IMMUNOME RESEARCH

RESEARCH Open Access

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

Eltaf Alamyar [1], Véronique Giudicelli [1], Shuo Li [2], Patrice Duroux [1], Marie-Paule Lefranc[1§]

ABSTRACT

Background

The number of antigen receptors, immunoglobulins (IG) or antibodies and T cell receptors (TR) of the adaptive immune response in vertebrates with jaws is almost unlimited (2.10^{12} per individual in humans). This huge diversity results from complex mechanisms in the synthesis of the variable (V) domains, that include DNA molecular rearrangements of the V, diversity (D) and joining (J) genes, N-diversity at the V-(D)-J junctions and, for IG, somatic hypermutations. The specificity of the V domains is conferred by the complementarity determining regions (CDR) and more particularly the CDR3. IMGT®, the international ImMunoGeneTics information system®, has developed online tools that provide a detailed and accurate sequence analysis of the V domains (IMGT/V-QUEST) and CDR3 (IMGT/JunctionAnalysis), based on IMGT-ONTOLOGY. However online analyses are limited to 50 sequences per batch. The challenge was to provide identical high-quality analysis for the huge number of sequences obtained by Next Generation Sequencing (NGS) high throughput and deep sequencing.

Results

IMGT® has developed IMGT/HighV-QUEST that analyses up to 150,000 IG or TR V domain sequences per batch and performs statistical analysis on the results of up to 450,000 sequences. IMGT/HighV-QUEST provides users with: (i) a friendly web interface for submission and results retrieval, (ii) high-quality detailed results of IMGT/V-QUEST and IMGT/JunctionAnalysis, based on the IMGT-ONTOLOGY concepts and IMGT Scientific chart rules, (iii) a standardized frame for NGS statistical analysis, based on 'Results category' ('1 copy', 'More than 1', 'single allele', 'several alleles (or genes)', (iv) detailed standardized statistical analysis tables and histograms (e.g., V, D and J usage, CDR3-IMGT lengths).

Conclusions

IMGT/HighV-QUEST has been freely available for use for academics on the IMGT® Home page (http://www.imgt.org) since October 2010. More than 123 million sequences were submitted during its first year. IMGT/HighV-QUEST is a key component for establishing reliable repertoires of IG and TR V domains. These repertoires will contribute to the individual immunoprofiles in diverse immune situations and on different B and T cell populations. They will also contribute to characterize potential therapeutic antibodies from combinatorial libraries.

BACKGROUND

The number of the antigen receptors, immunoglobulins (IG) or antibodies and T cell receptors (TR) of the adaptive immune response in vertebrates with jaws (or gnathostomata) is almost unlimited. In humans, the potential 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 [1,2]. This huge diversity is inherent to the particularly complex and unique molecular synthesis and genetics of the antigen receptor chains [3]. This includes biological mechanisms such as DNA molecular rearrangements of the variable (V), diversity (D) and joining (J) genes (combinatorial diversity) in multiple loci (three for IG and four for TR in humans) located on different chromosomes (four in humans), nucleotide deletions and insertions at the rearrangement junctions (or N-diversity) and, for IG, somatic hypermutations (for review, see [1,2]).

IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org) [3-5], was created in 1989 by Marie-Paule Lefranc, Laboratoire d'ImmunoGénétique Moléculaire LIGM (Université Montpellier 2 and CNRS) at Montpellier, France, in order to standardize and to manage the complexity of immunogenetics data. IMGT® has reached that goal through the building of a unique ontology, IMGT-ONTOLOGY [6-11], the first ontology in immunogenetics and immunoinformatics.

[1] 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, 34396, Montpellier, Cedex 05, France,
[2] Macfarlane Burnet Institute for Medical Research and Public Health, Melbourne,
Victoria, Australia

[§]Corresponding author
Email addresses:
EA: Eltaf.Alamyar@igh.cnrs.fr
VG: Veronique.Giudicelli@igh.cnrs.fr
SL: Shuo.Li@burnet.edu.au
M-PL: Marie-Paule.Lefranc@igh.cnrs.fr

IMGT-ONTOLOGY is now acknowledged as the global reference in immunogenetics and immunoinformatics, allowing IMGT® to bridge biological and computational spheres in bioinformatics [10]. IMGT-ONTOLOGY manages the immunogenetics knowledge through diverse facets that rely on the seven axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope: "IDENTIFICATION", "DESCRIPTION", "CLASSIFICATION", "NUMEROTATION", "LOCALIZATION", "ORIENTATION" and "OBTENTION" [9-11]. These axioms postulate that any object, any process and any relation has to be identified, described, classified, numbered, localized and orientated, and that the way it is obtained can be characterized [9-11]. From these axioms, concepts were generated that led to the IMGT Scientific chart rules (http://www.imgt.org): standardized keywords (concepts of identification) [12], standardized labels (concepts of description) [13], standardized gene and allele name and nomenclature (concepts of classification) [14], IMGT unique numbering and IMGT Colliers de Perles (concepts of numerotation) [15-17]. Owing to that standardization, IMGT® has become an internationally acknowledged high-quality integrated knowledge resource that comprises several databases for sequences (e.g., IMGT/LIGM-DB [18]), genes (IMGT/GENE-DB [19]), two-dimensional (2D) and three-dimensional (3D) structures (IMGT/2Dstructure-DB, IMGT/3Dstructure-DB [20-22]), monoclonal antibodies and fusion proteins for immune applications (FPIA) (IMGT/mAb-DB [23]), seventeen tools for analysis of nucleotide sequences (IMGT/V-QUEST [24-26], IMGT/JunctionAnalysis [27,28]), amino acid sequences (IMGT/DomainGapAlign [29]), genes and structure analysis [30], and >15,000 pages of web resources [3-5]. Among the nucleotide sequence analysis tool, IMGT/V-QUEST [24-26] is the most popular one, as it allows the standardized identification and very detailed description of any rearranged IG or TR sequence of human, mouse and rat. It constantly evolves with other species being added, following the IMGT® annotation of IG and TR loci of newly sequenced genomes. IMGT/V-QUEST provides a detailed and accurate characterization of the submitted IG and TR sequences entirely based on the IMGT-ONTOLOGY concepts of identification, description, classification and numerotation [12-17]. It identifies the V, D and J genes and alleles in rearranged V-J and V-D-J sequences by alignment with the germline IG and TR gene and allele sequences of the IMGT reference directory from IMGT/GENE-DB. It delimits the framework regions (FR-IMGT) and complementarity determining regions (CDR-IMGT) according to the IMGT unique numbering for V domain [15]. The tool describes the V-REGION mutations and identifies the hot spot positions in the closest germline V gene. It detects and accurately describes insertions and deletions in the submitted sequences by reference to the IMGT unique numbering [15]. IMGT/V-QUEST integrates IMGT/JunctionAnalysis [27,28] for a detailed analysis of the V-J and V-D-J junctions (it identifies the D-REGION if present, the nucleotides (nt) deleted as a result of exonuclease trimming and the nontemplated N-REGION nucleotides added at random by the terminal deoxynucleotide transferase TdT), and uses IMGT/Automat [31,32] for a full annotation of the V-J- and V-D-J-REGION.

In the context of Next Generation Sequencing (NGS) [33-38], computational power is required in order to be able to analyze huge amounts of data in less time with IMGT® tools. Although High Performance Computing (HPC) Systems are designed to solve advanced computational problems that are highly challenging, complex and time consuming, this means facing the problematic of the computational aspects, working with different HPC technologies and dealing constantly with changing hardware fronted by diverse operating systems. Even with a similar operating system the different aspects, like jobs queue, are not the same. The challenge for IMGT® in providing IMGT/V-QUEST high-quality results for the analysis of IG and TR sequences from NGS high throughput and deep sequencing was to create a friendly and system-neutral environment in which, from the user's point of view, the distributed character and heterogeneity of the computational system components is transparent. To reach that goal, IMGT® has developed IMGT/HighV-QUEST [39-41], the first web portal for IG and TR analysis from NGS high throughput and deep sequencing and a secure system destined to run a standalone version of IMGT/V-QUEST on remote computational resources.

RESULTS

IMGT/HighV-QUEST user friendly interface for analysis submission

IMGT/HighV-QUEST's users are from different scientific backgrounds. In order to let all users use conveniently the tool, a simple interface was developed. An analysis submission is an easy task: the user goes to IMGT/HighV-QUEST Search page [39-41] by clicking on the link on the menu bar. On the search page (Figure 1) the user gives a title for his analysis, selects the species, the antigen receptor type (IG or TR) (or the locus, for instance, IGH or TRB). The user uploads a file with the sequences to be analysed in FASTA format (up to 150,000 sequences per file). The user can choose to be notified by e-mail of the advancements of the analysis (when the analysis is queued, when it is submitted (dispatched) on computational resources and/or when it is completed). By clicking on 'Start', the analysis is performed with the default parameters.

Prior to submitting the analysis, the user may customize the results display options in 'Display results' (these options are identical to those of IMGT/V-QUEST [26]) (Figure 1). The 'Display results' comprises: 'A. Detailed view' for the display of the results of each analyzed sequence (with a choice of 13 different results displays) in individual result files. The user can choose to include them or not in the output. If included, the user can choose the 'Nb of nucleotides per line in alignments' (60 by default) and select among 13 results displays as mentioned above. 'B. Files in CSV' for the choice of the CSV files to be retrieved in the final outputs (Summary, nt-sequences and parameters are provided by default). For sophisticated queries or for unusual sequences, the users can modify the default values in 'Advanced parameters'. The customizable values, identical of those of

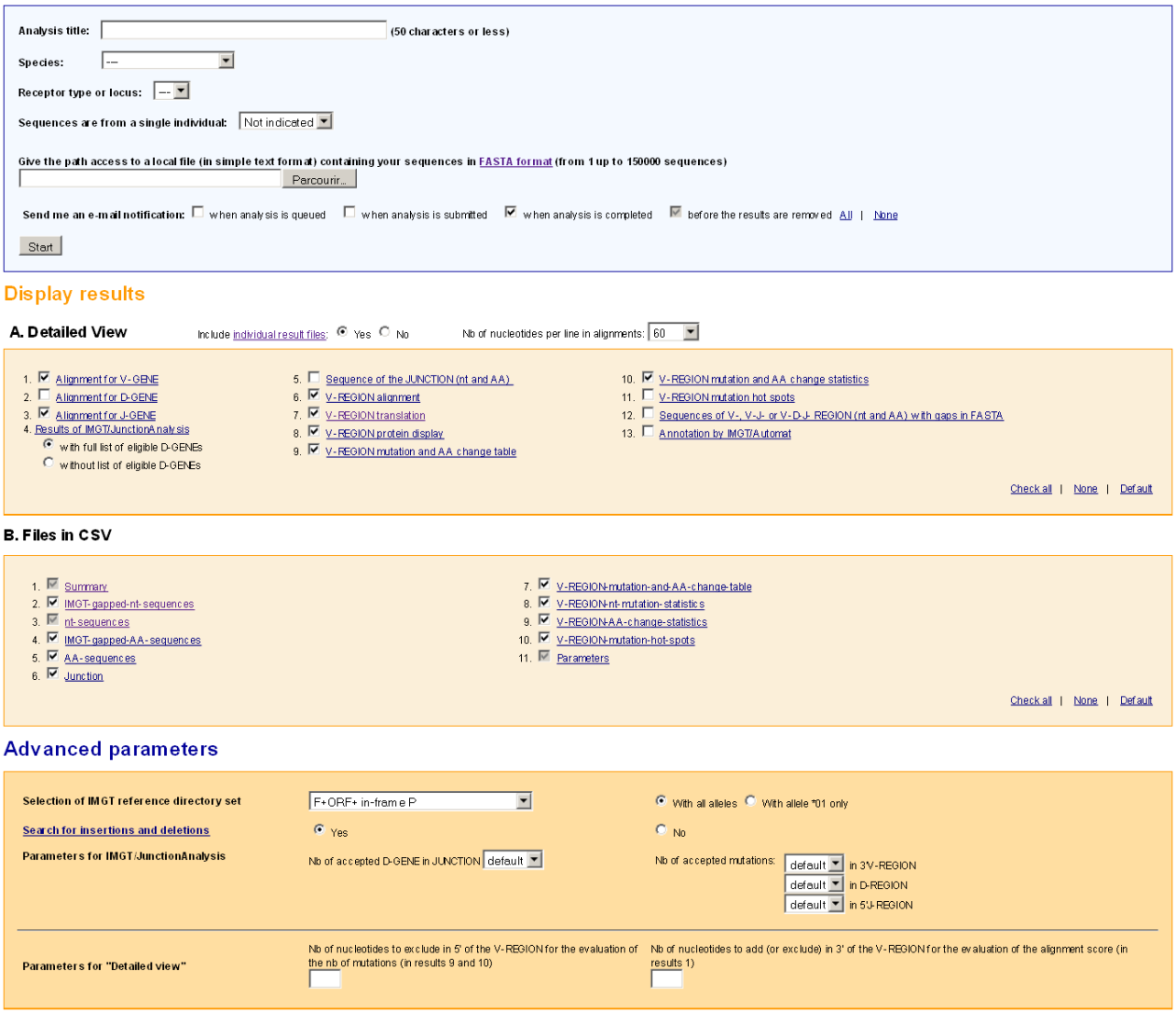


Figure 1 IMGT/HighV-QUEST analysis submission page.

IMGT/V-QUEST [26], are: (i) 'Selection of IMGT reference directory set' used for the V, D and J gene and allele identification and alignments with a choice of four sets ('F+ORF', 'F+ORF+ in-frame P' (by default), 'F+ORF including orphons' and 'F+ORF+ in-frame P including orphons', where F is functional, ORF is open reading frame and P is pseudogene). This allows sequences to be compared with only relevant gene sequences (e.g., orphon sequences are relevant for genomic but not for expressed repertoire studies). The selected set can also be chosen either 'With all alleles' or 'With allele *01 only'. (ii) 'Search for insertions and deletions in V-REGION' is selected by default ('Yes') and can be deactivated if the user does not want to take into account indels in alignment with germline genes and alleles. (iii) 'Parameters for IMGT/JunctionAnalysis': 'Nb of accepted D-GENE in JUNCTION' (provided for the IGH, TRB and TRD junctions) and 'Nb of accepted mutations' in 3'V-REGION, D -REGION and 5'J-REGION (default values are indicated per locus in the IMGT/V-QUEST Documentation and in [26]). (iv) 'Parameters for Detailed View': 'Nb of nucleotides to exclude in 5' of the V-REGION for the evaluation of the nb of mutations' (to avoid, e.g., counting primer specific nucleotides) and/or 'Nb of nucleotides to add (or exclude) in 3' of the V-REGION for the evaluation of the alignment score' (e.g., in case of low or high exonuclease activity).

IMGT/HighV-QUEST analysis life cycle

After submission, the required information is controlled and a popup message will appear if a required field is not filled. After the transfer of the FASTA file on the local web server, a syntax control is also performed in order to let the user correct the syntax before launching the analysis. This will save time for the user by preventing the analysis of syntactically incorrect sequences. The submitted analysis is kept in the local web server analysis queue and dispatched on a remote computational resource when a resource can accept it. This acceptation is based on different criteria, like the number of sequences, free resources, etc. Once the analysis by IMGT/V-QUEST on the remote resource is completed and the results are prepared, the user is notified by an e-mail (if the later has chosen to be informed), and the temporary files and folders are cleaned from the local and remote resources. The analysis results are then kept for 15 days after the analysis completion date and are removed afterwards. Five days before the expiration (or 10 days after the completion date), the user is notified by e-mail of the expiration in 5 days. When an analysis is deleted by the user or be-

cause of its expiration, all user data and results regarding that analysis are removed from the system. However, if the analysis has been chosen in a statistical analysis, it cannot be removed by the user or the system.

IMGT/HighV-QUEST analysis results outputs

The IMGT/HighV-QUEST analysis results outputs comprise a set of text files in two folders (Figure 2): the main folder and, if chosen, the individual result files folder, archived in a single ZIP file.

Figure 2 IMGT/HighV-QUEST analysis outputs.

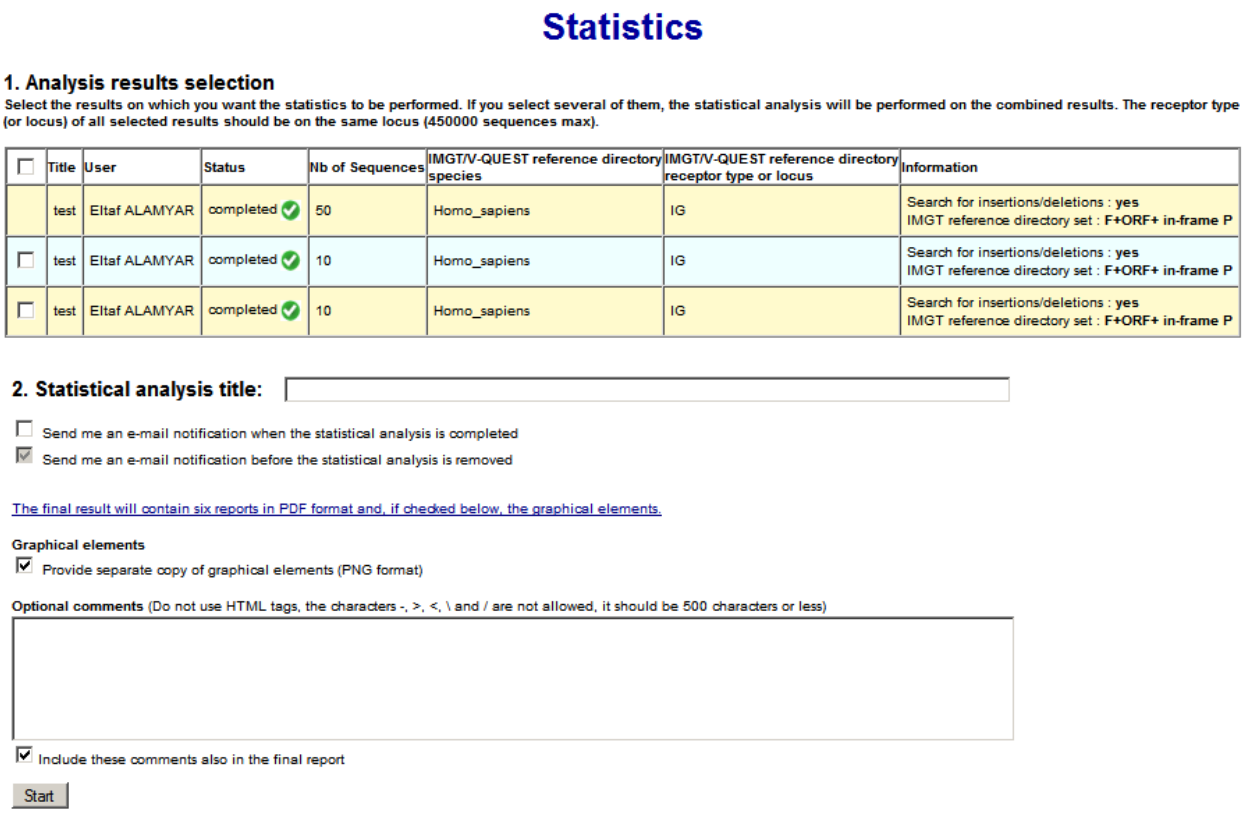
The IMGT/HighV-QUEST main folder includes eleven files (if all selected) in CSV format (results equivalent to those of the Excel file of IMGT/V-QUEST online [26]) that comprise: (i) the 'Summary' file provides the synthesis of the analysis (the sequence functionality, the names of the closest V, D and J genes and alleles with identity percentage, FR-IMGT and CDR-IMGT lengths, amino acid (AA) JUNCTION, the description of insertions and deletions if any), (ii) the 'IMGT-gapped-nt-sequences' file includes the nucleotide (nt) sequences of labels that have been gapped according to the IMGT unique numbering, (iii) the 'Nt-sequences' file includes the ungapped nt sequences of all described labels, (iv) the 'IMGT-gapped-AA-sequences' file includes the AA sequences of labels that have been gapped according to the IMGT unique numbering, (v) the 'AA-sequences' file includes the ungapped AA sequences of labels, (vi) the 'Junction' file includes the results of IMGT/JunctionAnalysis, (vii) the 'V-REGION-mutation-and-AA-change-table' file includes the list of mutations (nt mutations, AA changes, AA class identity (+) or change (-), total for the V-REGION and per FR-IMGT and CDR-IMGT), (viii) the 'V-REGION-nt-mutation-statistics' file includes the number (nb) of nt positions including IMGT gaps, the nb of nt, the nb of identical nt, the total nb of mutations, and then the nb of silent mutations, the nb of nonsilent mutations, the nb of transitions and the nb of transversions, total for the V-REGION and per FR-IMGT and CDR-IMGT, (ix) the 'V-REGION-AA-change-statistics' file includes the nb of AA positions including IMGT gaps, the nb of AA, the nb of identical AA, the total nb of AA changes, and then the nb of AA changes according to AAclassChangeType (e.g., +++) [26], and the nb of AA class changes according to AAclassSimilarityDegree (e.g., Very similar) [26], total for the V-REGION and per FR-IMGT and CDR-IMGT, (x) the 'V-REGION-mutation-hotspots' file indicates the localization of the hot spots motifs detected in the closest germline V-REGION with positions in FR-IMGT and CDR-IMGT, (xi) the 'Parameters' file includes the date of the analysis, the IMGT/V-QUEST version, and the parameters used for the analysis.

The IMGT/HighV-QUEST individual result files folder includes the individual files of all the sequences results (up to 150,000). They allow to visualize the results corresponding to 'Detailed view' for each analysed sequence (results identical to those of IMGT/V-QUEST online in Text; they have been detailed elsewhere [26] and are only briefly described here). Each file comprises: (i) the result summary that summarizes the main characteristics of the analysed sequence with the names of the closest V and J genes and alleles with their alignment score and the percentage of identity, the name of the closest D-gene and allele determined by IMGT/JunctionAnalysis with the D-REGION reading frame, the FR-IMGT and CDR-IMGT lengths and the AA JUNCTION sequence, and if selected, (ii) the Alignment for V, D, J genes and alleles, (iii) the detailed analysis of the JUNCTION by IMGT/JunctionAnalysis, (iv) different displays of the V-REGION, (v) the analysis of the mutations and AA changes, (vi) the localization of the mutation hot spots, and (vii) the annotation by IMGT/Automat.

IMGT/HighV-QUEST statistical analysis submission and life cycle

The IMGT/HighV-QUEST statistical analysis submission is performed by going to the 'Launch statistics' page for which the link is accessible from the menu bar. On this page (Figure 3), the tool gives the user a list of all his current analyses under '1. Analysis results selection'. The user can choose the different analysis results on which he wants to perform the statistical analysis. The analyses must answer the following criteria (available in the table): they should be completed without error or warnings, they should be on the same species and receptor type or locus (e.g., 'Homo sapiens''TRB') and analysed with the same IMGT reference directory set (e.g., 'F+ORF+in-frame P') and with the same indel option (e.g., 'Yes'). The user should verify by himself that the different analysis results were obtained with the same IMGT/HighV-QUEST version, IMGT/V-QUEST version and IMGT/V-QUEST reference directory release for consistency of the statistical analysis. Under '2. Statistical analysis title', the user chooses a title (required for job identification). Clicking on the option Graphical elements allows to obtain, in the outputs, separate graphical elements in PNG format, in order to use them in other documents. The user can add optional comments which will be included in the final reports (this functionality is added here as the output PDF reports are not editable).

Once the 'Start' button is clicked, the job is sent to the local web server.



Statistics

1. Analysis results selection

Select the results on which you want the statistics to be performed. If you select several of them, the statistical analysis will be performed on the combined results. The receptor type (or locus) of all selected results should be on the same locus (450000 sequences max).

	Title	User	Status	Nb of Sequences	IMGT/V-QUEST reference directory species	IMGT/V-QUEST reference directory receptor type or locus	Information
	test	Eltaf ALAMYAR	completed	50	Homo_sapiens	IG	Search for insertions/deletions : yes IMGT reference directory set : F+ORF+ in-frame P
	test	Eltaf ALAMYAR	completed	10	Homo_sapiens	IG	Search for insertions/deletions : yes IMGT reference directory set : F+ORF+ in-frame P
	test	Eltaf ALAMYAR	completed	10	Homo_sapiens	IG	Search for insertions/deletions : yes IMGT reference directory set : F+ORF+ in-frame P

2. Statistical analysis title:

Send me an e-mail notification when the statistical analysis is completed
Send me an e-mail notification before the statistical analysis is removed

The final result will contain six reports in PDF format and, if checked below, the graphical elements.

Graphical elements
Provide separate copy of graphical elements (PNG format)

Optional comments (Do not use HTML tags, the characters -, >, <, \ and / are not allowed, it should be 500 characters or less)

Include these comments also in the final report

Start

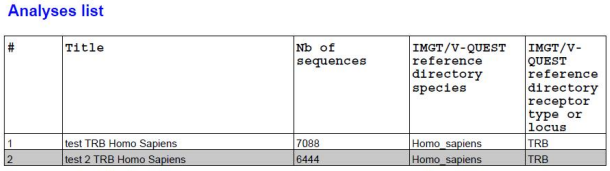
Figure 3 IMGT/HighV-QUEST statistical analysis 'Analysis results selection' table.
The 'Analysis results selection' table shows the list of the analyses with their status. In this table, the user selects the analyses on which he wants to perform the statistical analysis, respecting the criteria defined in the text.

The IMGT/HighV-QUEST statistical analysis life cycle is the following: once on the local web server, the job is queued until a free resource is available to perform the statistical analysis. Once the job is dispatched, it is monitored automatically and regularly until it is completed and then the statistical results are prepared. The completed statistical analysis is kept until 15 days after its completion date on the local web server after which it is removed from the system. Five days before the expiration (or 10 days after the completion date), the user is notified by e-mail of the expiration in 5 days.

IMGT/HighV-QUEST statistical analysis outputs

The IMGT/HighV-QUEST statistical analysis is automatically generated on the 'Summary' and 'Nt-sequences' CSV files of IMGT/HighV-QUEST results and contain the following output items:

1. Comments: if added by the user.
2. Analysis list table (Figure 4): this table recapitulates the list of the analysis results chosen by the user for the statistical analysis.

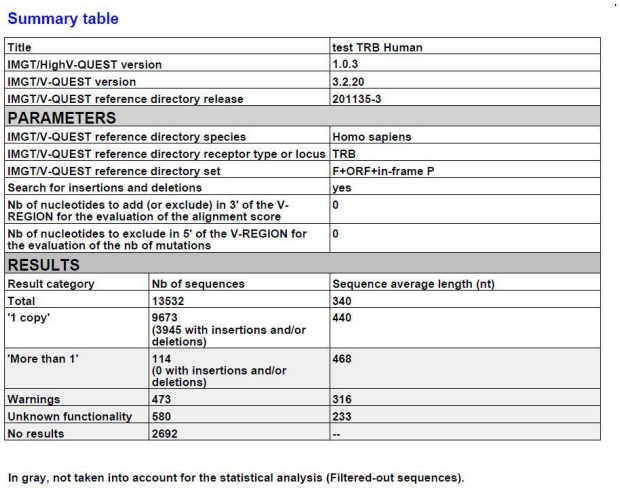


Analyses list

#	Title	Nb of sequences	IMGT/V-QUEST reference directory species	IMGT/V-QUEST reference directory receptor type or locus
1	test TRB Homo Sapiens	7088	Homo_sapiens	TRB
2	test 2 TRB Homo Sapiens	6444	Homo_sapiens	TRB

Figure 4 IMGT/HighV-QUEST statistical analysis 'Analyses list' table.
The 'Analyses list' table recapitulates the list of the analysis results chosen for the statistical analysis. For each of them, it recalls the Title, Nb of sequences, IMGT/V-QUEST reference directory species and IMGT/V-QUEST receptor type or locus.

3. Summary table (Figure 5): this table recalls, but only for the first analysis results in the Analysis list table, the IMGT/HighV-QUEST version, the IMGT/V-QUEST version and the IMGT/V-QUEST reference directory release and the 'PARAMETERS' used for the analyses. The 'RESULTS' section gives the general results of the statistical analysis that lead to 'Result category' with, for



Summary table

Title		test TRB Human
IMGT/HighV-QUEST version		1.0.3
IMGT/V-QUEST version		3.2.20
IMGT/V-QUEST reference directory release		201135-3
PARAMETERS		
IMGT/V-QUEST reference directory species		Homo sapiens
IMGT/V-QUEST reference directory receptor type or locus		TRB
IMGT/V-QUEST reference directory set		F+ORF+in-frame P
Search for insertions and deletions		yes
Nb of nucleotides to add (or exclude) in 3' of the V-REGION for the evaluation of the alignment score		0
Nb of nucleotides to exclude in 5' of the V-REGION for the evaluation of the nb of mutations		0
RESULTS		
Result category	Nb of sequences	Sequence average length (nt)
Total	13532	340
'1 copy'	9673 (3945 with insertions and/or deletions)	440
'More than 1'	114 (0 with insertions and/or deletions)	468
Warnings	473	316
Unknown functionality	580	233
No results	2692	--

In gray, not taken into account for the statistical analysis (Filtered-out sequences).

Figure 5 IMGT/HighV-QUEST statistical analysis 'Summary' table.
The IMGT/HighV-QUEST statistical analysis 'Summary table' indicates the title of the statistical analysis (as entered by the user), recalls the version of IMGT/HighV-QUEST and IMGT/V-QUEST, the IMGT/V-QUEST reference directory release and 'PARAMETERS' used for the analyses and provides in 'RESULTS', an overall view of the statistical analysis results that lead to 'Result category' (see details in Figure 6). This repartition in categories gives the user at first glance an idea of how much he/she can rely on his/her data.

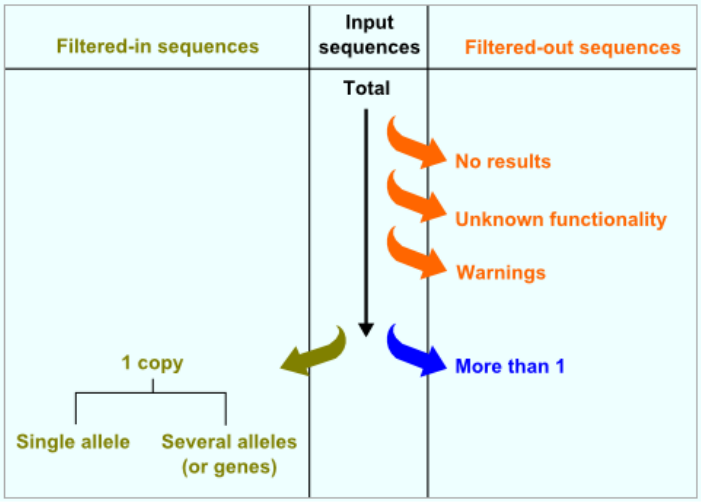


Figure 6 IMGT/HighV-QUEST statistical analysis Result categories.
'1 copy': sequences in one copy, and therefore different by their length and/or their sequence, and retained in 'filtered-in' sequences. For each set of identical sequences, only one copy is retained in '1 copy' and the other redundant sequences for that copy are put into 'More than 1'. The following four categories are excluded from statistical analysis (filtered-out sequences). 'More than 1': redundant identical sequences (after that one copy of each set of identical sequences has been retained in '1 copy'). 'Warnings': sequences with warnings for the V-REGION ('different CDR lengths' and/or 'id<85%'; 'different CDR lengths' means sequences with different AA lengths for CDR1-IMGT and/or CDR2-IMGT compared to the CDR1-IMGT and/or CDR2-IMGT lengths, respectively of the closest identified germline V gene and allele). Unknown functionality: sequences for which no functionality was detected. This category corresponds to the sequences for which the junction cannot be identified (no evidence of rearrangement, no evidence of junction anchors). No results: sequences for which IMGT/HighV-QUEST did not return any result. The statistical analysis is performed on the '1 copy' category divided in two sets, depending on the IMGT/HighV-QUEST result: 'single allele' (only one gene and allele identified by IMGT/HighV-QUEST), 'several alleles (or genes)' (several alleles (or genes) identified by IMGT/HighV-QUEST).

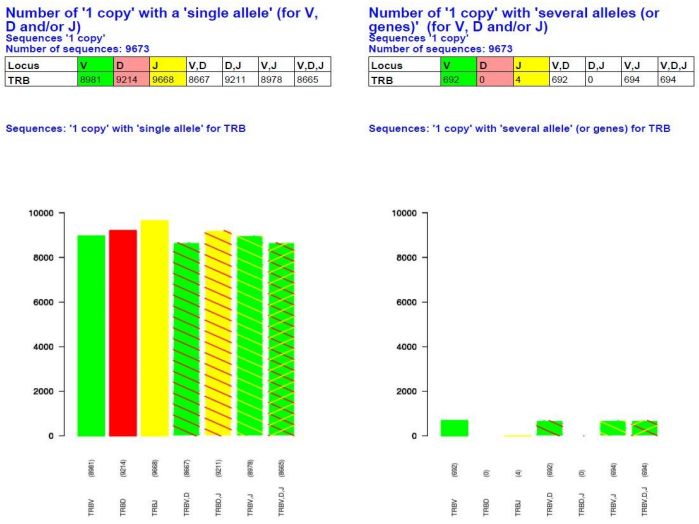


Figure 7 IMGT/HighV-QUEST statistical analysis Number of '1 copy' with 'single allele' and 'several alleles (or genes)' (for V, D and/or J) tables and histograms.
The results are provided for each V, D or J gene and for any combination of them, for 'single allele' (on the left hand side) and for 'several alleles (or genes)' (on the right hand side). Below the tables, histograms are provided, per gene, for each concerned locus. Color code for histograms: green for V genes, red for D genes, yellow for J genes, green with red hatchings for the combination of V and D genes, green with yellow hatchings for the combination of V and J genes, green with red and yellow hatchings for the combination of V, D and J genes.

TRBV gene and allele table

Sequences: '1 copy' with 'single allele' for TRBV
Number of sequences: 8981
Number of sequences with id=100%: 6308 (70.23717%)
Sequence average length: 450
V-REGION average length: 278

#	IMGT gene and allele	Total	Sequence average length	V-REGION average length	id=100%
1	Homsap TRBV10-1	3	451	279	3 (100.0%)
	Homsap TRBV10-1*01 F	3	451	279	3 (100.0%)
2	Homsap TRBV10-2	64	433	270	43 (67.19%)
	Homsap TRBV10-2*01 F	64	433	270	43 (67.19%)
3	Homsap TRBV10-3	369	403	242	219 (59.35%)
	Homsap TRBV10-3*01 F	368	443	208	219 (59.51%)
	Homsap TRBV10-3*04 [F]	1	363	276	0 (0.0%)
4	Homsap TRBV11-1	8	452	282	3 (37.5%)
	Homsap TRBV11-1*01 F	8	452	282	3 (37.5%)
5	Homsap TRBV11-2	177	429	270	129 (72.88%)
	Homsap TRBV11-2*01 F	177	429	270	129 (72.88%)
6	Homsap TRBV11-3	7	453	284	3 (42.86%)
	Homsap TRBV11-3*01 F	7	453	284	3 (42.86%)
7	Homsap TRBV12-1	1	470	285	0 (0.0%)
	Homsap TRBV12-1*01 P	1	470	285	0 (0.0%)
8	Homsap TRBV12-3	140	452	281	94 (67.14%)
	Homsap TRBV12-3*01 F	140	452	281	94 (67.14%)
9	Homsap TRBV12-4	311	452	281	221 (71.06%)
	Homsap TRBV12-4*01 F	311	452	281	221 (71.06%)

Figure 8 IMGT/HighV-QUEST statistical analysis gene and allele table for '1 copy' with 'single allele'.
The gene and allele table is provided per Group [14] (here TRBV) and shows a list of all genes (green lines) and alleles (white lines) found for '1 copy' with 'single allele'.

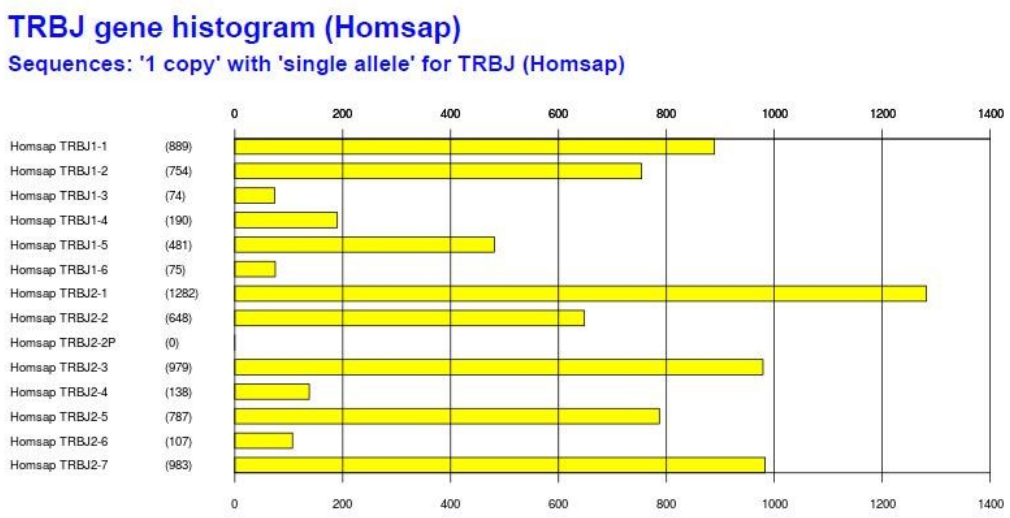


Figure 9 IMGT/HighV-QUEST statistical analysis gene histogram for '1 copy' with 'single allele'.

The gene histogram is provided per Group [14] (here TRBJ) for '1 copy' with 'single allele'. For each gene and allele table, a gene histogram is shown, localizing the gene in the locus, with the number of sequences found between parentheses.

each category, the nb of sequences and the sequence average length (in nt).

4. Terminology (Figure 6): this section provides the definition of the result categories and the terminology of the statistical analysis report (see details in Figure 6 legend).

5. Number of '1 copy' with 'single allele' and 'several alleles (or genes)' (for V, D and/or J) tables and histograms (Figure 7): two tables are provided depending if IMGT/V-QUEST found one 'single allele' for the identified gene (expected to be an unambiguous result) or 'several alleles (or genes)' (usually in the case of too short sequences). Each table shows, per locus, the number of sequences '1 copy' for each V, D and J gene, separately, and in combination. Histograms allow to visualize the information from the tables per locus.

6. IMGT/HighV-QUEST gene and allele tables for '1 copy' with 'single allele' (Figure 8): a table is provided per group of V, D and J genes (e.g., TRBV in Figure 8) [14]. Each table shows the list of identified genes, with for each gene the IMGT gene and allele name (with the taxonomy 6-letter abbreviation, for instance Homsap for *Homo sapiens*), the functionality (F, ORF or P) [12], number of '1 copy' sequences ('Total'), sequence average length (in nt) and in the column 'id=100%' the number (and between parentheses the percentage) of sequences with an identity percentage of 100% by comparison with the germline gene.

7. IMGT/HighV-QUEST gene histograms for '1 copy' with 'single allele' (Figure 9): a histogram is provided per V, D and J gene group [14]. The histograms vizualize (and between parentheses recapitulate) the number of sequences per gene in a given V, D or J group (e.g., TRBJ in Figure 9). The list of all V, D or J genes is provided according to their position in the concerned locus.

8. IMGT/HighV-QUEST CDR3-IMGT tables for '1 copy' and 'in-frame junction': three tables are provided for '1

CDR3-IMGT table

Sequences: '1 copy' with 'single allele' for TRBV and TRBJ
Junction: in-frame
Number of sequences: 8048
Sequence average length: 452
CDR3-IMGT average length (nt): 37
CDR3-IMGT average length (AA): 12

#	CDR3-IMGT length (nt)	CDR3-IMGT length (AA)	Total	Percent	Sequence average Length	Nb of sequences with different CDR3-IMGT		Nb of sequences (sets) with identical CDR3-IMGT	
						Nt	AA	Nt	AA
1	12	4	1	0.01	446	1	1	0 (0)	0 (0)
2	15	5	4	0.05	444	0	0	4 (1)	4 (1)
3	18	6	3	0.04	434	0	0	3 (1)	3 (1)
4	21	7	2	0.02	498	0	0	2 (1)	2 (1)
5	24	8	62	0.77	451	7	7	55 (17)	55 (17)
6	27	9	255	3.17	450	43	43	212 (58)	212 (58)
7	30	10	750	9.32	448	118	117	632 (177)	633 (177)
8	33	11	1495	18.58	448	303	293	1192 (343)	1202 (344)
9	36	12	1886	23.43	453	422	410	1464 (420)	1476 (422)
10	39	13	1741	21.63	453	369	355	1372 (417)	1386 (421)
11	42	14	1074	13.34	456	209	208	865 (252)	868 (252)
12	45	15	461	5.73	454	112	109	349 (100)	352 (100)
13	48	16	186	2.31	463	52	52	134 (42)	134 (42)
14	51	17	81	1.01	456	15	15	66 (24)	66 (24)
15	54	18	26	0.32	467	6	6	20 (6)	20 (6)
16	57	19	12	0.15	459	3	3	9 (2)	9 (2)
17	60	20	8	0.1	483	2	1	6 (1)	7 (1)
18	63	21	1	0.01	354	1	1	0 (0)	0 (0)

Figure 10 IMGT/HighV-QUEST statistical analysis CDR3-IMGT table.

The CDR3-IMGT tables are provided for '1 copy' and 'in-frame junctions'. The table shown is for '1 copy' with 'single allele' for both V and J, here TRBV and TRBJ.

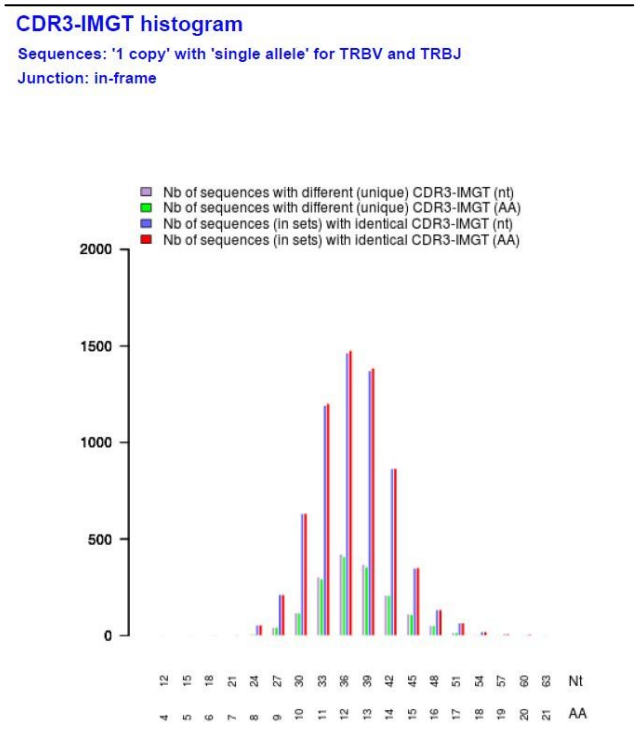


Figure 11 IMGT/HighV-QUEST statistical analysis CDR3-IMGT histogram.

The CDR3-IMGT histograms are provided for '1 copy' and 'in-frame junctions'. The histogram shown is for '1copy' with 'single allele' for both V and J, here TRBV and TRBJ.

copy' with 'single allele' for V and J genes (e.g., between TRBV and TRBJ in Figure 10), '1 copy' for 'several alleles (or genes)' for V and/or J genes (see 11. below), and the last one for all together. Each table gives, for each length of CDR3-IMGT observed in in-frame junctions between V and J, the length in nt and AA, the number ('Total') and percentage ('Percent') of sequences for each length, the sequence average length, the number of sequences with different CDR3-IMGT (in nt and AA) and the number of sequences (and between parentheses the number of sets) with identical CDR3-IMGT (in nt and AA).

9. IMGT/HighV-QUEST CDR3-IMGT histogram for '1 copy' and 'in-frame junction': three histograms are provided for '1 copy' with 'single allele' for V and J genes (e.g., between TRBV and TRBJ in Figure 11), '1 copy' for 'several alleles (or genes)' for V and/or J genes (see 11. below), and the last one for all together. Each histogram is the graphical illustration of the corresponding CDR3-IMGT table and gives, for each length of CDR3-IMGT (nt and AA) observed in in-frame junctions between V and J: Nb of sequences with different (unique) CDR3-IMGT (nt), Nb of sequences with different (unique) CDR3-IMGT (AA), Nb of sequences (in sets) with identical CDR3-IMGT (nt) and Nb of sequences (in sets) with identical CDR3-IMGT (AA).

10. IMGT/HighV-QUEST CDR3-IMGT sets (identical nt and AA) tables for '1 copy' and 'in-frame junction': three tables are provided for '1 copy' with 'single allele' for V and J genes (e.g., between TRBV and TRBJ in Figure 12), '1 copy' for 'several alleles (or genes)' for V and/or J genes (see 11. below), and the last one for all together. Each CDR3-IMGT sets table only shows lines from the corresponding CDR3-IMGT table (see 8. above) for which sets contain at least two identical sequences CDR3-IMGT in nt or AA (number between parentheses greater than 0). Below each recall line, details are provided with number of sets and number of sequences in the set (nt and AA) (Figure 12). For example, for the CDR3-IMGT length of 24 nt (8AA), 55 sequences belong to 17 sets (both at the nt and AA level) which correspond to 8 sets of 2 sequences, 5 sets of 3 sequences, 2 sets of 4 sequences and 2 sets of 8 sequences.

11. IMGT/HighV-QUEST gene and allele tables for '1 copy' with 'several alleles (or genes)' (Figure 13): these tables are provided with the same type of information as for '1 copy' with 'single allele' (see above in 6.). However these sequences are not shown as histograms owing to the uncertainty in the identification of the allele (or even gene), the region of interest in the sequences being too short to allow a correct analysis of the V-REGION by IMGT/V-QUEST. With the progress of the NGS sequencing methodology, the percentage of longer sequences should increase and this category should decrease. Despite the uncertain identification of the V and/or J alleles (or genes), these sequences are not excluded from CDR3-IMGT tables, histograms and sets (see 8., 9. and 10.).

12. IMGT/HighV-QUEST list of sequences in 'More than 1' (Figure 14): these sequences represent redundancies. These sequences are filtered-out and excluded from statistical analysis in order to obtain one and only one copy ('1 copy') of each sequence. The 'More than 1'sequences

CDR3-IMGT sets (identical nt and AA)

Sequences: '1 copy' with 'single allele' for TRBV and TRBJ

Junction: in-frame

CDR3-IMGT length (nt)	CDR3-IMGT length (AA)	Nb of sequences (sets) with identical nt CDR3-IMGT	Nb of sequences (sets) with identical AA CDR3-IMGT
15	5	4 (1)	4 (1)
Sets with identical CDR3-IMGT (nt)			
Nb of sequences in the set	4		
Nb of sets	1		
Sets with identical CDR3-IMGT (AA)			
Nb of sequences in the set	4		
Nb of sets	1		
18	6	3 (1)	3 (1)
Sets with identical CDR3-IMGT (nt)			
Nb of sequences in the set	3		
Nb of sets	1		
Sets with identical CDR3-IMGT (AA)			
Nb of sequences in the set	3		
Nb of sets	1		
21	7	2 (1)	2 (1)
Sets with identical CDR3-IMGT (nt)			
Nb of sequences in the set	2		
Nb of sets	1		
Sets with identical CDR3-IMGT (AA)			
Nb of sequences in the set	2		
Nb of sets	1		
24	8	55 (17)	55 (17)
Sets with identical CDR3-IMGT (nt)			
Nb of sequences in the set	2 3 4 8		
Nb of sets	8 5 2 2		
Sets with identical CDR3-IMGT (AA)			
Nb of sequences in the set	2 3 4 8		
Nb of sets	8 5 2 2		
27	9	212 (58)	212 (58)
Sets with identical CDR3-IMGT (nt)			
Nb of sequences in the set	2 3 4 5 6 7 8 9 11 13		
Nb of sets	24 14 7 4 2 3 1 1 1 1		
Sets with identical CDR3-IMGT (AA)			
Nb of sequences in the set	2 3 4 5 6 7 8 9 11 13		
Nb of sets	24 14 7 4 2 3 1 1 1 1		

Figure 12 IMGT/HighV-QUEST statistical analysis CDR3-IMGT sets (identical nt and AA) tables.

The CDR3-IMGT sets (identical nt and AA) tables show recall lines from the corresponding CDR3-IMGT tables for which sets contain at least two identical sequences CDR3-IMGT (nt or AA). Below each recall line details are provided with number of sets and number of sequences in the set (in nt and AA). That figure shows details for lines #2 to #6 (for which the number of sets is greater than 0) from Figure 10.

TRBV gene and allele table
Sequences: '1 copy' with 'several alleles (or genes)' for TRBV
Number of sequences: 692
Sequence average length: 316
V-REGION average length: 194

#	IMGT gene and allele	Total	Sequence average length	V-REGION average length	id=100%
1	Homsap TRBV10-3	40	311	190	25 (62.5%)
	Homsap TRBV10-3*01 F, or Homsap TRBV10-3*02 F	4	318	203	2 (50.0%)
	Homsap TRBV10-3*01 F, or Homsap TRBV10-3*02 F, or Homsap TRBV10-3*03 [F], or Homsap TRBV10-3*04 [F]	17	231	280	13 (76.47%)
	Homsap TRBV10-3*01 F, or Homsap TRBV10-3*02 F, or Homsap TRBV10-3*04 [F]	18	283	114	10 (55.56%)
	Homsap TRBV10-3*01 F, or Homsap TRBV10-3*04 [F]	1	413	164	0 (0.0%)
2	Homsap TRBV12-3, or Homsap TRBV12-4	39	257	144	24 (61.54%)
	Homsap TRBV12-3*01 F, or Homsap TRBV12-4*01 F	39	257	144	24 (61.54%)
3	Homsap TRBV12-4	1	490	285	0 (0.0%)
	Homsap TRBV12-4*01 F, or Homsap TRBV12-4*02 (F)	1	490	285	0 (0.0%)
4	Homsap TRBV14	1	209	98	1 (100.0%)
	Homsap TRBV14*01 F, or Homsap TRBV14*02 (F)	1	209	98	1 (100.0%)
5	Homsap TRBV2	1	319	213	1 (100.0%)
	Homsap TRBV2*01 F, or Homsap TRBV2*02 (F)	1	319	213	1 (100.0%)
6	Homsap TRBV20-1	204	278	159	112 (54.9%)
	Homsap TRBV20-1*01 F, or Homsap TRBV20-1*02 F	154	334	220	77 (50.0%)
	Homsap TRBV20-1*01 F, or Homsap TRBV20-1*02 F, or Homsap TRBV20-1*03 (F)	2	340	142	0 (0.0%)
	Homsap TRBV20-1*01 F, or Homsap TRBV20-1*02 F, or Homsap TRBV20-1*03 (F), or Homsap TRBV20-1*04 (F), or Homsap TRBV20-1*05 (F), or Homsap TRBV20-1*06 (F), or Homsap TRBV20-1*07 (F)	12	222	104	9 (75.0%)
	Homsap TRBV20-1*01 F, or Homsap TRBV20-1*02 F, or Homsap TRBV20-1*04 (F), or Homsap TRBV20-1*05 (F), or Homsap TRBV20-1*06 (F)	30	279	106	22 (73.33%)
	Homsap TRBV20-1*01 F, or Homsap TRBV20-1*02 F, or Homsap TRBV20-1*04 (F), or Homsap TRBV20-1*05 (F), or Homsap TRBV20-1*06 (F), or Homsap TRBV20-1*07 (F)	5	263	218	4 (80.0%)
	Homsap TRBV20-1*03 (F), or Homsap TRBV20-1*04 (F), or Homsap TRBV20-1*05 (F), or Homsap TRBV20-1*06 (F), or Homsap TRBV20-1*07 (F)	1	235	162	0 (0.0%)

Figure 13 IMGT/HighV-QUEST statistical analysis gene and allele table for '1 copy' with 'several alleles (or genes)'.

The gene and allele table is provided per Group [14] (here TRBV) and shows a list of all genes (gray lines) and alleles (white lines) found for '1 copy' with 'several alleles (or genes)'.

are listed below each corresponding '1 copy', with their sequence number and sequence ID.

13. IMGT/HighV-QUEST list of sequences with 'Warnings' (Figure 15): these sequences corresponds to sequences with warnings for the V-REGION ('different CDR lengths' and/or 'id<85 %'). These sequences are filtered-out and excluded from statistical analysis. The sequences with 'Warnings' are listed, with their sequence number and sequence ID.

14. Sequences with 'Unknown functionality': list of sequences for which no functionality was detected. This category corresponds to the sequences for which the junction cannot be identified (no evidence of rearrangement, no evidence of junction anchors).

15. Sequences with 'No results': list of sequences for which IMGT/HighV-QUEST did not return any result.

The IMGT/HighV-QUEST statistical analysis outputs are in PDF format (6 reports) and PNG (separate graphical elements), archived in a single ZIP file (Figure 16).

The content of the PDF reports is described below, with between parentheses, the results (paragraphes above) to which they refer:

1. IMGT report all: this report contains all results of the statistical analysis (1. to 15.)
2. IMGT report summary: contains Comments (from user, optional), Analysis results list, Summary table, Terminology and Number of '1 copy' with 'single allele' and 'several alleles (or genes)' tables and histograms (1. to 5.)

Filtered-out sequences (excluded from statistical analysis)
Sequences in 'More than 1'
Number of sequences: 114
Sequence average length: 468

Sequences in 'More than 1' (blue lines) are shown below the corresponding '1 copy' (green lines).

#	Sequence number	Sequence ID
1	25	GQMC0HM04IDEPT_length=473_xy=3316_0303_region=4_ru
	2263	GQMC0HM04I5M83_length=473_xy=3637_2917_region=4_ru
2	51	GQMC0HM04IR2K0_length=462_xy=3483_0386_region=4_ru
	1961	GQMC0HM04H4UEM_length=462_xy=3218_2140_region=4_ru
	3056	GQMC0HM04HYQVP_length=462_xy=3149_0259_region=4_ru
3	94	GQMC0HM04IAOOQ_length=463_xy=3285_0232_region=4_ru
	4668	GQMC0HM04H0MLK_length=463_xy=3170_2006_region=4_ru
4	101	GQMC0HM04JL4XI_length=484_xy=3825_2276_region=4_ru
	981	GQMC0HM04IAJ7W_length=484_xy=3283_2634_region=4_ru
5	111	GQMC0HM04IUI42_length=472_xy=3511_0468_region=4_ru
	1500	GQMC0HM04JTUA9_length=472_xy=3913_1315_region=4_ru
6	114	GQMC0HM04JPYY8_length=463_xy=3869_0962_region=4_ru
	665	GQMC0HM04IGK0U_length=463_xy=3352_0988_region=4_ru
7	183	GQMC0HM04I0LLF_length=462_xy=3580_0961_region=4_ru
	2806	GQMC0HM04I9O6X_length=462_xy=3683_3639_region=4_ru
8	278	GQMC0HM04I0U8N_length=456_xy=3583_1173_region=4_ru
	1334	GQMC0HM04I1YP6_length=456_xy=3595_3196_region=4_ru
	2224	GQMC0HM04HXT0T_length=456_xy=3138_2731_region=4_ru
9	288	GQMC0HM04JOZRP_length=464_xy=3858_0387_region=4_ru
	371	GQMC0HM04IA6RR_length=464_xy=3290_3189_region=4_ru
10	323	GQMC0HM04I7CT9_length=458_xy=3657_0815_region=4_ru
	355	GQMC0HM04I3K87_length=458_xy=3614_1225_region=4_ru
	1920	GQMC0HM04INPIE_length=458_xy=3433_1620_region=4_ru

Figure 14 IMGT/HighV-QUEST statistical analysis Sequences in 'More than 1'.

The table lists the sequences present in multiple copies. Green lines show the '1 copy' sequences that were taken into account for the detailed statistical analysis and blue lines show the 'More than 1' sequences that were filtered-out from the detailed statistical analysis. A green line and the blue lines below thus illustrate a single pool of identical sequences.

Filtered-out sequences (excluded from statistical analysis)
Sequences with Warnings
Number of sequences: 473
Sequence average length: 316

#	Sequence number	Sequence ID
1	33	GQMC0HM04IOTGQ_length=422_xy=3446_0152_region=4_ru
2	61	GQMC0HM04JI7IN_length=402_xy=3792_0829_region=4_ru
3	147	GQMC0HM04HYEKE_length=289_xy=3145_0684_region=4_ru
4	150	GQMC0HM04JELG3_length=357_xy=3739_2725_region=4_ru
5	210	GQMC0HM04IJQSP_length=444_xy=3388_0983_region=4_ru
6	284	GQMC0HM04IU77R_length=312_xy=3519_0197_region=4_ru
7	309	GQMC0HM04JMQ6X_length=358_xy=3832_2455_region=4_ru
8	364	GQMC0HM04HX9N1_length=410_xy=3143_2527_region=4_ru
9	417	GQMC0HM04IN85M_length=432_xy=3439_2504_region=4_ru
10	471	GQMC0HM04I3HIX_length=322_xy=3613_0487_region=4_ru
11	481	GQMC0HM04I8780_length=351_xy=3678_2162_region=4_ru
12	489	GQMC0HM04ILJW9_length=231_xy=3408_3467_region=4_ru
13	495	GQMC0HM04JR204_length=368_xy=3893_1222_region=4_ru
14	538	GQMC0HM04IGYJJ_length=413_xy=3356_2125_region=4_ru
15	566	GQMC0HM04IS5VQ_length=362_xy=3495_2164_region=4_ru
16	609	GQMC0HM04JT72Z_length=362_xy=3917_2777_region=4_ru
17	613	GQMC0HM04HY92F_length=345_xy=3155_0549_region=4_ru
18	703	GQMC0HM04IUBDE_length=368_xy=3508_2688_region=4_ru
19	734	GQMC0HM04I922L_length=265_xy=3688_1147_region=4_ru

Figure 15 IMGT/HighV-QUEST statistical analysis Sequences with 'Warnings'

The table lists the sequences with 'Warnings' with their position number in the input file and their identifier. 'Warnings' are defined as having an identity percentage of less than 85% in the alignment with the germline gene and/or 'different CDR lengths'. In order to decrease the doubt and to increase the reliability of the statistical analysis, the sequences with 'Warnings' are omitted from the analysis.

3.IMGT report 1 copy single-allele: contains '1 copy' with 'single allele' tables and histograms (6. and 7.)
4.IMGT report 1 copy several-alleles: contains '1 copy' with 'several alleles (or genes)' tables (11.)
5.IMGT report CDR3-IMGT: contains CDR3-IMGT tables, CDR3-IMGT histograms and CDR3-IMGT sets (identical nt and AA) tables (8. to 10.)
6.IMGT report filtered-out sequences: lists the four categories of filtered-out sequences (sequences in 'More than 1', sequences with 'Warnings', sequences with unknown functionality and sequences with no results) (12. to 15.)

DISCUSSION

The IMGT/HighV-QUEST web portal provides the high-quality results of IMGT/V-QUEST and IMGT/JunctionAnalysis [24-28] in the analysis of the antigen receptor (IG and TR) repertoire sequences generated from NGS high throughput and deep sequencing. High-quality results are based on the IMGT Scientific chart rules: standardized gene and allele nomenclature [1,2,14,19], standardized description and delimitation of labels [7,8,13], particularly the CDR-IMGT and FR-IMGT [15-17], and extensive and accurate analysis of the JUNCTION [27-28]. In contrast to computational software developed for handling huge amount of short

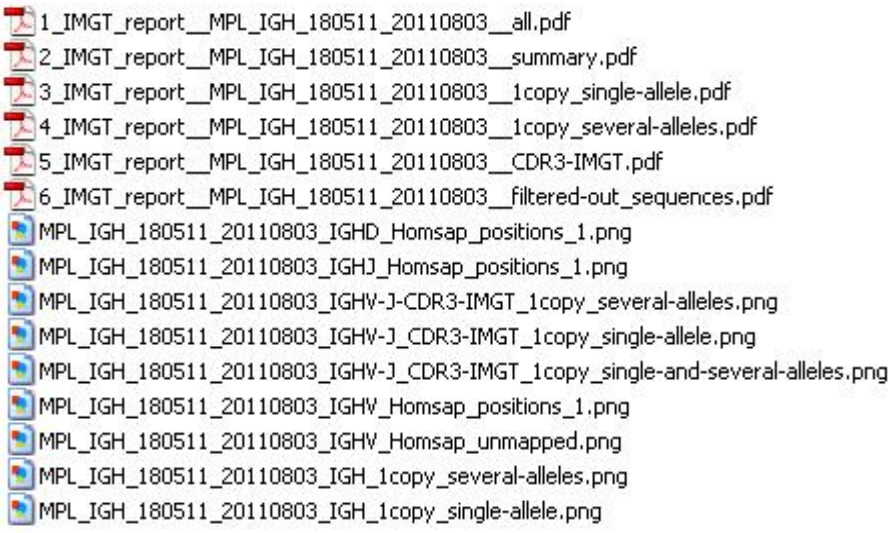


Figure 16 IMGT/HighV-QUEST statistical analysis Outputs.

The IMGT/HighV-QUEST statistical analysis outputs comprise six reports in PDF format. Separate graphical elements (figures in PNG format) are also included (as shown here) if the user made this choice during the submission.

sequences (e.g., from Illumina sequencing [42-49]), IMGT/HighV-QUEST works on longer sequences (from 454 Life Sciences sequencing) and, from the start, provides highly reliable results on sequences of good quality. Accessing the full immune antigen receptor repertoires requires sequences including the complete variable domains (~360 nt) for a reliable analysis, particularly for the IG. Currently the 454 Life Sciences technology [33] that provides the longer sequences is the more adapted for the analysis of the IG and TR repertoires. The 454 sequencing of variable domains has been done to analyse the IGH repertoire in zebrafish [50], to estimate the diversity of a combinatorial human antibody library [51], to monitor human B cell clonality in hematological lymphoid malignancies [52] and to analyse the TRA and TRB repertoires in humans with the comparison of eight T cell subsets in one healthy individual [53]. More specific analyses for human IGH sequences include individual polymorphic variations [54] and quantification of minimal residual disease in chronic lymphocytic leukemia [55] and, for human TRB sequences, analysis of specific rearrangements shared between clonotypes [56]. However, not only NGS methodologies still need to improve in particular to overcome/deal with sequencing errors [57,58], but great care should be taken in the obtaining of the sequences themselves to avoid biases that could lead to a skewed repertoire [59]. Thus 5'RACE (5' rapid amplification of cDNA ends) type protocols [60,61] should be favored over the use of multiplex PCR (polymerase chain reaction) amplification to avoid biases but also to obtain complete variable domains in 5'. In this context, IMGT/HighV-QUEST provides a unique standardized frame for the user to appreciate the quality of the experimental sequencing by comparing the percentage of the 'Result category' in the Statistical analysis: sequence length ('single allele' vs 'several alleles'), sequencing quality ('warnings' and the 'no results' categories should be as small as possible), amplification bias ('1 copy' vs 'More than 1', although this may indicate, for reliable data, relative clone expression). Moreover, IMGT/HighV-QUEST allows to compare, with the same criteria, results whatever the antigen receptor (IG or TR) and the species (human, mouse, rat, etc.). There is always the possibility to check visually unusual results with the IMGT/V-QUEST online as the same version and the same IMGT reference directory release are used and individual files are provided in the IMGT/HighV-QUEST results.

CONCLUSIONS

IMGT/HighV-QUEST, the high throughput version of IMGT/V-QUEST, has become the standard reference for the analysis of IG and TR V domain sequences generated from NGS high throughput and deep sequencing [40,41]. It analyses up to 150,000 nucleotide sequences per batch and performs statistical analysis on the results of up to 450,000 sequences. IMGT/HighV-QUEST provides users with: (i) a friendly web interface for submission and results retrieval, (ii) highly standardized and detailed IMGT/V-QUEST and IMGT/JunctionAnalysis results based on the IMGT-ONTOLOGY concepts and IMGT Scientific chart rules, (iii) a standardized frame for NGS statistical analysis based on 'Results category' ('1 copy', 'More than 1', 'single allele', 'several alleles (or genes)', (iv) detailed statistical analysis tables and histograms (e.g., V, D and J usage, CDR3-IMGT (nt and AA) lengths).

IMGT/HighV-QUEST has been freely available for use for academics on the IMGT® Home page (http://www.imgt.org) since October 2010. More than 123 million sequences were submitted during its first year. The jobs required 70,000 computational hours of resources and generated about three terabytes of results data. More than 83% of the sequences were submitted by users from USA, the others being submitted by users from the European Union (EU) for most, but also from China, Japan, Australia, Canada, Korea and Venezuela. Sequencing data are from both IG and TR [62,63] and from any vertebrate species for which the IMGT reference directory is available.

Beyond the complexity and diversity of the immune responses, it becomes possible, using IMGT/HighV-QUEST, to establish reliable repertoires of IG and TR V domains. These repertoires will contribute to the comparison of individual immunoprofiles in diverse immune situations (healthy vs disease-related repertoires, vaccination, autoimmunity, cancer, infections, immune reconstitution following bone marrow transplant, etc.) and on different B and T cell populations (e.g., characterized by their phenotype markers, differentiation state, activation state, etc.). They will also contribute to characterize potential therapeutic antibodies from combinatorial libraries.

METHODS

Distributed system

IMGT/HighV-QUEST is an automatic system. It provides users a web service to launch their analyses of up to 150,000 IG and TR nucleotide sequences. It must provide a constant web service to end users from scientific communities. There is likely to have resources that are not accessible for a period of time, planned or unplanned. The tool should continue to accept user sequences for the analysis. This reality introduces the ability to accept more than one HPC resource system in order to have at least one working resource when another resource is down. This simultaneous use of numerous HPC systems requires a distributed system with a generic nature. In this architecture the whole system is distributed on different resources (computers, servers, HPC resources) and the tasks are also shared amongst these resources. The tasks are distributed over different servers and resources. There are two main resources in action: a local web server and some remote computational resources. The local web server manages actually all the tasks related to IMGT/HighV-QUEST in scheduled tasks and offers services of user interaction via a web interface. The computational resources, on the other hand, are used to analyze user sequences using standalone version of analysis applications. At present, IMGT/HighV-QUEST uses computational resources on several HPC systems at Centre Informatique National de l'Enseignement Supérieur (CINES)

and at Institut de Génétique Humaine (IGH). It also provides some XML parameter files to add or remove resources in the list and its generic nature allows administrators to add as many resources as they want, and to configure performance related configurations used to perform a load balancing between the available resources.

A layered architecture

The distributed nature of the system requires an approach of internal management that is fully capable of facilitating the intervention of the developers for future extension and the characterization of errors and exceptions. For this reason the tasks are divided into three layers.

1. Web Service (WS) Layer: this layer is responsible of user interactions. The interface design is simple to let users with a minimum of knowledge in Internet to use easily the tool. The heterogeneity of the background system is not felt here. The user simply submits the analysis using a classical web interface and chooses whether he wants to be notified of the completion of the job by e-mail. He can download the results of completed analyses by one click and can at any time see the status of his submitted analyses. The interface lets also users to know whether there was an error during the execution of the analysis or there is a warning concerning its results. The web service is available via a simple HTTP connection and a HTTPS connection secured with SSL exchange.
2. Scheduled Tasks (ST) Layer: this layer is responsible of all jobs' management aspects of IMGT/HighV-QUEST. It runs periodically and each time accomplishes tasks by considering the current situation and the timers. It dispatches the analyses that are in the local queue, monitors previously launched analyses, prepares results of completed analyses, deletes expired analyses and notifies the concerned user if an analysis is going to be expired in five days. For the current configurations, it runs once each 60 seconds. It saves a historic of the date/time of its execution and other important information before exiting. This information is very important and useful for administrators and lets them to maximize the use of resources. Each exception that occurred in the ST layer is logged only once to the administrators and if it is localized, is saved in the short term memory in order to prevent it from reoccurring.
3. Computational Resources (CR) Layer: this layer is where the actual program (scientific program) is run in the standalone version to analyze the user sequences. IMGT/HighV-QUEST supports three types of launching programs and monitoring jobs. In simple mode, the program is launched using simple BASH scripts and monitored using the Process identifier (PID) of the concerned process. In PBS mode, the Portable Batch System (PBS) technology is used to launch and monitor jobs. Finally the IBM Load Leveler mode uses the IBM Load Lever (version 3.5.1.13) of IBM. All interactions with computational resources are done via SSH connections. This technology was used to enforce the security and integrity of transferred user data. The results of completed analyses are archived in a single file in ZIP format in order to let users of all operating systems to manipulate the outputs with the default archive tools. The results are transferred afterwards on the local server via SFTP connection to facilitate their download for users and the concerned user is notified by e-mail (if this latter has chosen to be informed).

System dynamics and reactivity

The system uses its configurations and uniformed exception objects to localize, resolve, if this is possible, or remember errors and exceptions in order to prevent their reoccurring. IMGT/HighV-QUEST has a system of short term memory that lets it tolerate breakdowns by remembering them, and prevent these failures in the future. It also logs administrators by e-mail of the occurred exceptions if it is necessary. This reactivity of the system lets the administrators find out the source of the problem and act in real time. This automatic routine combined with human interference makes IMGT/HighV-QUEST a system that is always stable and appropriate for load of lots of analyses running in parallel on different resources.

Error-prone system and error toleration

Errors are possible to occur during the program execution or management of jobs. No system is without errors. The abundance of errors and exceptions is directly proportionate to the heterogeneity of the background system. In IMGT/HighV-QUEST system a single core application is connected to multiple resources with different distributions of different operating systems and also with different configurations, and thus there may be exceptions that occur during the execution and management concerning the connections and configurations but also on the remote resources themselves. The system is able to manage these exceptions when they occurred in order to minimize loss of user data and time. On the other hand, it tries to compensate errors when they occurred if this is possible. The IMGT/HighV-QUEST system detects and localizes errors by means of a single exception class which uniforms the error detection and thus reduces the complexity of its localization and also its prevention.

Timers and optimization of resource use

A good system does not only perform its tasks but also accomplishes them in an appropriate time and in a good manner. The IMGT/HighV-QUEST system is responsible of three tasks simultaneously. The first task is the user interactions. As an analysis can take more than 24 hours to be completed and a user cannot wait 24 hours for his analysis to complete in front of his screen, and the user even closes the web browser some minutes after the analysis submission, so the user wants that all tasks be felt in real time. On the other hand there is a question of the performance of the tool, being reactive to users and administrators and more importantly to events (exceptions, errors, etc.). Another important issue is the accomplishment of scheduled tasks in a good time. For all these reasons, a system of timers is designed in IMGT/HighV-QUEST to synchronize the different types of tasks on heterogeneous systems. All computational resources do not have the same performance, they do not act similarly (the performance may go down if we have more than one job on a server using one core). To let a good use of these resources and to maximize the web services quality, for each task a specific timer is created. A timer is the time length between two times that a specified event (task) is triggered. In each trigger, a test is

performed whether it is useful to enter the core of the tasks. In case it is not useful, IMGT/HighV-QUEST exits from the tasks without doing anything. This sort of reaction is important in order to minimize the connections to remote resources. This is done for two reasons, the first reason is that interactions of automatic routines are performed via SSH connections and SSH connections demand some capacity of computing and transfer of data, in order to not use lots of resources for doing nothing, these timers are set. The second reason is that if a control is not set, the number of connections can increase dramatically and the more the number of connections the more likely an exception can occur, the server status is then set as 'down'. As said before, the performance of different resources is different. That is why each resource has a different timer for monitoring jobs, in order to optimize, for example, the number of connections and resource use.

IMGT/HighV-QUEST administration and parameterization

IMGT/HighV-QUEST is designed for the analysis of large amount of sequences. Multiple analyses of more than 100,000 sequences have been regularly running on the tool. Although the tool manages these jobs automatically, it is sometimes necessary to have a human intervention, especially for abnormal situations. To this end, an administration interface was also developed that comprises three levels of expertise.

1. Regular administration: in this level the administrator can see submitted analyses, online users, and information on the functionality of local and remote servers. He can cancel analyses, or in more urgent situations, cancel the jobs of an analysis directly on the remote server, hold queued analyses, etc. Using this level does not require special precautions.
2. Advanced administration: in this level the administrator can start/stop the scheduled tasks, backup database data in backup tables, tell the tool to hold all analyses that will be submitted after now. This part needs to be used with precautions.
3. Direct database and application context interaction: the SQL tool is designed to send SQL queries directly to the server and the direct context interaction facility serves to change context attributes of the application that are used for management purposes. This section has to be used only for very urgent situations and for experimentation (performance) and attention should be paid during the manipulations.

The parameterization of the tool lets administrator change the parameters that are rarely updated, adding a new resource, deleting a resource, adding logging e-mails, modifying the database connection parameters etc. The XML language was used for parameterization aspects in order to emphasize the simplicity and the generality.

Abbreviations

5'RACE: 5' rapid amplification of cDNA ends; AA: amino acid; BASH: Bourne-again shell; CDR: complementarity determining region; CR: Computational Resources; CSV: Comma-separated values; D: diversity; FPIA: fusion protein for immune applications; FR: framework region; HPC: High Performance Computing; HTTP: Hypertext Transfer Protocol; HTTPS: Hypertext Transfer Protocol Secure; IG: immunoglobulin; IMGT: IMGT®, the international ImMunoGeneTics information system®; J: junction; nb: number; NGS: Next Generation Sequencing; nt: nucleotide; ORF: open reading frame; PBS: Portable Batch System; PCR: polymerase chain reaction; PDF: Portable Document Format; PID: Process Identifier; PNG: Portable Network Graphics; SFTP: Secure File Transfer Protocol; SQL: Structured Query Language; SSH: Secure Shell; SSL: Secure Sockets Layer; ST: Scheduled Task; TR: T cell receptor; V: variable; WS: Web Service; XML: eXtensible Markup Language; ZIP: a file compression method and extension.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

EA developed IMGT/HighV-QUEST, SL provided NGS experimental data, VG, PD and M-PL initiated and supervised the project. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

We thank Gérard Lefranc for helpful comments. We are grateful to the IMGT® team for its constant motivation and expertise. IMGT® is Academic Institutional Member of the International Medical Informatics Association (IMIA). IMGT® is currently supported by the Ministère de l'Enseignement supérieur et de la Recherche (MESR), Infrastructures Biologie Santé et Agronomie (IBiSA), Centre national de la recherche scientifique (CNRS), Université Montpellier 2, Région Languedoc-Roussillon, Grand Plateau Technique pour la Recherche (GPTR) Sud de France, SFR BioCampus Montpellier, French National Research Agency/Agence Nationale de la Recherche ANR (BIOSYS-06-135457 and, under the program 'Investissements d'avenir', Grant agreement LabEx MAbImprove: ANR-10-LABX-53).
IMGT/HighV-QUEST was granted access to the HPC resources of Centre Informatique National de l'Enseignement Supérieur (CINES) under the allocations 2010-036029 and 2011-036029 made by GENCI (Grand Equipement National de Calcul Intensif).

REFERENCES

1. Lefranc M-P, Lefranc G: *The Immunoglobulin FactsBook*. London, UK: Academic Press; 2001, 1-458.
2. Lefranc M-P, Lefranc G: *The T cell receptor FactsBook*. London, UK: Academic Press; 2001, 1-398.
3. Lefranc M-P. **IMGT, the international ImMunoGeneTics information system.** *Cold Spring Harbor Protocols* 2011 Jun 1; 2011 (**6**):595-603. pii: pdb.top115. doi: 10.1101/pdb.top115.
4. Lefranc M-P: **IMGT, the international ImMunoGeneTics information system®: a standardized approach for immunogenetics and immunoinformatics**. *Immunome Research* 2005, Sept 20; **1**:3, doi:10.1186/1745-7580-1-3.
5. Lefranc M-P, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J, Regnier L, Ehrenmann F, Lefranc G, Duroux P: **IMGT®, the international ImMunoGeneTics information system®.** *Nucleic Acids Research* 2009, **37**:1006-1012.
6. Giudicelli V, Lefranc M-P: **Ontology for immunogenetics: IMGT-ONTOLOGY.** *Bioinformatics* 1999, **15**:1047-1054.
7. Lefranc M-P, Giudicelli V, Ginestoux C, Bosc N, Folch G, Guiraudou D, Jabado-Michaloud J, Magris S, Scaviner D, Thouvenin V, Combres K, Girod D, Jeanjean S, Protat C, Yousfi Monod M, Duprat E, Kaas Q, Pommié C, Chaume D, Lefranc G: **IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics, http://www.imgt.org.** *In Silico Biology*. Epub 2003, 0004, 22 Nov 2003; 2004, **4**:17-29.
8. Lefranc M-P, Clément O, Kaas Q, Duprat E, Chastellan P, Coelho I, Combres K, Ginestoux C, Giudicelli V, Chaume D, Lefranc G: **IMGT-Choreography for Immunogenetics and Immunoinformatics.** *In Silico Biology* Epub 2005, 5, 0006, 24 Dec 2004; 2005, **5**:45-60.
9. Duroux P, Kaas Q, Brochet X, Lane J, Ginestoux C, Lefranc M-P, Giudicelli V: **IMGT-Kaleidoscope, the formal IMGT-ONTOLOGY paradigm.** *Biochimie* 2008, **90**:570-583.
10. Lefranc M-P, Giudicelli V, Regnier L, Duroux P: **IMGT®, a system and an ontology that bridge biological and computational spheres in bioinformatics.** *Briefings in Bioinformatics* 2008, **9**(4):263-275. doi:10.1093/bib/bbn014.
11. Lefranc M-P: **IMGT-ONTOLOGY.** In: *Encyclopedia of Systems Biology*. Edited by Dubitzky W, Wolkenhauer O, Cho KH, Yokota H. New York: Springer; 2012 (in press).
12. Lefranc M-P: **From IMGT-ONTOLOGY IDENTIFICATION axiom to IMGT standardized keywords: for immunoglobulins**

(IG), T cell receptors (TR), and conventional genes. *Cold Spring Harbor Protocols*. 2011 Jun 1; 2011, **(6)**:604-613. pii: pdb.ip82. doi: 10.1101/pdb.ip82

13. Lefranc M-P: **From IMGT-ONTOLOGY DESCRIPTION axiom to IMGT standardized labels: for immunoglobulin (IG) and T cell receptor (TR) sequences and structures.** *Cold Spring Harbor Protocols* 2011 Jun 1; 2011, **(6)**:614-626. pii: pdb.ip83. doi: 10.1101/pdb.ip83
14. Lefranc M-P: **From IMGT-ONTOLOGY CLASSIFICATION axiom to IMGT standardized gene and allele momenclature: for immunoglobulins (IG) and T cell receptors (TR).** *Cold Spring Harbor Protocols* 2011 Jun 1; 2011, **(6)**:627-632. pii: pdb.ip84. doi: 10.1101/pdb.ip84
15. Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G: **IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains.** *Developmental and Comparative Immunology* 2003, **27**:55-77.
16. Lefranc M-P: **IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF.** *Cold Spring Harbor Protocols* 2011 Jun 1; 2011, **(6)**:633-642. pii: pdb.ip85. doi: 10.1101/pdb.ip85
17. Lefranc M-P: **IMGT Collier de Perles for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF.** *Cold Spring Harbor Protocols* 2011 Jun 1; 2011, **(6)**:643-651. pii: pdb.ip86. doi: 10.1101/pdb.ip86
18. Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, Lefranc M-P: **IMGT/LIGM-DB, the IMGT® comprehensive database of immunoglobulin and T cell receptor nucleotide sequences.** *Nucleic Acids Research* 2006, **34**:D781-784.
19. Giudicelli V, Chaume D, Lefranc M-P: **IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes.** *Nucleic Acids Research* 2005, **33**:D256-261.
20. Ehrenmann F, Kaas Q, Lefranc M-P: **IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF.** *Nucleic Acids Research* 2010, **38**:D301-307. Epub 2009 Nov 9; doi:10.1093/nar/gkp946.
21. Ehrenmann F, Lefranc M-P: **IMGT/3Dstructure-DB: querying the IMGT database for 3D structures in immunology and immunoinformatics (IG or antibodies, TR, MH, RPI, and FPIA)**. *Cold Spring Harbor Protocols* 2011 Jun 1; 2011, (**6**):750-761. pii: pdb.prot5637. doi: 10.1101/pdb.prot5637.
22. Lefranc M-P: **Antibody databases and tools: The IMGT® experience**. In: *Therapeutic monoclonal antibodies: from Bench to Clinic*. Edited by An Z. John Wiley & Sons, Inc; 2009, chapter **4**, 91-114.
23. Lefranc M-P: **Antibody nomenclature: From IMGT-ONTOLOGY to INN definition.** *MAbs* 2011, **3**(1):1-2. Epub 2011 Jan 1.
24. Giudicelli V, Chaume D, Lefranc M-P: **IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis.** *Nucleic Acids Research* 2004, **32**:W435-440.
25. Brochet X, Lefranc M-P, Giudicelli G: **IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis.** *Nucleic Acids Research* 2008, **36**:W503-508.
26. Giudicelli V, Brochet X, Lefranc M-P: **IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences.** *Cold Spring Harbor Protocols* 2011 Jun 1; 2011, **(6)**:695-715. pii: pdb.prot5633. doi: 10.1101/pdb.prot5633.
27. Yousfi Monod M, Giudicelli V, Chaume D, Lefranc M-P: **IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs.** *Bioinformatics* 2004, **20**:i379-385.
28. Giudicelli V, Lefranc M-P: **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 Harbor Protocols* 2011 Jun 1; 2011, **(6)**:716-725. pii: pdb.prot5634. doi: 10.1101/pdb.prot5634.
29. Ehrenmann F, Lefranc M-P: **IMGT/DomainGapAlign: IMGT standardized analysis of amino acid sequences of variable, constant, and groove domains (IG, TR, MH, IgSF, MhSF)**. *Cold Spring Harbor Protocols* 2011 Jun 1; 2011, (**6**):737-749. pii: pdb.prot5636. doi: 10.1101/pdb.prot5636.
30. Ehrenmann F, Giudicelli V, Duroux P, Lefranc M-P: **IMGT/Collier de Perles: IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains)**. *Cold Spring Harbor Protocols* 2011 Jun 1; 2011, (**6**):726-736. pii: pdb.prot5635. doi: 10.1101/pdb.prot5635.
31. Giudicelli V, Protat C, Lefranc M-P: **The IMGT strategy for the automatic annotation of IG and TR cDNA sequences: IMGT/Automat.** In: *Proceedings of the European Conference on Computational Biology (ECCB 2003)*. Edited by Christophe C, Lenhof H-P, Sagot M-F. Paris: INRIA (DISC/Spid) DKB-31; 2003, 103-104.
32. Giudicelli V, Chaume D, Jabado-Michaloud J, Lefranc M-P: **Immunogenetics Sequence Annotation: the Strategy of IMGT based on IMGT-ONTOLOGY.** *Studies in health technology and informatics* 2005, **116**:3-8.
33. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**:376-380.
34. David R B: **Whole-genome re-sequencing**. *Current Opinion in Genetics & Development* 2006, **16**:545-552.
35. Mardis ER: **Next-Generation DNA Sequencing Methods**. *Annual Review of Genomics and Human Genetics* 2008, **9**:387-402.
36. Ansorge WJ: **Next-generation DNA sequencing techniques**. *Nature Biotechnology* 2009, **25**:195–203.
37. Metzer ML: **Sequencing technologies - the next generation**. *Nature Rev Genet* 2010, **11**:31-46.
38. Meldrum C, Doyle MA, Tothill RW: **Next-Generation Sequencing for Cancer Diagnostics: a Practical Perspective**. *The Clinical Biochemist Reviews* 2011, **32**:177-195.
39. Shendure JA, Porreca GJ, Church GM, Gardner AF, Hendrickson CL, Kieleczawa J, Slatko BE: **Overview of DNA sequencing strategies.** *Current Protocols in Molecular Biology* 2011 Oct; Chapter **7**:Unit7.1.
40. Alamyar E., Giudicelli V, Duroux P, Lefranc M-P: **IMGT/HighV-QUEST 2011**. In: *JOBIM 2011 PARIS. Journées Ouvertes Biologie Informatique Mathématiques*. Edited by Institut Pasteur. Paris: Institut Pasteur; 2011, 384.
41. Alamyar E, Duroux P, Lefranc M-P, Giudicelli V: **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: *Methods in Molecular Biology – Immunogenetics*. Edited by Tait B and Christiansen F. Springer Humana Press; 2012, chapter **32** (in press).
42. Freeman JD, Warren RL, Webb JR., Nelson BH, Holt RA. **Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing**. *Genome Research* 2009, **19**:1817-1824.
43. Warren RL, Nelson BH, Holt RA. **Profiling model T-cell metagenomes with short reads**. *Bioinformatics* 2009, **25**:458-464.
44. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, Riddell SR, Warren EH, Carlson CS: **Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells**. *Blood* 2009, **114**:4099-4107.
45. Robins HS Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, Carlson CS, Warren EH: **Overlap and effective size of the human CD8+ T cell receptor repertoire**. *Science Translational Medicine* 2010, **2**:47ra64-47ra64.
46. Ravn U, Gueneau F, Baerlocher L, Osteras M, Desmurs M, Malinge P, Magistrelli G, Farinelli L, Kosco-Vilbois MH, Fischer N: **By-passing in vitro screening-next generation sequencing technologies applied to antibody display and in silico candidate selection**. *Nucleic Acids Research* 2010, **38**:e193.
47. Fischer N: **Sequencing antibody repertoires: the next generation**. *MAbs* 2011, **3**:17-20.
48. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, Webb JR, Holt RA: **Exhaustive T-cell repertoire sequencing of human peripheral blood samples revealssignatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes**. *Genome Research* 2011 May;**21**(5):790-7. Epub 2011 Feb 24.
49. Warren RL, Holt RA: **Targeted assembly of short sequence reads.** *PLoS One*. 2011 May 11;**6**(5):e19816.
50. Weinstein JA, Jiang N, White RA 3rd, Fisher DS, Quake SR: **High -throughput sequencing of the zebrafish antibody repertoire**. *Science* 2009, **324**:807-810.
51. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, Ni I, Mei L, Sundar PD, Day GMR, Cox D, Rajpal A, Pons J: **Precise determination of the diversity of a combinatorial antibody**

library gives insight into the human immunoglobulin repertoire. *Proceedings of the National Academy of Sciences* 2009, **106**:20216-20221.

52. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, Jones CD, Simen BB, Hanczaruk B, Nguyen KD, Nadeau KC, Egholm M, Miklos DB, Zehnder JL, Fire AZ: **Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing**. *Science translational medicine* 2009, **1**:12ra23.
53. Wang C. Sanders CM, Yang Q, Schroeder HW Jr, Wang E, Babrzadeh F, Gharizadeh B, Myers RM, Hudson JR Jr, Davis RW, Han J: **High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets**. *Proceedings of the National Academy of Sciences* 2010, **107**:1518-1523.
54. Boyd SD, Gaëta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, Jones CD, Simen BB, Hanczaruk B, Nguyen KD, Nadeau KC, Egholm M, Miklos DB, Zehnder JL, Collins AM: **Individual Variation in the Germline Ig Gene Repertoire Inferred from Variable Region Gene Rearrangements**. *The Journal of Immunology* 2010, **184**:6986-6992.
55. Logan AC, Gao H, Wang C, Sahaf B, Jones CD, Marshall EL, Buño I, Armstrong R, Fire AZ, Weinberg KI, Mindrinos M, Zehnder JL, Boyd SD, Xiao W, Davis RW, Miklos DB: **High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment**. *Proceedings of the National Academy of Sciences* 2011, **108**:21194 -21199
56. Venturi V, Quigley MF, Greenaway HY, Ng PC, Ende ZS, McIntosh T, Asher TE, Almeida JR, Levy S, Price DA, Davenport MP, Douek DC: **A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing**. *The Journal of Immunology* 2011 Apr 1;**186**(7):4285-4294. Epub 2011 Mar 7.
57. Nguyen P, Ma J, Pei D, Obert C, Cheng C, Geiger TL: **Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire**. *BMC Genomics* 2011, **12**:106.
58. Prabakaran P, Streaker E, Chen W, Dimitrov DS. **454 antibody sequencing - error characterization and correction.** *BMC Research Notes*. 2011, **4**:404.
59. Benichou J, Ben-Hamo R, Louzoun Y, Efroni S. **Rep-Seq: uncovering the immunological repertoire through Next Generation Sequencing**. *Immunology* 2012, **135**:183-191
60. Bertioli D: **Rapid amplification of cDNA ends**. *Methods in Molecular Biology* 1997 **67**:233-238.
61. Quigley MF, Almeida JR, Price DA, Douek DC: **Unbiased molecular analysis of T cell receptor expression using template-switch anchored RT-PCR.** *Current Protocols in Immunology* 2011, Chapter **10**:Unit10.33.
62. Prabakaran P, Chen W, Singarayan MG, Stewart CC, Streaker E, Feng Y, Dimitrov DS. **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* 2011 Dec 27. [Epub ahead of print]
63. Li S, Lefranc M-P, Corbin V, Freeman D, Giudicelli V, Alamyar E, Scheerlinck J-P, Cameron P, Frohman M, Plebanski M, Loveland B, Burrows SR, Papenfuss AT, Miles JJ, Gowans EJ. **High throughput sequencing and IMGT/HighV-QUEST analysis of natural regulatory and conventional T cell receptor repertoires following human H1N1 vaccination** (submitted).