Experimental and Clinical Immunogenetics

Exp Clin Immunogenet 1998;15:1–7

Received: January 7, 1998

Marie-Paule Lefranc

Laboratoire d'ImmunoGénétique
Moléculaire, CNRS, Université
Montpellier II, Montpellier,
France

IMGT Locus on Focus

A New Section of *Experimental and Clinical Immunogenetics*

Key Words
IMGT
Immunogenetics
Immunoglobulin
T cell receptor

Abstract
IMGT, the international ImMunoGeneTics database (http://imgt.cnusc.fr:8104) is an integrated database specialising in immunoglobulins (Ig), T-cell receptors (TcR) and major histocompatibility complex (MHC) of all vertebrate species, created by Marie-Paule Lefranc, CNRS, Montpellier II University, Montpellier, France (lefranc@ligm.crbm.cnrs-mop.fr). IMGT includes three databases: LIGM-DB for Ig and TcR, MHC/HLA-DB, and PRIMER-DB (the last two in development). The new section of *Experimental and Clinical Immunogenetics*, 'IMGT locus on focus', will give the more updated and extensive overview of the current status of the different Ig, TcR and MHC genes in IMGT, with standardization of gene nomenclature, functionality, and allele polymorphism according to the IMGT rules and unique numbering.

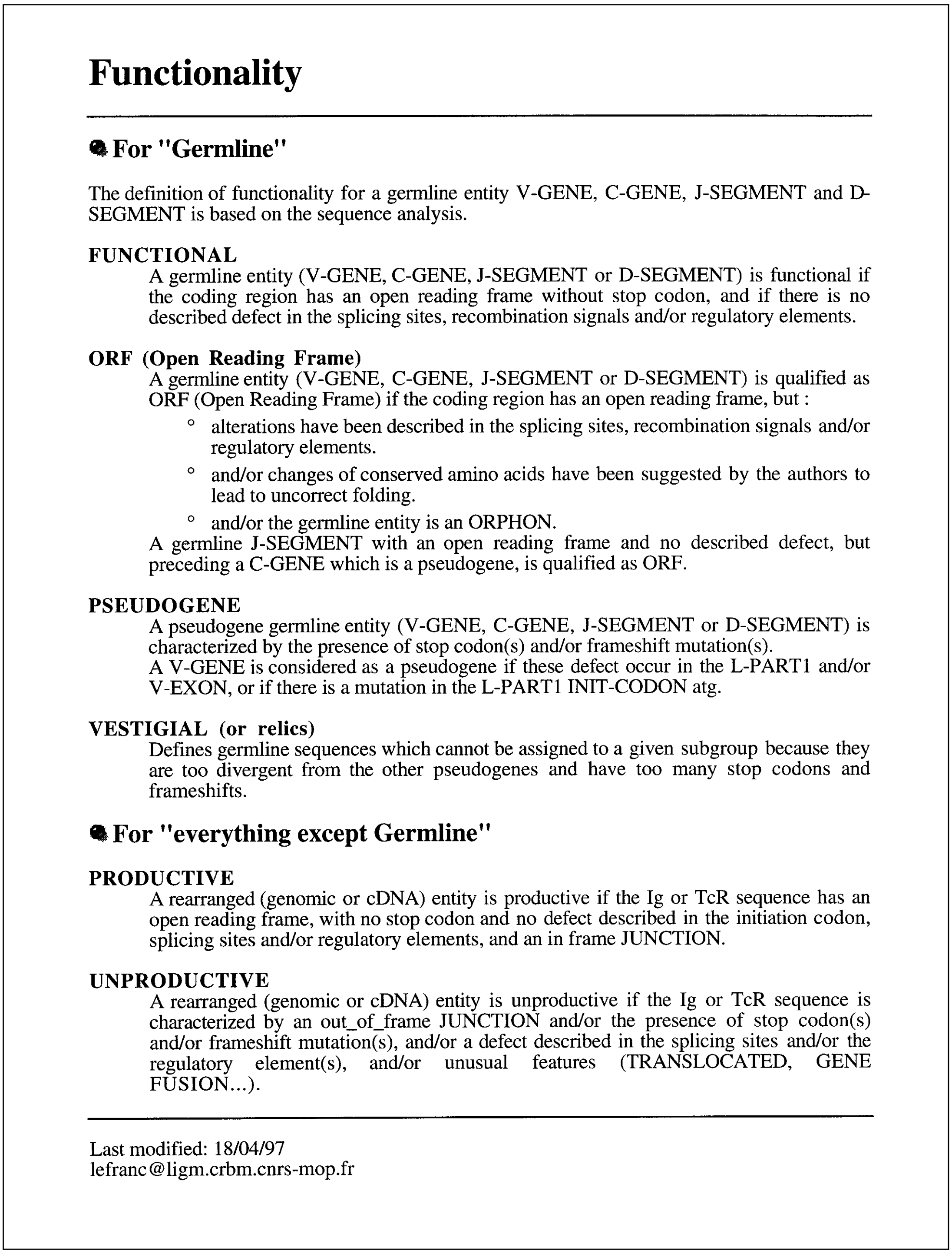
The molecular synthesis and genetics of the immunoglobulin (Ig) and T cell receptor (TcR) chains is particularly complex and unique since it includes biological mechanisms such as DNA molecular rearrangements in seven loci (three for Ig and four for TcR) located on four different chromosomes in human nucleotide deletions and insertions at the rearrangement junctions, and hypermutations in the Ig loci. The number of potential protein forms of Ig and TcR is almost unlimited. In order to deal with the complexity and the large number of published sequences, a specialized and integrated database, IMGT, the international ImMunoGeneTics database (http://imgt.cnusc.fr:8104) was created in 1992 by Marie-Paule Lefranc, CNRS, Université Montpellier II, Montpellier, France

KARGER

Fax + 41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

This article is also accessible online at:
http://BioMedNet.com/karger

Prof. Marie-Paule Lefranc, Laboratoire d'ImmunoGénétique Moléculaire
UMR 5535 (CNRS, Université Montpellier II), 1919, route de Mende
F–34293 Montpellier Cedex 5 (France), Tel. +33 (0)4 67 61 36 34
Fax +33 (0)4 67 04 02 31, E-Mail lefranc@ligm.crbm.cnrs-mop.fr
IMGT: http:/imgt.cnusc.fr:8104



Functionality

For "Germline"

The definition of functionality for a germline entity V-GENE, C-GENE, J-SEGMENT and D-SEGMENT is based on the sequence analysis.

FUNCTIONAL

A germline entity (V-GENE, C-GENE, J-SEGMENT or D-SEGMENT) is functional if the coding region has an open reading frame without stop codon, and if there is no described defect in the splicing sites, recombination signals and/or regulatory elements.

ORF (Open Reading Frame)

A germline entity (V-GENE, C-GENE, J-SEGMENT or D-SEGMENT) is qualified as ORF (Open Reading Frame) if the coding region has an open reading frame, but :

- ° alterations have been described in the splicing sites, recombination signals and/or regulatory elements.
- ° and/or changes of conserved amino acids have been suggested by the authors to lead to uncorrect folding.
- ° and/or the germline entity is an ORPHON.

A germline J-SEGMENT with an open reading frame and no described defect, but preceding a C-GENE which is a pseudogene, is qualified as ORF.

PSEUDOGENE

A pseudogene germline entity (V-GENE, C-GENE, J-SEGMENT or D-SEGMENT) is characterized by the presence of stop codon(s) and/or frameshift mutation(s).
A V-GENE is considered as a pseudogene if these defect occur in the L-PART1 and/or V-EXON, or if there is a mutation in the L-PART1 INIT-CODON atg.

VESTIGIAL (or relics)

Defines germline sequences which cannot be assigned to a given subgroup because they are too divergent from the other pseudogenes and have too many stop codons and frameshifts.

For "everything except Germline"

PRODUCTIVE

A rearranged (genomic or cDNA) entity is productive if the Ig or TcR sequence has an open reading frame, with no stop codon and no defect described in the initiation codon, splicing sites and/or regulatory elements, and an in frame JUNCTION.

UNPRODUCTIVE

A rearranged (genomic or cDNA) entity is unproductive if the Ig or TcR sequence is characterized by an out_of_frame JUNCTION and/or the presence of stop codon(s) and/or frameshift mutation(s), and/or a defect described in the splicing sites and/or the regulatory element(s), and/or unusual features (TRANSLOCATED, GENE FUSION...).

Last modified: 18/04/97
lefranc@ligm.crbm.cnrs-mop.fr

Document 1

(lefranc@ligm.crbm.cnrs-mop.fr) [1, 2]. IMGT comprises alignment tables and expertly annotated sequences and consists of three databases: LIGM-DB for Ig and TcR, MHC/HLA-DB, and PRIMER-DB (an Ig, TcR and MHC-related primer database), the two latter are currently being developed. The goals of IMGT are to establish a common data access to all immunogenetics data, including nucleotide and protein sequences, oligonucleotide primers, gene maps and other genetic data of Ig, TcR, and MHC molecules, and to provide a graphical user-friendly data access. IMGT has important implications in medical research (repertoire in autoimmune diseases, AIDS, leukemias, lymphomas), therapeutic approaches (antibody engineering), genome diversity and genome evolution studies.

Before the physical implementation of the database, the main and longest objective was to establish rules for the description of Ig and TcR sequences of any species. This was the major foundation of consistent expertise. These rules comprise:

Standardization of Key Words

IMGT key words for Ig and TcR include the following: (i) General key words: indispensable for sequence assignments, they are described in an exhaustive and nonredundant list, and are organized in a tree structure; (ii) Specific key words: they are more specifically associated to particularities of the sequences (orphon, transgene ...) or to diseases (leukemia, lymphoma, myeloma ...). The list is not definitive and new specific key words can easily be added if needed. IMGT/LIGM-DB standardized key words have been assigned to all entries.

Standardization of Sequence Annotation

Ig and TcR sequences have been analysed at the DNA and protein level in order to define a list of labels for the structural and functional motifs. 177 feature labels were shown to be necessary for an accurate annotation. The annotation is the most critical step and a very time-consuming process since about 50 sequences a week can be annotated by an experienced annotator. Levels of annotation have been defined, which allow the users to query sequences in IMGT/LIGM-DB even though they are not fully annotated. Resulting sequences can be obtained in different formats (FASTA, EMBL ...), with three reading frames, and with protein translation. This will be upgraded to set up a protein database for Ig and TcR.

Standardization of Ig and TcR Gene Designation

The objective is to provide immunologists and geneticists with a unique nomenclature per locus which will allow extraction and comparison of data for the complex B and T cell antigen receptor molecules, whatever the species. Data concerning the human Ig and TcR genes have been standardized, and maps of loci, tables of germline genes in the IMGT nomenclature, correspondence to other gene designations, and gene functionality (document 1) are available from http://imgt.cnusc.fr:8104.

The IMGT Unique Numbering

A uniform numbering system for Ig and TcR sequences of all species has been established by Marie-Paule Lefranc to facilitate sequence comparison and cross-referencing between experiments from different laboratories whatever the antigene receptor (Ig or TcR), the chain type or the species [3]. This numbering results from the analysis of more than 3,000 Ig and TcR variable region sequences of vertebrate species from fish to human. It takes into account and combines the definition of the framework (FR) and complementarity determining region (CDR) [4], structural data from X-ray diffraction studies [5], and the characterization of the hypervariable loops [6]. In the IMGT numbering, conserved amino acids from frameworks always have the same number whatever the Ig or TcR variable sequence, and whatever the species they come from. As examples: cysteine 23 (in FR1), tryptophan 41 (in FR2), leucine 89 and cysteine 104 (in FR3). Tables and graphs are available on the WWW interface at the IMGT Marie-Paule page from http://imgt.cnusc.fr:8104.

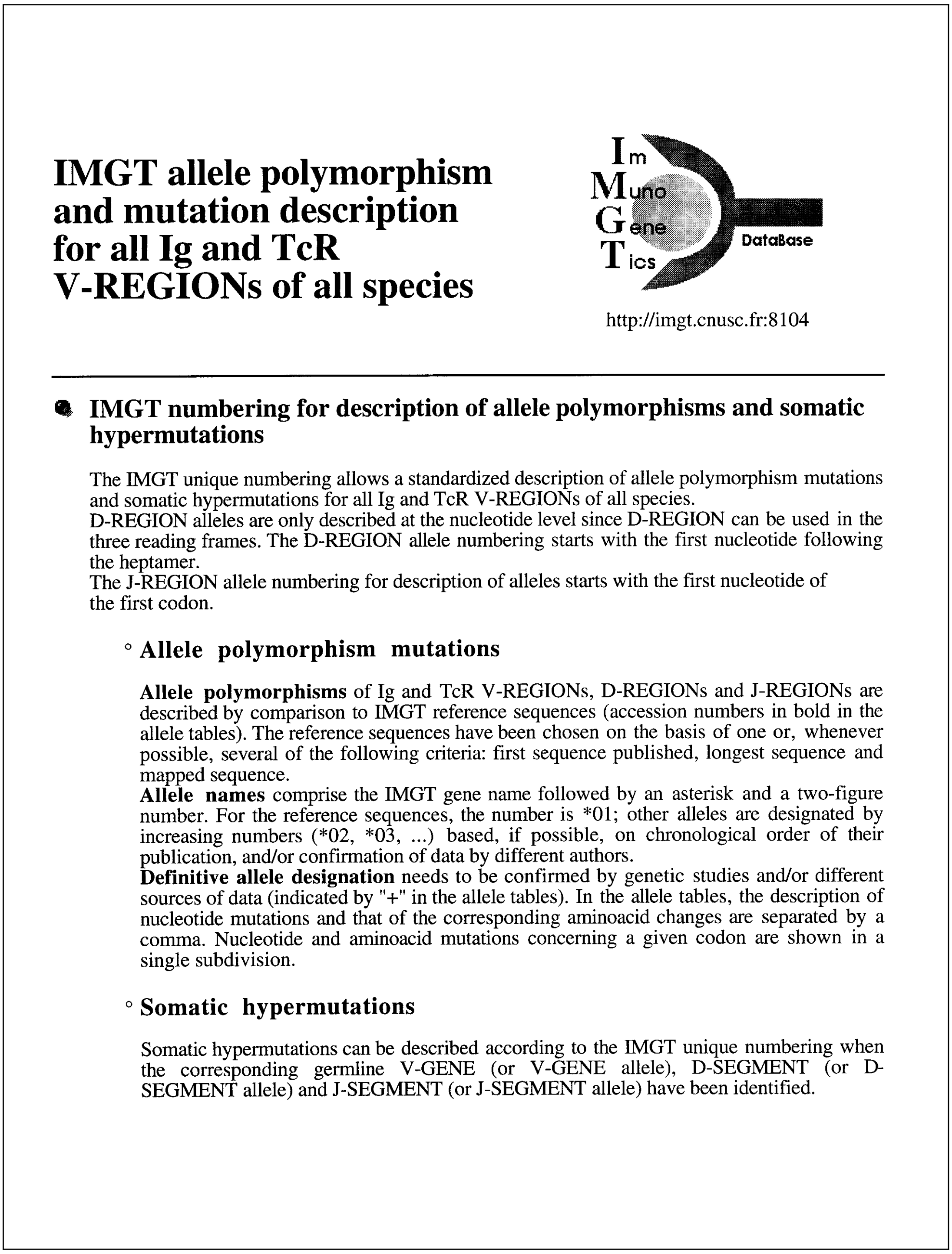
This IMGT unique numbering has several advantages:

- It has allowed to redefine the limits of the FR and CDR. The FR-IMGT and CDR-IMGT lengths become in themselves crucial information which characterize variable regions belonging to a group, a subgroup and/or a gene.
- Framework amino acids (and codons) located at the same position in different sequences can be compared without requiring sequence alignments. This also holds for amino acids belonging to CDR-IMGT of same length.
- The unique numbering is used as the output of the IMGT/DNAPLOT alignment tool. The aligned sequences are displayed according to the IMGT numbering and with the FR-IMGT and CDR-IMGT delimitations.
- The unique numbering has allowed a standardization of the description of the mutations and allelic polymorphisms of the variable regions (document 2). These mutations and allelic polymorphisms are described by comparison to the 'reference sequences' defined in IMGT.

By facilitating the comparison between sequences and by allowing the description of alleles and mutations, the IMGT unique numbering represents a big step forward in the analysis of the Ig and TcR sequences of all species. Moreover, it gives insight in the structural configuration of the variable domain and opens interesting views on the evolution of these sequences, since this numbering has been applied with success to all the sequences belonging to the V-set of the Ig superfamily, including nonrearranging sequences in vertebrates (CD4, CTX, ...) and in invertebrates (drosophila amalgam, drosophila fasciclin II, ...) (graphical representations available at the IMGT Marie-Paule page from http://imgt.cnusc.fr:8104).

In December 1997, IMGT/LIGM-DB contained more than 24,000 Ig and TcR from 81 species. IMGT sequences are identified by the EMBL accession number. Since August 1996, the IMGT/LIGM-DB content closely follows the Ig and TcR EMBL one, with the advantage of being deleted from sequences which have previously been wrongly assigned to Ig and TcR.

A new section of *Experimental and Clinical Immunogenetics*, 'IMGT Locus on Focus', will give the more updated and extensive overview of the current status of the different germline Ig and TcR genes of any species, as described and completed in IMGT. This issue contains the first report on the human immunoglobulin lambda IGLV genes and IGLJ segments.



IMGT allele polymorphism and mutation description for all Ig and TcR V-REGIONs of all species

http://imgt.cnusc.fr:8104

IMGT numbering for description of allele polymorphisms and somatic hypermutations

The IMGT unique numbering allows a standardized description of allele polymorphism mutations and somatic hypermutations for all Ig and TcR V-REGIONs of all species.
D-REGION alleles are only described at the nucleotide level since D-REGION can be used in the three reading frames. The D-REGION allele numbering starts with the first nucleotide following the heptamer.
The J-REGION allele numbering for description of alleles starts with the first nucleotide of the first codon.

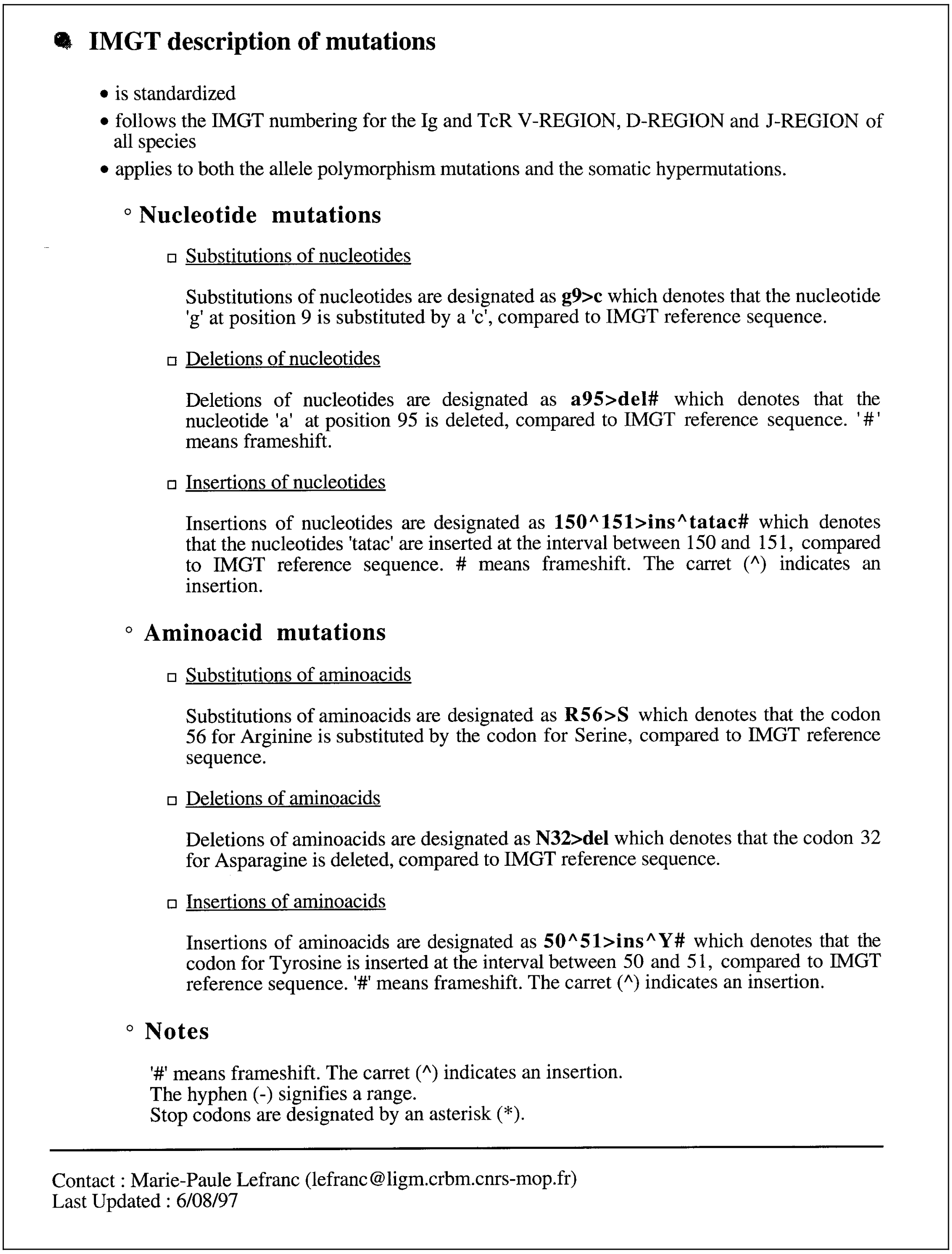
° Allele polymorphism mutations

Allele polymorphisms of Ig and TcR V-REGIONs, D-REGIONs and J-REGIONs are described by comparison to IMGT reference sequences (accession numbers in bold in the allele tables). The reference sequences have been chosen on the basis of one or, whenever possible, several of the following criteria: first sequence published, longest sequence and mapped sequence.
Allele names comprise the IMGT gene name followed by an asterisk and a two-figure number. For the reference sequences, the number is *01; other alleles are designated by increasing numbers (*02, *03, ...) based, if possible, on chronological order of their publication, and/or confirmation of data by different authors.
Definitive allele designation needs to be confirmed by genetic studies and/or different sources of data (indicated by "+" in the allele tables). In the allele tables, the description of nucleotide mutations and that of the corresponding aminoacid changes are separated by a comma. Nucleotide and aminoacid mutations concerning a given codon are shown in a single subdivision.

° Somatic hypermutations

Somatic hypermutations can be described according to the IMGT unique numbering when the corresponding germline V-GENE (or V-GENE allele), D-SEGMENT (or D-SEGMENT allele) and J-SEGMENT (or J-SEGMENT allele) have been identified.

Document 2 (continued next page)



IMGT description of mutations

- is standardized
- follows the IMGT numbering for the Ig and TcR V-REGION, D-REGION and J-REGION of all species
- applies to both the allele polymorphism mutations and the somatic hypermutations.

° Nucleotide mutations

□ Substitutions of nucleotides

Substitutions of nucleotides are designated as **g9>c** which denotes that the nucleotide 'g' at position 9 is substituted by a 'c', compared to IMGT reference sequence.

□ Deletions of nucleotides

Deletions of nucleotides are designated as **a95>del#** which denotes that the nucleotide 'a' at position 95 is deleted, compared to IMGT reference sequence. '#' means frameshift.

□ Insertions of nucleotides

Insertions of nucleotides are designated as **150^151>ins^tatac#** which denotes that the nucleotides 'tatac' are inserted at the interval between 150 and 151, compared to IMGT reference sequence. # means frameshift. The carret (^) indicates an insertion.

° Aminoacid mutations

□ Substitutions of aminoacids

Substitutions of aminoacids are designated as **R56>S** which denotes that the codon 56 for Arginine is substituted by the codon for Serine, compared to IMGT reference sequence.

□ Deletions of aminoacids

Deletions of aminoacids are designated as **N32>del** which denotes that the codon 32 for Asparagine is deleted, compared to IMGT reference sequence.

□ Insertions of aminoacids

Insertions of aminoacids are designated as **50^51>ins^Y#** which denotes that the codon for Tyrosine is inserted at the interval between 50 and 51, compared to IMGT reference sequence. '#' means frameshift. The carret (^) indicates an insertion.

° Notes

'#' means frameshift. The carret (^) indicates an insertion.
The hyphen (-) signifies a range.
Stop codons are designated by an asterisk (*).

Contact : Marie-Paule Lefranc (lefranc@ligm.crbm.cnrs-mop.fr)
Last Updated : 6/08/97

Document 2

Acknowledgements

IMGT is funded by the European Union's BIOTECH programme, the CNRS (Centre National de la Recherche Scientifique) and the MENRT (Ministère de l'Education Nationale, de la Recherche et de la Technologie). Subventions have been received from ARC (Association pour la Recherche sur le Cancer), ARP (Association de Recherche sur la Polyarthrite), FRM (Fondation pour la Recherche Médicale) and the Région Languedoc-Roussillon.

References

1 Giudicelli V, Chaume D, Bodmer J, Müller W, Busin C, Marsh S, Bontrop R, Lemaitre M, Malik A, Lefranc MP: IMGT, the international ImMunoGeneTics database. Nucleic Acids Res 1997;25:206–211.

2 Lefranc MP, Giudicelli V, Busin C, Bodmer J, Müller W, Bontrop R, Lemaitre M, Malik A, Chaume D: IMGT, the international ImMunoGeneTics database. Nucleic Acids Res 1998;26:297–303.

3 Lefranc MP: Unique database numbering system for immunogenetics. Immunol Today 1997;18:509.

4 Kabat EA, Wu TT, Perry HM, Gottesman KS, Foeller C: Sequences of Proteins of Immunological Interest. Bethesda, National Institute of Health, 1991.

5 Satow Y, Cohen GH, Padlan EA, Davies DR: Phosphocholine binding immunoglobulin Fab McPC603. J Mol Biol 1986;190:593–604.

6 Chothia C, Lesk AM: Canonical structures for the hypervariable regions of immunoglobulins. J Mol Biol 1987;196:901–917.