IMGT/Automat: the strategy for the annotation of human and mouse cDNA nucleotide sequences of IG and TR

Géraldine Folch, Joumana Jabado-Michaloud, Fatena Bellahcene, Laetitia Regnier, Véronique Giudicelli and Marie-Paule Lefranc

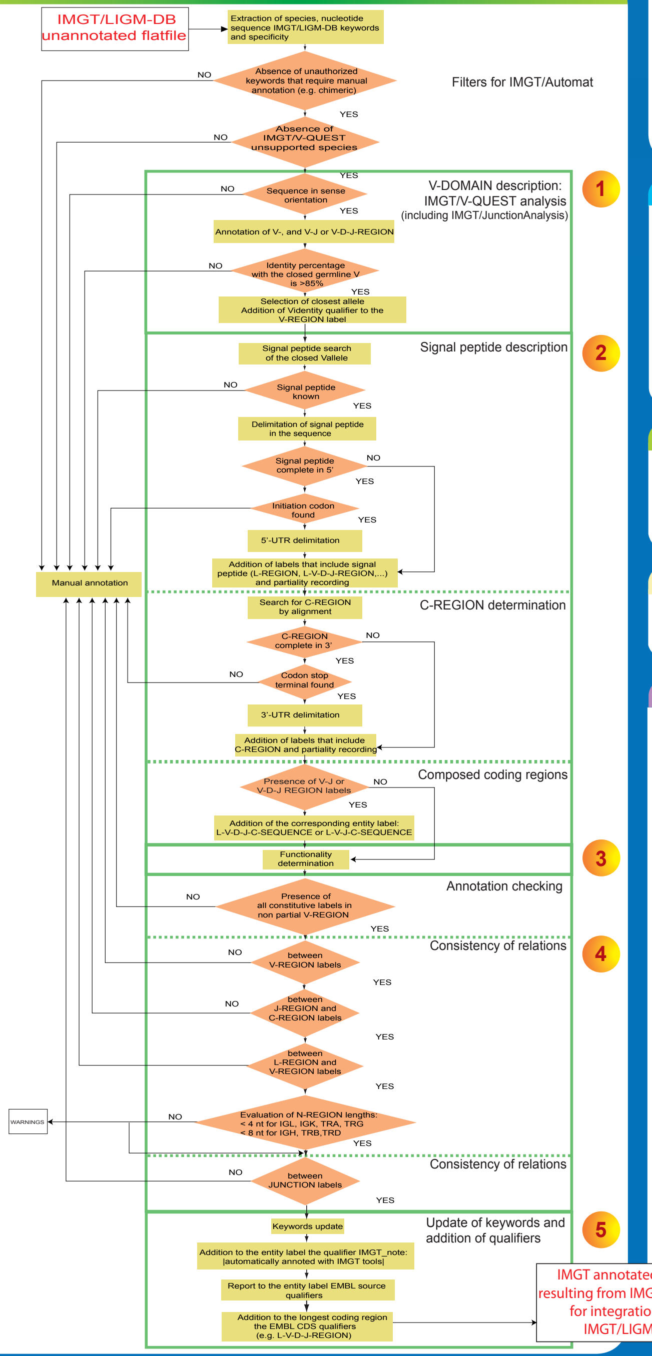
IMGT®, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université Montpellier 2, Institut de Génétique Humaine IGH, UPR CNRS 1142, 141 rue de la Cardonille, F-34396 Montpellier cedex 05, France

http://www.imgt.org

The cDNA sequences of immunoglobulins (IG) and T cell receptors (TR) represent more than one half of the sequences in the IMGT® nucleotide database IMGT/LIGM-DB [1] and 75% of them are from human and mouse. A few cDNA are germline but the great majority results from a V-D-J or V-J gene rearrangement, spliced to a C gene. The IG and TR genes have been studied extensively in IMGT® (http://www.imgt.org) [2], which allowed to set up their nomenclature and the corresponding germline reference sequences. These standardized reference directory sets (one for each group of each locus) and the IMGT-ONTOLOGY axioms and derived concepts [3] are the key elements indispensable to perform the annotation of IG and TR cDNA sequences. A Java program, IMGT/Automat [4], was developed by IMGT®, to automatically annotate the IG and TR cDNA sequences and to produce a totally automatic and complete annotation. More than 9,000 human and mouse cDNA have already been successfully automatically annotated. The quality of the cDNA automatic annotation is equivalent to the quality of the annotation achieved by a human expert. The IMGT® strategy is currently the only way, in the field of immunogenetics, to guarantee the annotation quality and the management of an always increasing number of IG and TR cDNA nucleotide sequences.

[1] Giudicelli V. et al. Nucleic Acids Res., 34, D781-784 (2006).
[2] Lefranc M.-P. et al. Nucleic Acids Res., 37, D1006-1012 (2009).
[3] Duroux P. et al. Biochimie, 90, 570-583 (2008).
[4] Giudicelli V. et al. Stud. Health Technol. Inform., 116, 3-8 (2005).
[5] Brochet X. et al. Nucleic Acids Res., 36, W503-508 (2008).
[6] Yousfi Monod M. et al. Bioinformatics, 20, i379-385 (2004).

IMGT/Automat main tasks


IMGT/LIGM-DB unannotated flatfile
Extraction of species, nucleotide sequence IMGT/LIGM-DB keywords and specificity
Absence of unauthorized keywords that require manual annotation (e.g. chimeric)
NO
YES
Filters for IMGT/Automat
Absence of IMGT/V-QUEST unsupported species
NO
YES
V-DOMAIN description: IMGT/V-QUEST analysis (including IMGT/JunctionAnalysis)
1
Sequence in sense orientation
NO
YES
Annotation of V-, and V-J or V-D-J-REGION
Identity percentage with the closed germline V is >85%
NO
YES
Selection of closest allele Addition of Videntity qualifier to the V-REGION label
Signal peptide description
2
Signal peptide search of the closed Vallele
Signal peptide known
NO
YES
Delimitation of signal peptide in the sequence
Signal peptide complete in 5'
NO
YES
Initiation codon found
YES
5'-UTR delimitation
Addition of labels that include signal peptide (L-REGION, L-V-D-J-REGION,...) and partiality recording
Manual annotation
C-REGION determination
Search for C-REGION by alignment
C-REGION complete in 3'
NO
YES
Codon stop terminal found
NO
YES
3'-UTR delimitation
Addition of labels that include C-REGION and partiality recording
Composed coding regions
Presence of V-J or V-D-J REGION labels
NO
YES
Addition of the corresponding entity label: L-V-D-J-C-SEQUENCE or L-V-J-C-SEQUENCE
3
Functionality determination
Annotation checking
Presence of all constitutive labels in non partial V-REGION
NO
YES
Consistency of relations
4
between V-REGION labels
NO
YES
between J-REGION and C-REGION labels
NO
YES
between L-REGION and V-REGION labels
NO
YES
WARNINGS
Evaluation of N-REGION lengths: < 4 nt for IGL, IGK, TRA, TRG < 8 nt for IGH, TRB,TRD
NO
YES
Consistency of relations
between JUNCTION labels
NO
YES
Update of keywords and addition of qualifiers
5
Keywords update
Addition to the entity label the qualifier IMGT_note: |automatically annoted with IMGT tools|
Report to the entity label EMBL source qualifiers
Addition to the longest coding region the EMBL CDS qualifiers (e.g. L-V-D-J-REGION)
IMGT annotated flat file resulting from IMGT/Automat for integration into IMGT/LIGM-DB

IMGT/Automat includes five main tasks:
In a first step IMGT/Automat implements IMGT/V-QUEST [5] . The description of the V-D-J and V-J junction is performed by the IMGT/JunctionAnalysis [6] tool. In a second step, IMGT/Automat delimits the signal peptide, the constant region and the composed coding regions (for example: L-V-D-J-C-REGION). In a third step, the functionality of the sequence (a concept of identification) is defined. The fourth step corresponds to a thorough annotation checking. In a fifth and final step, keywords are updated and qualifiers on biological origin and methodology used (concepts of obtention) are integrated, and the annotated flat file is generated.

1 V-DOMAIN description: IMGT/V-QUEST Analysis (including IMGT/JunctionAnalysis)

V-DOMAIN description (V-J-REGION and V-D-J-REGION is performed by IMGT/V-QUEST analysis. Detailed analysis of JUNCTION is performed by the integrated IMGT/JunctionAnalysis tool

Alignment for V-GENE and allele identification

```
                               <--------------------------------- FR1-IMG
BC024289                       gaggtgcagctggtggagtctgggga...ggcctggtcaagc
AB019439 IGHV3-21*01           ---------------------------...------------
M99658  IGHV3-21*02            --------a------------------...------------
M99675  IGHV3-48*01            ---------------------------...---t----ac--
AB019438 IGHV3-48*02           ---------------------------...---t----ac--
AJ879484 IGHV3-h*01(P)         ---------------------------...---t----a---
```

Alignment for J-GENE and allele identification

```
BC024289                       tctccgccagctaacttcctactggtacttcgatctctggggc
J00256 IGHJ2*01                .................-------------------------
M25625 IGHJ4*03                .......................-c-----t--cta------
```

Results of IMGT/JunctionAnalysis

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

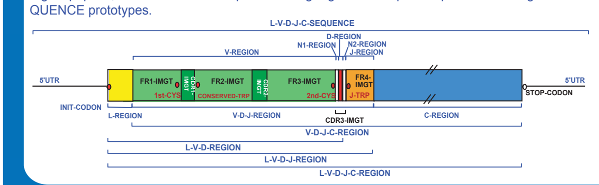
```
Input     V name      3'V-REGION   N1           D-REGION              N2
BC024289  IGHV3-21*01 tgtgcgagaga  t   .............tccgccagcta.......  acttc

Input          5'J-REGION          J name    D name       Vmut  Dmut  Jmut  Ngc
BC024289   ctactggtacttcgatctctgg  IGHJ2*01  IGHD3-10*01    0     4     0    3/7
```

IMGT/V-QUEST analysis provides:

- **Identification** of the sequence (chain type for ex: IG-Heavy)
- **Classification** of the V, D, J genes and alleles
- **Description** of the IG and TR specific constitutive motifs
- **Delimitation** of the framework regions (FR-IMGT) and complementarity determining regions (CDR-IMGT)
- **Numbering** of the codons

2 Signal peptide, C-REGION and composed coding regions

Signal peptide, C-REGION and composed coding regions description is performed using the L-V-J-C-SEQUENCE and L-V-D-J-C-SEQUENCE prototypes.



Composed coding regions for:
- Entity : L-V-D-J-C-SEQUENCE
L-REGION, L-V-D-REGION, L-V-D-J-REGION, L-V-D-J-C-REGION, V-D-J-C-REGION, C-REGION

Composed coding regions for:
- Entity : L-V-J-C-SEQUENCE
L-REGION, L-V-J-REGION, L-V-J-C-REGION, V-J-C-REGION, C-REGION

3 Functionality determination

The functionality of the sequence is defined according to the biological rules of the IMGT Scientific chart.

The sequence is PRODUCTIVE if the coding region 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.

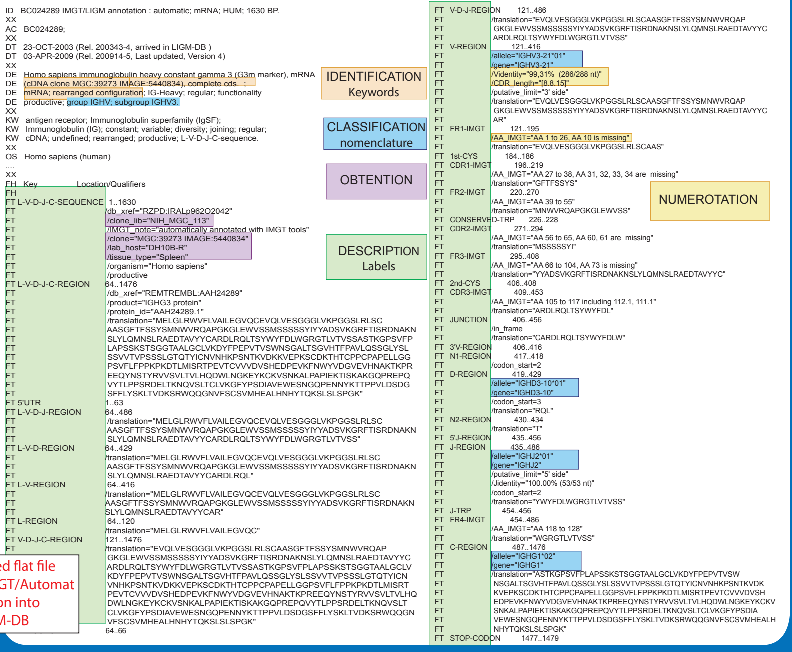
The sequence is UNPRODUCTIVE if the JUNCTION is out-of-frame and/or the presence of stop codon(s) and/or frameshift mutation(s), and/or a defect described in the splicing sites and/or the regulatory element(s), and/or unusual features (TRANSLOCATED, GENE FUSION...) and/or changes of conserved

4 Annotation checking

Annotation checking comprises several steps (see figure), for examples:

Presence of all constitutive labels by comparison with the prototype (e.g. I-REGION, V-REGION, D-REGION, ...)
Consistency of relations between labels (e.g. L-REGION adjacent_in_its_3_prime_with V-REGION, FR1-IMGT is_included_with_same_5_prime_in V-REGION)
.

5 Annotated IMGT FLAT-FILE resulting from IMGT/Automat


IDENTIFICATION
Keywords
CLASSIFICATION
nomenclature
OBTENTION
DESCRIPTION
Labels
NUMEROTATION

```
ID   BC024289 IMGT/LIGM annotation : automatic; mRNA; HUM; 1630 BP.
XX
AC   BC024289;
XX
DT   23-OCT-2003 (Rel. 200343-4, arrived in LIGM-DB )
DT   03-APR-2009 (Rel. 200914-5, Last updated, Version 4)
XX
DE   Homo sapiens immunoglobulin heavy constant gamma 3 (G3m marker), mRNA
DE   (cDNA clone MGC:39273 IMAGE:5440834), complete cds.  ;
DE   mRNA; rearranged configuration; IG-Heavy; regular; functionality
DE   productive; group IGHV; subgroup IGHV3.
XX
KW   antigen receptor; Immunoglobulin superfamily (IgSF);
KW   Immunoglobulin (IG); constant; variable; diversity; joining; regular;
KW   cDNA; undefined; rearranged; productive; L-V-D-J-C-sequence.
XX
OS   Homo sapiens (human)
....
XX
FH   Key                Location/Qualifiers
FH
FT L-V-D-J-C-SEQUENCE  1..1630
FT                     /db_xref="RZPD:IRALp962O2042"
FT                     /clone_lib="NIH_MGC_113"
FT                     /IMGT_note="automatically annotated with IMGT tools"
FT                     /clone="MGC:39273 IMAGE:5440834"
FT                     /lab_host="DH10B-R"
FT                     /tissue_type="Spleen"
FT                     /organism="Homo sapiens"
FT                     /productive
FT L-V-D-J-C-REGION    64..1476
FT                     /db_xref="REMTREMBL:AAH24289"
FT                     /product="IGHG3 protein"
FT                     /protein_id="AAH24289.1"
FT                     /translation="MELGLRWVFLVAILEGVQCEVQLVESGGGLVKPGGSLRLSC
FT                     AASGFTFSSYSMNWVRQAPGKGLEWVSSMSSSSSYIYYADSVKGRFTISRDNAKN
FT                     SLYLQMNSLRAEDTAVYYCARDLRQLTSYWYFDLWGRGTLVTVSSASTKGPSVFP
FT                     LAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSL
FT                     SSVVTVPSSSLGTQTYICNVNHKPSNTKVDKKVEPKSCDKTHTCPPCPAPELLGG
FT                     PSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEVKFNWYVDGVEVHNAKTKPR
FT                     EEQYNSTYRVVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQPREPQ
FT                     VYTLPPSRDELTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYKTTPPVLDSDG
FT                     SFFLYSKLTVDKSRWQQGNVFSCSVMHEALHNHYTQKSLSLSPGK"
FT 5'UTR               1..63
FT L-V-D-J-REGION      64..486
FT                     /translation="MELGLRWVFLVAILEGVQCEVQLVESGGGLVKPGGSLRLSC
FT                     AASGFTFSSYSMNWVRQAPGKGLEWVSSMSSSSSYIYYADSVKGRFTISRDNAKN
FT                     SLYLQMNSLRAEDTAVYYCARDLRQLTSYWYFDLWGRGTLVTVSS"
FT L-V-D-REGION        64..429
FT                     /translation="MELGLRWVFLVAILEGVQCEVQLVESGGGLVKPGGSLRLSC
FT                     AASGFTFSSYSMNWVRQAPGKGLEWVSSMSSSSSYIYYADSVKGRFTISRDNAKN
FT                     SLYLQMNSLRAEDTAVYYCARDLRQL"
FT L-V-REGION          64..416
FT                     /translation="MELGLRWVFLVAILEGVQCEVQLVESGGGLVKPGGSLRLSC
FT                     AASGFTFSSYSMNWVRQAPGKGLEWVSSMSSSSSYIYYADSVKGRFTISRDNAKN
FT                     SLYLQMNSLRAEDTAVYYCAR"
FT L-REGION            64..120
FT                     /translation="MELGLRWVFLVAILEGVQC"
FT V-D-J-C-REGION      121..1476
FT                     /translation="EVQLVESGGGLVKPGGSLRLSCAASGFTFSSYSMNWVRQAP
                       GKGLEWVSSMSSSSSYIYYADSVKGRFTISRDNAKNSLYLQMNSLRAEDTAVYYC
                       ARDLRQLTSYWYFDLWGRGTLVTVSSASTKGPSVFPLAPSSKSTSGGTAALGCLV
                       KDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYICN
                       VNHKPSNTKVDKKVEPKSCDKTHTCPPCPAPELLGGPSVFLFPPKPKDTLMISRT
                       PEVTCVVVDVSHEDPEVKFNWYVDGVEVHNAKTKPREEQYNSTYRVVSVLTVLHQ
                       DWLNGKEYKCKVSNKALPAPIEKTISKAKGQPREPQVYTLPPSRDELTKNQVSLT
                       CLVKGFYPSDIAVEWESNGQPENNYKTTPPVLDSDGSFFLYSKLTVDKSRWQQGN
                       VFSCSVMHEALHNHYTQKSLSLSPGK"
                       64..66

FT   V-D-J-REGION      121..486
FT                     /translation="EVQLVESGGGLVKPGGSLRLSCAASGFTFSSYSMNWVRQAP
FT                      GKGLEWVSSMSSSSSYIYYADSVKGRFTISRDNAKNSLYLQMNSLRAEDTAVYYC
FT                      ARDLRQLTSYWYFDLWGRGTLVTVSS"
FT   V-REGION           121..416
FT                     /allele="IGHV3-21*01"
FT                     /gene="IGHV3-21"
FT                     /Videntity="99,31%  (286/288 nt)"
FT                     /CDR_length="[8.8.15]"
FT                     /putative_limit="3' side"
FT                     /translation="EVQLVESGGGLVKPGGSLRLSCAASGFTFSSYSMNWVRQAP
FT                      GKGLEWVSSMSSSSSYIYYADSVKGRFTISRDNAKNSLYLQMNSLRAEDTAVYYC
FT                      AR"
FT   FR1-IMGT           121..195
FT                     /AA_IMGT="AA 1 to 26, AA 10 is missing"
FT                     /translation="EVQLVESGGGLVKPGGSLRLSCAAS"
FT   1st-CYS           184..186
FT   CDR1-IMGT          196..219
FT                     /AA_IMGT="AA 27 to 38, AA 31, 32, 33, 34 are  missing"
FT                     /translation="GFTFSSYS"
FT   FR2-IMGT           220..270
FT                     /AA_IMGT="AA 39 to 55"
FT                     /translation="MNWVRQAPGKGLEWVSS"
FT   CONSERVED-TRP      226..228
FT   CDR2-IMGT          271..294
FT                     /AA_IMGT="AA 56 to 65, AA 60, 61 are  missing"
FT                     /translation="MSSSSSYI"
FT   FR3-IMGT           295..408
FT                     /AA_IMGT="AA 66 to 104, AA 73 is missing"
FT                     /translation="YYADSVKGRFTISRDNAKNSLYLQMNSLRAEDTAVYYC"
FT   2nd-CYS           406..408
FT   CDR3-IMGT          409..453
FT                     /AA_IMGT="AA 105 to 117 including 112.1, 111.1"
FT                     /translation="ARDLRQLTSYWYFDL"
FT   JUNCTION           406..456
FT                     /in_frame
FT                     /translation="CARDLRQLTSYWYFDLW"
FT   3'V-REGION         406..416
FT   N1-REGION          417..418
FT                     /codon_start=2
FT   D-REGION           419..429
FT                     /allele="IGHD3-10*01"
FT                     /gene="IGHD3-10"
FT                     /codon_start=3
FT                     /translation="RQL"
FT   N2-REGION          430..434
FT                     /translation="T"
FT   5'J-REGION         435..456
FT   J-REGION           435..486
FT                     /allele="IGHJ2*01"
FT                     /gene="IGHJ2"
FT                     /putative_limit="5' side"
FT                     /Jidentity="100.00% (53/53 nt)"
FT                     /codon_start=2
FT                     /translation="YWYFDLWGRGTLVTVSS"
FT   J-TRP             454..456
FT   FR4-IMGT           454..486
FT                     /AA_IMGT="AA 118 to 128"
FT                     /translation="WGRGTLVTVSS"
FT   C-REGION           487..1476
FT                     /allele="IGHG1*02"
FT                     /gene="IGHG1"
FT                     /translation="ASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSW
FT                      NSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKVDK
FT                      KVEPKSCDKTHTCPPCPAPELLGGPSVFLFPPKPKDTLMISRTPEVTCVVVDVSH
FT                      EDPEVKFNWYVDGVEVHNAKTKPREEQYNSTYRVVSVLTVLHQDWLNGKEYKCKV
FT                      SNKALPAPIEKTISKAKGQPREPQVYTLPPSRDELTKNQVSLTCLVKGFYPSDIA
FT                      VEWESNGQPENNYKTTPPVLDSDGSFFLYSKLTVDKSRWQQGNVFSCSVMHEALH
FT                      NHYTQKSLSLSPGK"
FT   STOP-CODON         1477..1479
```

IMGT founder and director: Marie-Paule Lefranc (Marie-Paule.Lefranc@igh.cnrs.fr)
Bioinformatics manager: Véronique Giudicelli (Veronique.Giudicelli@igh.cnrs.fr)
Computer manager: Patrice Duroux (Patrice.Duroux@igh.cnrs.fr)
Interface design: Chantal Ginestoux (Chantal.Ginestoux@igh.cnrs.fr)

IMGT® is funded by CNRS, MESR and EU ImmunoGrid IST-2004-028069