015) IMGT®, the international ImMunoGeneTicsinformation system® 25 years on. Nucleic Acids Res 43: D413-422.

3. Lefranc M-P, Lefranc G (2001) The Immunoglobulin FactsBook, Academic Press: London, UK, 1-458.

4. Lefranc M-P, Lefranc G (2001) The T Cell Receptor FactsBook, Academic Press: London, UK, 1-398.

5. 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, NJ, USA 1-37.

6. 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, NJ, USA 1-23.

7. Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, et al. (2006) IMGT/LIGM-DB, the IMGT® comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. Nucleic Acids Res 34: D781-D784.

8. 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.

9. 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.

10. 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.

11. 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 6: 750-761.

12. Poiron C, Wu Y, Ginestoux C, Ehrenmann F, Duroux P, et al. (2010) IMGT/mAb-DB: the IMGT® database for therapeutic monoclonal antibodies. In: Proceedings of the 11èmes Journées Ouvertes de Biologie, Informatique et Mathématiques (JOBIM), Montpellier, France, September 2010, 7-9.

13. Giudicelli V, Chaume D, Lefranc M-P (2004) IMGT/V-QUEST, an integrated software for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. Nucleic Acids Res 32: W435-W440.

14. Giudicelli V, Lefranc M-P (2005) Interactive IMGT on-line tools for the analysis of immunoglobulin and T cell receptor repertoires. In: Veskler BA (ed) New Research on Immunology. Nova Science Publishers Inc, New York, NY, USA 77-105.

15. 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.

16. Giudicelli V, Lefranc M-P (2008) IMGT® standardized analysis of immunoglobulin rearranged sequences. In: Ghia P, Rosenquist R, Davi F (eds) Immunoglobulin gene analysis in Chronic Lymphocytic Leukemia. Wolters Kluwer Health Italy Ltd: Milan, Italy, Chapter 2, 33-52.

17. 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 6: 695-715.

18. 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. In: Christiansen FT, Tait BD (eds) Immunogenetics: Methods and Applications in Clinical Practice. Humana Press, Springer Science + Business Media: New York, NY, USA, Chapter 32, 569-604.

19. 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: i379-i385.

20. 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 6: 716-725.

21. Giudicelli V, Protat C, Lefranc M-P (2003) The IMGT strategy for the automatic annotation of IG and TR cDNA sequences: IMGT/Automat. In: Institut National de Recherche en Informatique et en Automatique, Paris, France. Proceedings of the European Conference on Computational Biology (ECCB 2003), Data and Knowledge Bases, ECCB 2003, Paris, France, 27-30 September 2003, 103-104.

22. 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 Inform1 16: 3-8.

23. Alamyar E, Giudicelli V, Shuo L, Duroux P, Lefranc M-P (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. Immunome Res 8: 1-15.

24. Li S, Lefranc M-P, Miles JJ, Alamyar E, Giudicelli V, et al. (2013) IMGT/HighV-QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. Nat Commun 4: 1-13.

25. 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 6: 737-749.

26. Ehrenmann F, Lefranc M-P (2012) IMGT/DomainGapAlign: The IMGT® tool for the analysis of IG, TR, MHC, IgSF and MhcSF domain amino acid polymorphism. In: Christiansen FT, Tait BD (eds) Immunogenetics: Methods and Applications in Clinical Practice. Humana Press, Springer Science + Business Media: New York, NY, USA, Chapter 33, 605-633.

27. 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 6: 726-736.

28. Lane J, Duroux P, Lefranc M-P (2010) From IMGT-ONTOLOGY to IMGT/LIGMotif: The IMGT® standardized approach for immunoglobulin and T cell receptor gene identification and description in large genomic sequences. BMC Bioinform 11: 1-16.

29. Pommié C, Levadoux S, Sabatier R, Lefranc G, Lefranc MP (2004) IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. J Mol Recognit 17: 17-32.

30. Lefranc M-P (2003) IMGT, the international ImMunoGeneTics information system. In: Bock G, Goode J (eds) Immunoinformatics: bioinformatic strategies for better understanding of immune function. Novartis Foundation Symposium. John Wiley and Sons: Chichester, UK, Volume 254: 126-142.

31. Lefranc M-P, Giudicelli V, Ginestoux C, Chaume D (2003) IMGT, the international ImMunoGeneTics information system: the reference in immunoinformatics. Stud Health Technol Inform 95: 74-79.

32. Lefranc M-P (2003) IMGT databases, web resources and tools for immunoglobulin and T cell receptor sequence analysis. Leukemia 17: 260-266.

33. Lefranc M-P (2004) IMGT, the international ImMunoGenetics information system®. In: Lo BKC (ed) Antibody Engineering Methods and Protocols, 2nd ed. Humana Press: Totowa, NJ, USA, 27-49.

34. Lefranc M-P (2004) IMGT-ONTOLOGY and IMGT databases, tools and Web resources for immunogenetics and immunoinformatics. Mol Immunol 40: 647-660.

35. Lefranc M-P (2005) IMGT, the international ImMunoGeneTics information system: A standardized approach for immunogenetics and immunoinformatics. Immunome Res 1: 1-11.

36. Lefranc M-P (2007) IMGT®, the international ImMunoGeneTics information system® for immunoinformatics. Methods for querying IMGT® databases, tools and Web resources in the context of immunoinformatics. In: Flower DR (ed) Immunoinformatics: Predicting Immunogenicity in Silico. Humana Press, Springer: Totowa, NJ, USA, Chapter 2, 19-42.

37. Lefranc M-P (2008) IMGT-ONTOLOGY, IMGT® databases, tools and Web resources for Immunoinformatics. In: Schoenbach C, Ranganathan S, Brusic V (eds) Immunoinformatics. Immunomics Reviews, Series of Springer Science and Business Media LLC; Springer: New York, NY, USA, Volume 1, Chapter 1, 1-18.

38. Lefranc M-P, Giudicelli V, Regnier L, Duroux P (2008) IMGT®, a system and an ontology that bridge biological and computational spheres in bioinformatics. Brief Bioinform 9: 263-275.

39. Lefranc M-P (2008) IMGT®, the international ImMunoGeneTics information system® for immunoinformatics. Methods for querying IMGT® databases, tools and Web resources in the context of immunoinformatics. Mol Biotechnol 40: 101-111.

40. Lefranc M-P (2009) Antibody databases and tools: The IMGT® experience. In: An Z (ed) Therapeutic Monoclonal Antibodies: From Bench to Clinic. John Wiley and Sons: Hoboken, NJ, USA, Chapter 4, 91-114.

41. Lefranc M-P (2009) Antibody databases: IMGT®, a French platform of world-wide interest (in French). Bases de données anticorps: IMGT®, une plate-forme française d'intérêt mondial. Med Sci 25: 1020-1023.

42. Ehrenmann F, Duroux P, Giudicelli V, Lefranc M-P (2010) Standardized sequence and structure analysis of antibody using IMGT®. In: Kontermann R, Dübel S (eds) Antibody Engineering. Springer-Verlab: Berlin, Heidelberg, Germany, 2010; Volume 2, Chapter 2, 11-31.

43. Lefranc M-P (2011) IMGT the international ImMunoGeneTics information system. Cold Spring Harb Protoc 6: 595-603.

44. Lefranc M-P (2013) IMGT® Information System. In: Dubitzky W, Wolkenhauer O, Cho K-H, Yokota H (eds) Encyclopedia of Systems Biology. Springer Science + Business Media, LLC012: New York, NY, USA 959-964.

45. Lefranc M-P (2014) Immunoglobulin (IG) and T cell receptor genes (TR): IMGT® and the birth and rise of immunoinformatics. Front Immunol 5: 1-22.

46. Lefranc M-P (2007) WHO-IUIS Nomenclature Subcommittee for immunoglobulins and T cell receptors report. Immunogenetics 59: 899-902.

47. 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.

48. World Health Organization. International Nonproprietary Names (INN) for Biological and Biotechnological Substances (a Review).

49. Lefranc M-P (2011) Antibody nomenclature: From IMGT-ONTOLOGY to INN definition. MAbs 3: 1-2.

50. Giudicelli V, Lefranc M-P (1999) Ontology for immunogenetics: IMGT-ONTOLOGY. Bioinformatics 15: 1047-1054.

51. Giudicelli V, Lefranc M-P (2012) IMGT-ONTOLOGY 2012. Frontiers in Bioinformatics and Computational Biology. Front Genet 3: 1-16.

52. Giudicelli V, Lefranc M-P (2013) IMGT-ONTOLOGY. In: Dubitzky W, Wolkenhauer O, Cho K-H, Yokota H (eds) Encyclopedia of Systems Biology. Springer Science + Business Media, LLC012: New York, NY, USA 964-972.

53. Giudicelli V, Lefranc M-P (2003) IMGT-ONTOLOGY: Gestion et découverte de connaissances au sein d'IMGT (in French). In: Hacid M-S, Kodratoff Y, Boulanger D (eds) Extraction et Gestion des Connaissances (EGC'2003). Hermès Science Publications, Lavoisier: Cachan, Paris, France 13-23.

54. Lefranc M-P, Giudicelli V, Ginestoux C, Bosc N, Folch G, et al. (2004) IMGT-ONTOLOGY for immunogenetics and immunoinformatics. In Silico Biol 4: 17-29.

55. Lefranc M-P, Clément O, Kaas Q, Duprat E, Chastellan P, et al. (2005) IMGT-Choreography for Immunogenetics and Immunoinformatics. In Silico Biol 5: 45-60.

56. Duroux P, Kaas Q, Brochet X, Lane J, Ginestoux C, et al. (2008) IMGT-Kaleidoscope, the formal IMGT-ONTOLOGY paradigm. Biochimie 90: 570-583.

57. Lefranc M-P (2011) From IMGT-ONTOLOGY IDENTIFICATION axiom to IMGT standardized keywords: For immunoglobulins (IG), T cell receptors (TR), and conventional genes. Cold Spring Harb Protoc 6: 604-613.

58. Lefranc M-P (2011) From IMGT-ONTOLOGY DESCRIPTION axiom to IMGT standardized labels: For immunoglobulin (IG) and T cell receptor (TR) sequences and structures. Cold Spring Harb Protoc 6: 614-626.

59. 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 6: 627-632.

60. Lefranc M-P (1997) Unique database numbering system for immunogenetic analysis. Immunol Today 18: 509.

61. Lefranc M-P (1999) The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. Immunologist 7: 132-136.

62. Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, 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.

63. Lefranc M-P, Pommié C, Kaas Q, Duprat E, Bosc N, 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.

64. Lefranc M-P, Duprat E, Kaas Q, Tranne M, Thiriot A, et al. (2005) IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. Dev Comp Immunol 29: 917-938.

65. 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 6: 633-642.

66. Ruiz M, Lefranc M-P (2002) IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. Immunogenetics 53: 857-883.

67. Kaas Q, Lefranc M-P (2007) IMGT Colliers de Perles: Standardized sequence-structure representations of the IgSF and MhcSF superfamily domains. Curr Bioinform 2: 21-30.

68. 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.

69. Lefranc M-P (2011) IMGT Collier de Perles for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. Cold Spring Harb Protoc 6: 643-651.

70. 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, 2nd ed. Elsevier Ltd: London, UK, 2014, Volume 1, Chapter 27, 481-514.

71. Robert R, Lefranc M-P, Ghochikyan A, Agadjanyan MG, Cribbs DH, et al. (2010) Restricted V gene usage and VH/VL pairing of mouse humoral response against the N-terminal immunodominant epitope of the amyloid β peptide. Mol Immunol 48: 59-72.

72. Ghia P, Stamatopoulos K, Belessi C, Moreno C, Stilgenbauer S, et al. (2007) ERIC recommendations on IGHV gene mutational status analysis in chronic lymphocytic leukemia. Leukemia 21:1-3.

73. Agathangelidis A, Darzentas N, Hadzidimitriou A, Brochet X, Murray F, et al. (2012) Stereotyped B-cell receptors in one third of chronic lymphocytic leukemia: Towards a molecular classification with implications for targeted therapeutic interventions. Blood 119: 4467-4475.

74. Kostareli E, Gounari M, Janus A, Murray F, Brochet X, et al. (2012) Antigen receptor stereotypy across B-cell lymphoproliferations: the case of IGHV4–59/IGKV3–20 receptors with rheumatoid factor activity. Leukemia 26: 1127-1131.

75. Xochelli A, Agathangelidis A, Kavakiotis I, Minga E, Sutton LA, et al. (2015) Immunoglobulin heavy variable (IGHV) genes and alleles: New entities, new names and implications for research and prognostication in chronic lymphocytic leukemia. Immunogenetics 67:61-66.

76. Jefferis R, Lefranc M-P (2009) Human immunoglobulin allotypes: Possible implications for immunogenicity. MAbs 1: 332-338.

77. Lefranc M-P, Lefranc G (2012) Human Gm, Km and Am allotypes and their molecular characterization: A remarkable demonstration of polymorphism. In: Christiansen FT, Tait BD (eds) Immunogenetics: Methods and Applications in Clinical Practice. Humana Press, Springer: New York, NY, USA, Chapter 34,635-680.

78. Dechavanne C, Guillonneau F, Chiappetta G, Sago L, Lévy P, et al. (2012) Mass spectrometry detection of G3m and IGHG3 alleles and follow-up of differential mother and neonate IgG3. PLoS One 7: e46097.

79. Magdelaine-Beuzelin C, Kaas Q, Wehbi V, Ohresser M, Jefferis R, et al. (2007) Structure-function relationships of the variable domains of monoclonal antibodies approved for cancer treatment. Crit Rev Oncol Hematol 64: 210-225.

80. Pelat T, Bedouelle H, Rees AR, Crennell SJ, Lefranc M-P, et al. (2008) Germline humanization of a non-human Primate antibody that neutralizes the anthrax toxin, by in vitro and in silico engineering. J Mol Biol 384: 1400-1407.

81. Pelat T, Hust M, Hale M, Lefranc M-P, Dübel S, et al. (2009) Isolation of a human-like antibody fragment (scFv) that neutralizes ricin biological activity. BMC Biotechnol. 9: 1-13.

82. Lefranc M-P, Ehrenmann F, Ginestoux C, Duroux P, Giudicelli V (2012) Use of IMGT® databases and tools for antibody engineering and humanization. In: Chames P (ed) Antibody Engineering, 2nd ed. Humana Press, Springer Science + Business Media LLC: New York, NY, USA, Chapter 1, 3-37.

83. Alamyar E, Giudicelli V, Duroux P, Lefranc M-P (2013) Antibody V and C domain sequence, structure and interaction analysis with special reference to IMGT®. In: Ossipow V, Fisher A (eds) Monoclonal Antibodies: Methods and Protocols, 2nd ed. Humana Press, Springer Science + Business Media LLC: New York, NY, USA, 2013, Chapter 21, 337-381.

84. Lefranc M-P (2013) Immunoinformatics of the V, C and G domains: IMGT® definitive system for IG, TR and IgSF, MH and MhSF. In: De RK, Tomar N (eds) Immunoinformatics: From Biology to Informatics, 2nd ed. Humana Press, Springer Science + Business Media LLC: New York, NY, USA, Chapter 4, 59-107.

85. Lefranc M-P (2014) How to use IMGT® for therapeutic antibody engineering. In: Dübel S, Reichert J (eds) Handbook of Therapeutic Antibodies, 2nd ed. Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, Volume1, Chapter 10, 229-263.

86. Shirai H, Prades C, Vita R, Marcatili P, Popovic B, et al. (2014) Antibody informatics for drug discovery. Biochim Biophys Acta 1844: 2002-2015.

87. http://bioportal.bioontology.org/ontologies/IMGT-ONTOLOGY

88. Wain HM, Bruford EA, Lovering RC, Lush MJ, Wright MW, et al. (2002) Guidelines for human gene nomenclature. Genomics 79: 464-470.

89. Bruford EA, Lush MJ, Wright MW, Sneddon TP, Povey S, et al. (2008) The HGNC database in 2008: A resource for the human genome. Nucleic Acids Res 36: D445-D448.

90. Letovsky SI, Cottingham RW, Porter CJ, Li PW (1998) GDB: The human Genome Database. Nucleic Acids Res 26: 94-99.

91. Maglott DR, Katz KS, Sicotte H, Pruitt KD (2000) NCBI's LocusLink and RefSeq. Nucleic Acids Res 28: 126-128.

92. Maglott D, Ostell J, Pruitt KD, Tatusova T (2007) Entrez Gene: Gene-centered information at NCBI. Nucleic Acids Res 35: D26-D31.

93. Stabenau A, McVicker G, Melsopp C, Proctor G, Clamp M, et al. (2004) The Ensembl core software libraries. Genome Res14: 929-933.

94. Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, et al. (2008) The vertebrate genome annotation (Vega) database. Nucleic Acids Res 36: D753-D760.

95. Blake JA, Eppig JT, Bult CJ, Kadin JA, Richardson JE, et al. (2006) The Mouse Genome Database (MGD): updates and enhancements. Nucleic Acids Res 34: D562-D567.

96. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, et al. (2005) The Universal Protein Resource (UniProt). Nucleic Acids Res 33: D154-D159.

97. International Human Genome Sequence Consortium (2001) Initial sequence and analysis of the human genome. Nature 409:860-921.

98. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291:1304-1351.

99. Herzig TA, Lefranc M-P, Baldwin CL (2010) Annotation and classification of the bovine T cell receptor delta genes. BMC Genomics 11:100.

100. Antonacci R, Mineccia M, Lefranc M-P, Ashmaoui HM, Lanave C, et al. (2011) Expression and genomic analyses of *Camelus dromedarius* T cell receptor delta (TRD) genes reveal a variable domain repertoire enlargement due to CDR3 diversification and somatic mutation. Mol Immunol 48: 1384-1396.

101. Castro R, Bernard D, Lefranc M-P, Six A, Benmansour A, et al. (2011) T cell diversity and TcR repertoires in teleost fish. Fish Shellfish Immunol 31: 644-654.

102. Schwartz JC, Lefranc M-P, Murtaugh MP (2012) Evolution of the porcine *(Sus scrofa domestica)* immunoglobulin kappa locus through germline gene conversion. Immunogenetics. 64: 303-311.

103. Schwartz JC, Lefranc M-P, Murtaugh MP (2012) Organization, complexity and allelic diversity of the porcine *(Sus scrofa domestica)* immunoglobulin lambda locus. Immunogenetics 64: 399-407.

104. Shi B, Ma L, He X, Wang X, Wang P, et al. (2014) Comparative analysis of human and mouse immunoglobulin variable heavy regions from IMGT/LIGM-DB with IMGT/HighV-QUEST. Theor Biol Med Model 11:30.

105. Lu J, Panavas T, Thys K, Aerssens J, Naso M, et al. (2014) IgG variable region and VH CDR3 diversity in unimmunized mice analyzed by massively parallel sequencing. Mol Immunol 57: 274-83.

106. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, et al. (2013) Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. Am J Hum Genet 92: 530-546.

107. Watson CT, Steinberg KM, Graves TA, Warren RL, Malig M, et al. (2015) Sequencing of the human IG light chain loci from a hydatidiform mole BAC library reveals locus-specific signatures of genetic diversity. Genes Immun 16: 24-34.

108. Prabakaran P, Chen W, Singarayan MG, Stewart CC, Streaker E, et al. (2012) Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations. Immunogenetics 64: 337-350.

109. Mroczek ES, Ippolito GC, Rogosch T, Hoi KH, Hwangpo TA, et al. (2014) Differences in the composition of the human antibody repertoire by B cell subsets in the blood. Front Immunol 5: 96.

110. O'Connell AE, Volpi S, Dobbs K, Fiorini C, Tsitsikov E, et al. (2014) Next generation sequencing reveals skewing of the T and B cell receptor repertoires in patients with Wiskott-Aldrich syndrome. Front Immunol 5: 340.

111. Hershberg U, Meng W, Zhang B, Haff N, St Clair EW, et al. (2014) Persistence and selection of an expanded B-cell clone in the setting of rituximab therapy for Sjögren's syndrome. Arthritis Res Therapy 16: R51.

112. Reiche S, Dwai Y, Bussmann BM, Horn S, Sieg M, et al. (2015) High inter-individual diversity of point mutations, insertions, and deletions in human influenza virus nucleoprotein-specific memory B cells. PLoS One 10: e0128684.

113. Zhang Y, Yuan T, Li J, Zhang Y, Xu J, et al. (2013) The potential of the human immune system to develop broadly neutralizing HIV-1 antibodies: implications for vaccine development. AIDS 27: 2529-2539.

114. Prabakaran P, Chen W, Dimitrov DS (2014) The antibody germline/maturation hypothesis, elicitation of broadly neutralizing antibodies against HIV-1 and cord blood IgM repertoires. Front Immunol 5: 398.

115. Li L, Wang X-H, Williams C, Volsky B, Steczko O, et al. (2015) A broad range of mutations in HIV-1 neutralizing human monoclonal antibodies specific for V2, V3, and the CD4 binding site. Mol Immunol 66: 364-374.