IMGT/HighV-QUEST & Clonotypes IMGT (AA) : Significativité Statistique de la Diversité et Expression par Gène pour les Immunoprofiles NGS des Immunoglobulines et Récepteurs T

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

[1] *IMGT®, the international ImMunoGeneTics information system®, Institut de Génétique Humaine, UPR CNRS 1142, Université de Montpellier, Montpellier, France, (safa.aouinti@igh.cnrs.fr, veronique.giudicelli@igh.cnrs.fr, patrice.duroux@igh.cnrs.fr, sofia.kossida@igh.cnrs.fr, marie-paule.lefranc@igh.cnrs.fr)*

[2] *École Supérieure de la Statistique et de l'Analyse de l'Information de Tunis, Unité Modélisation et Analyse Statistique et Economique, Tunisie, (dhafer.malouche@me.com)*

[3] *École Nationale d'Ingénieurs de Tunis, Tunisie*

Résumé. Les réponses immunitaires adaptatives de l'espèce humaine et d'autres espèces de vertébrés à mâchoires (gnathostomata) sont caractérisées par les cellules B et T et de leurs récepteurs d'antigènes spécifiques, les immunoglobulines (IG) ou anticorps et les récepteurs de cellules T (TR) (jusqu'à 2.10^{12} différents IG et TR par individu). IMGT®, the international ImMunoGeneTics information system® (http ://www.imgt.org) basé sur IMGT-ONTOLOGY a été créé pour gérer cette diversité. IMGT/HighV-QUEST est l'unique portail web pour l'analyse des séquences IG et TR '*big data*' obtenues par *next generation sequencing* (NGS). L'une de ses caractéristiques majeures est l'identification des clonotypes IMGT (AA) et en particulier de leur diversité et expression. Nous présentons une procédure statistique normalisée pour l'analyse des résultats d'IMGT/HighV-QUEST. Nous évaluons la significativité des différences en proportions de la diversité et l'expression des clonotypes IMGT (AA) entre deux lots de résultats par gène d'un groupe donné. Dans cet objectif, des tests d'hypothèses multiples sont appliqués avec ajustement des p-valeurs via les procédures de contrôle du *FWER* (Bonferroni, Holm, Hochberg, ŠidákSS et ŠidákSD) et du *FDR* (BH et BY).

Mots-clés. IMGT, IMGT/HighV-QUEST, IMGT-ONTOLOGY, immunoglobuline, anticorps, récepteur T, *big data*, *next generation sequencing (NGS)*, significativité statistique, différence en proportions, test d'hypothèses multiples.

Abstract. The adaptive immune responses of humans and other jawed vertebrate species (gnathostomata) 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) built on IMGT-ONTOLOGY was created to manage this huge diversity. IMGT/HighV-QUEST, the first web portal, and so far the only one, for the next generation sequencing (NGS) analysis of IG and TR big data sequences. One of its main features is the identification of IMGT clonotypes (AA) and in particular their diversity and expression. We present a standardized statistical procedure to analyze. We assess the significance of differences in proportions of diversity and expression of clonotype IMGT (AA), per gene of a given group, between two batches. Multiple tests

of hypotheses are conducted with p-values adjustment via FWER control procedures (Bonferroni, Holm, Hochberg, SidákSS and SidákSD) and FDR (BH and BY).

Keywords. IMGT, IMGT/HighV-QUEST, IMGT-ONTOLOGY, immunoglobulin, antibody, receptor T, big data, next generation sequencing (NGS), statistical significance, differences in proportions, multiple hypothesis testing.

1 Introduction

IMGT®, the international ImMunoGeneTics information system® (`http://www.imgt.org`) créé par Marie-Paule Lefranc en 1989, est l'unique système d'information intégré en immunogénétique et immunoinformatique [Lefranc et al. (2015)]. IMGT®, basé sur IMGT-ONTOLOGY [Giudicelli et al. (2012)], est à l'origine d'une nouvelle science, l'immunoinformatique [Lefranc et al. (2014)].

Les réponses immunitaires adaptatives de l'espèce humaine et d'autres espèces de vertébrés à mâchoires (gnathostomata) sont caractérisées par les cellules B et T et de leurs récepteurs d'antigènes spécifiques, les immunoglobulines (IG) ou anticorps [Lefranc and Lefranc (2001a)] et les récepteurs de cellules T (TR) [Lefranc and Lefranc. (2001b)] (jusqu'à 2.10^{12} différents IG et TR par individu). Les IG sont constituées de deux chaînes lourdes (H) et deux chaînes légères (L) où la chaîne L est une chaîne kappa ou lambda. Chaque chaîne est constituée par un domaine variable V et un ou plusieurs domaines constants C. Le domaine V est constitué des régions hypervariables ou CDR (*complementarity determining region*) qui déterminent le site de reconnaissance et de liaison à l'antigène et des régions dites charpentes ou FR (*framework region*).

Le domaine V résulte du réarrangement de trois gènes V, D et J (V-D-J-gene code le domaine VH des chaînes lourdes) ou de deux gènes V et J (V-J-gene code le domaine VL des chaînes légères kappa ou lambda). Ce sont ces réarrangements qui créent la diversité combinatoire lors de la synthèse des chaines d'IG (à laquelle s'ajoutent la N-diversité de la jonction et pour les IG, les mutations somatiques).

IMGT/HighV-QUEST est le portail de référence pour l'analyse des séquences d'immunoglobulines (IG) et récepteurs T (TR) obtenues par les technologies de séquençage next generation sequencing (NGS) [Alamyar et al. (2012)]. IMGT/HighV-QUEST permet l'analyse des répertoires d'anticorps ou des récepteurs T dans les réponses immunitaires en situation normale (vaccination) ou pathologique (infections, maladies autoimmunes ou cancer). IMGT/HighV-QUEST permet la caractérisation des clonotype IMGT (AA) et en particulier l'analyse de leur diversité et expression. Un clonotype désigné par 'IMGT clonotype (AA)' est défini par un réarrangement V-(D)-J unique, des ancres conservées (C104, W ou F118), une jonction CDR3-IMGT (AA) *in-frame* unique [Alamyar et al. (2012)]. Chaque 'IMGT clonotype (AA)' est caractérisé par une séquence représentative unique. Pour la première fois dans l'analyse des données NGS des récepteurs d'antigènes, l'approche standardisée d'IMGT permet une distinction claire entre la diversité des clonotypes (nombre des clonotypes IMGT (AA) par V, D ou J gène), et l'expression des clonotypes (nombre de séquences assignées à un clonotype IMGT (AA) donné par un V, D ou J gène, sans ambiguïté.) [Li et al. (2013)]. Dans ce travail présenté en détails dans l'article [Aouinti et al. (2015)] nous analysons les résultats d'IMGT/HighV-QUEST pour évaluer la significativité des différences en proportions de la diversité et l'expression des clonotypes IMGT (AA), par gène d'un groupe donné, entre deux lots ('sets') de résultats.

2 Méthodologie

2.1 Résultats d'IMGT/HighV-QUEST

La présentation de la méthodologie est basée sur un exemple de huit sets de données [Li et al. (2013)] analysés par IMGT/HighV-QUEST : deux populations de cellules T ($CD4^-$ et $CD4^+$) à quatre points temporels (avant vaccination H1N1 (Pre), jour 3 (d3), jour 8 (d8) et jour 26 (d26) après vaccination) d'un même individu.

2.2 Procédure statistique

2.2.1 Différence en proportions

Le but est de comparer, entre les deux sets de données, les différences en proportions des clonotypes IMGT (AA) par gène d'un groupe donné (la diversité des clonotypes).

2.2.2 Significativité des différences en proportions

Soit m le nombre de gènes (58 TRBV gènes, 2 TRBD gènes et 13 TRBJ gènes indexés par k=1, ..., m) et les 2 sets comparés indexés par $\{i, j\}_{i\neq j}$.
Soient $(X_r)_{r=1..n_i}^k \sim \mathcal{B}(p_i^{(k)})$ et $(Y_s)_{s=1..n_j}^k \sim \mathcal{B}(p_j^{(k)})$ (i.i.d.) ; n_i, n_j le nombre total de clonotypes dans chaque set :

$$(X_r)_{r=1..n_i}^k = \begin{cases} 1, \text{ si le } r^{ème} \text{ clonotype IMGT (AA) dans le set } i \text{ a le gène } (k) \\ 0, \text{ sinon} \end{cases}$$

$$\text{et } (Y_s)_{s=1..n_j}^k = \begin{cases} 1, \text{ si le } s^{ème} \text{ clonotype IMGT (AA) dans le set } j \text{ a le gène } (k) \\ 0, \text{ sinon} \end{cases}$$

$$\begin{cases} \widehat{p}_i^{(k)} = \frac{\text{Nombre des clonotypes IMGT (AA) dans le set } i \text{ ayant le gène } (k)}{\text{Nombre total des clonotypes IMGT (AA) dans le set } i} = \frac{1}{n_i} \sum_{r=1}^{n_i} X_r \\ \widehat{p}_j^{(k)} = \frac{\text{Nombre des clonotypes IMGT (AA) dans le set } j \text{ ayant le gène } (k)}{\text{Nombre total des clonotypes IMGT (AA) dans le set } j} = \frac{1}{n_j} \sum_{s=1}^{n_j} Y_s \end{cases}$$

Le test de comparaisons multiples pour une différence de deux proportions a été appliqué sous ces conditions : $n_i\widehat{p}_i^{(k)} \geq 5$, $n_i(1\text{-}\widehat{p}_i^{(k)}) \geq 5$ et $n_j\widehat{p}_j^{(k)} \geq 5$, $n_j(1\text{-}\widehat{p}_j^{(k)}) \geq 5$ et sous ces hypothèses :

$$\begin{cases} H_0^{(k)} : p_i^{(k)} = p_j^{(k)} = p \rightarrow p_i^{(k)} - p_j^{(k)} = 0 \\ H_1^{(k)} : p_i^{(k)} \gtrless p_j^{(k)} \rightarrow p_i^{(k)} - p_j^{(k)} \gtrless 0 \end{cases}$$

La statistique du test $z^{(k)} = \frac{\widehat{p}_i^{(k)} - \widehat{p}_j^{(k)}}{\widehat{\sigma}_{\widehat{p}_i^{(k)} - \widehat{p}_j^{(k)}}} = \frac{\widehat{p}_i^{(k)} - \widehat{p}_j^{(k)}}{\sqrt{\widehat{p}^{(k)}(1-\widehat{p}^{(k)})(\frac{1}{n_i}+\frac{1}{n_j})}} \sim \mathcal{N}(0,1)$ a été calculé avec $\widehat{p}^{(k)} = \frac{n_i\widehat{p}_i^{(k)} + n_j\widehat{p}_j^{(k)}}{n_i+n_j}$ pour accorder plus de poids à l'échantillon de plus grande taille.

2.2.3 Contrôle de l'inflation du risque d'erreur

Dans le cas de plusieurs tests d'hypothèses menés simultanément, un ajustement des p-valeurs doit être effectué par le biais des procédures de tests multiples. Les procédures Bonferroni, Holm, Hochberg, ŠidákSS and ŠidákSD, BH and BY ont été appliquées (voir Table 1). Ces procédures suivent l'un des deux types de stratégies pour contrôler l'inflation du risque d'erreur [Dudoit et al. (2008)] :

— *Family-wise error rate (FWER)* qui est la probabilité de rejeter à tort au moins une hypothèse vraie parmi les m hypothèses testées , i.e.,

$$FWER = \mathbb{P}(V \geq 1)$$

avec V le nombre de faux positifs.

— *False discovery rate (FDR)* de Benjamini & Hochberg qui est l'espérance de la proportion des erreurs de Type I parmi les hypothèses rejetées, i.e.,

$$FDR = \mathbb{E}(Q)$$

avec Q la variable aléatoire définie comme la proportion des erreurs accomplies par rejet à tort des hypothèses nulles, i.e., $Q = \begin{cases} \frac{V}{R} \text{ si } R > 0 \\ 0 \text{ si } R = 0 \end{cases}$ avec R le nombre d'hypothèses nulles rejetées.

Procédures	Contrôle	Structure	Dépendance (p-valeurs)$_{H_0}$	Propriétés
Bonferonni	FWER	*Single-step*	Générale/ignorance	la plus conservative
Šidák	FWER	*Single-step*	Indépendance	moins conservative que Bonferroni
Holm	FWER	*Step-down*	Générale/ignorance	moins conservative que Bonferroni
Step-down Šidák	FWER	*Step-down*	Dépendance	très similaire à Holm
Hochberg	FWER	*Step-up*	Indépendance	*step-up* de Holm
Benjamini & Hochberg (BH)	FDR	*Step-up*	Indépendance	la moins conservative
Benjamini & Yekutieli (BY)	FDR	*Step-up*	Générale/ignorance	plus conservative que BH

TABLE 1 – Propriétés des procédures de tests multiples

3 Résultats

3.1 Normalisation des clonotypes IMGT (AA)

La première étape de la procédure est de représenter, sous forme de diagrammes en barres juxtaposées, les proportions normalisées pour 10000 clonotypes IMGT (AA) par gène et par groupe (TRBV, TRBD ou TRBJ). La différence en proportions et l'intervalle de confiance (IC) à 95% ont été calculés pour chaque gène k d'un groupe donné (TRBV, TRBD et TRBJ) entre les deux populations de cellules T ($CD4^-$ et $CD4^+$) à 4 points temporels (Pre, d3, d8 et d26).

3.2 Significativité en utilisant les procédures de tests multiples

Afin d'évaluer la significativité des différences en proportions, les procédures de test multiples citées dans la Table 1 ont été appliquées. Les p-valeurs non ajustées et ajustées suite à l'application des ces procédures ont été calculées. Deux types de graphes ont été générés pour visualiser les résultats obtenus suite à l'ajustement des p-valeurs (Figure 1).

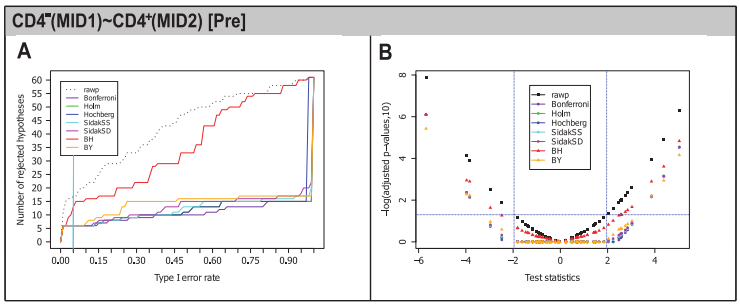

FIGURE 1 – Graphes des procédures de tests multiples. (A) : Graphique en courbes affichant le nombre d'hypothèses nulles rejetées en fonction des taux d'erreur de Type I. (B) : Nuage de points des logarithmes décimaux négatifs des p-valeurs non ajustées et ajustées en fonction des statistiques du test z-scores.

Le graphique en courbes qui permet de visualiser le nombre d'hypothèses nulles rejetées pour un seuil de significativité choisi (α) sous chaque procédure et le graphique des

nuage de points qui affiche les logarithmes décimaux négatifs des p-valeurs non ajustées et ajustées en fonction des z-scores.

3.3 Graphes de synthèse

Pour faciliter la comparaison avec les résultats expérimentaux, des graphes de synthèse qui combinent les diagrammes en barres normalisés des proportions et les différences en proportions avec les intervalles de confiances (IC) à 95% ont été générés (Figure 2). Les significativités obtenues en appliquant les procédures de test multiples sont reportées sur les graphes de différences en proportions. Les IC qui correspondent aux différences en proportions déclarées comme significatives par toutes les procédures de tests multiples sont colorés en bleu, par deux ou plusieurs procédures sont colorés en rose et uniquement par la procédure BH sont colorés en vert. Ce type de graphe permet d'avoir une vue d'ensemble sur l'évolution des proportions des clonotypes IMGT (AA) au cours du temps et entre les différentes cellules étudiées.

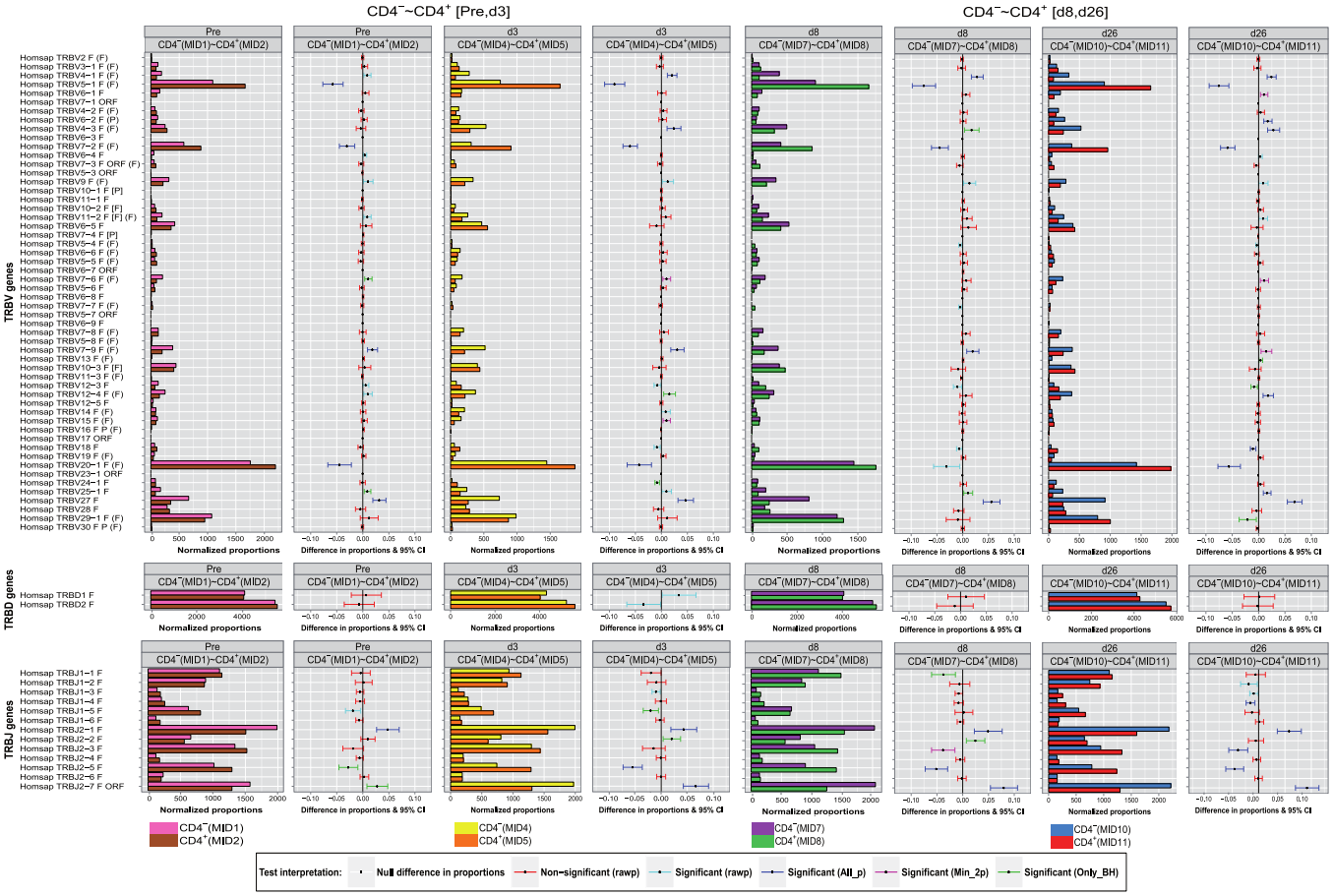

FIGURE 2 – Graphe de synthèse. Ce graphe affiche les clonotypes IMGT (AA) ayant un gène d'un groupe donné (TRBV, TRBD ou TRBJ) entre deux populations de cellules T ($CD4^-$ et $CD4^+$) à quatre points temporels (Pre, d3, d8 et d26).

Pour cet exemple étudié, un résultat potentiellement important en comparant les populations $CD4^-$ et $CD4^+$ est le fait que la vaccination n'a pas changé le signe des différences en proportions significatives trouvés avant et après vaccination. En d'autres termes, si la proportion des clonotypes IMGT (AA) ayant un gène k bien défini dans $CD4^-$ est supérieure à celle des $CD4^+$ avant vaccination, ce signe positif reste inchangé jusqu'au jour

26 (d26) après vaccination. Les caractéristiques du répertoire en terme de diversité des clonotypes IMGT (AA) par gène sont donc maintenues à un point temporel quelconque et les différences en proportions observées sont maintenues entre les deux populations de lymphocytes T ($CD4^-$ et $CD4^+$).

4 Discussion

Pour contrôler l'inflation du taux d'erreur dans le cas des tests multiples, nous avons appliqué les procédures contrôlant le FWER (Bonferroni, Holm, Hochberg, **Š**idákSS et **Š**idákSD) et le FDR (BH et BY). Nous avons observé que les procédures FWER, surtout Bonferroni, Holm et Hochberg, ont donné des p-valeurs ajustées très similaires (i.e., même nombre d'hypothèses nulles rejetées). Ceci est probablement dû à la structure d'indépendance des p-valeurs non ajustées. Nous avons également noté que les procédures FDR (BY et surtout BH) ont donné plus d'hypothèses nulles rejetées confirmant qu'elles sont plus puissantes et moins conservatrices que les procédures FWER. Les procédures FWER ont été critiquées pour les études génomiques comme étant des méthodes très strictes et conduisant à la perte d'informations dans certains cas traités. Notre étude confirme que le contrôle FDR est plus puissant que FWER, avec la procédure BH étant encore moins conservatrice que BY. Pour cette raison, une attention particulière devrait être accordée aux résultats obtenus par la procédure BH. L'analyse de ces répertoires est incontournable pour la comparaison des profils immunitaires en situation protectrice (vaccination, infections, cancers) ou pathogénique (auto-immunité, troubles lymphoprolifératifs). Les résultats ainsi présentés, qui prennent en compte les différentes procédures, sont primordiaux pour l'interprétation biologique. Notre méthodologie est adaptée pour évaluer la significativité de la différence de la diversité et l'expression pour un gène d'un groupe donné, ce qui permet de détecter des changements significatifs des profils immunitaires des répertoires à des temps différents entre populations B ou T d'un même individu ou entre individus.

Bibliographie

[1] Lefranc, M-P. et al. (2015) IMGT®, the international ImMunoGeneTics information system® 25 years. Nucleic Acids Res., 43 :D413-22.
[2] Giudicelli, V. and Lefranc, M-P. (2012) IMGT-ONTOLOGY 2012. Front. Genet., 3 :79.
[3] Lefranc, M-P. (2014) Immunoglobulin (IG) and T cell receptor genes (TR) : IMGT® and the birth and rise of immunoinformatics. Front Immunol. 5 :22.
[4] Lefranc, M.-P., Lefranc, G. (2001a). The Immunoglobulin FactsBook. Academic Press, London, UK, (458 pages).
[5] Lefranc, M-P. and Lefranc, G. (2001b) The T cell receptor FactsBook. Academic Press, London, UK (398 pages).
[6] Alamyar, E. et al. (2012) 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. Immunome Res., 8 :1 :2.
[7] Li, S. et al. (2013) IMGT/HighV-QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. Nature Comm., 4 :2333.
[8] Aouinti, S. et al. (2015) IMGT/HighV-QUEST statistical significance of IMGT clonotype (AA) diversity per gene for standardized comparisons of next generation sequencing immunoprofiles of immunoglobulins and T cell receptors. PloS ONE, 10(11) : e0142353.
[9] Dudoit, S., van der Laan MJ. (2008) Multiple testing procedures with application to genomics. Springer Series in Statistics.