# Department retreat Molecular Bases of Human Diseases,

# IGH, UPR CNRS 1142,

# Mèze, November 5-6 2015

## ORAL PRESENTATIONS

## IMGT®, the international ImMunoGeneTics information system®

Sofia Kossida

IMGT®, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université de Montpellier, Institut de Génétique Humaine, IGH, CNRS UPR 1142, 141 rue de la Cardonille, F-34396 Montpellier cedex 5, France

IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org) is the global reference in immunogenetics and immunoinformatics [1]. By its creation in 1989 by Marie-Paule Lefranc (Montpellier University and CNRS), IMGT® marked the advent of immunoinformatics, which emerged at the interface between immunogenetics and bioinformatics [2]. IMGT® is specialized in the immunoglobulins (IG) or antibodies, T cell receptors (TR), major histocompatibility (MH) and proteins of the IgSF and MhSF superfamilies. IMGT® is built on the IMGT-ONTOLOGY axioms and concepts, which bridged the gap between genes, sequences and 3D structures [3]. The concepts include the IMGT® standardized keywords (concepts of identification), IMGT® standardized labels (concepts of description), IMGT® standardized nomenclature (concepts of classification), IMGT unique numbering, and IMGT Colliers de Perles (concepts of numerotation). IMGT® comprises seven databases, 15,000 pages of web resources, and 17 tools, and provides a high-quality and integrated system for the analysis of the genomic and expressed IG and TR repertoire of the adaptive immune responses. Tools and databases are used in basic, veterinary, and medical research, in clinical applications (mutation analysis in leukemia and lymphoma) and in antibody engineering and humanization. They include, for example IMGT/V-QUEST and IMGT/JunctionAnalysis for nucleotide sequence analysis and their high-throughput version IMGT/HighV-QUEST for next-generation sequencing (500,000 sequences per batch), IMGT/DomainGapAlign for amino acid sequence analysis of IG and TR variable and constant domains and of MH groove domains, IMGT/3Dstructure-DB for 3D structures, contact analysis and paratope/epitope interactions of IG/antigen and TR/peptide-MH complexes and IMGT/mAb-DB interface for therapeutic antibodies and fusion proteins for immune applications (FPIA).

In my presentation, I will introduce the IMGT® current projects, and the oral presentations and posters contributed by the IMGT® team members.

[1] Nucleic Acids Res. 43:D413-22 (2015)
[2] Front Immunol (2014) http://journal.frontiersin.org/article/10.3389/fimmu.2014.00022/abstract
[3] Front Genet (2012) http://journal.frontiersin.org/article/10.3389/fgene.2012.00079/abstrac

# Gene and NGS at IMGT® in 2015

Véronique Giudicelli

IMGT®, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université de Montpellier, Institut de Génétique Humaine, IGH, CNRS UPR 1142, 141 rue de la Cardonille, F-34396 Montpellier cedex 5, France

Genes and NGS at IMGT® in 2015 will be presented in terms of IMGT/GENE-DB developments and IMGT/HighV-QUEST novel functionalities.

IMGT/GENE-DB [1] is the gene database of IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org) [2]IMGT/GENE-DB manages the reference sequences of the functional, ORF and in-frame pseudogenes IG and TR genes, once these sequences have been numbered and gapped according to the IMGT unique numbering. Thus, IMGT/GENE-DB represents the repository of the references directories used internally by all IMGT® analysis tools that perform nucleotide or amino acid sequence comparison and worldwide by HGNC, NCBI, genome browsers (EBI Ensembl and Vega), NCBI tool IgBlast, sequence analysis tool developers, academic and pharmaceutical societies. The most recent human genome assembly, GRCh38, is built from the haploid genome from an hydatiform mole which allows for the first time the description of a full haplotype for each locus. The schema of IMGT/GENE-DB has been improved to manage the localizations of human genes in NCBI assemblies, and the positions and the orientation of IGH and TRB genes in GRCh38 are currently available. In addition, several haplotypes have been recently highlighted in the IGH locus, with structural polymorphisms by gene copy number variations (CNV) [3]. The ongoing developments of IMGT/GENE-DB include in particular the management, the characterization and the representation of the CNV in the IG and TR loci. The IMGT/GENE-DB database is also regularly enriched with new IG and TR genes resulting from IMGT genome biocuration projects. In 2015, the human IGHV, IGKV and IGLV groups have been checked and updated (381 genes (of which 14 are new) and 760 alleles (of which 38 are new), and genes from other species have been integrated (the IGHC genes from dog and the complete IGH locus of the platypus, Ornithorhynchus *anatinus,* which together correspond to 98 new genes and 107 alleles). In October 2015, IMGT/GENE-DB includes 3570 IG and TR genes and 5267 alleles from 22 species, of which there were 718 genes and 1478 alleles for *Homo sapiens* and 869 genes and 1319 alleles for *Mus musculus*.

IMGT/HighV-QUEST [4-6], 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 (up to 1 million outputs compared), with the same degree of resolution and high-quality results as IMGT/V-QUEST and IMGT/JunctionAnalysis. Indeed IMGT/HighV-QUEST uses the same algorithm and runs against the same IMGT reference directory. IMGT/HighV-QUEST functionalities are basically the same as IMGT/ V-QUEST, the IMGT® online tool for the analysis of nucleotide sequences of the IG and TR V-DOMAIN. IMGT/HighV-QUEST numbers the user sequences according to the IMGT unique numbering 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 by comparison with the IMGT/V-QUEST reference directory. The tool

integrates IMGT/JunctionAnalysis for the detailed characterization of the V-D-J or V-J junctions, IMGT/Automat for a complete sequence annotation with the delimitation of the IMGT labels of description. IMGT/HighV-QUEST includes a module dedicated to statistical analysis. After the filter of the sequences reliable for statistics, the module characterizes the IMGT clonotypes (AA) with the evaluation of the diversity and the expression per V, D and J genes and performs the comparison between the batches selected for a same statistical analysis. In October 2015, more than 7.66 billion of sequences were analysed by IMGT/HighV-QUEST, by 1499 users from 41 countries (46% users from USA, 34% from EU, 20% from the remaining world).

[1] Giudicelli et al. Nucl. Acids Res. 33, D256-261 (2005). [2] Lefranc et al. Nucl. Acids Res. 43, D413-422 (2015). [3] Watson et al. Am. J. Hum. Genet. 92, 530-546 (2013). [4] Alamyar et al. Immunome Res. 20, 1-15 (2012). [5] Li et al. Nat. Commun. 4:2333 (2013). [6] Giudicelli et al. Autoimmun. Infec. Dis. 1(1) (2015).

Acknowledgment. IMGT/HighV-QUEST was granted access to HPC@LR and to the High Performance Computing (HPC) resources of Centre Informatique National de l'Enseignement Supérieur (CINES) and to Très Grand Centre de Calcul (TGCC) of the Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) under the allocation 036029 (2010-2015) made by GENCI (Grand Equipement National de Calcul Intensif).

## IMGT® informatics challenges in 2015

Patrice Duroux

IMGT®, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université de Montpellier, Institut de Génétique Humaine, IGH, CNRS UPR 1142, 141 rue de la Cardonille, F-34396 Montpellier cedex 5, France

IMGT®, the international ImMunoGeneTics information system® [1] is at the birth and rise of immunoinformatics, a new science at the interface between bioinformatics and immunogenetics [2]. Behind the scene, informatics represents many challenges as it requires that the data and system management respects the IMGT-ONTOLOGY concepts [3] and constantly adapts itself to the changes in infrastructure and resources, and above all that the system, available worldwide on Internet, be operational 24 hours every day. In 2015, IMGT® comprises 7 databases, 17 online tools and more than 15 000 pages of web resources. Despite the computing heterogeneity (due to the data, chronological or practical reasons), the informatics system maintains the data coherence, whatever the data types, for example nucleotides sequences in IMGT/LIGM-DB (177 050 entries from 351 vertebrate species), IMGT/CLL-DB (IG V domain sequences from CLL patients) and IMGT/PRIMER-DB (primer sequences), genes and alleles in IMGT/GENE-DB (3570 genes, 5267 alleles), amino acid (AA) sequences in IMGT/2Dstructure-DB (562 entries), and three-dimensional structures in IMGT/3Dstructure-DB (3257 entries )

The development of IMGT/2Dstructure-DB is worth noting as it was created as an extension of IMGT/3Dstructure-DB to describe and analyse AA sequences of chains and domains for which no 3D structures were available. IMGT/2Dstructure-DB uses the IMGT/3Dstructure-DB informatics frame and interface which allow one to analyse, manage and query immunoglobulins (IG) or antibodies, T cell receptors (TR) and major histocompatibility (MH)

proteins, as well as other proteins of the IgSF and MhSF superfamilies, as polymeric receptors made of several chains, in contrast to the IMGT/LIGM-DB sequence database that analyses and manages sequences individually. The AA sequences are analysed with the IMGT® criteria of standardized identification, description, nomenclature and numerotation. These databases are in constant development as they bridge the gap between sequences and 3D structures and are queried by academics and by pharmaceutical societies.

Therapeutic antibodies being the fastest growing domain in drug discovery, in order to provide an easy access to the antibody data, we developed a new database and interface, IMGT/mAb-DB with links to IMGT/2Dstructure-DB (antibody AA sequences) and to IMGT/3Dstructure-DB (if 3D structures are available). IMGT/mAb-DB provides information from the World Health Organization WHO/International Nonproprietary Names INN programme. IMGT/mAb-DB (516 entries) includes monoclonal therapeutic antibodies (mAb, INN suffix –mab) (a –mab is defined by the presence of at least an IG variable domain), fusion proteins for immune applications (FPIA, INN suffix –cept) (a –cept is defined by a receptor fused to a Fc). IMGT/mAb-DB also includes a few CPCA (e.g. protein or peptide fused to a Fc for only increasing their half-life, identified by the INN prefix ef–) and some RPI used, unmodified, for clinical applications. The informatics challenge of integrating a new database and its applications in IMGT® is to follow the rules and criteria of IMGT-Choreography for maintaining the coherence of the system [4].

An unexpected recent breakthrough in biology was the NGS. In 2010 we created IMGT/HighV-QUEST a portal online, available worldwide on Internet, dedicated to the analysis of rearranged nucleotide sequences of IG and TR (up to 500.000 sequences per run) obtained by high-throughput technologies. Since 2014, it includes a statistical post-treatment done on different batches (up to 1 million sequences). All this system requires the availability of high performance computing resources obtained since 2009 by applying to the Grand Équipement National de Calcul Intensif (GENCI).

I will illustrate on three axes (annotation process, antibody engineering and high-throughput analysis) how informatics meets and answers the challenges of those large developments.

[1] Lefranc M-P et al. Nucl. Acids Res. 43, D413-422 (2015). [2] Lefranc M-P. Front Immunol Feb 05;5:22. (2014). [3] Giudicelli V and Lefranc M-P. Front. Genet. 3:79 (2012). [4] Lefranc M-P et al. *In Silico* Biology, 5, 45-60 (2005).

# POSTERS

## Correlation between IMGT® Biocuration, IMGT/LIGM-DB and IMGT/GENE-DB

**Géraldine Folch*, Joumana Michaloud*, Marine Peralta, Mélanie Arrivet, Imène Chentli, Mélissa Cambon, Pascal Bento, Patrice Duroux, Véronique Giudicelli, Marie-Paule Lefranc, Sofia Kossida**

* Equal contribution

IMGT®, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université de Montpellier, Institut de Génétique Humaine, IGH, CNRS UPR 1142, 141 rue de la Cardonille, F-34396 Montpellier cedex 5, France

IMGT®, the international ImMunoGeneTics information system®, http://www.imgt.org, has developed a biocuration pipeline for immunoglobulin (IG) and T cell receptor (TR) sequence annotation. The expert annotation and added standardized knowledge are based on the seven IMGT-ONTOLOGY axioms: IDENTIFICATION, CLASSIFICATION, DESCRIPTION, NUMEROTATION, LOCALIZATION, ORIENTATION and OBTENTION [1-3]. IMGT/LIGMotif is the tool for genomic DNA sequences analysis [4], and IMGT/Automat is the tool for automatic annotation of rearranged cDNA sequences [5, 6].

IMGT expert biocurators check the annotation tool results for consistency, both manually and by using IMGT® tools (IMGT/NTI to VALD, IMGT/V-QUEST, IMGT/BLAST…). These annotated sequences are integrated into IMGT/LIGM-DB, the comprehensive and largest IMGT® database of IG and TR nucleotide sequences from human and other vertebrate species. For a given entry, nine types of display are available, including the IMGT flat file, the translation of the coding regions and the analysis by the IMGT/V-QUEST tool. They include the sequence identification, the gene and allele classification, the constitutive and specific motif description, the codon and amino acid numbering and the sequence obtaining information IMGT/LIGM-DB keywords (identification), labels (description), nomenclature (classification) and specificity (obtention) allow data retrieval from IMGT/LIGM-DB, and also from other IMGT® databases which have links to IMGT/LIGM-DB accession numbers.

The main source of IG and TR gene and allele knowledge is stored in IMGT/GENE-DB [7], the comprehensive IMGT® genome database and in the IMGT reference directory. IMGT/GENE-DB provides a search of IG and TR genes by locus, group and subgroup. An IMGT/GENE-DB entry displays accurate gene data related to genome, allelic polymorphisms, gene expression, proteins and structures. IMGT/GENE-DB manages the IMGT reference directory used by the IMGT tools for gene and allele comparison and assignment, and by the IMGT databases for gene data annotation. IMGT/GENE-DB is the official repository of all the IG and TR genes and alleles, IMGT® gene and allele names had been approved by HGNC and endorsed by WHO/IUIS, the World Health Organization (WHO)/International Union of Immunological Societies (IUIS) Nomenclature Subcommittee for IG and TR. Reciprocal links exist between IMGT/GENE-DB and HGNC, NCBI (Gene) and the genome browsers Ensembl and Vega (EBI).

IMGT® is used in very diverse domains: fundamental and medical research, veterinary research, repertoire analysis, biotechnology related to antibody engineering, diagnostics and therapeutical approaches.

[1] Giudicelli, V. and Lefranc, M.-P., Bioinformatics, 15, 1047-1054 (1999).
[2] Giudicelli, V. and Lefranc, M-P, Front Genet, 3:79 (2012).
[3] Giudicelli, V. and Lefranc, M.-P., Encycl Systems Biology, 964-972 (2013).
[4] Lane L., Duroux P.,and Lefranc M.-P. BMC Bioinformatics, 11:223 (2010).
[5] Giudicelli, V.et al. Stud. Health Technol. Inform, 116, 3-8 (2008).
[6] Giudicelli V, Protat C, Lefranc M-P. Data and Knowledge Bases, Poster DKB_31, ECCB pp. 103–104 (2003).
[7] Giudicelli V. et al. Nucleic Acids Res, 33, D256-261 (2005).

# From IMGT/mAb-DB to IMGT/3Dstructure-DB: highlighting the antigen and antibody contact amino acids of the target and mAb binding sites

**Typhaine Paysan-Lafosse, Souphatta Sasorith, Patrice Duroux, 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 and CNRS, Montpellier (France)

IMGT®, the international ImMunoGeneTics information system® http://www.imgt.org, is the global reference in immunogenetics and immunoinformatics specialized in the immunoglobulins (IG) or antibodies, T cell receptors (TR), major histocompatibility (MH) of human and other vertebrate species, and in the immunoglobulin superfamily (IgSF), MH superfamily (MhSF) and related proteins of the immune system (RPI) of vertebrates and invertebrates.

IMGT/mAb-DB is a unique resource of expertized annotations on monoclonal antibodies (mAbs, suffix -mab), fusion proteins for immune applications (FPIA, suffix -cept), composite proteins for clinical applications (CPCA) and relative proteins of the immune system (RPI) with diagnostic or therapeutic indications. The database includes information on antibody identification, format description, specificity, clinical indications, publications, clinical trials and regulations of monoclonal antibodies. IMGT/mAb-DB includes 413 entries from the World Health Organization (WHO)/International Nonproprietary Name (INN) Programme.

IMGT/2Dstructure-DB and IMGT/3Dstructure-DB contain information about amino acid (AA) sequences, contact analysis, paratope and epitope interactions and IMGT/Colliers de Perles. These annotations are based on the IMGT standards (IMGT gene and allele names, IMGT unique numbering, CDR-IMGT).

286 INN antibodies from IMGT/mAb-DB are provided with their AA sequences and are annotated in IMGT/2Dstructure-DB.

For each antibody, whose experimental 3D structures are available, the corresponding structures are also described and fully annotated in IMGT/3Dstructure-DB using the IMGT standards. In addition, if the antibody has been co-crystallized with an antigen, the user can characterize the complex; for example by identifying the AA involved in the antibody-antigen interaction (using Contact analysis, IMGT/Collier de Perles and the Paratope and epitope description) or by visualizing the complex highlighted by the IMGT labels and standard colors (with the molecular viewer Jmol).

# IMGT/HighV-QUEST and IMGT Clonotype (AA): Identification and Statistical Significance Diversity per Gene for NGS Immunoprofiles of Immunoglobulins and T Cell Receptors

**Safa Aouinti[1,3], Dhafer Malouche[2,3], Véronique Giudicelli[1], Patrice Duroux[1], Arthur Lavoie[1], Sofia Kossida[1] , Marie-Paule Lefranc[1]**

[1] IMGT®, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire (LIGM), Institut de Génétique Humaine (IGH), UPR CNRS 1142, Montpellier University and CNRS, Montpellier (France)

[2] Higher School of Statistics and Information Analysis, University of Carthage, Tunis (Tunisia)

[3] National Schools of Engineers of Tunis, Laboratory U2S, University of Tunis El-Manar, Tunis (Tunisia)

The adaptive immune responses of humans and other jawed vertebrate species (gnasthostomata) are characterized by the B and T cells and their specific antigen receptors, the immunoglobulins (IG) or antibodies and the T cell receptors (TR) (up to $2.10^{12}$ different IG and TR per individual). IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org), was created in 1989 by Marie-Paule Lefranc (Montpellier University and CNRS) to manage the huge and complex diversity of these antigen receptors [1]. IMGT® built on IMGT-ONTOLOGY [2] concepts of identification (keywords), description (labels), classification (gene and allele nomenclature) and numerotation (IMGT unique numbering) is at the origin of immunoinformatics, a science at the interface between immunogenetics and bioinformatics. IMGT/HighV-QUEST [3-6], the first web portal, and so far the only one, for the next generation sequencing (NGS) analysis of IG and TR, is the paradigm for immune repertoire standardized outputs and immunoprofiles of the adaptive immune responses. It provides the identification of the variable (V), diversity (D) and joining (J) genes and alleles, analysis of the V-(D)-J junction and complementarity determining region 3 (CDR3) and the characterization of the 'IMGT clonotype (AA)' (AA for amino acid) diversity and expression. IMGT/HighV-QUEST compares outputs of different batches, up to one million nucleotide sequences for the statistical module. These high throughput IG and TR repertoire immunoprofiles are of prime importance in vaccination, cancer, infectious diseases, autoimmunity and lymphoproliferative disorders, however their comparative statistical analysis still remains a challenge. We present a standardized statistical procedure to analyze IMGT/HighV-QUEST outputs for the evaluation of the significance of the IMGT clonotype (AA) diversity differences in proportions, per gene of a given group, between NGS IG and TR repertoire immunoprofiles. The procedure is generic for evaluating significance of the IMGT clonotype (AA) diversity and expression per gene, and suitable for any IG and TR immunoprofiles of any species.

*[1] Lefranc M-P, Front Immunol, 5:22, 2014. [2] Giudicelli V and Lefranc M-P, Front Genet, 3:79, 2012. [3] Alamyar E et al. Mol Biol 882:569-604, 2012. [4] Alamyar E et al. Immunome Res 8(1):26, 2012. [5] Li S et al. Nat. Commun. 4:2333, 2013. [6] Giudicelli V et al. AutoImmun Infec Dis 1(1), 2015. [7] Dudoit S, van der Laan MJ. Multiple testing procedures with application to genomics. Springer Series in Statistics; 2008*