IMGT/HighV-QUEST 2011

Eltaf Alamyar, Véronique Giudicelli, Patrice Duroux and Marie-Paule Lefranc

Laboratoire d'ImmunoGénétique Moléculaire (LIGM), Université Montpellier 2,
Institut de Génétique Humaine (IGH), UPR CNRS 1142, Montpellier (France)
{Eltaf.Alamyar,Veronique.Giudicelli,Patrice.Duroux,Marie-Paule.Lefranc}@igh.cnrs.fr

ImMunoGeneTics Information system®

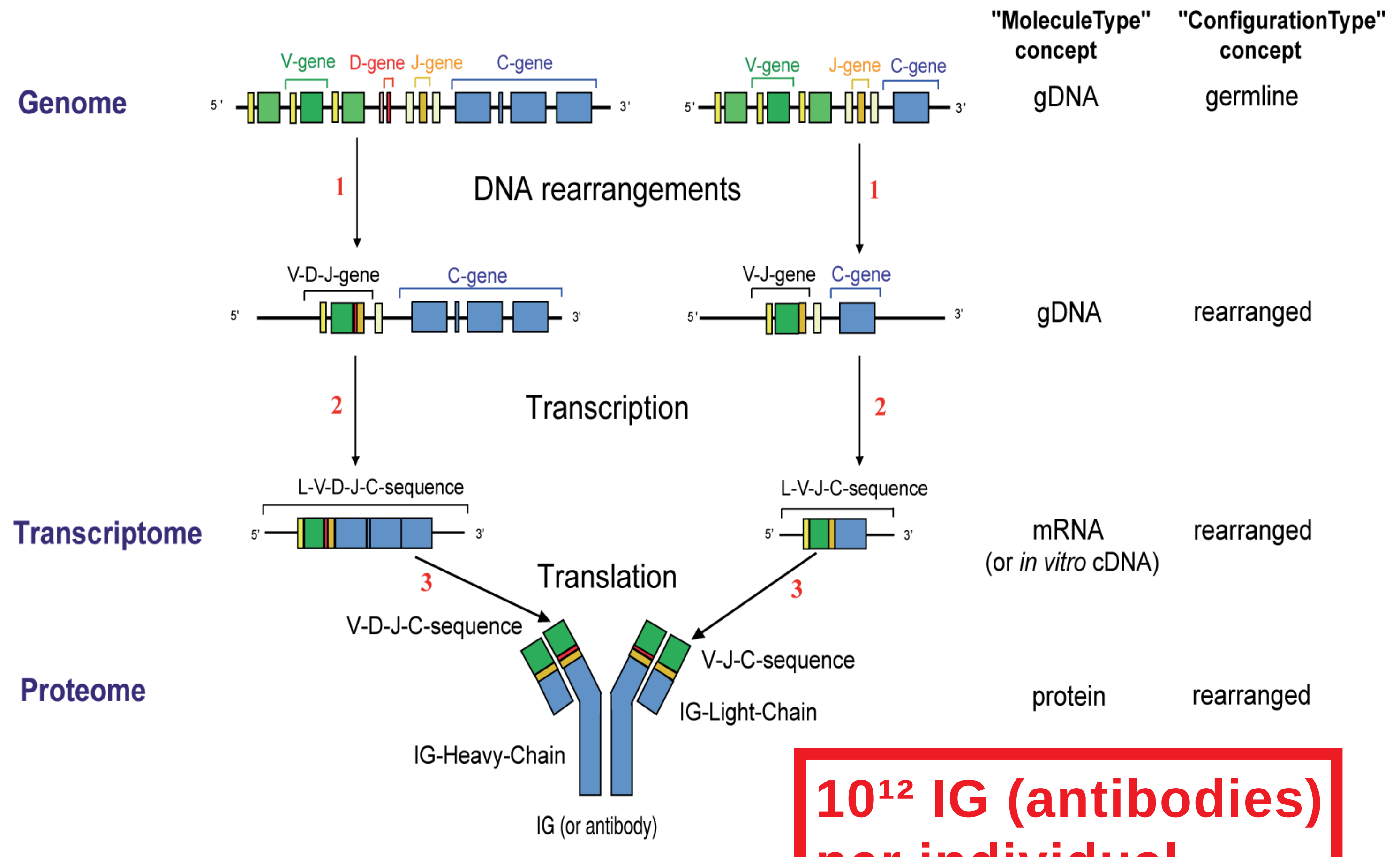
http://www.imgt.org

The analysis of expressed repertoires of antigen receptors - immunoglobulins (IG) or antibodies and T cell receptors (TR) - represents a huge challenge for the study of the adaptive immune response in normal and disease-related situations, such as viral infections. To answer that need, IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org) [1] has developed IMGT/HighV-QUEST [2]. IMGT/HighV-QUEST is devoted to the analysis of large repertoires of IG and TR sequences that result from Next Generation Sequencing technologies. IMGT/HighV-QUEST, a high throughput version of IMGT/V-QUEST [3], analyses up to 150,000 sequences per run. It identifies the IG and TR variable (V), diversity (D) and joining (J) genes and alleles by alignment with the germline IG and TR gene and allele sequences of the IMGT reference directory. It describes the V-REGION mutations and identifies the hot spot positions in the closest germline V gene. The analysis is based on the IMGT-ONTOLOGY concepts of description, classification and numerotation [4, 5]. New functionalities have been developed that comprise the introduction of statistical analysis on results estimated as reliable based on selected criteria.

[1] Lefranc, M.P. et al., Nucleic Acids Res., 37,1006-1012, 2009.
[2] Alamyar, E. et al., Proceedings of the 11th JOBIM, P27 pp. 156, 2010.
[3] Brochet, X. et al., Nucleic Acids Res., 36:W503-508, 2008.
[4] Giudicelli, V. and Lefranc, M.-P., Bioinformatics, 15:1047-1054, 1999.
[5] Duroux, P. et al., Biochimie, 90:570-583, 2008.

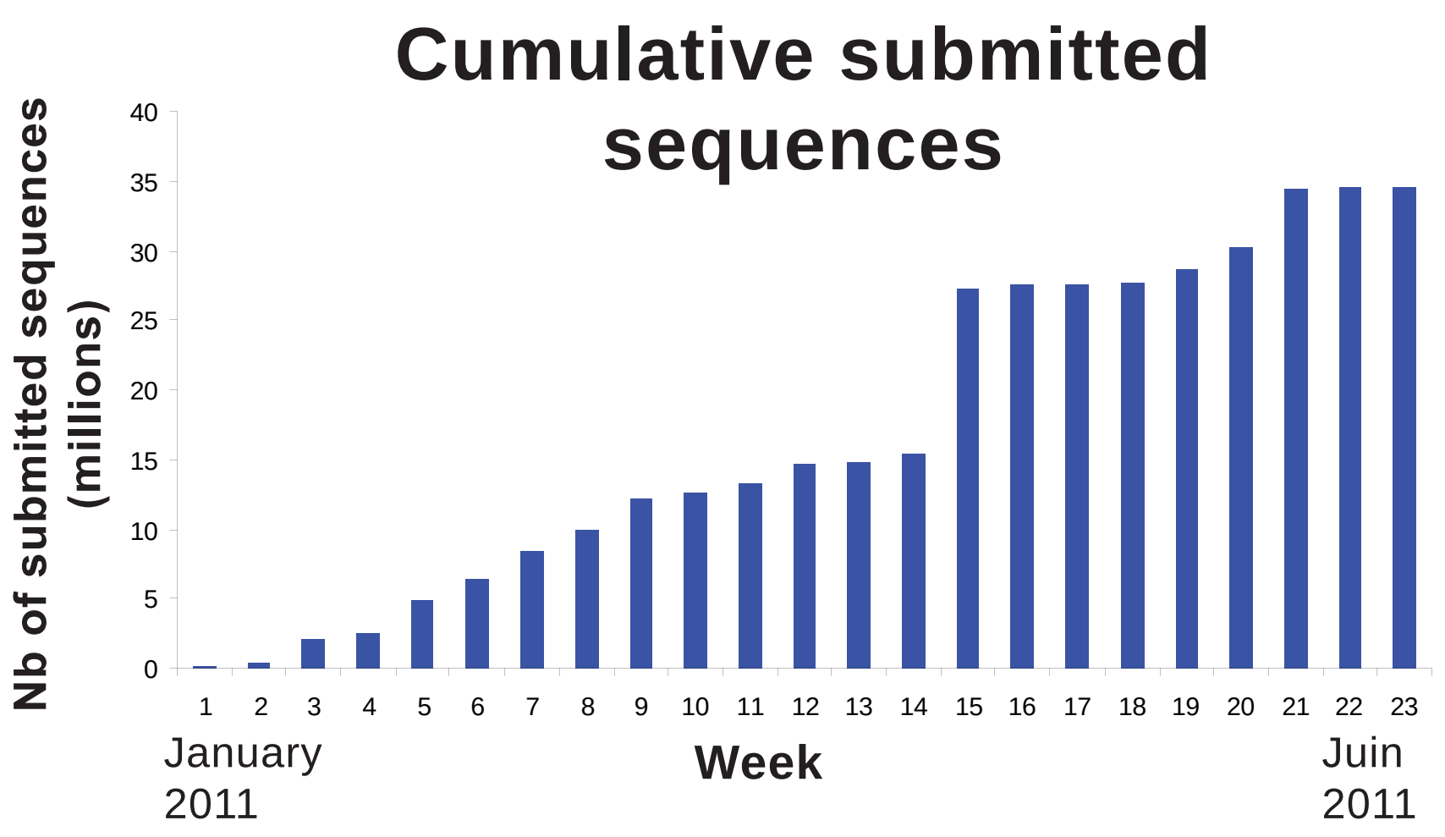
Biological Context

The adaptive immune response is characterized by an extreme diversity of the specific antigen receptors that comprise the immunoglobulins (IG) or antibodies and the T cell receptors (TR) (10^{12} different IG and 10^{12} different TR per individual, in humans). The complex molecular mechanisms (DNA rearrangements, N-diversity, and for IG, somatic hypermutations) that occur in B cells and T cells are at the origin of that huge diversity.

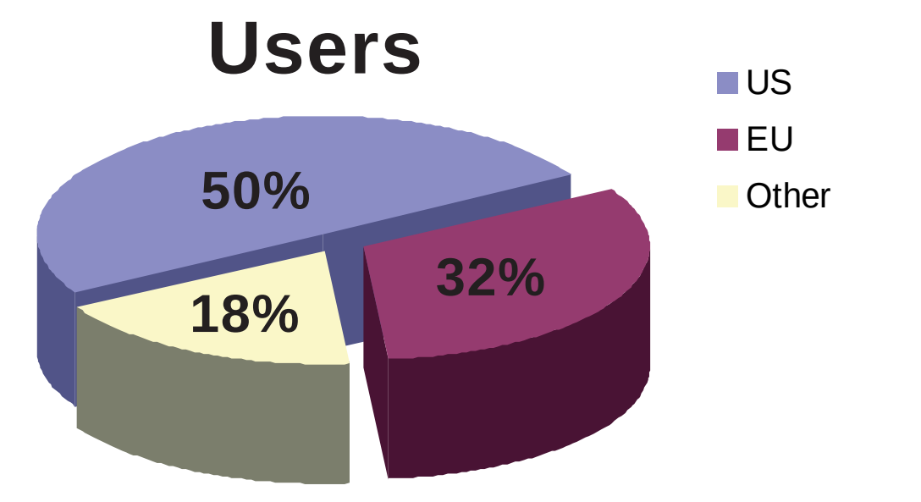


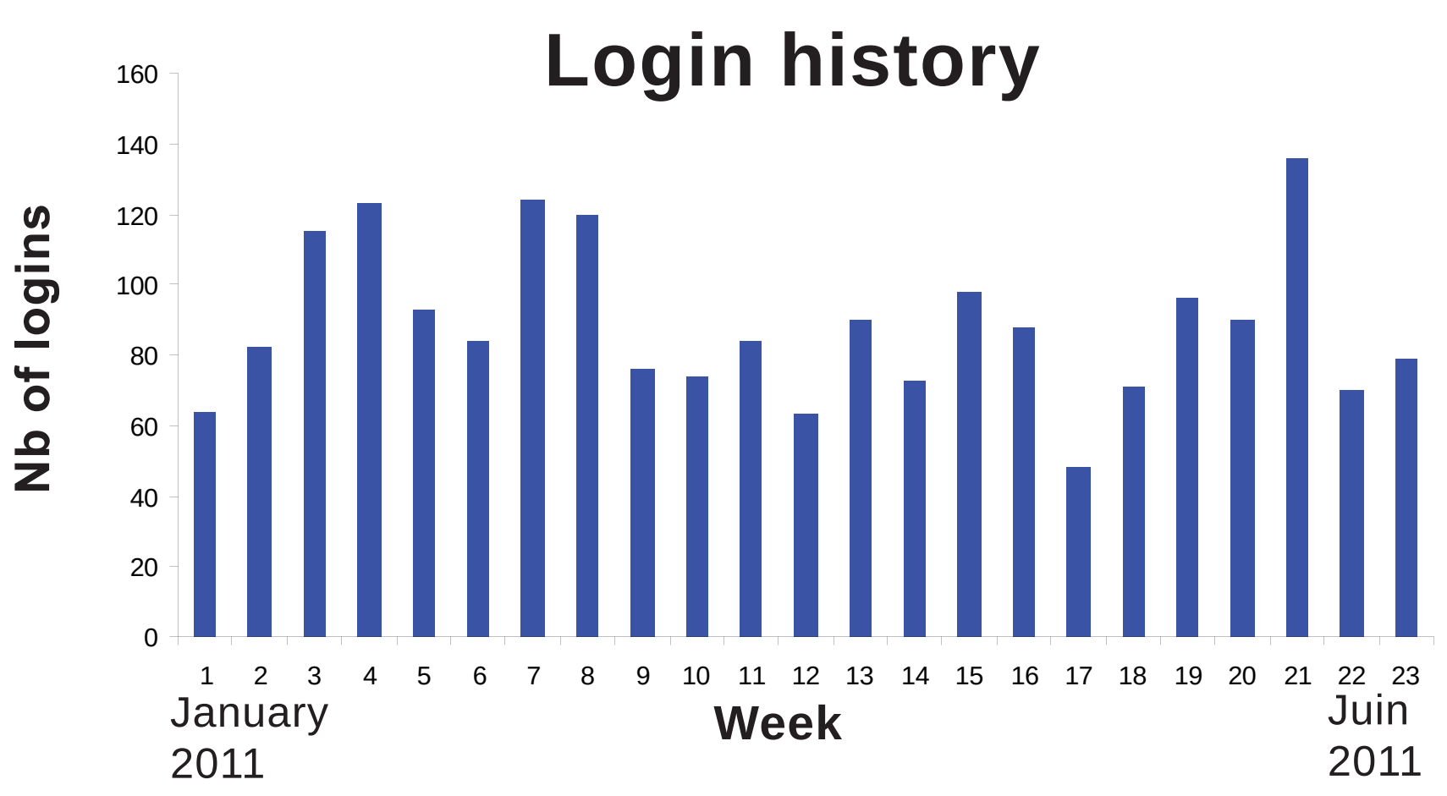
10^{12} IG (antibodies) per individual

Users and Analyses

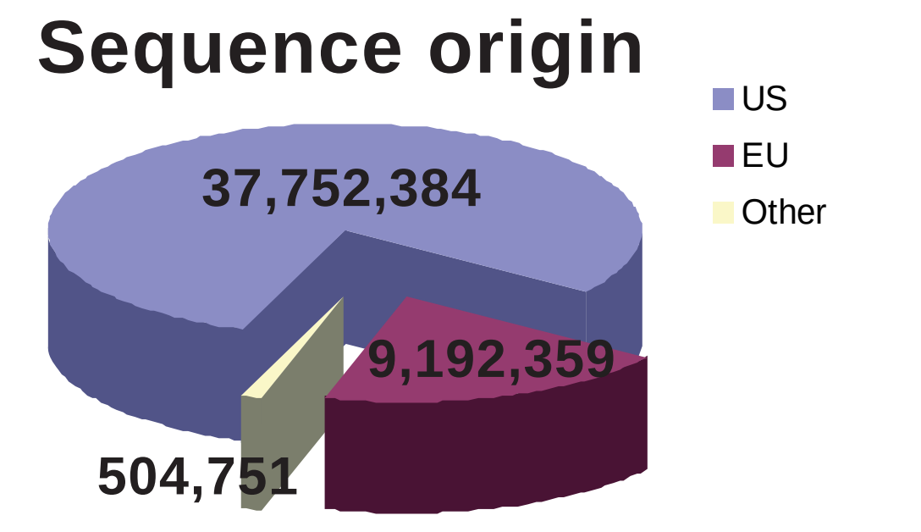


Since the availability of IMGT/HighV-QUEST in October 2010, >47 millions of sequences have been submitted (July 2011). They required >20,000 hours of computational resources. More than one terabyte of results was generated.





Half of IMGT/HighV-QUEST users are from USA, the others being from EU for most, but also from China, Japan, Australia, Canada, Korea, Mexico, Israel and Venezuela.



Countries	Nb of submitted sequences	Computational resources (hours)	Size of generated files (Gbyte)
United States	37752384	16254	900,09
Germany	6323015	2635	150,75
United Kingdom	1081055	450	25,77
France	851180	355	20,29
Spain	469374	196	11,19
Denmark	408460	170	9,74
Australia	174534	73	4,16
Japan	100624	42	2,40
Mexico	90414	38	2,16
Austria	49390	21	1,18
Venezuela	47556	20	1,13
Canada	46716	19	1,11
China	44227	18	1,05
Netherlands	6262	3	0,15
Finland	3417	1	0,08
Israel	650	0	0,02
Belgium	193	0	0,00
Korea	30	0	0,00
Italy	10	0	0,00
Sweden	3	0	0,00
Total	**47449494**	**20295**	**1131,28**

Users from USA submitted 79% of the sequences, users from EU submitted 19%, while the remaining sequences were submitted by users from other countries.

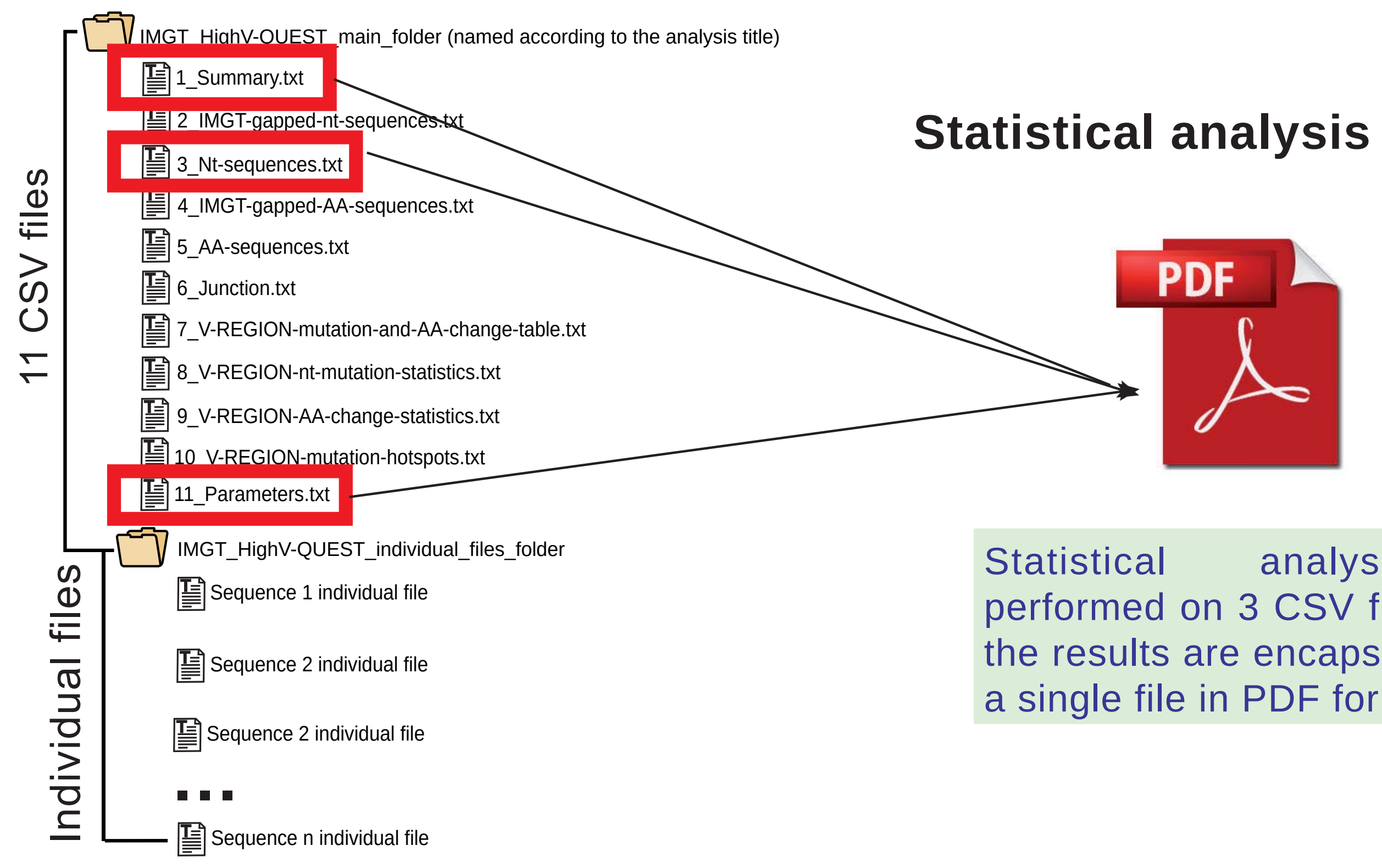
Statistics in 2011 show an increasing number of IMGT/HighV-QUEST users and a growing analysis demand compared with 2010 (50% increase in the number of submitted sequences and 30% increase in user registration in time average).

Outputs

Analysis output

The outputs are archived in a single file in ZIP format which comprises:

- **11 CSV files** equivalent to the eleven sheets of the 'Excel files' of IMGT/V-QUEST
- for each analysed sequence, the 'Detailed view' **individual files** that allows one to visualize the individual detailed results


11 CSV files
IMGT_HighV-QUEST_main_folder (named according to the analysis title)
1_Summary.txt
2_IMGT-gapped-nt-sequences.txt
3_Nt-sequences.txt
4_IMGT-gapped-AA-sequences.txt
5_AA-sequences.txt
6_Junction.txt
7_V-REGION-mutation-and-AA-change-table.txt
8_V-REGION-nt-mutation-statistics.txt
9_V-REGION-AA-change-statistics.txt
10_V-REGION-mutation-hotspots.txt
11_Parameters.txt
Individual files
IMGT_HighV-QUEST_individual_files_folder
Sequence 1 individual file
Sequence 2 individual file
Sequence 2 individual file
...
Sequence n individual file
PDF

Statistical analysis output

Statistical analysis is performed on 3 CSV files and the results are encapsulated in a single file in PDF format

Statistical Analysis

1. Selection of results for statistical analysis

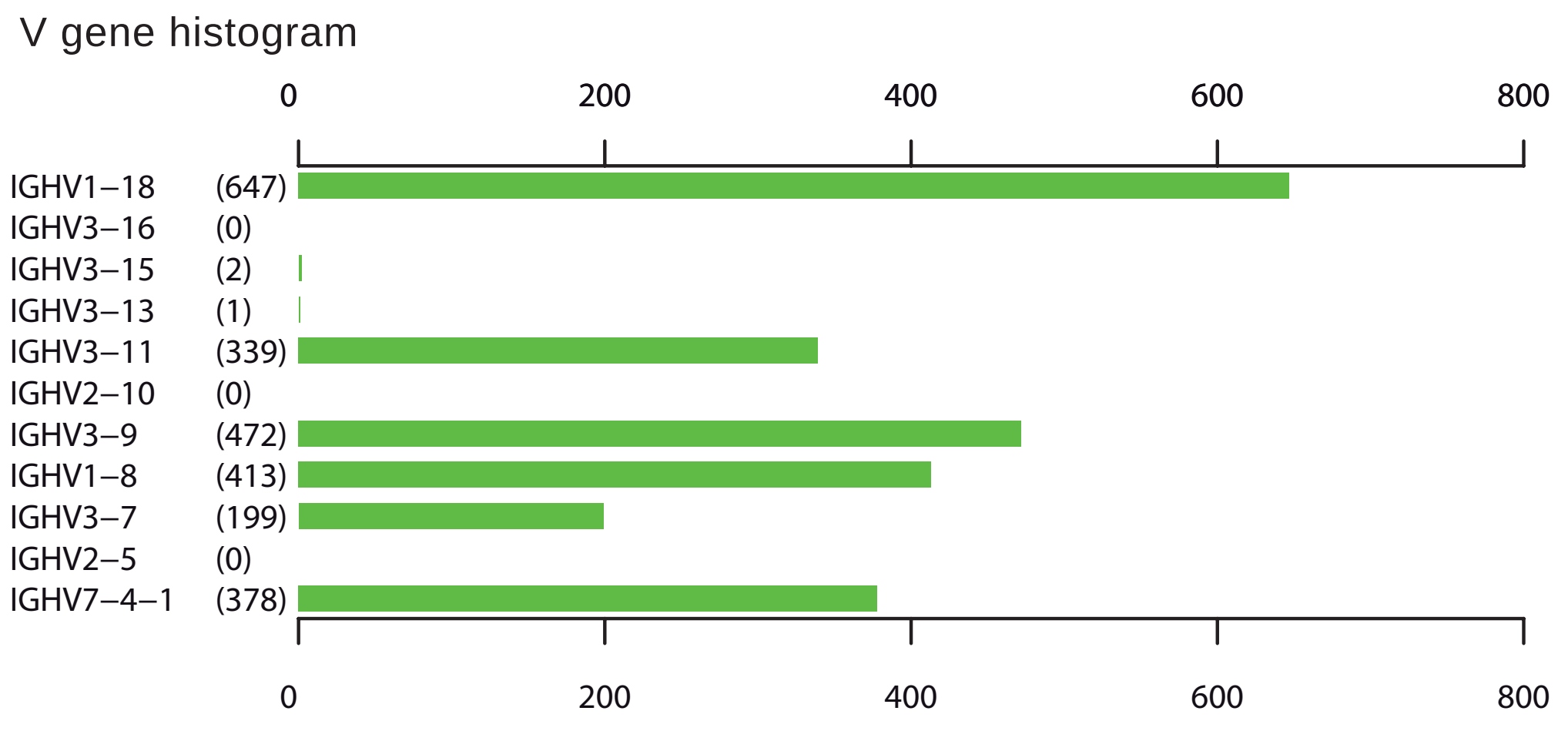
Statistical analyses are performed on results selected as ‘1 copy’ (redundancies are recorded but not processed), and with quality criteria (identification of a single gene/allele, known functionality, absence of IMGT/V-QUEST warnings regarding the CDR1-IMGT and CDR2-IMGT lengths and the percentage of identity).

2. Tables and histograms for each gene (V, D and J)

For each gene, number of sequences, average sequence length, average V-, D-, J-REGION length, and number of sequences with an identity percentage of 100% by comparison with the germline, are provided.

V gene and allele table

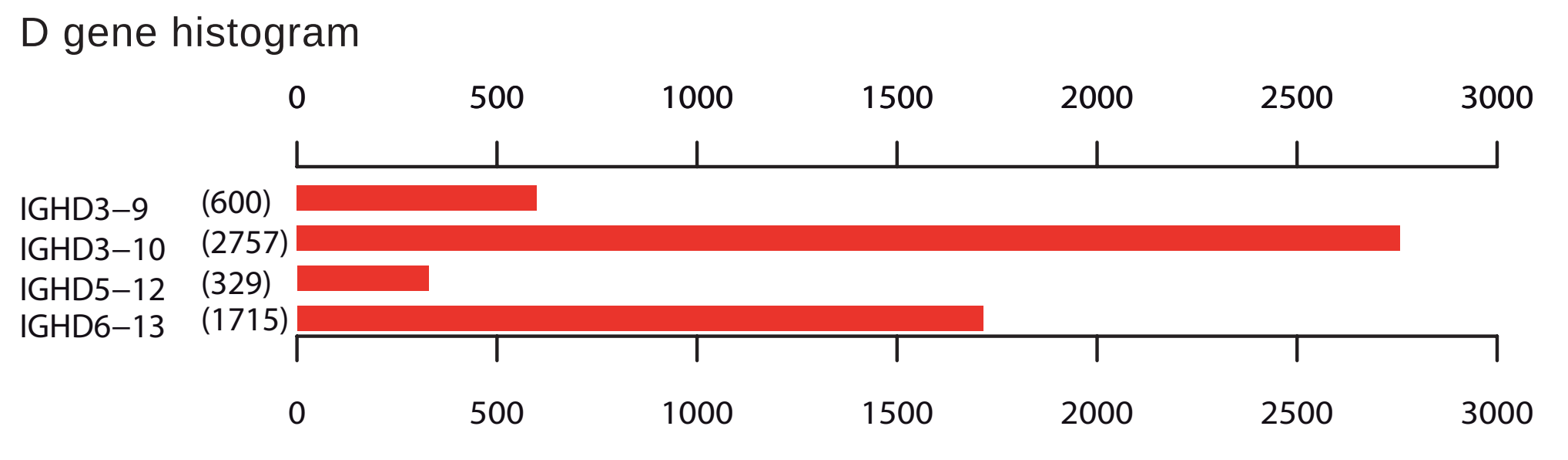
#	IMGT gene and allele	Total	Average sequence length	Average V-REGION length	id=100% nb (%)
1	IGHV1-18	647	243	166	455 (70.32%)
	IGHV1-18*01	647	243	166	455 (70.32%)
9	IGHV3-11	339	242	166	253 (74.63%)
	IGHV3-11*01	339	242	166	253 (74.63%)
10	IGHV3-13	1	223	158	1 (100.0%)
	IGHV3-13*01	1	223	158	1 (100.0%)
11	IGHV3-15	2	266	173	1 (50.0%)
	IGHV3-15*04	1	283	173	0 (0.0%)
	IGHV3-15*07	1	248	173	1 (100.0%)



D gene and allele table

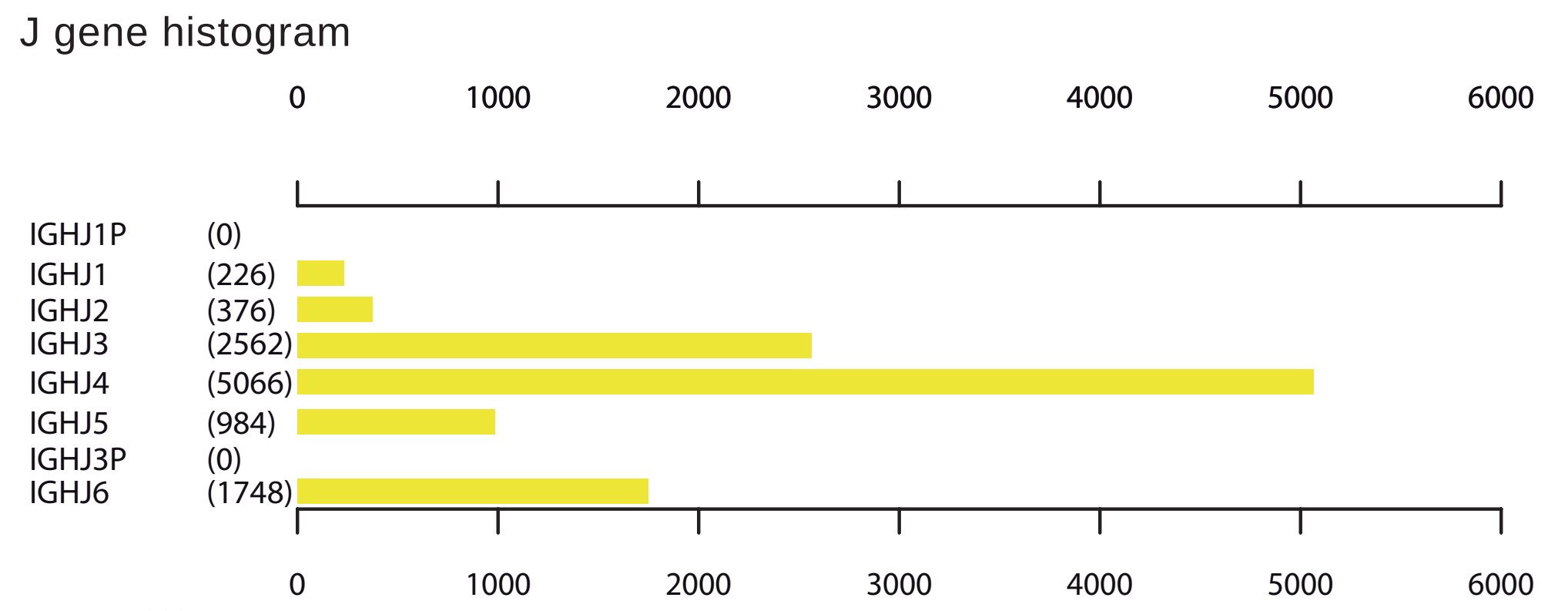
#	IMGT gene and allele	Total	Average sequence length	Average D-REGION length
10	IGHD3-10	2757	243	17
	IGHD3-10*01	2693	244	15
	IGHD3-10*02	64	242	19
14	IGHD3-9	600	246	19
	IGHD3-9*01	600	246	19
18	IGHD5-12	329	238	14
	IGHD5-12*01	329	238	14
21	IGHD6-13	1715	239	15
	IGHD6-13*01	1715	239	15

Colored lines illustrate results per gene and white lines under each gene illustrate the results per allele, individually. In the histograms, genes are ordered according to their positions from 5’ to 3’ in the locus.



J gene and allele table

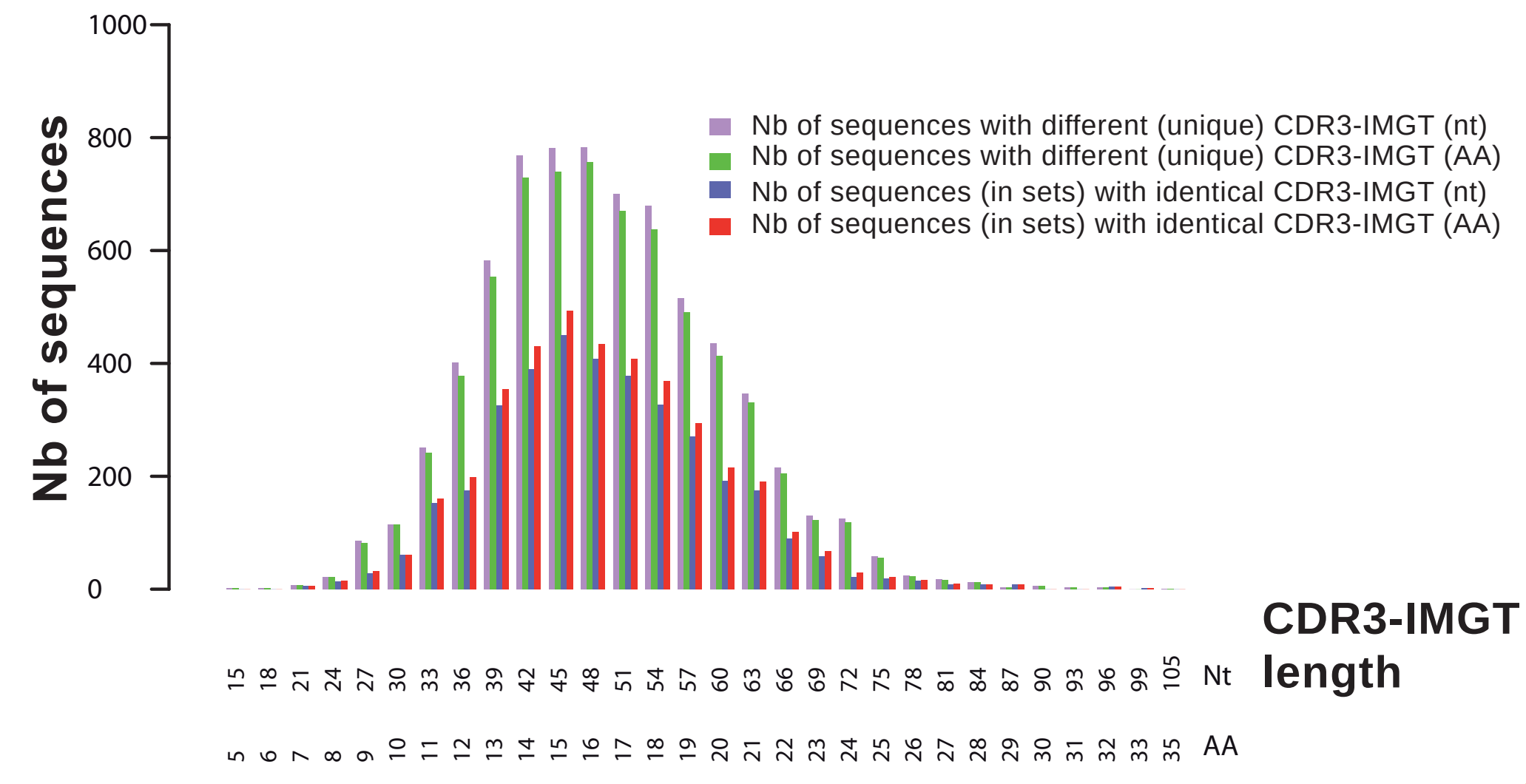
#	IMGT gene and allele	Total	Average sequence length	Average J-REGION length	id=100% nb (%)
2	IGHJ2	414	243	50	0 (0.0%)
	IGHJ2*01	414	243	50	0 (0.0%)
3	IGHJ3	2685	244	44	0 (0.0%)
	IGHJ3*01	36	245	41	0 (0.0%)
	IGHJ3*02	2649	243	48	0 (0.0%)
4	IGHJ4	5795	240	41	754 (13.01%)
	IGHJ4*01	5	239	46	3 (60.0%)
	IGHJ4*02	5708	238	33	751 (13.16%)
	IGHJ4*03	82	242	43	0 (0.0%)



3. CDR3-IMGT length analysis

Statistics provide the histogram of different and identical CDR3-IMGT sequences for each CDR3-IMGT length in nucleotides (nt) and amino acids (AA).
Results are shown as:

- 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)
- Nb of sequences (in sets) with identical CDR3-IMGT (AA)



Acknowledgments: this work was granted access to the HPC resources of CINES under the allocation 2010-036029 made by GENCI (Grand Equipement National de Calcul Intensif).

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)
Webmaster: Chantal Ginestoux (Chantal.Ginestoux@igh.cnrs.fr)