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

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

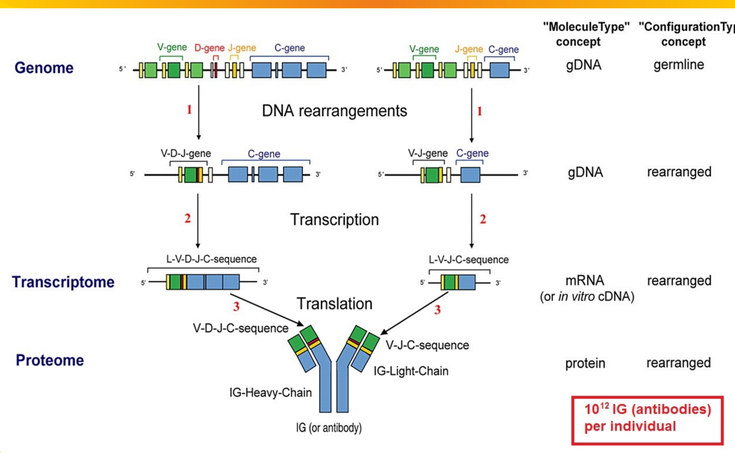
[1]IMGT®, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire (LIGM), Institut de Génétique Humaine (IGH), UPR CNRS 1142, Montpellier University and CNRS, Montpellier (France)
[2]Higher School of Statistics and Information Analysis, University of Carthage, Tunis (Tunisia)
[3]National Schools of Engineers of Tunis, Laboratory U2S, University of Tunis El-Manar, Tunis (Tunisia)

http://www.imgt.org

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

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

Biological Context



IMGT Clonotypes (AA)

In IMGT®, the clonotype designated as 'IMGT clonotype (AA)' is defined by a unique V-(D)-J rearrangement (IMGT genes and alleles determined at the nucleotide level), conserved anchors (C104, W or F 118), and a unique CDR3-IMGT AA in frame junction [4]. Each 'IMGT clonotype (AA)' is characterized by a selected unique representative sequence.

ID		IMGT clonotype (AA) definition						IMGT clonotype (AA) representative sequence			Nb			IMGT clonotypes (nt)
#	Exp. ID	V gene and allele	D gene and allele	J gene and allele	CDR3-IMGT length (AA)	CDR3-IMGT sequence (AA)	Anchor 104,118	V %	Sequence length	Sequence ID	Total nb of '1 copy'	Total nb of 'More than 1'	Total	Sequences file ('1 copy')
Homsap IGHV1-2*02 F														
1	137-mid5	Homsap IGHV1-2*02 F	Homsap IGHD2-2*01 F	Homsap IGHJ6*03 F	22 AA	ARDLYCSSTSCYGGWYYYYMDV	C,W	95.14	425	GJNZTB402H8X9K_length=425	1	0	1	Sequences file
2	157-mid5	Homsap IGHV1-2*02 F	Homsap IGHD6-6*01 F	Homsap IGHJ6*03 F	22 AA	ARERVGRSIAARRAPDYYYYMDV	C,W	97.92	426	GJNZTB402H4DK_length=426	1	0	1	Sequences file
3	305-mid5	Homsap IGHV1-2*02 F	Homsap IGHD3-22*01 F	Homsap IGHJ4*02 F	21 AA	ARGPYHRPTYYYDSSGYYGDY	C,W	96.15	374	GJNZTB402FSHFI_length=374	1	0	1	Sequences file

Statistical Procedure

Data and study purpose

8 sets of NGS sequences analyzed by IMGT/HighV-QUEST [5]: 2 T cell populations : $CD4^-$ and $CD4^+$ at 4 time points : Pre, day 3 (d3), day 8 (d8) and day 26 (d26) post-vaccination against H1N1 influenza virus of a single individual. The purpose is to make comparisons of the immunoprofiles of the two T cell populations ($CD4^-$/$CD4^+$) at a given time point (Pre, d3, d8, d26) and of a given T cell population between two time points (Pre/d3, d3/d8, d3/d26, d8/d26, Pre/d8, Pre/d26).

Differences in proportions

- m: the nb of genes analysed in a given group of a locus.
- k (k=1,...,m): the index of each gene of a given group identified by its IMGT® gene name.
- Sets to be compared indexed by $\{i,j\}_{(i\neq j)}$.
- n_i, n_j nb of IMGT clonotype (AA) per group in the set i and j, respectively.
- p_i^k and p_j^k vary in $\Theta=[0,1]^2$ and represent the probabilities of finding at least one IMGT clonotype (AA) with the gene k in the set i and the set j, respectively. $\bar{p}_i^k$ and $\bar{p}_j^k$ are the estimators of p_i^k and p_j^k, respectively.
- Approximate 95% confidence intervals (CI) calculated for reference to provide a magnitude of the true difference in proportions $p_i^k - p_j^k$ as follows:

$$(\bar{p}_i^k - \bar{p}_j^k) \pm z_{(1-\frac{\alpha}{2})}.\sigma_{\bar{p}_i^k-\bar{p}_j^k}$$

where $z_{(1-\frac{\alpha}{2})}$ is the 1-$\frac{\alpha}{2}$ percentile of a standard normal distribution and $\sigma_{\bar{p}_i^k-\bar{p}_j^k}$ is the standard deviation $\sigma_{\bar{p}_i^k-\bar{p}_j^k} = \sqrt{\frac{\bar{p}_i^k(1-\bar{p}_i^k)}{n_i} + \frac{\bar{p}_j^k(1-\bar{p}_j^k)}{n_j}}$

Test and error rate control

- Tested hypotheses:

$$\begin{cases} H_0^k : p_i^k - p_j^k = 0 \\ H_1^k : p_i^k - p_j^k \neq 0 \end{cases}$$

- Test statistics (z-scores):

$$z^k = \frac{\bar{p}_i^k - \bar{p}_j^k}{\sqrt{\bar{p}_{ij}^k(1-\bar{p}_{ij}^k)(\frac{1}{n_i}+\frac{1}{n_j})}} \sim \mathcal{N}(0,1)$$

where $\bar{p}_{ij}^k$) is the weighted pooled proportion $\bar{p}_{ij}^k = \frac{n_i\bar{p}_i^k + n_j\bar{p}_j^k}{n_i+n_j}$

When multiple tests of hypotheses are conducted simultaneously, more than 5% of them are very likely to be statistically significant purely by chance. An adjustment of the p-values must be made through a multiple testing procedure by the two strategies:

- Family-wise error rate (FWER) is the probability to make one or more false positives. The Bonferroni, Sidak, Holm, Sidak (single step and step down) and Hochberg FWER procedures were applied.
- False discovery rate (FDR) is the expected proportion of false positives. The Benjamini & Hochberg (BH) and Benjamini & Yekutieli (BY) FDR procedure were applied.

IMGT-ONTOLOGY Main Concepts

1. DESCRIPTION

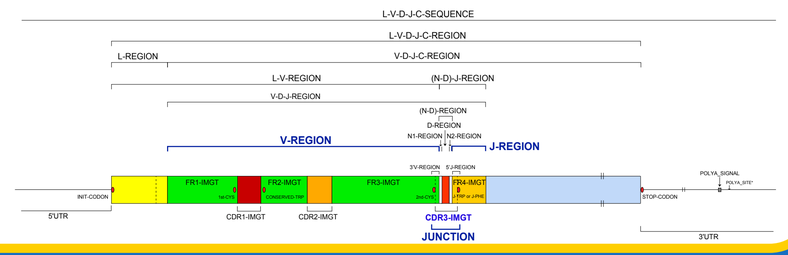
The concepts of description correspond to IMGT® standardized labels. They are more than 560 standardized labels (available in the IMGT Scientific chart), 277 for the nucleotide sequences and 285 for the 3D structures.

2. CLASSIFICATION

The concepts of classification allowed to classify and name the human IG and TR genes and alleles which were approved by HGNC and endorsed by WHO-IUIS.They provide the frame for the standardized IG and TR nomenclature of jawed vertebrates.

3. NUMEROTATION

The concepts of numerotation comprise the 'IMGT unique numbering' and 'IMGT Collier de Perles'.



Statistical Significance of IMGT Clonotype (AA) Diversity

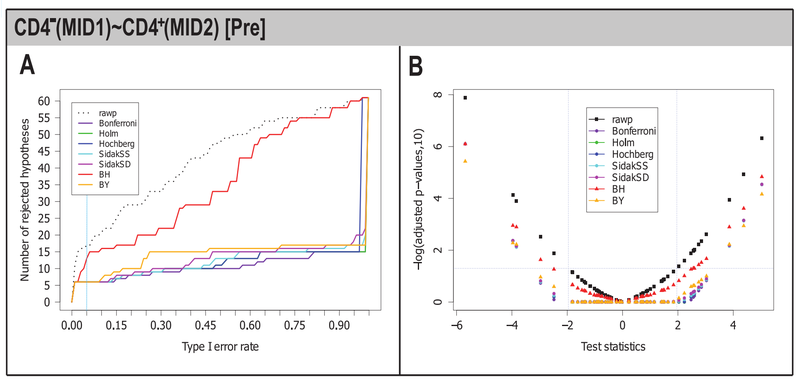


Figure 1. Multiple testing procedures plots [7] displayed for comparison of the differences in proportions for IMGT clonotypes (AA) with a gene of a given group (TRBV, TRBD or TRBJ), between two T cell populations ($CD4^-$ and $CD4^+$) at Pre. The following procedures: Bonferroni, Holm, Hochberg, SidakSS and SidakSD, BH and BY were applied. Left panel (A): Line graphs showing the number of rejected null hypotheses against the Type I error rate. Dotted lines represent unadjusted p-values (rawp) whereas colored lines represent ajusted p-values of the seven procedures. A vertical line corresponds to a Type I error rate at 5%. Right panel (B): Negative decimal logarithms of unadjusted p-values and adjusted p-values against the test statistics z-scores. Two areas in the scatter plot (top left and top right) correspond to significant differences in proportions and they are delimited at a significance level of 5% by -log 10 (p-values)$>$ 1.3 (horizontal line) and by z-scores ($<$-1.96 for negative differences or $>$1.96 for positive differences) (vertical line).

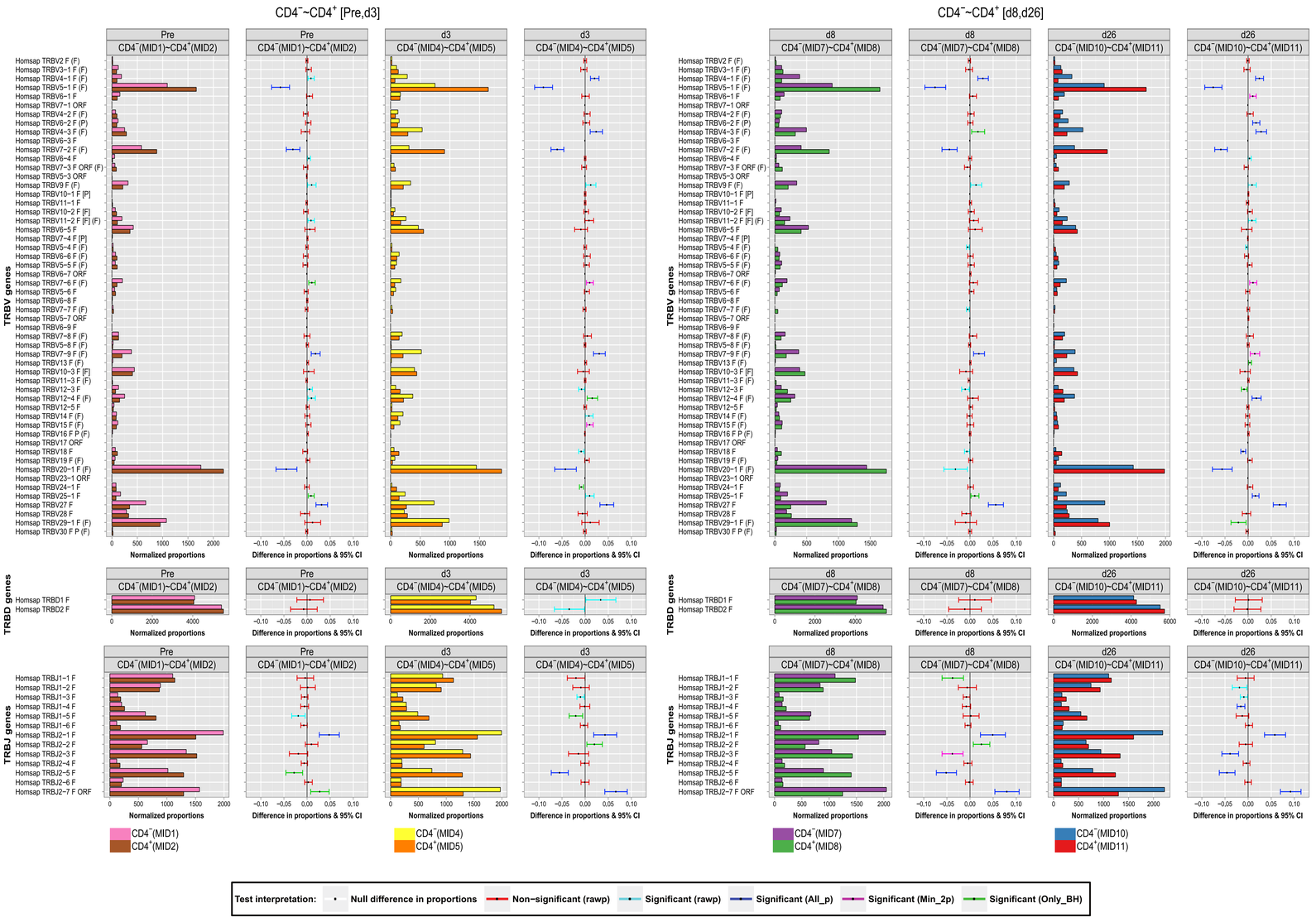


Figure 2. Synthesis graph. The synthesis graph is displayed for *Homo sapiens* TRB IMGT clonotypes (AA) with a gene of a given group (TRBV, TRBD or TRBJ) between two T cell populations ($CD4^-$ and $CD4^+$) at two time points (d8 and d26), with for each time point, two panels. Left panel: normalized bar graph (IMGT clonotypes (AA) proportions normalized for 10,000 IMGT clonotypes (AA) per group), with juxtaposed colored bars corresponding to $CD4^-$ (top) and $CD4^+$ (bottom). Right panel: difference in proportions graph with significance and confidence interval (CI) bars. CI bar colors correspond to the test interpretation before adjustment of *p*-values (rawp) (non-significant: red, significant: light blue) and after adjustment by the multiple testing procedures (significant differences validated by the seven procedures (All_p): dark blue, by two or more multiple testing procedures (Min_2p): pink, and only by BH (Only_BH): green.

Acknowledgments: this work was granted access to the HPC resources of HPC@LR and of CINES and TGCC-CEA under the allocation 036029-(2010-2015) made by GENCI (Grand Equipement National de Calcul Intensif).

IMGT® director: Sofia Kossida (Sofia.Kossida@igh.cnrs.fr)
IMGT® founder and executive director emeritus: Marie-Paule Lefranc (Marie-Paule.Lefranc@igh.cnrs.fr)
Bioinformatics manager: Véronique Giudicelli (Veroniqe.Giudicelli@igh.cnrs.fr)
Computer manager: Patrice Duroux (Patrice.Duroux@igh.cnrs.fr)