Caractérisation de la diversité du répertoire TCR par modélisation de données de séquençage haut-débit

Wahiba Chaara

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité

BIOLOGIE DES SYSTÈMES

École doctorale Complexité du Vivant

Laboratoire Immunologie, Immunopathologie, Immunothérapie

Caractérisation de la diversité du répertoire TCR par modélisation de données de séquençage haut-débit

Par Wahiba CHAARA

Dirigée par les Pr. Adrien SIX et Pierre-André CAZENAVE

Présentée et soutenue publiquement le 27 septembre 2016

Devant un jury composé de :

M Gilles **FISCHER**	Président du jury
M Adrien **SIX**	Directeur de thèse
M Pierre-André **CAZENAVE**	Directeur de thèse
M Mathieu **GIRAUD**	Rapporteur
M Christophe **FERRAND**	Rapporteur
Mme Claudine **LANDES-DEVAUCHELLE**	Examinatrice
M François **ARTIGUENAVE**	Examinateur

À mon père, Allah y rahmo
À ma mère, Allah y hafda

Remerciements

Je suis arrivée au laboratoire il y a 8 ans, pensant y rester quelques mois…. Outre les choix que l'on fait, ce sont les personnes que l'on rencontre qui font évoluer nos vies. Je saisi donc cette opportunité pour remercier toutes les personnes qui ont fait mon quotidien pendant toutes ces années pour leur bienveillance, leur amitié et surtout pour l'impact qu'ils ont eu sur ma vie.

Je remercie en premier lieu ***David Klatzmann,*** *non seulement pour m'avoir engagé mais surtout pour sa confiance. J'ai beaucoup appris grâce à toi, tes attentes et ton exigence m'ont permis d'évoluer et de grandir au cours des années. Il est rare de trouver un environnement de travail comme le labo I3 et c'est à toi qu'on le doit.*
Merci à ***Adrien Six****. Il est des moments qui changent tout, ton arrivée au laboratoire en fait partie pour moi. J'ai appris mon métier avec toi. Ton exigence m'a permis de dépasser mes limites. L'autonomie et la liberté que tu m'as données m'ont fait gagner en assurance. Merci pour tout cela et pour tout le reste.*
J'adresse toute ma gratitude à ***Pierre-André Cazenave*** *pour avoir accepté la codirection de cette thèse et pour sa bienveillance. J'espère avoir été à la hauteur de vos attentes.*

Merci aux membres du jury d'avoir accepté de prendre part à ma soutenance et pour le temps que vous avez consacré à mon travail.
Merci à ***Pierre Boudinot*** *pour avoir participé à mon comité de thèse et pour les discussions enrichissantes que nous avons pu avoir au cours des années.*

Merci à ma « ***Husson Mourrier family*** *» :*

- ***Nicolas****, nous avons commencé ensemble, traversé ces années ensemble, cela semble naturel que l'on termine ce chapitre ensemble. Nous sommes différents sur pleins de points et pourtant on se rejoint sur l'essentiel. Que dire de plus à la personne avec laquelle on a partagé son quotidien pendant 8 ans à part tu vas me manquer…*
- ***PHP****, la personne qui trouve toujours le mot qu'il faut pour m'agacer ;-). Il y a un truc chez toi qui fait que malgré ça je n'arrive jamais à t'en vouloir. Merci pour ton esprit critique, tes blagues foireuses et ta gentillesse.*
- ***Djam****, entre nous ça passait ou ça cassait ; non seulement c'est passé mais tu es le petit frère que je n'ai pas eu. Pas sûre que tu lises ces mots donc je ne vais pas m'appesantir mais ces dernières années n'auraient pas été les mêmes sans toi (et ton rire cartoonesque ;-). J'ai hâte de voir les grandes choses que te réserve l'avenir…*
- ***Valentin****, le dernier arrivé des 4 affreux ! Avec ta bonne humeur et tes punchlines fulgurantes, tu t'es tout de suite intégré au groupe. Éternel insatisfait de l'analyse avec tes requêtes extravagantes, maintenant que tu prends la relève, tous nos espoirs reposent sur tes épaules ;-)*
- ***Encarnita****, ma seule team-mate au féminin dans ce monde de mecs ! Je ne saurais pas dire depuis combien de temps tu es au labo, tellement j'ai l'impression que l'on se connaît*

depuis toujours. Je ne te remercierai jamais assez pour ton investissement dans ma thèse mais aussi pour tes conseils et ton soutien à chaque fois que j'en ai eu besoin.

- ***Monsieur Claude Bernard**, les effets que tu as eus sur mon cerveau sont irréversibles ; je suis condamnée à voir des jeux de mots là où il n'y en a pas. Nos discussions et ton regard bienveillant vont me manquer.*
- ***Claire** et **Roberta**, merci pour votre bonne humeur et vos gentilles attentions.*
- ***Walid,** membre officiel mais à temps (très) partiel de la family. Merci pour les bons moments passés ensemble. You're the next !!*
- ***Iannis,** our discussions meant a lot for me. Thank you for your advices and for encouraging me to look forward.*

J'admire et je respecte chacun d'entre vous. Nous avons eu la chance de partager quelque chose de rare. Vous avoir autour de moi m'a permis de traverser ces dernières années particulièrement difficiles, c'est ce dont je suis le plus reconnaissante.

***Gwladys**, une des rares personnes de mon entourage qui est arrivée au laboratoire depuis plus longtemps que moi. Tu es un pilier de ce laboratoire. Ton homme et toi êtes certainement les personnes les plus adorables que je connais.*

*Merci à **Véronique**, pour les discussions enrichissantes que nous avons eues, ton intérêt pour mon travail et ta gentillesse.*

*Merci à **Sophie Miller** pour ta gentillesse, ta disponibilité et ton aide précieuse.*

*Merci à l'équipe **TriPoD**, j'espère que le succès du projet sera à la hauteur du travail fourni par chacun d'entre vous.*

*Merci aux collègues du **CERVI** (passé et présent) : AnnSo, Audrey, Fabien, Bertrand, Tristan, Laura, Caroline, Rachel, Férial, Thomas, Cornelia, Michelle, Claude Baillou et tous les autres que je n'oublie pas.*

*Merci à **Sonia** (ma petite sœur américaine), **Soumia**, **Rayane** pour vos encouragements ; vous êtes la preuve que les amitiés qui naissent au labo perdurent.*

***Marion,** merci pour ton soutien. Qui aurait cru qu'une conversation enclenchée à l'âge de 15 ans durerait aussi longtemps… et ce n'est pas prêt de s'arrêter ;-)*

***Laetitia**, on partage tout depuis plus de 10 ans. Merci d'être là pour me rappeler à l'ordre lorsque je déraille et m'encourager quand le besoin s'en fait sentir.*

*Merci à **ma famille** pour TOUT. Vous êtes pour beaucoup dans ce que j'ai pu accomplir. Vous avez toujours cru en moi et m'avez encouragé même si ce que je faisais vous semblez obscur. J'espère qu'après ma soutenance certaines choses s'éclaireront (ou tout du moins que vous admettrez que je ne suis pas technicienne de saisie ou tout autre métier que vous avez pu m'attribuer depuis toutes ces années). Je vous aime.*

Table des matières

Liste des Figures

Liste des Tables

Liste des Annexes

Liste des Abréviations

aa	acide aminé
Ac	Anticorps
ACP	Analyse en composante principale
ADN	Acide désoxyribonucléique
ADNc	Acide désoxyribonucléique complémentaire
Ag	Antigène
amTreg	Lymphocyte T régulateur activée mémoire
ARN	Acide ribonucléique
BCR	B-cell receptor
BLN	Brachial lymph node
CD4	Cluster de différenciation 4
CD8	Cluster de différenciation 8
CDR3	Complementary Determining Region 3
CMH	Complexe Majeur d'Histocompatibilité
CPA	Cellule Présentatrice d'Antigène
Ds	Indice quadratique de Simpson
exp(H)	Exponentielle de l'entropie de Shannon
exp(Ha)	Exponentielle de l'entropie de Rényi
Ig	Immunoglobuline
ILN	Inguinal lymph node
LB	Lymphocyte B
LT	Lymphocyte T
MH	Indice de Morisita-Horn
MLN	Mesenteric lymph node
NOD	Non-obese diabetic
nTregs	Lymphocyte T régulateur naif
PALN	Para-aortic lymph node
pCMH	Complexe peptide/CMH
PCR	Polymerase chain reaction
PLN	Pancreatic lymph node
RACE	Rapid amplification of cDNA-ends
RLN	Renal Lymph node
SPL	Spleen
TCR	T-cell receptor
Teff	Lymphocyte T CD4+ effectrice
TRA	Chaine α du TCR
TRB	Chaine β du TCR
TRBJ	Gène J codant pour la chaine β du TCR
TRBV	Gène V codant pour la chaine β du TCR
TRBVBJ	Combinaison de gènes TRBV et TRBJ
Treg	Lymphocyte T régulateur
Za	paramètre α d'une distribution Zipf

INTRODUCTION

A. Système immunitaire adaptatif

Le système immunitaire comprend deux composantes majeures : le système inné et le système adaptatif. Alors que le premier permet à l'ensemble des organismes une réaction défensive immédiate et rapide lors de l'intrusion d'un élément étranger, le système immunitaire adaptif est apparu au cours de l'évolution chez les vertébrés, les munissant ainsi d'une seconde ligne de défense, plus longue à mettre en place. Combinée à l'immunité innée, elle a vocation à procurer une couverture spécifique à large spectre (Kourilsky, 2014) permettant de faire face aux nombreux types d'infection et autres agressions (cancer, allergie, autoimmunité…).

1) Les acteurs de l'immunité adaptative

L'immunité adaptative cellulaire repose sur 3 piliers : (i) les lymphocytes B, (ii) les lymphocytes T et (iii) les cellules présentatrices d'antigènes, qui vont interagir pour mettre en place une réponse immunitaire adaptée à la situation rencontrée. Cette thèse ayant pour objet biologique d'intérêt le répertoire lymphocytaire T, une plus grande emphase sera accordée à cette population dans cette introduction, particulièrement à celle des lymphocytes CD4+.

Les lymphocytes représentent une proportion variable des leucocytes, communément appelés globules blancs, produits dans la moelle osseuse des vertébrés et circulant dans leur organisme.

Les **lymphocytes B** (LB) se différencient et sont sélectionnés dans un premier temps dans la moelle osseuse. Arrivés à maturation, ils expriment à leur surface un récepteur, appelé anticorps (Ac) membranaires, Immunoglobulines (Ig) ou BCR pour *B-cell receptor*, impliqué dans la reconnaissance des antigènes (Ag). Il est à noter que la plupart des molécules peuvent être des antigènes. Une fois en périphérie, chaque cellule lymphocytaire B exprime à sa surface environ 10^5 récepteurs identiques intégrant le même site de liaison à l'antigène (Kourilsky, 2014).

Les **lymphocytes T** (LT) quant à eux, se différencient dans le thymus suite à la migration de leurs précurseurs, dits progéniteurs ou pro-thymocytes, depuis la moelle osseuse. De manière comparable aux lymphocytes B, les lymphocytes T matures expriment à leur surface un récepteur membranaire, appelé TCR pour *T-cell receptor* (Tonegawa, 1983). Néanmoins,

contrairement aux LB qui reconnaissent l'antigène dans sa forme tridimensionnelle native, les LT reconnaissent les antigènes sous forme de peptides présentés par les molécules du **Complexe Majeur d'Histocompatibilité** (CMH). En effet, avant de pouvoir être reconnu par un TCR, l'antigène natif doit avoir été capturé, dénaturé par les cellules nucléées qui vont ensuite associer ses dérivés peptidiques aux molécules de CMH exprimé à leur surface. Les molécules du CMH, très polymorphes, sont distinguées en deux classes : les molécules de classe I (CMH-I) exprimées par la quasi-totalité des cellules nucléées et les molécules de classe II (CMH-II) uniquement exprimées par les cellules dites présentatrices d'antigène (CPA) (Marrack and Kappler, 1986). Cette dénomination englobe les lymphocytes B, les cellules dendritiques et les macrophages. Chaque molécule de CMH peut s'associer à des dizaines de milliers de peptides différents.

Lors de leur différenciation, les lymphocytes T se polarisent (revue par Singer et al., 2008; Thomas-Vaslin et al., 2008) pour finalement constituer deux grandes populations cellulaires caractérisées par l'expression à leur surface de deux glycoprotéines appelées CD4 et CD8 (CD : cluster de différenciation) (Marrack and Kappler, 1986; Murphy, 2012). Les **LT CD4+**, aussi appelés LT auxiliaires, jouent un rôle fondamental dans l'orchestration de la réponse immunitaire contre l'antigène. Les **LT CD8+**, ou LT cytotoxiques, ont, quant à eux, une action plus directe en détruisant des cellules cibles (Cantor and Boyse, 1975). Les marqueurs CD4 et CD8 conditionnent l'interaction du TCR avec le complexe peptide/CMH (pCMH) en interagissant avec le CMH lors de la présentation et vont également permettre la transmission du signal d'activation à la cellule (Irving and Weiss, 1991). Ainsi, les LT CD4+ reconnaissent le peptide antigénique présenté par les molécules de CMH-II à la surface d'une CPA alors que les LT CD8+ vont reconnaître un Ag présenté par les molécules CMH-I exprimées par l'ensemble des cellules nucléées (**Figure 1**) (Blackman et al., 1986; Wang and Reinherz, 2002). Dans ce contexte, un TCR doit rencontrer un antigène dans les circonstances adéquates pour induire un signal d'activation à la cellule T porteuse afin qu'elle prolifère et se différencie en cellules T effectrices nécessaires à la réponse immunitaire.

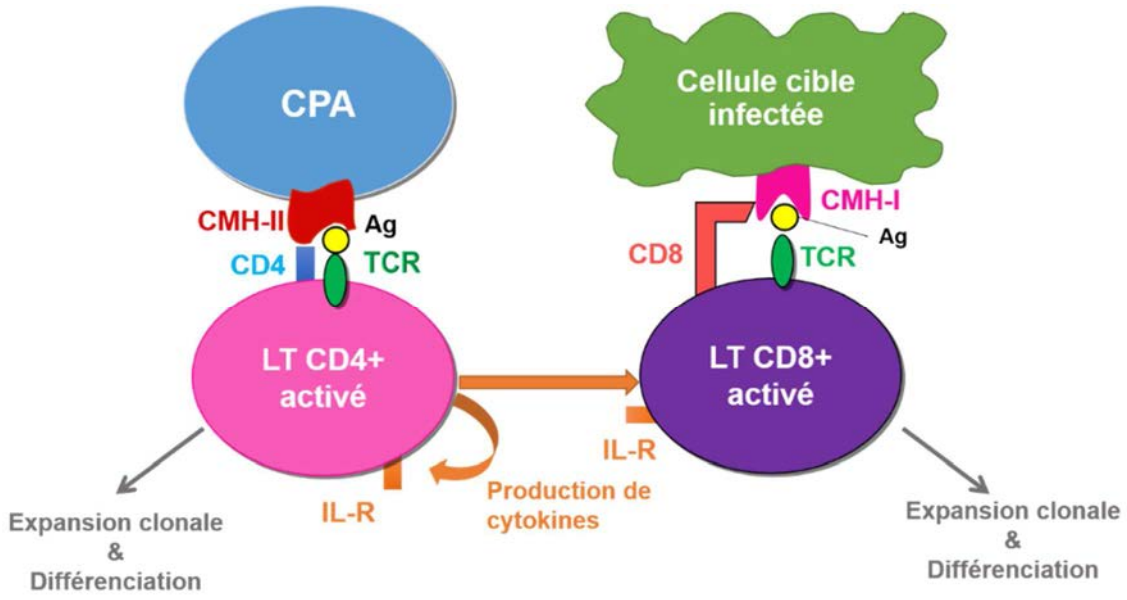


Figure 1 : Activation des lymphocytes dans le cadre d'une infection – Les cellules infectées, exprimant à leur surface des complexes pCMH-I, activent les LT CD8+ dont le TCR reconnaît le peptide antigénique. En parallèle, les cellules présentatrices d'antigènes (CPA) ayant internalisé l'antigène présentent à leur surface des complexes pCMH-II. Les LT CD4+ exprimant un TCR spécifique de complexes pCMH s'activent et sécrètent des interleukines (ou cytokines) qui vont stimuler l'expansion clonale et la différenciation des LT CD4+ mais surtout celles des LT CD8+ qui vont ainsi pouvoir détruire la cellule cible via les récepteurs aux interleukines (IL-R). (Inspiré de Chatenoud, 2008)

Un LT CD4+ activé, suite à l'interaction de son TCR avec un pCMH-II spécifique, va enclencher, via la sécrétion d'interleukines[1], les mécanismes de réponse immunitaire adéquate à l'élimination de l'antigène. En effet, la cellule LT CD4+ activée peut soit stimuler l'activation et l'expansion clonale de LT CD8+, déjà activés par le complexe pCMH-I exprimé par les cellules infectées, qui élimineront ces cellules (Wong and Pamer, 2003), soit contribuer à l'activation de LB, eux-mêmes entrés en contact avec l'antigène via leur immunoglobulines membranaires, et ainsi induire leur division et différenciation en LB effecteurs, dits plasmocytes, qui en secrétant leurs anticorps dans le sang permettront de neutraliser « la menace ».

Ainsi, on distingue deux branches de l'immunité adaptative : la branche dite humorale dont les lymphocytes B sont considérés comme les acteurs clés et la branche cellulaire au sein de laquelle se trouvent les LT.

[1] Groupe de cytokines, ainsi nommées car les premières observations semblaient montrer qu'elles étaient exprimées par les globules blancs (leucocytes, d'où -leukin) en guise de moyen de communication (d'où inter-).

On dénombre entre 1 à 2.10^8 LT au total chez la souris (Casrouge et al., 2000) et 10^{12} chez l'homme (Arstila et al., 1999). Ces cellules se distribuent en de nombreuses populations lymphocytaires, en général définies par leur profil d'expression de cytokines, facteurs de transcription et par leur phénotype. En 1995, S. Sakaguchi et son équipe ont identifié une population représentant 5 à 10% des LT CD4+, appelée **LT régulateurs** (Tregs) (Powrie & Mason, 1990 ; Fowell & Mason, 1993 ; Sakaguchi et al., 1995). Cette appellation est due à l'activité régulatrice de cette petite population LT CD4 + de l'activité effectrice des LT, en particulier les LT autoréactifs. En effet, lorsque cette population fait défaut chez la souris, on observe l'apparition spontanée de lésions systémiques dues à des troubles auto-immuns (Fontenot et al., 2003; Hori et al., 2003a). De plus, les patients atteints d'un syndrome IPEX[2] présentent une mutation du gène foxp3, gène caractérisant la population Tregs, qui entraîne une déficience des Tregs (Hori et al., 2003b). Bien que leurs mécanismes de sélection thymique soient toujours méconnus, de nombreuses études s'accordent sur le rôle clé du TCR dans la sélection des Tregs (Burchill et al., 2008; Lio and Hsieh, 2008) ainsi que dans leur activité suppressive (Levine et al., 2014; Zhu and Shevach, 2014). De plus, Hsieh et ses collègues ont montré que les TCR des Tregs en périphérie semblent interagir avec plus d'affinité avec les complexes CMH-II présentant un peptide du soi (Hsieh et al., 2004, 2006).
L'immunité adaptative résulte donc de la coopération de ses acteurs, les lymphocytes T et B, avec des entités initialement impliquées dans l'immunité innée (macrophages, cellules dendritiques…), travaillant ensemble pour permettre une réponse plus rapide et efficace de l'organisme lors d'une agression.

2) Une approche systémique pour un système complexe

Le système immunitaire regroupe donc un ensemble structuré d'entités, directement ou indirectement, interdépendantes. Ce réseau d'interconnections contribue à la robustesse du système immunitaire en assurant son bon fonctionnement en dépit des perturbations, généralement aléatoires, qui se produisent dans son environnement extérieur ou dans son milieu intérieur (Kourilsky, 2014). C'est l'imprévisibilité potentielle - non calculable *a priori*, des comportements du système qui lui confère un haut niveau de complexité.
Face à cette constatation, la nécessité d'une approche systémique prend tout son sens. En effet, combiner l'étude des entités et de leurs interactions semble essentiel à la bonne

[2] Syndrome d'Immunodérégulation, Polyendocrinopathie, Entéropathie auto-immune lié au chromosome X

compréhension des comportements physiologiques et pathologiques caractéristiques de ce système. L'approche systémique, théorisée par Ludwig von Bertalanffy en 1968 (Bertalanffy, 1993) fut longtemps ignorée par les immunologistes qui lui préféraient une approche réductionniste, notamment suite à l'avènement de la biologie moléculaire dans les années 1970. Or, l'approche réductionniste, bien que cruciale pour la description des entités (cellulaires et moléculaires) et des mécanismes impliqués dans les processus de réponses immunitaires, ne permet pas une analyse intégrative de ces processus. Ainsi, l'approche systémique a fini par prendre sa juste place en immunologie au début des années 2000 (Benoist et al., 2006; Gardy et al., 2009) permettant à d'autres champs de prendre de l'essor tels que la modélisation. Germain et ses collègues, dans une revue publiée en 2011, détaillent comment une liste exhaustive de questions immunologiques ont pu être abordées et trouver des réponses grâce à la modélisation informatique et/ou statistique (Germain et al., 2011). Bien que communément utilisée en biologie par le passé, notamment en immunologie, la modélisation n'a trouvé son sens qu'avec l'émergence de l'approche systémique. Au moyen de nouvelles technologies, les modélisateurs ont eu accès à des données riches en information, qui combinées aux connaissances acquises, leur permettent de construire des modèles solides, capables de reproduire et d'expliquer l'émergence de propriétés jusqu'à lors mal comprises.

> *Systems biology is an emerging discipline that combines high-content, multiplexed measurements with informatics and computational modelling methods to better understand biological function at various scales.*
>
> *(Germain et al., 2011)*

Les différentes populations lymphocytaires interagissant au sein du système immunitaire ainsi que les collections de cellules qui les composent sont elles-mêmes des systèmes complexes dont il est nécessaire d'étudier les comportements. Ainsi, capturer de manière systémique une image représentative de leur **composition** et de leur **topologie** permettrait une meilleure compréhension de leur dynamique et pourrait même servir à monitorer le statut physiopathologique des individus (Thomas et al., 2013a).

B. Qu'est-ce qu'un répertoire lymphocytaire ?

Un répertoire est défini comme un ensemble d'entités présentant des caractéristiques communes. Ainsi, le répertoire lymphocytaire décrit l'**ensemble des lymphocytes d'un individu, que ce soit en termes de populations cellulaires, de clones ou de fonctions**. De manière plus abstraite, le concept de répertoire immunitaire a été proposé pour décrire la diversité des lymphocytes impliqués dans le système immunitaire d'un individu dans un contexte physiopathologique donné (Boudinot et al., 2008). Si l'on considère un lymphocyte T capable de reconnaître un peptide antigénique donné, il exprimera un TCR en surface qui déterminera sa spécificité pour cet antigène. L'ensemble des lymphocytes T d'un individu peut ainsi être considéré comme une collection de lymphocytes « presque » tous différents au niveau de leur TCR, ayant chacun une capacité propre à reconnaître un ensemble de peptides antigéniques donnés. Ainsi, **le répertoire lymphocytaire disponible à un instant donné conditionne le répertoire des antigènes reconnus**.

1) La diversité du répertoire lymphocytaire

En 1901, Paul Ehrlich proposait la théorie selon laquelle la réponse immunitaire serait centrée sur l'interaction entre les antigènes et des récepteurs exprimés par les cellules. Cinquante ans plus tard, Niels Jerne (1955) fut le premier à décrire un mécanisme de sélection des anticorps, générés aléatoirement, en fonction de leur affinité pour les antigènes. Cette hypothèse fut développée ensuite par Franck Burnet qui la fit évoluer en ce qui aujourd'hui est considéré comme un paradigme fondamental de l'immunité adaptative : la **théorie de la sélection clonale** (Burnet, 1962, 1976). Ainsi, Burnet énonce que, suite à l'activation spécifique des lymphocytes B naïfs par un antigène[3], ceux-ci prolifèrent et se différencient en cellules effectrices qui éliminent l'agent pathogène, et en cellules mémoires, sorte de réserves prêtes à agir en cas de seconde réponse immunitaire immédiate, qui permettent de maintenir l'immunité.

Du fait de la diversité d'origine et de structure des antigènes que peut rencontrer un organisme, la diversité de son répertoire lymphocytaire se doit d'être d'une ampleur équivalente (Jerne, 1972). Ainsi, si chaque cellule exprime un seul et unique récepteur spécifique, cette théorie implique qu'il faudrait que l'organisme produise autant de

[3] À cette époque, seuls les lymphocytes B, producteurs d'anticorps, étaient décrits, pas les lymphocytes T

lymphocytes que d'antigènes potentiels. Or, les expériences de Karl Landsteiner ayant démontré au début du XIXème siècle qu'il n'était pas possible que l'animal puisse posséder dans son génome « fini » les informations nécessaires pour produire un nombre infini d'anticorps (Cziko, 1997), la théorie de la sélection clonale, bien que reconnue comme cruciale s'avérait incohérente. Finalement, ce paradoxe fut expliqué par S. Tonegawa qui décrivit le mécanisme de **recombinaison somatique** à l'origine de la diversité des anticorps (1974) puis des TCR (Chien et al., 1984).

> *The immune system does not attempt to predict the antibody structure that will bind with an antigen, but rather uses a type of "shotgun" approach that sends in a diverse army to meet the invaders.*
>
> *Cziko, 1997*

Ainsi, chaque LT exprime à sa surface des milliers copies du même récepteur TCR. Cet hétérodimère est formé par l'appariement de deux chaînes polypeptidiques transmembranaires, appartenant à la superfamille des immunoglobulines (Chothia et al., 1988; Hannum et al., 1984): une chaîne lourde et une chaîne légère. Chez l'homme et la souris, 95% des cellules T expriment un TCR composé d'une chaîne α (légère) et d'une chaîne β (lourde), les 5% restant exprimant un TCRγδ. Chaque chaîne du TCR possède un domaine N-terminal variable (V) et un domaine C-terminal constant (C). La partie variable des deux chaînes α et β possède trois domaines hypervariables (dits *complementarity determining regions*, CDR) (Chothia et al., 1988): les régions CDR1 et CDR2 vont (plutôt) interagir avec le CMH alors que les **CDR3** seront directement en contact avec le peptide (**Figure 2**). C'est donc la conformation du domaine Variable de chaque chaîne qui va déterminer la spécificité du TCR à un type d'antigène.

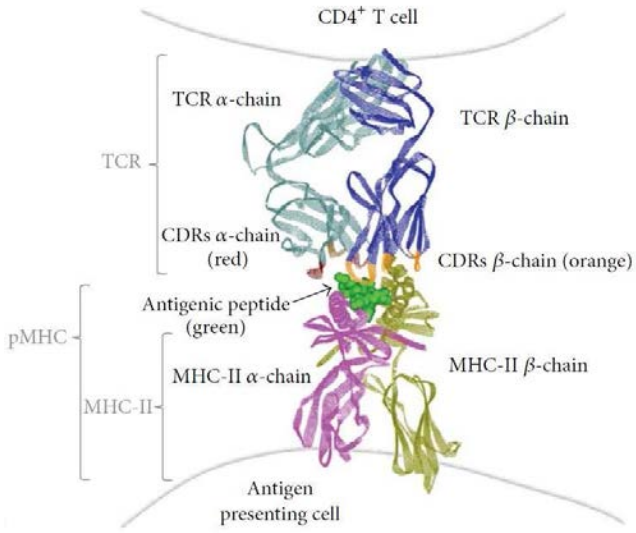


Figure 2 : Structure tridimensionnelle d'un TCR de LT CD4+ – La structure du récepteur des cellules T (TCR) et son ligand, le complexe peptide/CMH (pMHC). Le TCR est composé d'une chaîne alpha (bleu clair) et bêta (bleu foncé) liées par un pont disulfure. Les régions CDR des deux chaînes (en rouge et orange) permettent la reconnaissance du peptide antigénique (vert) présentés par les molécules CMH-II elles-mêmes composées d'une chaîne alpha (rose) et d'une chaîne bêta (vert clair) (Adaptée de Gonzalez, 2013).

Les chaînes composant le TCR ne sont pas codées *a priori* dans le génome. En effet, au sein des précurseurs des lymphocytes T, les locus tra et trb codant ces chaînes sont composés de gènes regroupés en trois familles : V (Variable), D (Diversité - pour les locus des chaînes β et δ uniquement) et J (Jonction) (Arden et al., 1985). Les gènes V, D, J et C codant pour les chaînes α et β sont localisés sur deux chromosomes différents (chromosomes 14 et 7 chez l'homme ; chromosomes 14 et 6 chez la souris). Suite à des séries de réarrangements somatiques ayant lieu lors de la différenciation lymphocytaire décrites **Figure 3** (Davis and Bjorkman, 1988), toutes les chaînes du TCR ont une structure primaire similaire : un gène V, un gène D (dans le cas des chaînes β et δ), un gène J et un gène constant C. Il en va de même pour l'architecture des chaînes composant les anticorps. Les régions de jonction entre les gènes V, (D) et J correspondent au CDR3 alors que les CDR1 et CDR2 sont intégralement codées au sein du gène V. Ce phénomène, spécifique du système immunitaire adaptatif, va donc concentrer l'essentiel de la diversité des chaînes de TCR et BCR au niveau de la région CDR3.

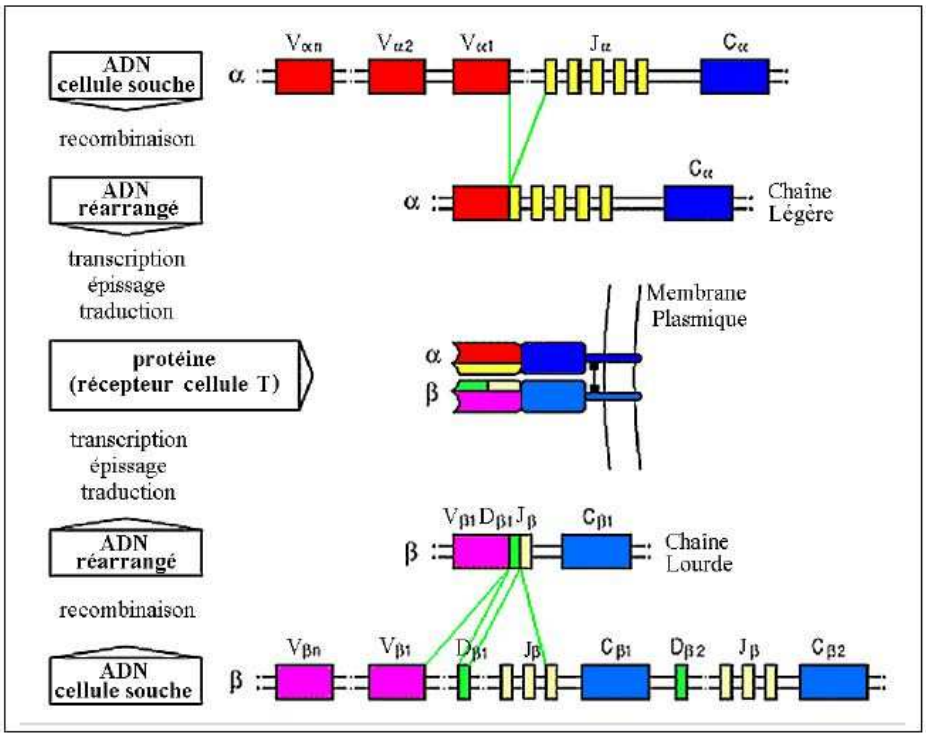


Figure 3 : Recombinaison somatique des gènes V(D)J pour la formation d'un TCR – En haut, le locus tra comprend en 5' le cluster de gènes V suivi du cluster central et de l'unique gène C (TRAC). Après recombinaison V-J et épissage VJC, un transcrit TRA productif composé d'un gène V, un J et d'un C adjacents est obtenu. En bas, le locus trb est composé d'un cluster de gènes V en 5' suivi de deux séries de clusters D, J et C. La recombinaison VDJ se déroule en deux étapes. La première consiste en la juxtaposition du gène D1 avec un des gènes J1 ou J2, du gène D2 avec un des gènes J2 puis la jonction obtenue est recombinée à un des gènes V. Après épissage, le transcrit TRB productif obtenu est composé d'un gène V, un D, un J et d'un C adjacents. (Tempel, 2007)

Comme illustré **Figure 3**, les gènes V sont, en général, situés en amont des gènes J (et D) qui précèdent les gènes C. Le nombre de gènes par famille varie en fonction des espèces et la variabilité combinatoire qui en découle est la première source de diversité du répertoire des TCR. Ainsi, comme résumé par la **Table 1**, le nombre de gènes V et J présents au locus de la chaîne α étant 3 fois plus grand que celui du locus β, la **diversité combinatoire** potentielle de la chaîne α est plus élevée.

Table 1 : Gènes codant la région variable des chaînes α et β du TCR chez la souris

Récepteur des cellules T αβ		
Nombre de gènes germinaux		
	Chaîne α	*Chaîne β*
V	79	21
D	0	2
J	38	11
Nombre de combinaisons possibles		
Jonction V-(D)-J	79x38= 3.10^3	21x2x11= $4.6.10^2$
Diversité combinatoire potentielle		$1.4.10^6$

Inspirée de Immunologie, 6ème édition © W.H. Freeman and Company, 2007

La seconde source de diversité du TCR est la **diversité jonctionnelle**. En effet, la juxtaposition et la ligature des gènes Vα-Jα et Vβ-Dβ-Jβ qui vont constituer le CDR3 se font par l'intervention d'une collection d'enzymes, appelées *VDJ recombinase*, parmi lesquelles des nucléases permettent l'excision au hasard d'un certain nombre de nucléotides de chaque extrémité (V et J) et la *Terminal Deoxynucleotidyl Transferase* (TDT) qui en rajoute de nouveaux de façon aléatoire à la fois en nature et en nombre (nucléotides N) (Motea and Berdis, 2010). A ces derniers, peuvent également s'ajouter des nucléotides P ainsi nommés car ils forment des séquences palindromiques aux extrémités des gènes. Ces mécanismes contribuent à la diversification en séquence et en longueur de jonction V(D)J impliquant les mêmes gènes. Néanmoins, cette diversité est considérablement réduite par la présence, au sein des gènes V et J, de séquences consensuelles assurant la présence respective d'un codon d'initiation et de terminaison de la traduction. Or, 2/3 des jonctions produites ne sont pas en phase ouverte de lecture ; la sélection thymique permettant de ne garder que les lymphocytes T disposant d'au moins d'une chaîne β puis α fonctionnelles pour former leur TCR.

> *Les hasards de l'évolution ont engendré et continuent d'engendrer, dans le monde du vivant une immense diversité, dont la gestion et le maintien supposent une spécificité commensurable.*
>
> *Kourilsky, 2014*

Le troisième élément contribuant à la diversité du TCR est l'**appariement des chaînes α et β**. Si le réarrangement des gènes V(D)J peut théoriquement avoir lieu sur les deux allèles du chromosome les portant, un même lymphocyte pourrait potentiellement produire deux

transcrits pour chacune des chaînes et donc exprimer plusieurs TCR différents à sa surface si ces transcrits sont productifs. Cependant, lors de la différenciation thymique, un processus d'exclusion allélique empêche tout réarrangement secondaire des chaînes β puis α dès le moment où le premier réarrangement est sélectionné positivement. Ainsi, lorsque les progéniteurs T reçoivent le signal de se différencier, ils amorcent le réarrangement d'un 1er allèle du locus codant pour la chaîne β (**Figure 4**). Si celui-ci est productif, la cellule reçoit un signal de survie, prolifère et amorce le réarrangement de sa chaîne α, sinon le second allèle β est recombiné. La chaîne β, étant la première à se réarranger, a donc un effet limitant sur la composition du TCR. En effet, à ce stade non seulement le réarrangement de l'allèle α doit être productif, mais en plus la chaîne générée doit être compatible à la chaîne β préexistante. Toutefois, l'exclusion allélique de la recombinaison α semble incomplète (Blüthmann et al., 1988). Ainsi, environ 1/3 des LT périphériques exprimeraient à leur surface deux TCR différents, associant une même chaîne β à deux chaînes α différentes suite à la recombinaison des deux allèles α (Marolleau et al., 1988 ; Padovan et al., 1993).

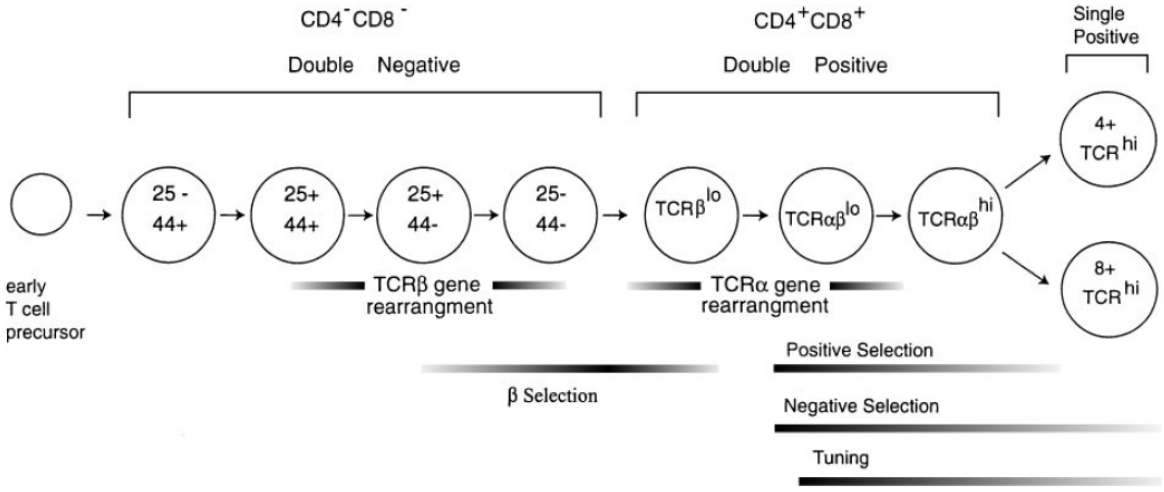


Figure 4 : Différenciation thymique des lymphocytes T – Les progéniteurs T commencent leur différenciation en n'exprimant ni CD4/CD8 ni TCR : c'est le stade Double Négatif. Ils procèdent au réarrangement de leur chaîne β au cours de cette étape. Les cellules dont le réarrangement β est productif sont sélectionnées et passent au stade Double Positif pendant lequel la chaîne α réarrange. Une fois les chaînes associées et le TCR complet exprimé, deux vagues de sélections, dites positive et négative, permettent d'éliminer les lymphocytes autoréactifs. (Sebzda et al., 1999)

Les différents processus de diversification décrits précédemment, notamment les diversités combinatoires et jonctionnelles, font que la diversité potentielle des TCRαβ est estimée à 10^{15} chez la souris (Davis and Bjorkman, 1988) et 10^{18} chez l'homme (Murphy, 2012). Elle est donc largement supérieure à la capacité cellulaire du compartiment lymphocytaire T périphérique

qui compte environ 2.10^8 lymphocytes T. Cette réduction de la diversité s'explique, entre autres, par les phénomènes de sélection que subissent les thymocytes lors de leur différenciation (Klein et al., 2014). Chaque clone de lymphocyte T se caractérisant par un TCR, la diversité des cellules T découle donc directement de la diversité des TCR (Pannetier et al., 1993). Or les TCR étant formés de manière stochastique sont théoriquement aussi bien capables de reconnaître des épitopes « étrangers », dits du non-soi, que des épitopes du « soi » (on parle dans ce dernier cas de lymphocytes autoréactifs). Les processus de sélection thymique permettent notamment d'éliminer ces cellules (**Figure 4**). Ainsi, seuls **5% des lymphocytes matures produits par jour dans le thymus sont sélectionnés** et rendus disponibles en périphérie (Thomas-Vaslin et al., 2008). Le répertoire périphérique sera ensuite modulé, en particulier au cours des réponses immunitaires.

Ainsi, alors que pour Jerne, le répertoire lymphocytaire englobe à la fois l'ensemble du potentiel de la diversité permise par les ressources génétiques de l'espèce et l'ensemble des récepteurs disponibles exprimés dans un tissu à un instant t (Jerne, 1971), on peut distinguer quatre niveaux d'organisation du répertoire lymphocytaire T : 1) le **répertoire génétique (ou potentiel)** , défini par les différentes formes alléliques des gènes germinaux V, D, J et C; 2) le **répertoire émergent**, produit par les mécanismes de réarrangements somatiques ayant lieu lors des phases précoces de différenciation des lymphocytes T dans le thymus ; 3) le **répertoire disponible**, résultant des mécanismes de sélection positive et négative au cours de la différenciation thymique ; la diversité TCRαβ disponible est estimée à 2.10^6 chez la souris (Casrouge et al., 2000) et 2.10^7 chez l'homme (Arstila et al., 1999; Naylor et al., 2005), enfin 4) le **répertoire effecteur**, modelé par le contexte antigénique en périphérie (Attaf et al., 2015; Harty and Badovinac, 2008; Nikolich-Zugich et al., 2004).

2) Spécificité de la reconnaissance par les récepteurs spécifiques d'antigènes

Si les lymphocytes B et T sont les piliers de la réponse immunitaire adaptative, leurs récepteurs, vecteurs de la reconnaissance antigénique et surtout de sa spécificité, en sont la clé de voûte.

Bien que structurellement similaires, ce sont leurs différences qui caractérisent le mieux le rôle de ces deux types récepteurs. Contrairement aux LB qui sécrètent les anticorps pour « attaquer » les agents pathogènes, les récepteurs des LT restent à leur surface et influent sur

les voies de signalisation cellulaires. De plus, comme expliqué plus tôt, les récepteurs des LB et LT ont des modes de reconnaissance de l'antigène différents puisque le premier reconnaît la forme native alors que le second ne peut l'identifier que s'il est dénaturé sous forme peptidique et présenté par une molécule de CMH. Toutefois, la différence majeure entre ces deux molécules reste la modulation de leur affinité. En effet, alors que l'affinité des récepteurs des LB connaît une « maturation » par le biais d'hypermutations, générant ainsi des LB de hautes affinités pour leur Ag, celle des TCR pour leur antigène est conditionnée de manière définitive par la séquence de leur CDR3.

Toutefois, le principe d'unicité de reconnaissance des TCR énoncé par la théorie de sélection clonale semble improbable. Ainsi, Don Mason, par son concept de **dégénérescence de la reconnaissance par le TCR** (Mason, 1998), énonce qu'*un même TCR doit pouvoir reconnaître différents complexes peptide/CMH pour permettre le large panel de reconnaissance des lymphocytes T et potentiellement la capacité d'une même population clonale à répondre à différents stimulus antigéniques.*

En 2005, Melvin Cohn a redéfini le principe de dégénérescence de la reconnaissance par le TCR en le distinguant de sa spécificité : *« La dégénérescence est un concept essentiellement fondé sur des interactions chimiques alors que la spécificité se traduit par la capacité d'une cellule à répondre biologiquement lors de son interaction avec un épitope, dépendant tout de même des interactions chimiques ».* Le terme de polyspécificité, défini plus tard, permet de combiner ces deux aspects : la spécificité de la reconnaissance du récepteur et sa capacité à reconnaître plusieurs ligands (Wucherpfennig et al., 2007).

De nombreuses études ont montré la capacité des TCR à reconnaître différents peptides antigéniques que leurs séquences soient grandement homologues ou complétement différentes, qu'ils soient ou non associés au même complexe CMH (Kersh and Allen, 1996; Wilson et al., 2004; revue par Wucherpfennig et al., 2007). Il a également été prouvé qu'un même complexe pCMH peut être reconnu par différents TCR (Pacholczyk et al., 2006; Wong et al., 2007)

Don Mason a estimé à 10^6 le nombre de complexes pCMH pouvant être reconnu par un seul TCR. Cependant, il suppose ici que la reconnaissance des peptides est libre de contraintes ce qui rend cette estimation erronée. En effet, la géométrie adoptée par le TCRαβ en s'amarrant au complexe pCMH va conditionner l'interaction spatiale entre les CDR1 et CDR2 des chaînes α et ß avec le peptide présenté par le CMH, la limitant à un nombre restreint d'acides aminés

situés au cœur du peptide (**Figure 5** - Degauque et al., 2016). Ainsi, la diversité de ces points d'ancrage étant faible, la spécificité de reconnaissance des TCR s'en trouve beaucoup moins forte et le degré de polyspécificité variable en fonction des TCR est possiblement bien plus élevé qu'anticipé (Su and Davis, 2013). De plus, certaines études tendent à montrer que l'étendue du panel de reconnaissance des LT semble être amplifiée par le fait que certains LT peuvent exprimer des TCR différents composés de deux chaînes α différentes (Heath and Miller, 1993; Padovan et al., 1993; Petrie et al., 1993).

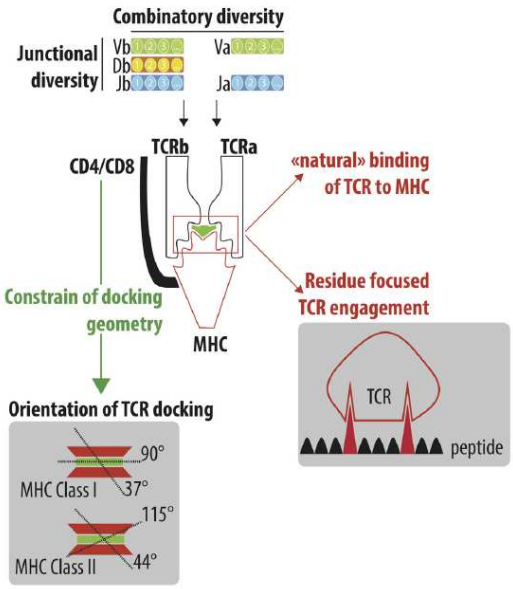


Figure 5 : Cross-réactivité du TCR – L'interaction entre le TCR et le pCMH est conditionnée par l'angle d'ancrage. L'angle médian du TCR est de 63.2° (min–max 37–90°) avec un CMH-I et de 76.4° (min–max 44–115°) avec un CMH-II. La zone d'interaction TCR-pCMH, qui se limite à quelques acides aminés, va donc varier en fonction du TCR (Degauque et al., 2016).

Ainsi, dans son article *« Why must T cells be cross-reactive? »*, Andrew K. Sewell argumente en faveur de la théorie de la polyspécificité qu'il qualifie de « vue systémique de la reconnaissance TCR » et évoque la possibilité que ce phénomène puisse être à l'origine de maladies autoimmunes en permettant aux LT porteurs de TCR autoréactifs de contourner les mécanismes de sélections thymiques et de s'échapper vers la périphérie (Sewell, 2012).

3) Plus qu'un objet biologique, un concept

Comme décrit précédemment, les récepteurs membranaires des lymphocytes B et T permettent à ces cellules de moduler les attributs caractéristiques de l'immunité adaptative : spécificité, diversité, mémoire et discrimination du soi et du non-soi (Kourilsky, 2014).

Un concept est une idée, une représentation de l'esprit qui abrège et résume une multiplicité d'objets empiriques (ou mentaux) par abstraction et généralisation de traits communs identifiables.

La description de la composition et de l'architecture de ces populations lymphocytaires permettrait une meilleure compréhension des mécanismes de l'immunité. La notion de répertoire permet de structurer cette description en utilisant des paramètres plus ou moins abstraits communs à toutes les cellules lymphocytaires. Ainsi, un répertoire lymphocytaire peut être défini en termes de populations ou de fonctions cellulaires mais également en termes de niveau d'organisation (potentiel, émergent, disponible, effecteur). De plus, le même répertoire peut être décrit à différents niveaux de granularité : usage des gènes V/D/J, longueur ou séquence du CDR3… Toutes ces abstractions font du concept de répertoire lymphocytaire un outil précieux pour la description des populations lymphocytaires lors du développement du système immunitaire ou dans le cas de pathologies telles que les maladies auto-immunes ou les cancers par exemple.

C. Description de la diversité du répertoire

1) Principe de diversité en biologie

La notion de « diversité » est définie comme l'état de ce qui est divers, ou en d'autres termes la pluralité. Le concept général est celui de la **répartition d'une quantité en un certain nombre de catégories bien définies**. Cette quantité pouvant être sous la forme de ressources, d'investissement, de temps, d'énergie, d'abondance… Une prise de conscience s'est opérée au cours du temps, plaçant la diversité, ou plutôt sa caractérisation, à une place prépondérante dans le cadre de nombreux domaines tels que la biologie, l'écologie, la physique, l'économie, la gestion ou la sociologie (diversité biologique, culturelle, génétique, diversité des climats…). Ainsi, par exemple, la diversité culturelle est considérée comme la constatation de l'existence de différentes cultures, au même titre que la diversité biologique et la diversité génétique sont les constatations de l'existence de multiples formes de vie dans la nature (Blondel, 2005).

Le terme « **biodiversité** » a été proposé en 1986 par Walter Rosen, un membre du National Research Council américain, lors du premier forum américain sur la diversité biologique. L'apparition du néologisme et la popularisation de ce concept a coïncidé avec la prise de conscience de la société de l'extinction d'espèces et ce concept a, au fil des années, fini par déborder de la sphère des sciences de la vie pour envahir celle des sciences humaine et sociale. À ce jour, il en existe une centaine de définitions (DeLong (1996) en recensait déjà 85

dans les dix premières années de littérature) gravitant autour de la notion de variation des formes de vie dans un écosystème donné. En 1987, l'*U.S. Office of Technologies Evaluation* a instauré une nouvelle définition étendant la biodiversité à la *« variabilité des organismes vivants de toute origine y compris les écosystèmes terrestres, marins, et autres écosystèmes aquatiques, et les complexes biologiques dont ils font partie »* (Magurran, 2004). Ainsi, cette nouvelle vision permet de mettre l'accent sur les dimensions intra- et inter-espèces de la diversité.

Selon Jacques Blondel (2005), ancien président de la Commission Scientifique de l'Institut Français de la Biodiversité, trois conceptions de la biodiversité peuvent être formulées, chacune étant associée à une définition du mot qui lui est propre :

- La biodiversité est un **concept abstrait** plus ou moins synonyme de « variété de la vie » dans sa totalité mais potentiellement organisable en champs thématiques aux contours plus précis. Néanmoins, d'après Blondel, un tel concept est tellement vaste qu'il en devient inapplicable à la connaissance scientifique et donc inaccessible à la recherche.

- La biodiversité peut également être considérée en tant qu'**entité** ou **hiérarchie d'entités objectives et mesurables au moyen d'outils appropriés.** Définie de cette manière, elle est désormais directement accessible à la méthode scientifique et ce, quelle que soit l'échelle du système étudié. Cette conception rend non seulement compte de ces entités hiérarchisées mais également des processus qui les relient et font fonctionner le système. Elle implique également une notion de dynamique de la biodiversité en considérant des échelles d'espace et de temps. Cette perspective fait donc de la biodiversité un concept lié aux différences et aux variations quantitatives mais aussi qualitatives entre entités biologiques.

- La biodiversité peut enfin être vue comme une **construction sociale**, économique et politique dont les enjeux relèvent des interactions étroites qui existent entre cette biodiversité et les sociétés humaines. Ces interactions peuvent être prises en compte dans l'optique du partage des avantages et des biens procurés par la biodiversité, mais aussi comme outil pour sa gestion et sa conservation. Ce type de biodiversité s'intéresse aux interactions entre l'homme et son milieu dans un contexte écologique, évolutif, mais aussi socioculturel.

Ces trois visions reflètent les multiples facettes que couvre la biodiversité, englobant une telle amplitude de champs et de domaines qu'elles rendent impossible la délimitation d'une discipline scientifique unitaire dotée de ses propres théories et lois. Aussi, personne ne peut prétendre définir complétement la biodiversité en la réduisant à son propre domaine d'intérêt

et de compétence car aucune variable ne peut « encapsuler » tous ses aspects. Il incombe au scientifique de trouver quelle conception convient le mieux à son sujet.

2) Comment mesurer la diversité ?

Un système biologique est régi par des processus opérant à trois grandes échelles (le temps, l'espace et les interactions). Aussi, étudier la diversité de ce système permet d'en comprendre sa genèse mais aussi les mécanismes assurant son maintien et son renouvellement. Du point de vue du biologiste, l'étude de la biodiversité peut être appliquée à une large gamme d'échelles spatiales et d'organisation tels que la génétique, la taxonomie ou la communauté, de manière à considérer d'un regard nouveau des disciplines classiques de la biologie.

En 1972, Robert H. Whittaker a décrit trois dimensions dans la mesure de la biodiversité (d'après la revue de Manley and Schlesinger, 2001) :

- La diversité **alpha** s'attache à la diversité au sein d'un même système. Elle est, en général, exprimée par sa richesse spécifique (c'est-à-dire le nombre d'espèces observées dans le système), un système étant défini par un ensemble d'entités et l'environnement physique qu'il occupe à un temps donné.
- La diversité **bêta** consiste à mesurer quantitativement le taux de variation en composition d'espèces entre les systèmes ou le long de gradients environnementaux par exemple. Elle reflète donc la modification de la diversité alpha entre les systèmes.
- La diversité **gamma** correspond au taux d'addition de nouvelles espèces quand on échantillonne le même « habitat » en différents endroits. C'est donc une notion plus globale et un indicateur plus tributaire des phénomènes globaux que des phénomènes locaux qui influent sur les diversités alpha et bêta.

Le biologiste s'intéresse habituellement à la diversité de trois entités biologiques : les gènes (diversité génétique), les espèces (diversité taxonomique) et les écosystèmes (diversité écosystémique). Dans ce contexte, la compréhension des mécanismes régulant la biodiversité implique de s'intéresser également au rôle fonctionnel des entités étudiées. En effet, établir les relations entre la diversité fonctionnelle d'un système et sa biodiversité est un enjeu majeur pour évaluer dans quelle mesure la perte de diversité peut influer sur les processus fonctionnels d'un système (Blondel, 2005).

Afin d'obtenir un maximum d'informations sur la diversité d'un système, il est nécessaire de disposer de sa description complète en termes de nombre d'entités différentes (c'est-à-dire

des gènes, des espèces et des fonctions écologiques), de leurs abondance et caractéristiques. Une telle description se décline en différentes distributions statistiques complexes. Pour la comparaison de deux systèmes, ou pour la description de l'évolution d'un même système au fil du temps, il est indispensable de condenser cette information en une valeur facile à calculer et à interpréter, bien que ce procédé sous-entende certainement une perte d'information (Baumgärtner, 2006). Ainsi, toutes les informations pertinentes quant à la diversité d'un système sont souvent condensées en un seul nombre réel, communément appelé « mesure de la diversité » ou « indice de diversité ».

3) Les indices de mesure de la Diversité

Un indice de diversité (aussi nommé indice de variabilité) est une mesure couramment utilisée afin de déterminer la variation des données catégorielles. En écologie, un indice de diversité est une statistique, qui est destinée à mesurer la biodiversité d'un écosystème. Plus généralement, les indices de diversité peuvent être utilisés pour évaluer la diversité d'une population dans laquelle chaque membre appartient à une unique espèce. Comme il existe de nombreux indices de diversité, fondés sur des approches différentes, et du fait de la complexité et des multiples paramètres caractérisant un même système étudié, il est essentiel d'être conscient des aspects de l'information réellement pris en compte dans le calcul de ces indices, notamment de ce qui minimisé, voire totalement ignoré afin d'évaluer leur pertinence d'application. Par exemple, alors que les indices utilisés en écologie tiennent compte de l'abondance, les indices économiques décident de l'ignorer (Baumgärtner, 2006). De plus, il est important de garder à l'esprit qu'un indice de diversité n'est pas nécessairement lui-même une valeur de la diversité. De nombreux indices sont en réalité des entropies qu'il faut nécessairement convertir en nombre effectif d'espèces pour pouvoir obtenir une interprétation homogène et intuitive de la diversité. Une recherche bibliographique dans les différents domaines s'intéressant à la « diversité » m'a permis de définir le concept en lui-même, comme présenté dans la première partie de cette section, mais également d'établir un inventaire des indices de mesure de la diversité existants.

Il existe différentes catégories d'indices de diversité, fondées sur :

- La richesse spécifique
- L'abondance relative des espèces
- La similarité des espèces

- ❖ Les caractéristiques des espèces
- ❖ La distribution des espèces

J'ai choisi de présenter dans cette section une liste non exhaustive des trois premières catégories d'indices qui sont les plus couramment cités dans la littérature.

Richesse spécifique

Bien que la mise au point d'indicateurs ou d'inventaires des fonctions des écosystèmes suscite beaucoup d'intérêt, la richesse en espèces (ou richesse spécifique) reste le moyen communément employé pour faire la synthèse des informations disponibles car facile à comprendre dans le contexte de méthodes d'évaluation par exemple. Il s'agit simplement d'un inventaire systématique du nombre d'espèces observées dans un système. C'est la méthode la plus fréquemment employée pour les études rapides d'impact relatives à l'évolution de la diversité.

Une richesse spécifique peut s'exprimer en richesse totale ou en richesse moyenne :

- La richesse totale correspond au nombre total d'espèces présentes dans le système
- La richesse moyenne correspond au nombre moyen d'espèces observées dans les échantillons du système étudié.

Équations 1.1 et 1.2 : Richesse Spécifique

$$\boldsymbol{R = S} \text{ ou } \quad \boldsymbol{R = \frac{S}{\sqrt{N}}}$$

avec S : le nombre d'espèces observées et N : la taille de l'échantillon

Deux hypothèses sont toutefois supposées par l'utilisation de cet indice :

- Les espèces sont bien connues : en écologie, compter le nombre d'espèces a peu de sens si la phylogénie n'est pas bien établie.
- Les espèces sont équidistantes : la richesse augmente d'une unité quand on rajoute une espèce, que cette espèce soit proche des précédentes ou extrêmement originale.

Les indices incluant l'abondance relative des espèces

L'abondance de différentes espèces dans un écosystème est décrite par la distribution des abondances absolues des individus à travers différentes espèces. Toutefois, plutôt que de ne prendre en compte que l'abondance absolue d'une espèce, il est plus intéressant de considérer l'abondance relative de cette espèce par rapport à toutes les autres espèces (Baumgärtner, 2006), une espèce représentée abondamment ou par un seul individu n'apportant pas la même contribution à l'écosystème. À nombre d'espèces égal, la présence

d'espèces très dominantes entraîne mathématiquement la rareté de certaines autres, on comprend donc assez intuitivement que le maximum de diversité sera atteint quand les espèces auront une répartition très régulière. Les variations d'abondance permettent de prédire les variations de richesse spécifique. En formulant des indices de diversité dans lesquels la contribution de chaque espèce est pondérée par son abondance relative dans l'écosystème, les écologistes ont, d'après Baumgärtner, trouvé le moyen d'intégrer la notion de rôle fonctionnel des espèces, considérant qu'une espèce rare va, par définition, moins contribuer à la biodiversité qu'une espèce commune.

- L'entropie de Shannon (incorrectement appelé indice de Shannon–Weaver ou encore de Shannon–Wiener) est l'un des indices de diversité utilisé pour mesurer la diversité dans les données catégorielles. Il a été introduit par Claude Shannon (Shannon, 1948) dans le contexte de la théorie de la communication, discipline qui étudie les principes de la transmission de l'information et les méthodes par lesquelles elle est livrée (comme l'impression, la radio ou la télévision…). L'entropie de Shannon est donc une fonction mathématique qui correspond à la quantité d'information contenue ou délivrée par une source d'information. Cette source pouvant être une langue, un signal électrique, ou un fichier informatique quelconque.

Équation 2 : Entropie de Shannon d'une variable aléatoire discrète X à n valeurs possibles

$$\boldsymbol{H(X) = -\sum_{i=1}^{S} P(X = xi)\,.\log\, P(X = xi)}$$

avec P(X=xi), la probabilité d'observer un symbole i

Ainsi, en considérant les espèces comme des symboles et en remplaçant leur probabilité par leur fréquence relative des espèces au sein d'une population, cet indice est le plus utilisé en écologie pour quantifier la biodiversité d'un milieu d'étude et observer son évolution au cours du temps. Sensible aux espèces rares, il prend en compte le nombre d'espèces et leur régularité au sein de la population. H sera donc minimal (=0) si tous les individus de la population appartiennent à une même espèce, ou si, dans une population, chaque espèce est représentée par un seul individu, à l'exception d'une, représentée par tous les autres individus de la population. Ainsi, l'entropie de Shannon seule ne permet pas de distinguer entre richesse et distribution inégale. L'indice est maximal quand tous les individus sont répartis d'une façon égale sur toutes les espèces.

- L'indice de Shannon est souvent accompagné de l'indice E de Piélou (Pielou, 1966).

Équation 3 : Indice d'équitabilité de Piélou

$$\boldsymbol{E = \frac{H}{Hmax}}$$

avec Hmax= log(S) et S, le nombre d'espèces dans l'échantillon

C'est une mesure de l'uniformité de la distribution des espèces au sein d'une population, normalisant la diversité observée à une distribution uniforme des espèces. Cet indice varie entre 0 (population comprenant des espèces dominantes) et 1 (très bon équilibre entre les espèces). Insensible à la richesse spécifique, il est très utile pour comparer des échantillons.

- L'entropie de Rényi a été proposée (1961)comme généralisation de l'entropie de Shannon. C'est une fonction mathématique paramétrique qui correspond à la quantité d'information contenue dans la probabilité d'observation d'une variable aléatoire.

Équation 4 : Entropie de Rényi d'ordre α d'une variable aléatoire discrète X à S valeurs possibles

$$\boldsymbol{H_{\alpha}(X) = \frac{1}{1-\alpha}\log\sum_{i=1}^{S} P(X = xi)^{\alpha}}$$

avec un paramètre réel α tel que α > 0 et ≠ 1.

L'entropie de Shannon est obtenue à partir de cette équation quand α tend vers 1. Comme elle, la richesse spécifique (α = 0), l'indice de Simpson (α = 2), et l'indice de Berger-Parker (α = +∞), décrits ci-dessous, sont considérés comme des cas particuliers de l'entropie de Rényi.

- L'indice de Simpson, introduit en 1949 par Edward Hugh Simpson, mesure la probabilité que deux individus, sélectionnés au hasard, appartiennent à la même espèce.

Équation 5.1 : Indice de concentration de Simpson

$$\boldsymbol{D = \sum_{i=1}^{S} P(X = xi)^2}$$

Néanmoins, du fait de ses valeurs contre-intuitives (Dmin=1 ; Dmax =0), il est le plus souvent utilisé dans une version indiquant la probabilité que deux individus, choisis au hasard au sein d'un échantillon, appartiennent à des espèces différentes.

Équation 5.2 : Indice de Simpson

$$\boldsymbol{Ds = 1 - \sum_{i=1}^{S} P(X = xi)^2}$$

Il est défini dans l'intervalle [0 ; 1[, aura une valeur de 0 si une seule espèce est observée, et atteindra sa valeur maximale (1 - 1/S) si les S espèces observées sont équivalemment représentées. La valeur de 1 est atteinte pour un nombre infini d'espèces.

Cet indice donne davantage de poids aux espèces abondantes qu'aux espèces rares puisque le fait d'ajouter des espèces rares à un échantillon ne modifie pratiquement pas la valeur de l'indice de diversité.

Une troisième version permet également de surmonter le problème de la nature contre-intuitive de l'indice de Simpson.

Équation 5.3 : Indice de diversité quadratique

$$Q = {}^{1}/_{D}$$

Cette redéfinition, appelée indice de diversité quadratique, rend l'indice plus sensible. La valeur minimale (Q= 1) correspond à une communauté contenant une seule espèce. Plus la valeur est élevée, plus grande est la diversité. La valeur maximale est le nombre d'espèces (ou d'une autre catégorie utilisée) dans l'échantillon. Par exemple, s'il y a cinq espèces dans l'échantillon, la valeur maximale sera de 5. Cette formule lie l'indice de Simpson à l'entropie de Rényi pour une valeur de α=2 tel que :

$$H_2(X) = \log \frac{1}{\sum_{i=1}^{S} P(X = xi)^2}$$

L'indice de Simpson est comparable à celui de Shannon pour la mesure de la diversité locale (α). Toutefois, moins sensible aux espèces rares, l'indice de Simpson est considéré comme étant plus significatif en termes d'écologie lorsqu'il est utilisé seul.

– L'indice de diversité de Hill mesure de l'abondance de manière proportionnelle, en associant les indices de Shannon et de Simpson (Hill, 1973).

Équation 6 : Indice de diversité de Hill

$$Hill = \frac{Q}{exp(H)}$$

L'indice de diversité de Hill permet une évaluation encore plus précise de la diversité observée. L'indice de diversité quadratique (1/D) va permettre la mesure du nombre effectif d'individus très abondants alors que « exp(H) » va en revanche mesurer le nombre effectif d'individus abondants et rares. Plus l'indice de Hill s'approche de la valeur 1, et plus la diversité est faible. Afin de faciliter l'interprétation, il est alors possible d'utiliser l'indice « 1-Hill ». L'indice de Hill semble le plus pertinent pour la comparaison de populations différentes. Toutefois, il peut être utile d'utiliser les trois indices conjointement afin d'en extraire un maximum d'informations et de mieux comprendre la structure des communautés.

➢ L'indice de dominance de Berger-Parker utilise une information partielle sur l'abondance relative des différentes espèces puisqu'il ne considère que l'abondance relative l'espèce la plus commune du système (Berger and Parker, 1970).

Équation 7 : Indice de Berger-Parker

$$\boldsymbol{BP = \frac{1}{\max P(X = xi)}}$$

L'indice prend seulement en compte la dominance relative de l'espèce la plus commune en négligeant les autres.

Ainsi, on remarquera que plus α est élevée dans l'équation de Rényi, plus l'indice obtenu met de poids sur les espèces les plus abondantes du système tout en étant moins sensible aux différences minimes d'abondance et de richesse des espèces dans la globalité. Pour α = 1, on élimine l'abondance relative des espèces les plus fréquentes ; pour α = 2, on élimine les secondes plus fréquentes… (Baumgärtner, 2006). Par exemple, si deux communautés ne diffèrent que sur le nombre et la composition d'espèces très rares, l'indice de Berger-Parker sera complétement insensible à ces différences, tout comme l'indice de Simpson. L'indice de Shannon-Wiener (α = 1) est plus sensible aux petites différences dans l'abondance relative que l'indice de Simpson, mais seule la richesse des espèces (α = 0) prend pleinement en compte le plus grand nombre des espèces très rares (**Figure 6**).

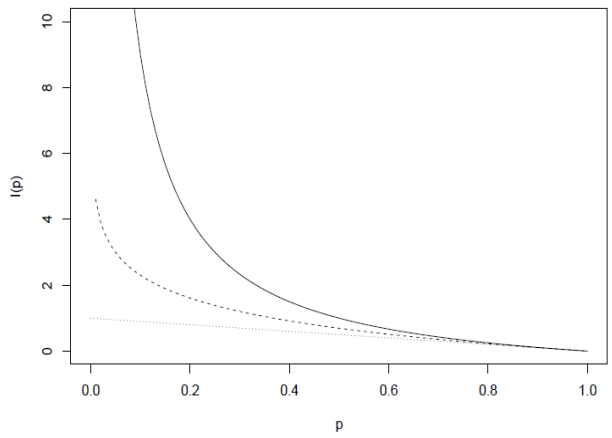


Figure 6 : Fonctions d'informations des espèces rares – Quantité d'information apportée par la richesse spécifique (trait plein), l'indice de Shannon (pointillés longs) et l'indice de Simpson (pointillés) en fonction de la fréquence des espèces. L'information apportée par l'observation d'espèces rares décroît de la richesse spécifique à l'indice de Simpson. (Marcon, 2015)

Comme précisé précédemment, les indices appartenant à la famille de l'entropie de Rényi doivent être convertis en nombre effectif d'espèces pour être interprétés correctement. Les modalités de conversion sont résumées dans la table ci-dessous.

Table 2 : Conversion des entropies en vraie diversité – Adaptée **de Jost, 2006**

pour pi= P(X=xi)	Formule	Diversité	Diversité en termes de p_i
Richesse Spécifique	$R = \sum_{i=1}^{S} p_i^0$	R	$\sum_{i=1}^{S} p_i^0$
Entropie de Shannon	$H = -\sum_{i=1}^{S} p_i \ln p_i$	exp(H)	$\exp(-\sum_{i=1}^{S} p_i \ln p_i)$
Concentration de Simpson	$D = \sum_{i=1}^{S} p_i^2$	1/D	$1/\sum_{i=1}^{S} p_i^2$
Indice de Simpson	$Ds = 1-\sum_{i=1}^{S} p_i^2$	1/(1-D)	$1/\sum_{i=1}^{S} p_i^2$
Entropie de Rényi	$Hq = (-\ln \sum_{i=1}^{s} p_i^q)/(q-1)$	exp(Hq)	$(\sum_{i=1}^{S} p_i^q)^{1/(1-q)}$

Les indices de similarité

Ce type d'indice sert à évaluer la ressemblance entre deux échantillons en faisant le rapport entre les espèces communes aux deux et les espèces propres à chacun.

– L'indice de Sørensen (1948) est une mesure très simple de la biodiversité β, dont la valeur varie de 0 quand il n'y a pas d'espèces communes entre les deux communautés, à 1 lorsque les mêmes espèces existent dans les deux communautés.

Équation 8 : Indice de similitude de Sørensen

$$S = \frac{2\,|A \cap B|}{|A| + |B|}$$

– L'indice et la distance de Jaccard (1901) sont deux métriques utilisées en statistiques pour comparer la similarité et la diversité entre des échantillons. Ils sont nommés d'après le botaniste suisse Paul Jaccard et sont notamment utilisés pour étudier la diversité lexicale de textes.

L'indice de Jaccard (ou coefficient de Jaccard) est le rapport entre la taille de l'intersection des ensembles considérés et celle de l'union des ensembles. Initialement développé pour évaluer la similarité entre deux ensembles, son application peut être étendue à *n* ensembles.

Équations 9.1 et 9.2 : Indices de Jaccard

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad \text{ou} \quad J(S_1, \ldots, S_n) = \frac{|S_1 \cap S_2 \cap \ldots \cap S_n|}{|S_1 \cap S_2 \cup \ldots \cup S_n|}$$

La distance de Jaccard mesure la dissimilarité entre les ensembles.

Équation 9.3 : Distance de Jaccard

$$\boldsymbol{J_\delta\,(A,B) = 1 - J(A,B)}$$

Uniquement basés sur l'observation des espèces, ces indices de similarité sont particulièrement sensibles à la taille des échantillons et ne tiennent pas compte de l'abondance relative des espèces dans ces échantillons, négligeant de fait les espèces rares. C'est pourquoi Anne Chao (2006) a proposé une version ajustée des deux précédents indices.

– L'indice de Morisita-Horn permet d'évaluer le niveau de similarité de structures entre 2 communautés. Initialement proposé par Morisita en 1959, ce coefficient de chevauchement a été redéfini par Horn en 1966.

Équation 10 : Indice de Morisita-Horn

$$MH = \frac{2\,.\sum_{i=1}^{S} \frac{a_i}{A}\frac{b_i}{B}}{\left(\frac{\sum_{i=1}^{S} a_i^2}{A^2} + \frac{\sum_{i=1}^{S} b_i^2}{B^2}\right)}$$

avec a_i : occurrence de l'espèce i dans l'échantillon A ; b_i : occurrence de l'espèce i dans l'échantillon B ; S : le nombre d'espèces uniques à travers les 2 échantillons.

À l'exception de l'indice de Morisita-Horn, tous les indices de similarité sont influencés par la taille de l'échantillon et sa richesse en espèces (Wolda, 1981). Toutefois, un inconvénient de l'indice de Morisita-Horn (MH) est qu'il est très sensible à l'abondance des espèces les plus abondantes. Ainsi, une généralisation, le NESS (*Normalized Expected Species Shared*), proposée par Grassle and Smith (1976; Wolda, 1983) évalue la similarité de deux populations sur la base du nombre d'espèces partagées entre des échantillons de taille fixe *m* tirés aléatoirement dans chacune d'entre elles. Pour m = 1, NESS est équivalent à l'indice de Morisita.

Équation 11 : Indice NESS entre les populations A et B

$$\boldsymbol{NESS(A,B,m) = \frac{2\ \sum_{i=1}^{S} \mu_{iA}(m)\,.\,\mu_{iB}(m)}{\sum_{i=1}^{S}[\mu_{iA}(m)]^2 + \sum_{i=1}^{S}[\mu_{iB}(m)]^2}}$$

avec $\mu_i(m) = 1 - (1 - p_i)^m$; m, la taille de l'échantillon et S, le nombre d'espèces au travers des deux populations.

Toutefois, afin de ne pas introduire un biais par un tirage aléatoire unique, on pourrait envisager de calculer cet indice à travers plusieurs itérations de tirages pour s'assurer de sa robustesse.

De nombreux outils implémentent un panel plus ou moins large de ces indices de diversité : les packages R diveRsity (Keenan et al., 2013), vegan (Oksanen, 2011), entropy (Strimmer, 2014) ou entropart (Marcon and Herault, 2016) mais aussi BiodivR (Hardy, 2010) ou EstimateS (Colwell, 2005).

Les indices décrits ici sont des outils objectifs et informatifs sur la diversité des espèces d'un système. Cette démarche peut être appliquée à l'étude de la diversité du répertoire immunitaire.

4) Un répertoire ou un écosystème ?

L'écologie peut être définie comme l'étude des interactions qu'ont les organismes les uns avec les autres et avec leur environnement. Elle s'intéresse à la distribution et à l'abondance de ces organismes au sein et entre les écosystèmes mais également la relation entre la diversité d'un écosystème et la stabilité de celui-ci. Un écosystème est défini comme un système formé par un environnement et par l'ensemble des espèces qui y résident et y interagissent. C'est un système organisé au sein duquel chaque espèce joue un rôle dans le maintien de l'équilibre global. Il constitue une unité fonctionnelle caractérisée par sa dynamique qui s'adapte en fonction de l'évolution de l'environnement et des conditions extérieures.

Le répertoire lymphocytaire répond également à ces critères. En effet, il se définit non seulement par les cellules qui le composent (les espèces) mais également par le contexte physiopathologique et l'organe – plus largement, l'organisme – dans lequel il est observé (environnement). Par ailleurs, il existe une multitude de définitions du concept d'espèce en écologie, qui varient en fonction des critères pris en compte par l'observateur. Ainsi, on peut définir une espèce morphologique (individus avec les mêmes caractéristiques structurales), phylogénétique (individus présentant une combinaison unique de caractères diagnostiques), écologique (individus occupant la même niche écologique), phénétique (individus se « ressemblant » plus entre eux qu'à d'autres ensembles équivalents). Similairement, un répertoire immunitaire donné peut être décomposé d'après de nombreux facteurs (chaîne, usage des gènes…), ce qui lui confère une structure caractéristique. Les lymphocytes occupant un espace limité, la taille des clones dépendent les unes des autres de manière à conserver l'homéostasie entre les espèces. Ainsi, l'étude de sa diversité permet d'évaluer le statut physiopathologique de l'individu.

Enfin, que ce soit en écologie ou en immunologie, la diversité d'un système est généralement perçue assez positivement car considérée comme une sorte d'assurance contre l'imprévu. On voit là l'intérêt de monitorer cette diversité afin de s'assurer de son maintien et la restaurer le cas échéant (Blondel, 2005).

Pour ce faire, les trois dimensions de la mesure de la biodiversité peuvent s'appliquer à l'analyse de la diversité lymphocytaire. La diversité alpha permettant d'évaluer la diversité au sein d'un répertoire donné, la diversité bêta permettra de comparer les diversités alpha de deux populations cellulaires ou de la même population avant et après traitement par exemple. Enfin, la diversité gamma, quant à elle, permettra d'exprimer les différences de diversité observées dans la même population cellulaire au travers de plusieurs organes par exemple. Ainsi, de plus en plus d'études empruntent des outils à l'écologie pour décrire la diversité du répertoire immunitaire (Laydon et al., 2015; Thomas et al., 2014; Wu et al., 2015).

5) Évolution de l'approche en parallèle de la technologie

De nombreuses approches ont été développées pour estimer la diversité des répertoires immunitaires (Six et al., 2013). Le répertoire TCR étant un système multi-entitaires, il peut être observé par différentes approches dont la granularité a varié avec les progrès technologiques tel que revu par (Boudinot et al., 2008).

Le répertoire TCR peut être décrit au niveau protéique par **cytométrie en flux** en utilisant des anticorps monoclonaux spécifiques des familles V de la chaîne β (BV) mesurant ainsi leur fréquence d'utilisation au sein d'une population lymphocytaire donnée (Ciupe et al., 2013; Salaün et al., 1990; Thomas-Vaslin et al., 2012). Bien que cette approche permette une analyse qualitative et quantitative du domaine Variable des TCR, elle est limitée par la disponibilité des anticorps monoclonaux et ne permet pas d'évaluer la diversité de la jonction CDR3. D'autres approches protéomiques plus sensibles telles que le PANAMA-blot (Nobrega et al., 1993) furent développées pour l'analyse du répertoire des lymphocytes B mais ne sont pas applicables pour le TCR.

L'analyse moléculaire du répertoire TCR peut se faire tant quantitativement que qualitativement. Ainsi, de nombreuses stratégies de **PCR** ont été mises au point pour cibler spécifiquement les combinaisons V-C des régions variables permettant une évaluation quantitative de leur usage (Matsutani et al., 1997; VanderBorght et al., 1999). Une autre approche fut d'utiliser des **puces à ADN**, supports de sondes ciblant par exemple les gènes

TRAV et/ou TRBV (Matsutani et al., 1997; Ogle et al., 2003). Alternativement et de manière plus qualitative, de nombreuses études ont eu recours à l'analyse de la diversité des transcrits TR par **clonage bactérien et séquençage de Sanger** (Moss and Bell, 1995; Moss et al., 1992; Rosenberg et al., 1992). Toutefois, ces techniques ne permettent pas une caractérisation globale de la diversité du répertoire. C'est pourquoi l'approche **Immunoscope®** (INSERM, Paris), aussi appelée spectratypage du CDR3, a été mise au point par l'équipe de Philippe Kourilsky au début des années 1990 (Cochet et al., 1992; Pannetier et al., 1993, 1995). Cette technique permet de décrire la diversité des répertoires lymphocytaires complets par l'analyse de la distribution des longueurs de la région hypervariable CDR3. Brièvement, les ARNm sont rétro-transcrits et amplifiés par des PCR en série ciblant les réarrangements V-C ou V-J puis séparés sur la base de la longueur des transcrits (Pannetier et al., 1993) générant ainsi pour chaque amplicon un spectratype (**Figure 7**). Chaque pic du spectratype représente une longueur de CDR3 mais peut correspondre à plusieurs séquences de CDR3 différentes. Dans des conditions physiologiques, la distribution de ces longueurs est visualisable par un profil de six à huit pics ayant l'allure d'une gaussienne caractérisant une réponse polyclonale (Pannetier et al., 1995). En cas de perturbation, un ou plusieurs pics prédominants seront le reflet de d'expansions monoclonales ou oligoclonales des lymphocytes.

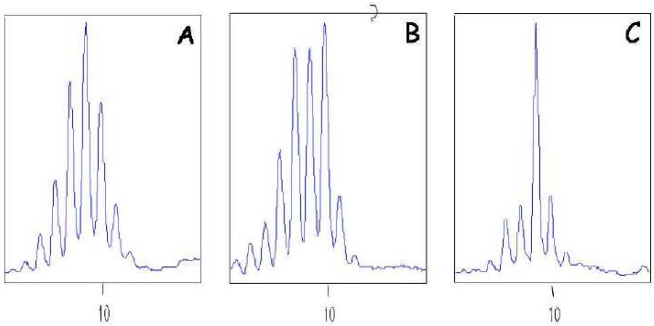


Figure 7 : Exemples de profils des distributions de longueurs de CDR3 – **(A)** Profil de type gaussien dans conditions physiologiques. **(B)** et **(C)** Profils perturbés reflétant respectivement une expansion oligoclonale ou monoclonale (d'après des résultats obtenus au sein du laboratoire dans le cadre du projet de neuropaludisme expérimental chez la souris).

L'étude du répertoire a subi des vagues successives d'engouement et de désintérêt. Le développement d'Immunoscope® a ouvert la voie à des études à plus grande échelle cherchant à identifier des expansions clonales en réponse à des peptides antigéniques particuliers (Bousso et al., 1999; Casrouge et al., 2000) ou à caractériser le développement lymphocytaire (Arstila et al., 1999; Regnault et al., 1994) mais aussi à évaluer la modulation de réponse immunitaire T en contexte pathologique (infectieux, tumoral…). Comme il en sera

discuté dans le prochain chapitre, les technologies de séquençage d'ADN ont connu de grands progrès depuis le début du XXIème siècle (Voelkerding et al., 2009) créant un nouvel engouement pour l'étude du répertoire T comme illustré **Figure 8**.

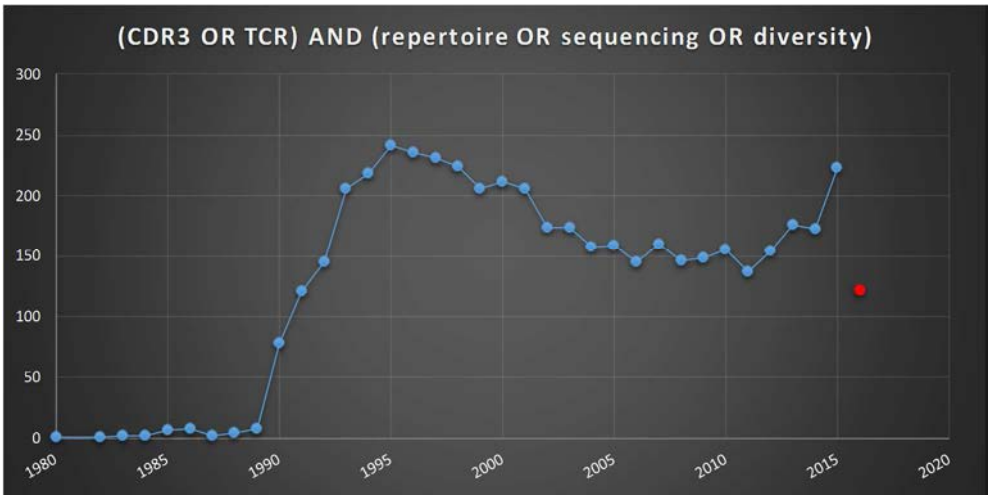


Figure 8 : Evolution du nombre de publications traitant du répertoire TCR de 1980 à nos jours (Source: Pubmed)

En 2009, alors que Weinstein *et al.* produisaient une description à grande échelle du répertoire B des poissons zèbre, les équipes de Robins et de Holt cherchaient à évaluer la diversité du répertoire TCRβ chez l'homme (Freeman et al., 2009; Robins et al., 2009; Warren et al., 2009; Weinstein et al., 2009). La résolution de cette approche est telle qu'elle a permis à de nombreuses équipes de décrire de manière inédite des processus fondamentaux tels que la polarisation des lymphocytes T vers les différents compartiments cellulaires (Cebula et al., 2013; Föhse et al., 2011; Qi et al., 2014; Sherwood et al., 2011), les processus de diversification du TCR (Murugan et al., 2012; Srivastava and Robins, 2012) ou la similarité clonale interindividuelle (Prabakaran et al., 2012; Robins et al., 2010). Par ailleurs, de nombreuses études ont eu recours au séquençage à haut débit du répertoire TCR, dit **RepSeq** ou immunoséquençage, à des fins cliniques avec pour ambition d'identifier des biomarqueurs utilisables comme outils diagnostic (revue par Woodsworth et al., 2013).

En parallèle, du fait du niveau de détails, sans précédent, sur la composition du répertoire TCR, le RepSeq a engendré un besoin d'outils adaptés permettant d'extraire de manière exhaustive et standardisée l'information, et de caractériser la diversité de manière objective et facile à interpréter pour les immunologistes. Ainsi, de nouveaux champs de recherche se sont développés s'intéressant la modélisation informatique et statistique (Covacu et al., 2016; Murugan et al., 2012) ou à la caractérisation de la diversité du répertoire (Bolotin et al., 2013;

Brochet et al., 2008; Giraud et al., 2014; Plessy et al., 2015; Thomas et al., 2013b). Toutes ces approches ont leurs avantages et leurs inconvénients. Toutefois, comme décrit dans notre revue (Six et al., 2013) jointe en Annexe 5, chacune d'elles apporte un éclairage particulier sur ce même objet biologique ; à titre d'exemple, l'analyse par cytométrie en flux offre une description quantitative de la diversité du répertoire en l'abordant du point de vue de l'expression des grandes familles TRV, au niveau protéique à la surface du lymphocyte T, avec la limite de la disponibilité des anticorps monoclonaux. Le spectratypage du CDR3 permet de caractériser la diversité du répertoire d'un point de vue qualitatif en prenant en compte toutes les combinaisons TRVJ et la longueur de la jonction, en niveau des réarrangements génomiques ADN ou des transcrits correspondants. Ces deux exemples sont caractéristiques des niveaux de granularités étudiés en fonction des approches ; la pertinence de leur application va donc dépendre de la question biologique d'intérêt et les combiner permet d'obtenir des informations complémentaires, comme par exemple dans les études de Bergot et al. ou Wu et al. (Bergot et al., 2015; Wu et al., 2015).

D. Qu'est-ce que le séquençage à haut débit du TCR ?

1) Technologie

La méthode de Sanger, utilisée dans le cadre du *Human Genome Project*, a été considérée pendant plus de 30 ans comme un standard pour le séquençage. En 2005, fut lancée la première plate-forme de séquençage dit à haut-débit. Cette expression désigne un ensemble de méthodes parallèles permettant de séquencer simultanément des millions de fragments d'ADN. Ces méthodes, en s'affranchissant des étapes de clonage et de constitution de banques génomiques, permettent de réduire considérablement les temps et coûts de séquençage.

Les deux principales plates-formes de séquençage à haut débit (appelé NGS pour *next generation sequencing* ou HTS pour *high-throughput sequencing*) utilisées dans l'analyse de répertoire étaient **Roche/454** et **Illumina/MiSeq ou HiSeq** (voir Metzker 2011 pour une revue complète).

La technologie de Roche/454 Life Sciences combine la PCR en émulsion (Tawfik and Griffiths, 1998) et le pyroséquençage (Nyren et al., 1993; Ronaghi et al., 1996, 1998) comme décrit **Figure 9A**. Cette technologie permet de produire des séquences relativement longues (400-

700 paires de bases en moyenne) avec des performances variant en fonction des générations de séquenceurs entre 10^5 et 3.10^6 séquences par expérience de séquençage.

En 2006, Illumina fit l'acquisition d'une technologie de séquençage initialement conceptualisée par les fondateurs de Solexa. Cette technique combine l'amplification en phase solide et le séquençage par synthèse (mesure de fluorescence) (**Figure 9B**). Les séquences produites étant courtes (35-150 pb), le nombre de séquences par séquençage est bien plus élevé que celui des séquenceurs 454 : de 15.10^6 à plus de 3.10^9 en fonction des séquenceurs.

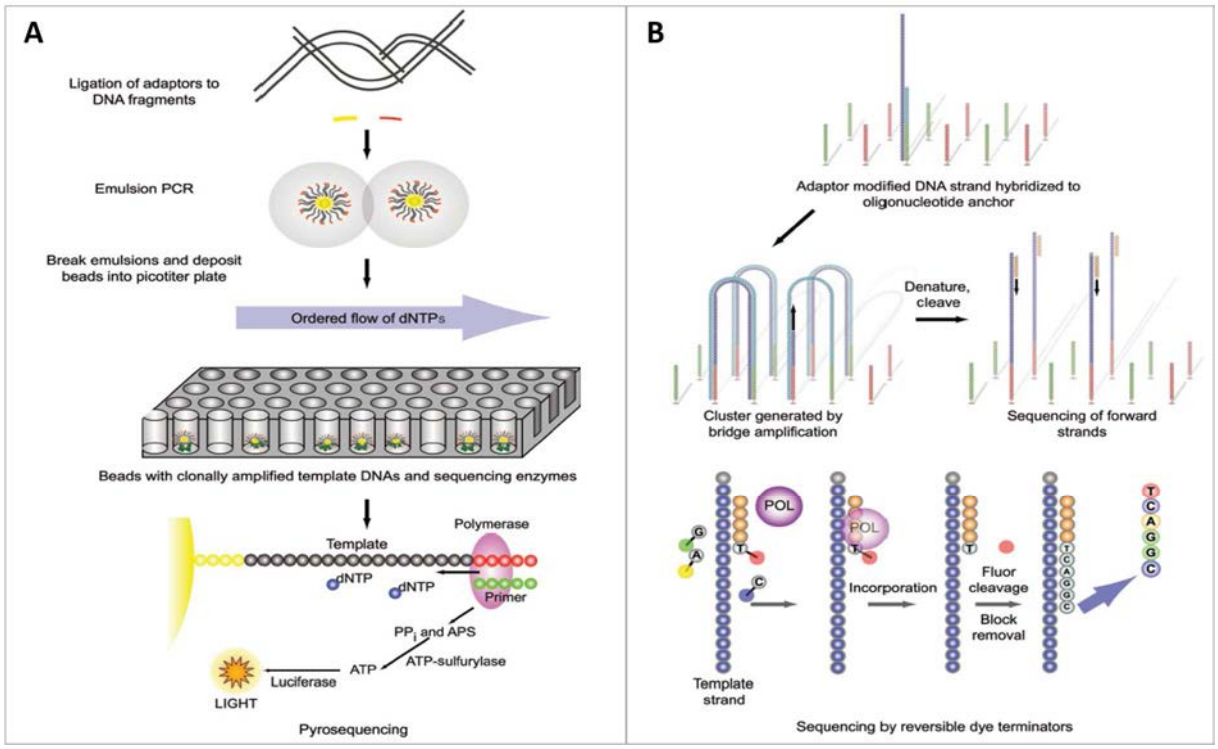


Figure 9 : Technologies de séquençage – **A)** Une « librairie » est préparée par ligature de fragments ADN (300 à 800 paires de bases) avec des adaptateurs oligonucléotidiques ou par amplification avec des amorces incluant les adaptateurs. Les fragments ADN de la librairie sont ensuite dénaturés et individuellement liés via leur adaptateur à des billes. Chaque bille est ensuite isolée dans une microvésicule où a lieu la PCR en émulsion de la molécule d'ADN qu'elle porte. Après amplification, les billes ainsi recouvertes de millions de copies de leur fragment ADN d'origine sont ensuite distribuées sur une plaque recouverte de puits ne pouvant capturer qu'une seule bille. C'est au sein de ces puits que vont avoir lieu itérativement les réactions de pyroséquençage. **B)** Chaque fragment d'ADN de la librairie est dénaturé et fixé à un adaptateur puis immobilisé sur un support tapissé d'amorces complémentaires aux adaptateurs. Un brin complémentaire est synthétisé pour chaque fragment. Le nouvel ADN double-brin est dénaturé et chaque brin fixé forme alors un pont en s'hybridant localement avec l'amorce complémentaire de l'autre extrémité. Le brin complémentaire est synthétisé. L'opération est répétée un grand nombre de fois : il y a formation d'amas (cluster) du même fragment d'ADN. Ces clusters sont ensuite dénaturés et clivés. Le séquençage est initié avec l'addition d'amorces, d'une polymérase (POL) et d'un des 4 nucléotides « terminateurs réversibles » fluorescents (une couleur par base). Le signal de fluorescence est enregistré lors de l'incorporation

des nucléotides puis le « terminateur » est retiré avant le prochain cycle de synthèse. (Voelkerding et al., 2009)

En termes de précision, il apparaît que bien que son taux d'erreur global soit de l'ordre 1%, la technique de pyroséquençage de 454 montre des difficultés à gérer les longues (> 6) répétitions du même nucléotide (homopolymère), ce qui peut entraîner des insertions et des délétions. A l'inverse, même si le taux de délétions/insertions de la technologie Illumina est faible, son taux de substitution est plus élevé que celui de 454, de l'ordre de 3% (Bolotin et al., 2012; Mardis, 2011).

En 2012, Luo *et al.* ont comparé ces plates-formes dans le cadre de l'étude métagénomique d'une communauté planctonique d'eau douce. Ainsi, un échantillon d'ADN obtenu à partir d'un prélèvement fut divisé en deux aliquotes, l'un séquencé par le Roche 454 FLX Titanium et l'autre par Illumina Genome Analyzer II. Malgré les différences de protocoles, de longueur de séquences et surtout de nombre de séquences générées, les deux technologies semblent restituer une représentation similaire de la diversité de la communauté, avec un chevauchement d'environ 90% en termes d'espèces et une forte corrélation de leur abondances (R^2>0.9) (Luo et al., 2012). Toutefois, une étude similaire, comparant le séquençage d'échantillons d'ARN par des séquenceurs Illumina et Ion Torrent – une troisième technologie NGS développée par Life technologies 2010, a démontré que le choix de la plate-forme de séquençage pouvait potentiellement biaiser l'évaluation de la diversité d'une communauté (Salipante et al., 2014).

2) Séquençage appliqué au répertoire TCR

La principale différence de protocole entre un séquençage « classique » et une expérience d'immunoséquençage réside dans la préparation de la librairie de fragments d'ADN qui serviront de matrice pour le séquençage. Dans le cadre d'un séquençage TCR, cette librairie est enrichie en fragments d'ADN génomique ou complémentaire codant le domaine variable de la chaîne d'intérêt, grâce à l'utilisation de combinaisons d'amorces ciblant les gènes TRA ou TRB lors des cycles d'amplifications (**Figure 10**). En effet, le séquençage des deux chaînes du TCR se fait généralement séparément. Bien que de plus en plus performants, les protocoles permettant de capturer simultanément les séquences des deux chaînes exprimées par les cellules ne sont pas encore devenus routiniers du fait de leur complexité d'exécution et

d'analyse (DeKosky et al., 2013; Howie et al., 2015). Parmi les protocoles d'amplification mis au point, la technique *Rapid Amplification of cDNA-ends* ou RACE implique la fixation par PCR d'une ancre dont la séquence connue est incorporée à l'une des extrémités de la région d'intérêt, ici en 5' au niveau du gène TRC. Puis, la région est amplifiée en combinant une amorce complémentaire à l'ancre et une amorce ciblant le gène d'intérêt. Cette approche, bien que particulièrement efficace et appréciée, reste peu utilisée par les biologistes car elle est difficile à implémenter du fait de la sensibilité de certaines étapes du protocole et du prix onéreux du kit proposé par ThermoFisher©. Toutefois, de plus en plus d'équipes arrivent à mettre en place avec succès cette procédure (Bolotin et al., 2012; Freeman et al., 2009; Quigley et al., 2010; Warren et al., 2009).

Une autre approche, appelée *TCR gene-capture*, utilise une librairie ADN composée de séquences connues, telles que les gènes TRA et/ou TRB référencées par la base IMGT (*International Immunogenetics Information System*) (Lefranc et al., 2005). Comme décrit par Linneman *et al.* (2013), ces séquences vont ensuite « capturer » par complémentarité les séquences ARN d'intérêt présentes dans l'échantillon pour les amplifier. Cette technique a l'avantage de permettre de cibler les deux chaînes simultanément mais ne semble pas optimale pour assurer l'amplification de toutes les molécules TR en présence.

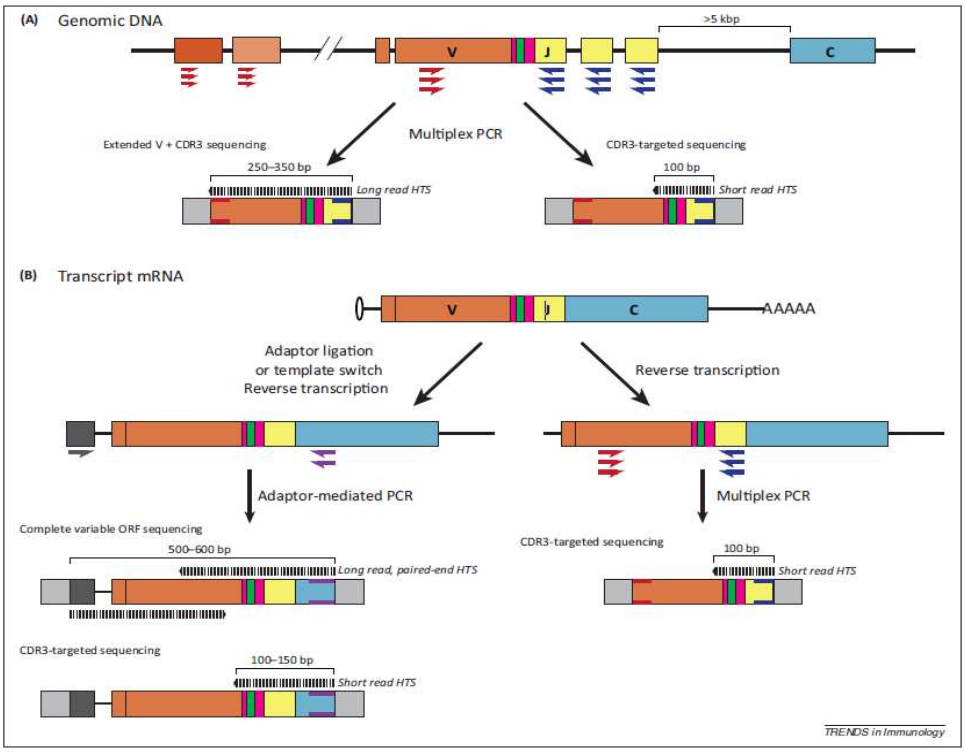


Figure 10 : Exemples de stratégies de préparation des librairies d'ADN – A) Amplification par PCR et préparation de librairie à partir d'ADN génomique. B) Amplification par RT-PCR et préparation de

librairie à partir d'ARN messager. Dans les deux cas, l'amplification se fait via une stratégie de PCR multiplexe utilisant un mélange d'amorces complémentaires aux différents gènes V (en rouge) qui peut éventuellement être combiné à un mélange d'amorces ciblant les gènes J (en bleu). L'incorporation d'adaptateurs (en gris) permet de capturer les séquences. Dans le cas de l'ARNm, une amorce unique ciblant la région C (violet) peut être utilisée.

Du fait de la dominance d'Illumina sur le marché du séquençage, le choix de la plate-forme n'est plus d'actualité. Toutefois, ce non-choix reste discutable en ce qui concerne l'immunoséquençage. En effet, les séquences produites par 454 étant plus longues, elles couvrent la totalité de la région CDR3 ainsi que les segments V et J, ce qui n'était pas le cas des séquences Illumina qui doivent être appariées pour couvrir la région d'intérêt (on parle de séquençage « *pair-ended* »), entraînant un risque d'erreur (Hou et al., 2016). De plus, alors que dans le cadre d'un séquençage génomique ou transcriptomique « classique », il est possible de corriger une erreur de séquençage sur la base de séquences « références ». Lorsque l'on s'intéresse à une séquence variable telle que le CDR3 pour laquelle, par définition, on ne dispose pas de séquences de référence, il n'est pas possible d'identifier les séquences erronées. Ainsi, le taux d'erreur global plus faible de la technologie 454 aurait semblé plus pertinent pour minimiser le taux de « faux-positif ». Malgré cela, la grande différence de performance entre les deux technologies rend la technologie 454 beaucoup moins intéressante d'autant plus qu'un grand effort a été fourni par de nombreuses équipes pour développer des algorithmiques permettant la correction *a posteriori* des erreurs de séquençage (Salmela and Schröder, 2011; Shugay et al., 2014; Thomas et al., 2013b).

Récemment, Brown et son équipe ont analysés le répertoire TCR à partir de données de séquençages transcriptomiques complets (Brown et al., 2015). Ainsi, en analysant les données de RNAseq de 6738 échantillons tumoraux, disponibles dans la base de données TCGA (*The Cancer Genome Atlas*), ils ont évalué qu'avec un processus adéquat de traitement et d'extraction des données, des séquences TR pouvaient être identifiées avec un rendement de 1 pour 10 millions de séquences. Leur méthodologie leur a ainsi permis d'identifier des séquences CDR3 α et β qu'ils qualifient de « spécifiques de tumeur ». D'après les auteurs, cette approche mettrait donc « en échec » le séquençage spécifique du CDR3 pour permettre une analyse systémique intégrant non seulement la « diversité » du répertoire TCRαβ et le contexte transcriptomique global. Or, bien que très attractive, cette approche reste néanmoins problématique. En effet, le rendement de détection des séquences TR exige une profondeur de séquençage bien supérieure à la norme actuelle. De plus, et c'est là le problème

majeur, si l'identification de quelques clones est suffisante pour caractériser le profil TCR de tumeurs dont le répertoire perturbé contient des (quelques) expansions clonales majeures (Jang et al., 2015), il ne semble pas envisageable d'appliquer la même approche dans un contexte plus polyclonal.

Comme revu par Calis et Rosenberg (Calis and Rosenberg, 2014), les différentes stratégies de séquençage du CDR3 présentent toutes des avantages et des inconvénients (Bolotin et al., 2012) qu'il est nécessaire d'avoir à l'esprit lors de la définition des stratégies expérimentale et analytique d'une expérience.

3) Gestion et analyse standardisées des données

L'exécution d'un projet de séquençage se décompose en 4 partie : 1) le design du plan expérimental et la production des échantillons biologiques, 2) le séquençage des échantillons, 3) la gestion et le traitement des données produites, 4) l'analyse des données (**Figure 11**). Alors que la dernière décennie a permis le développement et l'amélioration des technologies de séquençage, ce sont les étapes de gestion et d'analyse de données qui mobilisent le plus les ressources désormais (Sboner et al., 2011).

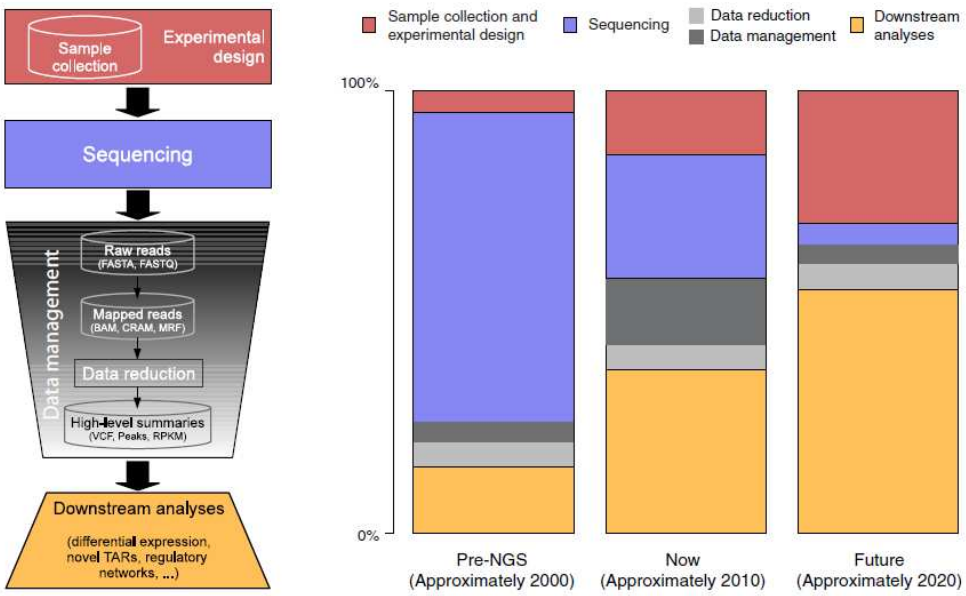


Figure 11 : Décomposition d'un projet de séquençage – A gauche, les quatre étapes nécessaires au bon déroulement d'un projet de séquençage. À droite, l'évolution de l'impact financier de ces étapes à travers le temps. (Sboner et al., 2011)

La quantité massive de données produites par l'immunoséquençage induit un besoin de standardisation du traitement, d'intégration et d'analyse des données de manière similaire aux développements mis en place pour le transcriptome.

Les données brutes de séquençage sont traditionnellement stockées dans un format plus standard : FASTQ (Cock et al., 2010). Ce format texte permet de stocker à la fois les séquences nucléotidiques et les scores de qualité Phred (Ewing and Green, 1998; Ewing et al., 1998) associés à chaque base, encodés en ASCII. Le score de qualité Phred Q est lié de façon logarithmique à la probabilité d'erreur P de séquençage ($Q= -10\ log_{10}\ P$) ; un score de 30 correspondant à une précision d'identification de la base à 99.9% est généralement le seuil minimal utilisé. Les fichiers de sortie des séquenceurs Illumina sont directement produits au format FASTQ alors que ceux de Roche/454 Life Sciences, au format SFF (*Standard Flowgram Format*), nécessitent une étape de conversion. Chaque plate-forme dispose d'un premier processus de contrôle qualité et de nombreux outils sont désormais disponibles pour permettre aux expérimentateurs d'évaluer la qualité des données produites avec, entre autres critères, le score de Phred mais également sur la distribution d'usage des nucléotides ou le taux de GC, dit coefficient de Chargaff qui peut influer sur le taux d'erreur de séquençage (Dohm et al., 2008).

Une expérience de séquençage peut être exécutée selon deux modalités : 1) un seul échantillon biologique est séquencé à la fois ; le contenu du fichier fastq est homogène et peut être pris en charge tel quel. 2) Plusieurs librairies d'ADN (identifiées par des codes-barres spécifiques) sont mélangées pour le séquençage. Le fichier fastq produit doit être divisé en autant de fichiers que d'échantillons mélangés.

Chaque fichier fastq peut ensuite être analysé pour l'identification des séquences de TCR. La plupart des outils d'analyse prennent également en compte le score Phred pour l'extraction des données. Chaque séquence est alignée à une série de séquences de référence, la plupart du temps extraites de la base IMGT (Lefranc et al., 2005), et est annotée par un ensemble d'attributs décrivant la composition de la séquence en termes de gènes utilisés (V, D, J) et la séquence du CDR3. Le niveau de description des séquences TR identifiées ainsi que le format de stockage de ces informations varient en fonction de l'outil d'analyse utilisé.

E. Les défis de l'immunoséquençage à haut-débit

De nombreuses questions fondamentales en immunologie sont intimement liées à la diversité du répertoire des lymphocytes T. Certaines sont descriptives : quelle est la forme du répertoire T ? (Correia-Neves et al., 2001), quelle est la diversité des cellules T mémoires (Kedzierska et

al., 2006) ou des Tregs (Ferreira et al., 2014; Hsieh et al., 2006; Pacholczyk et al., 2006; Wong et al., 2007) ? D'autres questions sont plus mécanistiques : quelle diversité T est nécessaire pour répondre de manière appropriée à une infection virale ? (Naumov et al., 2003; Pewe et al., 2004)

Ces questions ont pu trouver des éléments de réponse grâce au séquençage à haut débit du répertoire T qui reste l'objet de nombreux défis (Benichou et al., 2012; Robins, 2013). En effet, de nombreuses étapes techniques et analytiques séparent l'échantillon biologique d'intérêt des données RepSeq qui vont permettre d'étudier son répertoire TCR. Comme pour la plupart des procédures expérimentales, chacune de ces étapes peut influencer l'interprétation biologique des résultats obtenus en fin d'expérience. De nombreux efforts restent à faire quant à la définition de bonnes pratiques de production de données RepSeq fiables et facilement exploitables.

1) Optimisation du protocole expérimental

L'analyse de répertoire par immunoséquençage cherche à déterminer la quantité relative des clones constituant la population cellulaire étudiée. Un premier challenge est de minimiser les biais de distributions pouvant être introduits par les multiples étapes d'amplification précédant le séquençage. De nombreuses stratégies, utilisant, entre autres, des techniques de PCR multiplexe[4], emboîtée[5] ou RACE, ont été développées pour pallier ce problème, tant pour les analyses génomiques (Boyd et al., 2009; Carlson et al., 2013; Robins et al., 2009; Sherwood et al., 2011) que transcriptomiques (Mamedov et al., 2011; Wang et al., 2010; Warren et al., 2011; Zhu et al., 2001).

Comme précisé précédemment, une seconde problématique expérimentale est la mise au point d'un protocole performant permettant de séquencer simultanément les deux chaînes du TCR (Cukalac et al., 2015; DeKosky et al., 2013; Howie et al., 2015; Turchaninova et al., 2013). Cette approche prend notamment tout son sens dans l'optique d'une application clinique, permettant ainsi d'identifier les paires de chaînes des TCR impliquées dans une réponse antigénique d'intérêt et de développer une thérapie ciblant les cellules exprimant ces TCR par exemple (Benichou et al., 2012; Brusko et al., 2010).

[4] Technique permettant l'amplification simultanée de plusieurs amplicons par l'utilisation de plusieurs amorces
[5] PCR se déroulant en plusieurs étapes successives utilisant des combinaisons d'amorces différentes

2) Représentativité des échantillons

L'immunoséquençage est un « jeu de nombres » (Benichou et al., 2012) ce qui le rend particulièrement sensible à l'échantillonnage (**Figure 12**). En effet, comme établi précédemment, l'analyse de répertoire cherche à décrire la diversité d'une population lymphocytaire. Or, à l'exception de certaines populations cellulaires très petites, il est très rare que la totalité des cellules d'une population d'intérêt soient prélevées lors de la collecte des échantillons. Aussi, il est essentiel que la quantité de cellules prélevées soit assez grande pour assurer de la représentativité statistique de l'échantillon par rapport à la population d'origine ; on parle d'**échantillonnage biologique**. Ce facteur est particulièrement problématique pour les études faites chez l'homme, chez qui la majorité des études se font sur des échantillons de sang périphérique (environ 2,5% de LT parmi les cellules circulantes) alors que chez la souris, les organes lymphoïdes, riches en LT, peuvent être facilement prélevés (Greiff et al., 2015a). Ainsi, Warren et son équipe ont pu observer que l'analyse du répertoire TCRβ de deux échantillons de sang du même individu prélevés à 1 semaine d'intervalle ne présentait que peu de similitudes (environ 10% de séquences partagées par les deux échantillons), concluant à un fort sous-échantillonnage biologique (Warren et al., 2011).

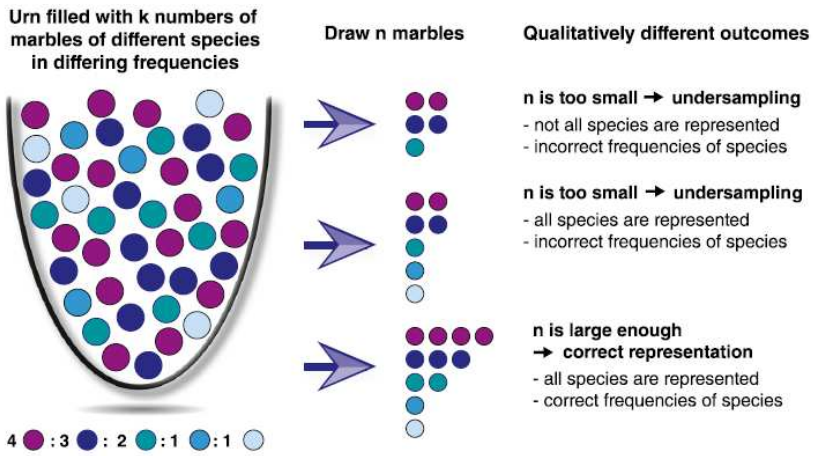


Figure 12 : Impact de l'échantillonnage sur la représentativité de l'observation – L'effet de l'échantillonnage sur l'observation de la diversité peut s'illustrer par analogie avec une urne pleine de billes de couleur. Si la taille d'un échantillon prélevé au hasard est trop petite, on observe deux cas : 1) toutes les espèces existantes dans l'urne (définies par la couleur des billes) ne sont pas observées et les fréquences de celles observées ne reflètent pas leur fréquence d'origine 2) toutes les espèces sont observées mais les fréquences d'observation sont erronées. Un échantillon est considéré représentatif, lorsque sa taille permet d'observer toutes les espèces existantes tout en respectant leur fréquence relative dans l'urne d'origine. (d'après Greiff et al., 2014b)

De plus, Sepúlveda a démontré dans ses travaux que le nombre de clonotypes observés dans un échantillon est loin de représenter la richesse totale de la population, la probabilité

d'observer de nouveaux clonotypes en augmentant la taille de l'échantillon étant comprise entre 45% et 55% en fonction de la population LT observée (de Sepúlveda, 2009).

L'observation d'échantillons aléatoires implique donc que de nombreuses espèces restent invisibles ou non détectées (Fisher et al., 1943). Une évaluation précise du nombre d'espèces uniques (richesse spécifique) dans une population importante et complexe nécessite d'estimer ces espèces invisibles sur la base de la richesse observée (Robins et al., 2009). Différentes méthodes ont été développées pour estimer la diversité globale d'une population en utilisant les données d'échantillonnage (Chao, 1987; Colwell et al., 2012; Efron and Thisted, 1976; Palmer, 1990). Une technique est d'utiliser des estimateurs non paramétriques pour extrapoler la richesse spécifique globale du répertoire d'une population cellulaire à partir des données produites sur un échantillon de cette population (Oksanen, 2011). Cette approche permet de s'assurer que les observations faites à partir des données d'échantillonnage sont représentatives du comportement global, et ne sont pas biaisées par l'effet d'échantillonnage.

Un second facteur pouvant affecter l'interprétation des résultats est la profondeur de séquençage (Bashford-Rogers et al., 2014). En effet, un nombre insuffisant de séquences produites ne permettra pas d'évaluer correctement la diversité moléculaire de l'échantillon ; on parle d'**échantillonnage technique**. De nombreuses études ont montré que le nombre de séquences observées est corrélé positivement avec la taille de l'échantillonnage (Madi et al., 2014; Robins et al., 2010; Shugay et al., 2013). Ce biais se ressent notamment dans la recherche de séquences partagées par les individus puisqu'une séquence observée chez un seul individu à un certain niveau d'observation pourrait être détectée chez les autres si la profondeur de séquençage était augmentée (Venturi et al., 2013). Ainsi, l'immunoséquençage se confronte à une problématique connue en écologie, à savoir que seule l'observation d'une espèce peut être concluante car l'absence d'une espèce peut être soit réelle soit l'effet d'un sous-échantillonnage (Magurran, 1998).

Dans sa revue, Greiff énonce donc deux règles à considérer lorsque l'on réalise une expérience d'immunoséquençage (Greiff et al., 2015a) :

- Le nombre de séquences produites doit être au moins équivalent à la richesse clonale de la population.
- Plus un clone est rare, plus grande doit être la profondeur de séquençage pour pouvoir le détecter.

Toutefois, la profondeur de séquençage optimale va surtout dépendre des échantillons utilisés et des questions biologiques posées (Bashford-Rogers et al., 2014). Une population cellulaire très peu diverse, du fait d'un nombre de cellules réduit ou de la présence de clones prédominants, ne nécessitera pas la même profondeur de séquençage qu'une population hautement polyclonale pour observer tous les clones en présence. De plus, une perturbation majeure du répertoire, suite à une infection par exemple, ne nécessitera pas le même degré d'analyse que la recherche de TCR spécifiques d'antigènes. Cependant, un séquençage trop profond peut aussi être nuisible car un nombre trop important de séquences implique une plus grande proportion de « fausses » séquences TR. Nguyen *et al.* estiment à 1-6% le pourcentage de séquences erronées du fait des erreurs de séquençage (Nguyen et al., 2011). Tout séquençage implique donc de trouver un compromis entre une détection des clonotypes satisfaisante et un « bruit de fond » limité, la clonalité des populations cellulaires variant en fonction de leur nature (Estorninho et al., 2013; Tipton et al., 2015). Ainsi, la recherche de cet équilibre reste un défi clé dans l'analyse de répertoire TCR par séquençage à haut-débit, qui peut être notamment testé par modélisation des distributions des fréquences des clones (Greiff et al., 2014).

3) Identification des séquences TR

Il existe un grand nombre d'outils d'annotation dédiés à l'identification de séquences TR : clonotypeR (Plessy et al., 2015), Decombinator (Thomas et al., 2013b), IgBLAST (Ye et al., 2013), IMGT/V-QUEST (Aouinti et al., 2015; Brochet et al., 2008), IMEX (Schaller et al., 2015), MiTCR (Bolotin et al., 2013), pRESTO (Vander Heiden et al., 2014), VDJSeq-Solver (Paciello et al., 2015) et Vidjil (Giraud et al., 2014). Ces outils diffèrent par leurs performances mais surtout par leur démarche d'annotation et de correction des erreurs de séquençage (Bolotin et al., 2015; Greiff et al., 2014).

Pour qu'une séquence soit identifiée en tant que séquence TR, ou clonotype, elle doit remplir un certain nombre de critères relatifs à la longueur et au degré de similitude de son alignement avec les séquences références, à la qualité Phred des portions de séquences alignées, à la détection et à l'identification non ambiguë des gènes V(D)J recombinés… Ces critères vont changer en fonction des outils d'annotations, rendant leur sélection plus ou moins stricte ce qui peut interférer avec l'analyse de la diversité observée.

Par ailleurs, la quantité d'information fournie pour décrire un clonotype varie de l'identification de la séquence CDR3 seule à la description de l'ensemble des régions FR (*framework*) et CDR constituant le domaine variable de la chaîne étudiée (Bolotin et al., 2015). Cette hétérogénéité rend donc difficile la comparaison de ces données. Une standardisation, telle que la mise en place de nomenclatures d'annotation par exemple (Yassai et al., 2009), faciliterait donc grandement l'exploitation des données d'immunoséquençage. Toutefois cette démarche est loin d'être triviale.

En 2009, M. Yassai proposait la définition suivante :

> *A TCR clonotype is a unique nucleotide sequence that arises during the gene rearrangement process for that receptor. The combination of nucleotide sequences for the surface expressed receptor pair would define the T cell clonotype.*

Malgré sa simplicité, cette définition peut tout de même être sujette à débat. Beaucoup préfèrent étudier la diversité de séquence du CDR3 au niveau protéique considérant que : 1) plusieurs séquences nucléotidiques pouvant produire la même séquence d'acides aminés, la traduction permet de minimiser une variabilité non pertinente, 2) ce sont les différences de séquences protéiques des CDR3 qui vont conditionner leurs interactions avec les Ag.

Une autre variante de définition concerne la délimitation du CDR3. En effet, celui-ci est délimité par un résidu Cystéine C-terminal conservé dans tous les gènes V et un résidu Phénylalanine (ou Tryptophane) inclus dans un motif F/W-G-x-G-T conservé dans tous les gènes J. En fonction des outils, ces résidus seront ou non inclus dans la séquences des CDR3 identifiés ; certains outils, tels que MiTCR, proposent à l'utilisateur de choisir ce qu'il préfère. Il est à noter que les régions hypervariables du TCR ont été identifiées par analogie à celles des immunoglobulines décrites par Kabat (1988). D'après la nomenclature Kabat, la séquence du CDR3 commence 3 résidus après la Cystéine et se termine 2 résidus avant le motif J. Or, suite à l'initiative de standardisation de la nomenclature d'appellation des gènes TR en 2002, cette définition a également changé faisant commencer et s'arrêter la séquence du CDR3 immédiatement après et avant les résidus conservés.

Comme précisé plus tôt, la description des clonotypes diffère en fonction des outils utilisés pour identifier les séquences TR. Toutefois, quel que soit l'outil d'annotation, trois paramètres sont toujours décrits pour chaque clonotype : le gène V, le gène J et la séquence protéique du CDR3. La combinaison de ces paramètres est donc pour moi ce qui définit un clonotype TR.

Les processus d'amplification PCR et de séquençage sont tous deux sources d'erreurs potentielles. L'impact des erreurs de PCR est considéré comme négligeable car le faible taux d'erreur associé aux polymérases rend peu probable de multiples erreurs sur la même séquence pendant le même cycle. Or, les erreurs de PCR se propagent par cycle. Ainsi, une erreur de PCR n'affectera en moyenne qu'un seul nucléotide (Nguyen et al., 2011; Robins, 2013). Le taux et surtout le type d'erreur de séquençage est spécifique de la technologique.

Une première façon de minimiser l'impact de ces erreurs est d'intégrer à chaque molécule ADN ou ARN de la librairie un identifiant unique permettant de la reconnaître. Ainsi, grâce à des outils informatiques adaptés, toutes les séquences portant le même identifiant sont associées au même clonotype, peu importe les éventuelles différences de séquences dues aux erreurs de séquençage (Bolotin et al., 2013; Shugay et al., 2014; Vander Heiden et al., 2014; Vollmers et al., 2013). Cette méthode est particulièrement efficace : la méthode UMI (*Unique Molecular Identifier*), développée par Shugay et ses collègues, par exemple semble réduire le taux d'erreur jusqu'à 100 fois.

La correction d'erreurs de séquençage peut également se faire *a posteriori* en filtrant les clonotypes détectés à de très faibles fréquences (Becattini et al., 2015; Greiff et al., 2014) considérant la redondance de détection comme un critère de confiance. Toutefois, cette approche stricte implique une perte d'information conséquente notamment pour l'observation des espèces rares ou l'étude de répertoires très polyclonaux. Aussi, une alternative efficace est l'assimilation de clonotypes uniques à des clonotypes plus abondants. En phylogénétique, la parcimonie cherche à identifier l'arbre qui minimise le nombre de « pas » (insertion/délétion/substitution) pour passer d'une séquence à l'autre. De manière similaire, certains outils intègrent un algorithme qui identifie des paires de clonotypes dont la séquence ne diffère que d'un nucléotide. Lorsque dans une paire, la fréquence de l'un des clonotypes est très supérieure à celle de l'autre, le clonotype « minoritaire » est assimilé au clonotype « majoritaire » (Bolotin et al., 2013; Robins et al., 2009). Cette approche permet donc de minimiser le nombre de « faux » clonotypes dus à des erreurs de PCR ou de séquençage.

Enfin, un autre problème que soulève l'identification des clonotypes est le degré de pertinence des annotations qui est également très variable en fonction des outils. En effet, à l'exception de certains outils, l'identification des gènes TR recombinés de chaque clonotype se fait par alignement des séquences produites avec des séquences génomiques TR de

référence. Outre la fiabilité de ces références, qui sont souvent incomplètes et hétérogènes en termes de sources, les alignements peuvent être plus ou moins exigeants en fonction des algorithmes utilisés. Ces facteurs rendent donc l'annotation des séquences produites plus ou moins fiables, notamment celle des séquences courtes. Toutefois, ces algorithmes sont en constantes évolution et leur amélioration peut notamment se faire par l'utilisation de jeux de données synthétiques dont la diversité est maîtrisée et donc dont la précision d'annotation peut être évaluée (Bolotin et al., 2012; Safonova et al., 2015).

F. Problématique

La prise en charge et l'analyse de données d'immunoséquençage soulèvent donc de nombreuses problématiques méthodologiques et conceptuelles, de manière similaire aux développements induits par l'émergence du transcriptome.

Ma thèse a eu pour objet de proposer une approche permettant d'exploiter de manière optimale les données RepSeq produites par le laboratoire.

Pour ce faire, j'ai mis au point une méthodologie originale répondant aux besoins et limites des outils actuels et permettant une caractérisation approfondie et facilement interprétable de la diversité des répertoires TCR analysés. Cette approche permet la décomposition de ces répertoires et l'identification de paramètres spécifiques à différents états physiopathologiques étudiés. Du fait de la complexité et de la richesse d'information des données RepSeq, leur analyse est particulièrement gourmande en temps et en ressources. L'objectif de ce travail a donc été de fournir aux biologistes un outil standardisé leur permettant une représentation robuste du répertoire TCR de leurs échantillons, et ainsi de tester, voire de construire, en toute confiance des hypothèses, sans gaspiller leurs ressources dans des analyses infructueuses.

Un second axe de travail a été d'évaluer la représentativité de données produites. En effet, malgré l'engouement croissant pour l'immunoséquençage, de nombreuses questions quant à la fiabilité de cette approche restent sans réponse, notamment quant à l'impact de la profondeur de séquençage sur la mesure de la diversité d'un répertoire. Ainsi, afin de nous assurer de la vraisemblance de nos résultats, il semble crucial de tester certaines limites de cette approche.

METHODOLOGIE

Ces travaux ont été réalisés sous R. Les spécifications fonctionnelles du code développé pour réaliser les analyses présentées sont disponibles en Annexe 1.

A. Production des données

1) Données expérimentales

Cette thèse s'est déroulée dans le cadre d'un projet ERC Advanced « *TRiPoD - Deciphering the regulatory T cell repertoire : towards biomarkers and biotherapies for autoimmune diseases* ». Ce projet a, entre autres, pour objectif d'explorer le(s) répertoire(s) TCR de la population LT régulatrices (Tregs) en situation physiologique et de caractériser les changements impliqués dans l'apparition d'un diabète de type 1 chez la souris.

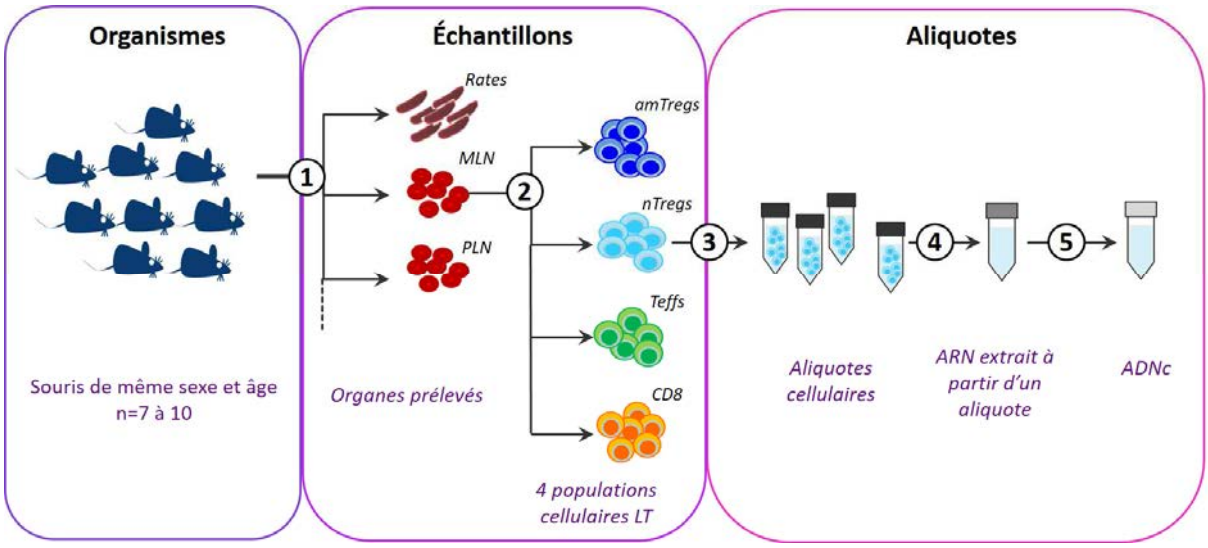


Figure 13: Schéma expérimental des expériences TriPoD – Chaque expérience réalisée dans le cadre du projet TriPoD comprend cinq étapes : 1) 7 à 10 (=n) souris de fonds génétique NOD-foxp3GFP ou C57BL/6-foxp3GFP, de sexe et d'âge identiques, sont sacrifiées. Sept organes ou tissus sont prélevés : rate (Rates), les ganglions mésentériques (MLN), ganglions pancréatiques (PLN), ganglions brachiaux, inguinaux, rénaux et para-aortiques. 2) Les cellules isolées des n organes d'un même type sont mélangées puis triées en quatre populations lymphocytaires T par cytométrie en flux : amTregs, nTregs, Teffs, CD8 (voir **Figure 14**). 3) Les cellules triées sont stockées sous forme d'aliquotes cellulaires avec un maximum de cinq millions de cellules par aliquotes. 4) Lors d'une expérience d'immunoséquençage, l'ARN d'un aliquote est extrait et utilisé pour 5) synthétiser l'ADN complémentaire (ADNc) qui servira à la préparation de la librairie d'ADN à séquencer.

Comme schématisé sur la **Figure 13**, lors de chaque expérience effectuée dans le cadre de ce projet, plusieurs souris (n=7 à 10 en fonction des expériences) de mêmes fonds génétique

(C57BL/6 ou NOD[6], les deux lignées de souris étant foxp3-GFP[7]), d'âge et de sexe sont sacrifiées pour prélever leur **rate**, ganglions brachiaux et inguinaux (dits **ganglions périphériques**), ganglions pancréatiques, para-aortiques et rénaux (dits **ganglions profonds**) et **ganglions mésentériques**. Alors que les ganglions prélevés sont gardés dans leur intégralité, seul un tiers de la rate est utilisé pour la suite. Afin d'assurer une quantité suffisante de matériel pour le séquençage (notamment pour les petits organes et populations), les cellules de toutes les souris sont mises en commun par organe et triées par cytométrie en flux (FACS ARIA Becton Dickinson) en quatre populations lymphocytaires T : LT CD4+ effectrices (**Teff**), LT **CD8**+ et deux sous-populations LT régulatrices dites naïves (**nTregs**) et activées mémoires (**amTregs**), comme décrit **Figure 14**.

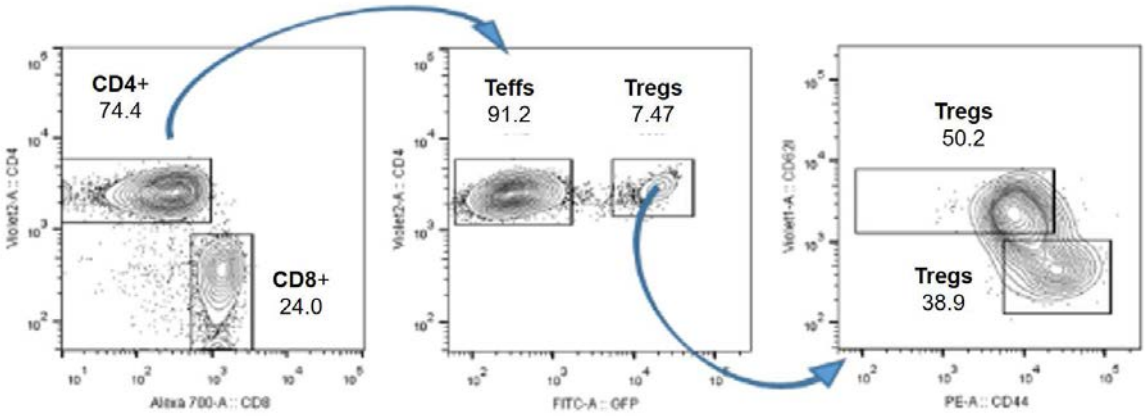


Figure 14 : Stratégie de tri des populations cellulaires – Les cellules de rate ou des six ganglions prélevés chez les souris sont mises en commun pour chaque organe. Chaque « pool » cellulaire est trié sur le marqueur CD3 caractérisant les LT. Les cellules CD3+ sont ensuite séparées et triées en populations CD8+ (CD8+), Teff (CD4+GFP-), amTregs (CD4+GFP+CD44HiCD62Llo/-) et nTregs (CD4+GFP+CD44lo/-CD62L+). La pureté des tris est supérieure à 99%.

Une fois le tri terminé, les cellules sont lavées en PBS1X puis lysées en tampon de lyse (RNAAquous, Ambion) selon les recommandations du fournisseur. Chaque tube généré à la fin de la collecte (aliquote) est décrit de façon précise dans une « base de données » en séparant les informations en 3 catégories : expérience, échantillon et aliquote. Ces 3 catégories sont liées de manière hiérarchique selon une nomenclature d'identification et une incrémentation de la numérotation continue (détails dans les Annexes 2 et 4). Ainsi, chaque aliquote d'un échantillon obtenu par une expérience donnée est nommé de la façon suivante :

[6] Non-Obese Diabetic : modèle de souris développant spontanément un diabète de type 1 vers l'âge de 2 ans environ.

[7] Souris mutante portant un gène rapporteur (Green Fluorescent Protein, GFP) sous le promoteur du gène foxp3 qui permet ici de trier les populations Tregs de manière précise.

IDexpérience_IDéchantillon_NUMEROaliquot (exemple : Dans l'expérience TRiPoD_06, l'échantillon 116 était constitué de 3 aliquotes > TriPoD_06_119_1, TriPoD_06_119_2 et TriPoD_06_119_3).

TriPoD_06 : Cette expérience a conduit à la préparation de 28 échantillons cellulaires obtenus par tri des sous-populations de LT de la rate et des 6 ganglions précédemment décrits prélevées chez 9 souris NOD mâles âgées de 9 semaines. L'ARN de ces échantillons a été extrait au laboratoire et envoyé pour séquençage à iREPERTOIRE Inc. dont la stratégie de préparation des librairies repose sur une série de PCR emboîtées ciblant les gènes TRBV et TRBJ (Wang et al., 2010 ; Annexe 3). Le séquençage TRB a été effectué sur un séquenceur Illumina MiSeq et une moyenne de $1{,}5.10^6$ séquences (+/- 20%) ont été produits par aliquote d'ARN (voir Annexe 4 pour le descriptif détaillé). Ces jeux de données seront utilisés dans les deux sections suivantes.

TriPoD_38_1070 : Échantillon de cellules Teff triées à partir de rates prélevées chez 7 femelles C57BL/6 âgées de 24 à 26 semaines dans le cadre de l'expérience TriPoD_38. Trois aliquotes de 3.10^6 cellules ont été préparés et les cellules suspendues en tampon de lyse. L'ARN de chaque aliquote a été extrait et divisé en 3 sous-aliquotes d'ARN équivalents pour le séquençage de leur répertoire TRB. Les 3 sous-aliquotes de l'aliquote TriPoD_38_1070_5 ont été envoyés à iREPERTOIRE Inc. pour être séquencés en parallèle sur un séquenceur Illumina HiSeq 2500. En moyenne, $8(+/- 1).10^6$ de séquences ont été produites par aliquote. Les 3 sous-aliquotes de l'aliquote TriPoD_38_1070_6 ont été séquencés en parallèle au laboratoire sur un séquenceur Roche/454 Life Sciences GS Junior après préparations de librairies par PCR multiplexes ciblant les gènes TRBV et TRBC (Bergot et al., 2015; Annexe 6). En moyenne, $1.6 (+/- 0.2).10^5$ de séquences ont été produites par aliquote (Voir Annexe 4 pour le descriptif détaillé). Ces jeux de données seront utilisés dans la section REPRESENTATIVITE DE LA DIVERSITE OBSERVEE.

2) Répertoire TR artificiel

Afin d'évaluer avec « certitude » la représentativité de la diversité observée, il est nécessaire de connaître a priori la diversité attendue pour un répertoire analysé. J'ai ainsi mis en place une méthode permettant de produire artificiellement des répertoires de clonotypes auxquels il est possible d'assigner une distribution et donc une diversité souhaitée (et connue). Pour ce faire, j'ai tiré parti du package tcR développé par Nazarov (2015), qui propose une fonction

permettant de générer de manière aléatoire des clonotypes TR humains, en intégrant le modèle de Murugan et al. (2012). Ces derniers ont utilisé les données produites par Robins et al (Robins et al., 2009, 2010) pour inférer les probabilités de recombinaisons entre les différentes familles de gènes V, D et J chez l'Homme. L'analyse des séquences non productives leur a, en effet, permis de construire un modèle prédisant la diversité potentielle des séquences de CDR3 (~1014 séquences nucléotidiques) et la contribution des insertions (~58%), des délétions (~25%) et de l'association des gènes à (~17%) avant la sélection thymique. Leur modèle permet donc de prédire la probabilité de générer une séquence CDR3 donnée, démontrant aussi qu'une même séquence CDR3 peut être produite, en moyenne, par 32 scenarios de recombinaisons différents.

Ainsi, sur la base de cette fonction, le module que j'ai développé permet de simuler un répertoire TR de la **richesse** clonotypique, **profondeur** et **diversité** souhaitées. À noter que combiné à une base de séquences, de qualité et longueur très variables, ce module permet de créer des fichiers fastq artificiels pouvant servir à évaluer les performances des outils d'annotations par exemple. Ce point ne sera toutefois pas abordé ici.

Une fois la richesse et la taille du jeu de données déterminées, le module propose cinq lois différentes pour la distribution des comptes :

- *Uniforme* : les valeurs de comptage assignées à chaque clonotype suivent une loi uniforme et leur somme est égale à la profondeur souhaitée.
- *Équiprobabilité* : tous les clonotypes se voient attribués le même compte à savoir une valeur arrondie du ratio entre le nombre de clonotypes et la profondeur souhaitée.
- *Log-normale*
- *Exponentielle* dont les valeurs de μ et σ^2 sont à fixer par l'utilisateur
- *Zipf-Mandelbrot* (Evert, 2004), généralisation de la loi de Zipf dont les paramètres sont à fixer par l'utilisateur

En particulier, la loi de Zipf, définie dans le cadre de l'analyse des fréquences des mots dans un texte, lie la fréquence d'occurrence f(n) d'un élément à son rang n dans l'ordre de l'ensemble des fréquences d'observation. Elle établit la probabilité p d'observer le $i^{ème}$ élément d'un ensemble infini d'objets par tirage aléatoire telle que $pi = {A}/{(1 + Bi)^{\alpha}}$ avec A une constante et les deux paramètres de la loi α et B tels que $\alpha \in [0 ; +\infty[$ et $B \in]0 ; 1]$.

L'exposant α détermine la pente de décroissance de la distribution des fréquences : $\alpha>1$ assure la régularité de la distribution, $\alpha>2$ autorise une estimation finie de la moyenne et $\alpha>3$ celle de la variance (Pearson, 2010).

Ainsi, alors que Sepúlveda suggère que la loi log-normale permet de modéliser la distribution clonale dans un contexte polyclonale (Sepúlveda et al., 2010), plusieurs études ont établi que c'est une loi de Zipf qui émerge notamment lors de fluctuations du système (Burgos and Moreno-Tovar, 1996; Greiff et al., 2015b; Mora and Walczak, 2016; Schwab et al., 2014). De plus, Greiff a démontré la corrélation entre le paramètre α d'une loi Zipf imputée à un répertoire clonal et le profil de diversité de ce dernier (Greiff et al., 2014).

B. Gestion des données

Comme décrit dans l'Annexe 2, j'ai mis en place une procédure de stockage et de traitement des données d'immunoséquençage produites pour faciliter et standardiser la gestion des données produites. Les données brutes sont stockées au format FASTQ. Une description des scores de qualité et des distributions nucléotidiques à travers les séquences est produite pour chaque fichier en utilisant FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) sous Linux puis les séquences sont annotées avec clonotypeR (Plessy et al., 2015). Le fichier de séquences annotées (format TSV) décrit pour chaque séquence TR identifiée : i) l'identifiant de la séquence ii) les segments de gènes V et J utilisés iii) un score d'alignement, variant avec la longueur de la région de concordance entre la séquence testée et le gène V, et le score évalué par l'algorithme Burrow-Wheeler (Li and Durbin, 2010) reflétant la spécificité et la précision de l'alignement de la séquence avec le gène TRV, et iv) la région CDR3 identifiée sous trois formats différents : la séquence nucléotidique, sa séquence Phred et sa traduction en séquence peptidique. Chaque fichier d'annotation est nommé par la concaténation de l'identifiant de l'aliquote correspondant et de l'identifiant du passage sur séquenceur à partir duquel il a été produit.

La « base d'échantillons » Excel décrite précédemment permet d'inventorier et de décrire toutes les expériences, échantillons, aliquotes et jeux de données RepSeq produits dans le cadre du projet. À ce jour, 29 expériences regroupant plus de 1000 aliquotes cellulaires ont été produits et sont en attente de séquençage.

C. Pré-traitement des données

Comme décrit précédemment, avant d'être analysées les données RepSeq peuvent être traitées de différentes manières pour minimiser l'impact des erreurs de séquençage. La méthodologie développée ici permet de choisir d'ignorer les clonotypes observés une seule fois (singletons) ou de les assimiler aux clonotypes plus abondants dont la séquence du CDR3 ne diffère que d'un seul nucléotide. Pour ce faire, après annotation des séquences brutes, les clonotypes identifiés, définis par leur combinaison unique V-CDR3-J, sont d'abord catégorisés en fonction de leur combinaison V-J. Au sein de chaque catégorie V-J. Les singletons sont séparés des autres clonotypes dits « non-singletons ». Une **distance de Levenshtein**[8] est ensuite calculée entre les séquences peptidiques des CDR3 de chacun des deux groupes de clonotypes. Pour chaque clonotype « non-singleton », si la distance de Levenshtein entre son CDR3 et celui d'un singleton est égale à 1, leurs séquences nucléotidiques respectives sont comparées. Si ces deux séquences nucléotidiques ont également une distance de Levenshtein de 1, la séquence du singleton est considérée comme erronée et est remplacée (ou « assimilée ») par celle du clonotype « non-singleton ». Par ailleurs, les séquences TR peuvent également être filtrées de façon à ignorer les séquences avec les plus faibles scores d'alignement, d'après le score fourni par *clonotypeR*.

La **Figure 15** met en parallèle la composition d'un même jeu de données RepSeq en fonction du traitement appliqué : sans traitement (*Original*), retrait des singletons (*Original w/o single*), assimilation des singletons (*Merged*), assimilation et retrait des singletons (*Merged w/o single*) et filtrage des séquences TR dont les annotations sont les moins fiables (*Filtered*).

[8] Distance mathématique mesurant la similarité entre deux séquences. Elle est égale au nombre minimal de délétions, insertions ou substitutions nécessaires pour passer d'une séquence à l'autre et est pondérée par la longueur des séquences comparées

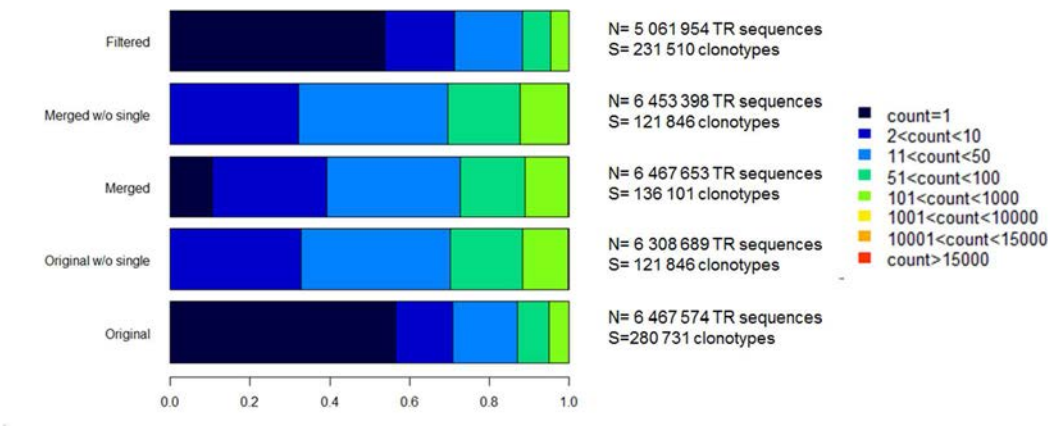


Figure 15 : Distributions des occurrences de séquences TR – Pour chaque jeu de données, les clonotypes sont catégorisés en fonction de leur occurrence dont la distribution est représentée en histogramme cumulé. Chaque couleur correspond à une tranche d'occurrences tel qu'indiqué dans la légende. Pour chaque jeu de données, les nombres de séquences TR (N) et de clonotypes (S) sont indiqués.

Dans le jeu de données d'origine, la proportion de singletons représente près de 60% des clonotypes identifiés. L'assimilation de ces clonotypes permet de réduire considérablement cette proportion à 10%. Ainsi, la richesse observée est réduite de près de 50%. Les singletons « assimilés » viennent majoritairement enrichir les sections de clonotypes les plus enrichies par observé entre 2 et 10 fois et entre 100 et 1000 fois, tout en conservant une valeur de similarité Morisita-Horn calculée entre les jeux « originaux » et les jeux « assimilés » étant égal à 0,9999299. Ce traitement est donc plus conservatif que le retrait de singletons qui réduit la richesse de près de 60% des clonotypes (dont ici 95% sont « assimilables ») avec un MH= 0,9996751.

Le filtre des séquences TR en fonction du score d'alignement peut avoir un effet non négligeable sur le jeu de données. Les séquences sont filtrées si leur score d'alignement est inférieur au premier quartile de la distribution des scores à travers toutes les séquences annotées. Ici, le nombre de séquences filtrées représente 22% du jeu d'origine ce qui réduit à la richesse de 18%, ce qui signifie que la majorité des séquences filtrées sont des singletons. Toutefois, la similarité entre ce jeu et le jeu original est de 0,9905596 ; la diversité du jeu est donc un peu plus impactée par cette méthode. Il est à noter que cette méthode est sensible aux longueurs de séquences et peut notamment être utile pour s'assurer de la crédibilité de l'annotation des séquences courtes.

Ces corrections vont s'avérer plus ou moins nécessaires en fonction des conditions expérimentales de séquençage. La proportion des singletons varie notamment en fonction de

la profondeur de séquençage relativement à la taille de l'échantillon cellulaire analysé. Il a été observé au laboratoire que le séquençage très profond (de l'ordre de plusieurs millions de séquences) de du répertoire de populations comptant 10^4 à 10^5 cellules par exemple, produit une proportion très grande de singletons ce qui affecte la richesse de clonotypes observés qui apparaît bien supérieure au nombre de cellule initial. Ce type de biais nécessite donc une correction adaptée en fonction de l'objectif de l'analyse : si l'on souhaite identifier les clonotypes prédominants, le retrait des singletons n'aura pas d'impact sur le résultat alors que la mesure de la diversité globale d'un répertoire peut être fortement sous-estimée si les singletons sont retirés, particulièrement dans le cas d'un répertoire polyclonal. Ce dernier cas va donc nécessiter l'assimilation des singletons. Les analyses de données expérimentales présentées dans les sections suivantes visent à caractériser la diversité de répertoires TR plus ou moins polyclonaux. Toutes les données expérimentales ont donc été ici normalisées par « assimilation des singletons » (**Table 3**).

Table 3 : Distribution de la similarité MH des 28 jeux de données de TriPoD_06 avant et après assimilation des singletons.

Minimum	1er Quartile	Médiane	Moyenne	3ème Quartile
0,9821	0,9983	0,9991	0,9983	0,9999

MODELISATION DE LA DIVERSITE DU REPERTOIRE TR

Un jeu de données RepSeq décrit au niveau moléculaire une population lymphocytaire. Le niveau de détails fournis par ces données est sans précédent et doit être exploité aux mieux. Mon but a donc été de mettre au point une stratégie pertinente permettant la description de la diversité d'un répertoire lymphocytaire T en combinant différents indices de diversité qui, du fait de leurs différences conceptuelles, seront employés différemment en fonction de la problématique abordée. À l'instar de ce qui est communément effectué en Écologie, la diversité de l'inventaire des molécules TRA ou TRB analysées peut être observée à différents niveaux de granularité en fonction de ce que l'on définit comme espèce. De manière classique, on considère chaque clonotype TR comme une espèce plus ou moins abondante au sein du répertoire observé. Toutefois, un même jeu de données peut être décomposé en termes d'expression des gènes TR (V ou J), de séquences ou de longueurs de CDR3 mais aussi en combinant ces paramètres. Ces catégories peuvent être analysées indépendamment les unes des autres et par différentes méthodes complémentaires : estimation de richesse, distribution de fréquence relative, distribution des longueurs de CDR3, propriétés physico-chimiques des séquences protéiques... de manière à capturer une image globale de la diversité du répertoire observé.

L'approche proposée ici a donc eu pour premier objectif l'identification des paramètres caractérisant le mieux la diversité du répertoire lymphocytaire pour ensuite les combiner et les utiliser comme indicateurs du statut physiopathologique des individus (souris ou patients) dont les répertoires sont analysés.

A. Exploration individuelle des répertoires TR

La caractérisation approfondie de chaque répertoire TR est essentielle à une compréhension précise du phénomène immunologique d'intérêt pour le biologiste. Pour permettre cela, la méthodologie que j'ai développée produit systématiquement, pour chaque jeu de données, une série de 17 fichiers permettant d'explorer la composition, mais aussi la structure du répertoire de TR analysé (voir Annexe 1 pour plus de détails sur le code et les fichiers de sortie). Ces analyses reposent sur l'application des indices de diversité α décrits en introduction à différents niveaux de granularité. Le défi a été d'établir une stratégie

cohérente, prêtant attention au choix des indices adoptés et au choix des paramètres à utiliser, ainsi qu'à l'iconographie pour la représentation des résultats.

Le jeu de données utilisé pour la démonstration de la méthodologie individuelle a pour identifiant *TriPoD_06_235_1_22_iREP002*. Il a été obtenu par séquençage du répertoire TRB de la population amTregs des ganglions pancréatiques (PLN-amTregs) prélevés lors de l'expérience TriPoD_06.

1) Statistiques descriptives

Un premier niveau de description est permis par l'utilisation de compteurs résumant le contenu du jeu de données (**Table 4**) mais aussi de métriques objectives (**Figure 16**) telles que :

- Le ratio entre le nombre de clonotypes observés et le nombre de séquences TR analysées, appelé « ***clonality*** » $CL = \frac{\#\ clonotypes}{\#\ TR\ sequences}$
- Le facteur **D50**, introduit par iREPERTOIRE sur leur plate-forme d'analyse (https://irweb.irepertoire.com/nir/#); c'est la proportion de clonotypes nécessaires pour atteindre une fréquence cumulée de 50%.
- Les indices de **Piélou**, **Simpson Ds** et **Hill** tels que décrits précédemment.

Table 4 : Statistiques descriptives du jeu de données.

	TriPoD_06_235_1_22
Nb séquences TR	1 198 777
% productive	92,44%
Nb séquences TR productives	1 108 156
Nb TRBV	20
Nb TRBJ	12
Nb TRBVBJ	191
Nb clonotypes	10 811
Nb CDR3nt	18 169
Nb CDR3pep	10 334

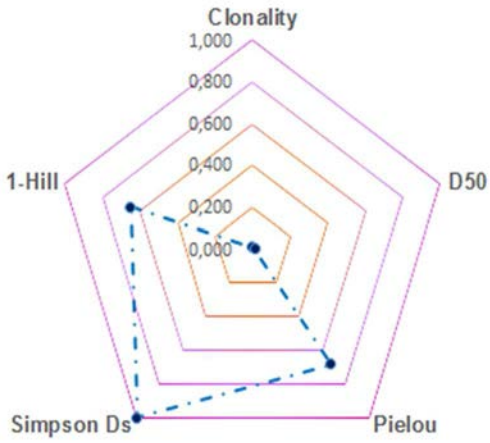


Figure 16 : Métriques descriptives de la composition globale du répertoire TRB analysé – Chaque axe de la représentation en radar indique la valeur d'une des cinq métriques calculées sur la base de la distribution de fréquences des 10 811 clonotypes observés.

Comme résumé **Table 4**, 92% des séquences TR identifiées dans le jeu de données sont des séquences productives (séquence du CDR3 en phase de lecture et sans codon stop). Ces 1 108 156 séquences TR se distribuent en 10 811 clonotypes (V-CDR3pep-J). Le profil descriptif

(**Figure 16**) indique que le nombre de clonotypes est très inférieur à la taille du jeu de données (*Clonality*) ; ces clonotypes ne semblent pas équivalemment représentés comme le suggère les indices D50 (=0.012), Piélou (=0.679) et Hill (=0.351). Toutefois, l'indice quadratique de Simpson, qui ignore les espèces les moins abondantes, est élevé ce qui indique une faible proportion de clonotypes rares dans le jeu de données. Ces valeurs suggèrent que le jeu de données est assez large pour permettre une observation représentative des clonotypes en présence et que le répertoire observé est dominé par quelques clonotypes plus abondants.

2) Composition en gènes V et J

La **Table 4** indique que les 10 811 clonotypes observés expriment 20 des 24 gènes TRBV possibles ainsi que la totalité des gènes TRBJ et se distribuent en 191 combinaisons TRBVBJ. Le **profil d'usage TRBV** à travers les clonotypes observés représenté **Figure 17** montre une utilisation préférentielle des membres de la famille TRBV13 et de TRBV5 dans cet échantillon alors que les deux familles TRBJ semblent équivalemment représentées.

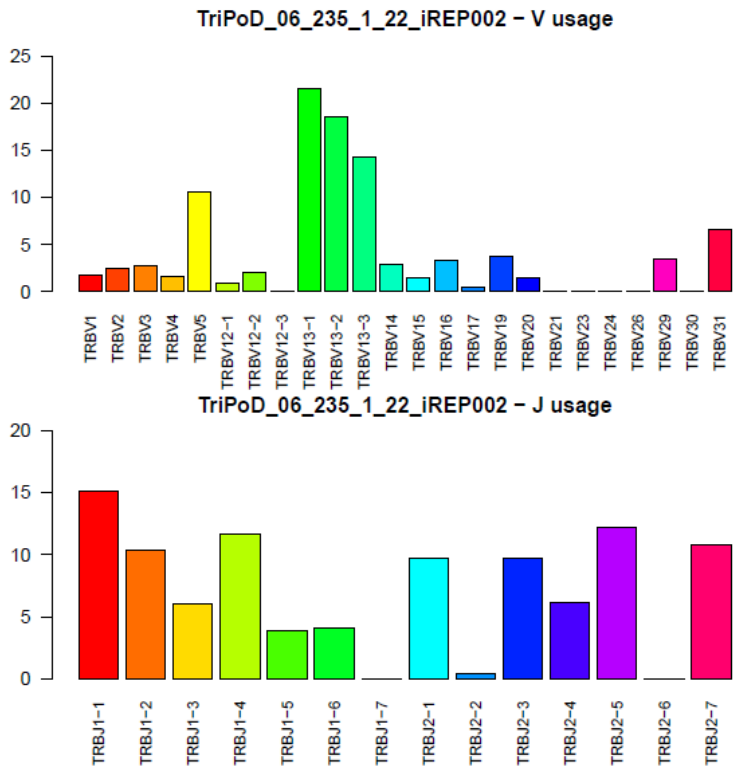


Figure 17 : Usage des gènes TRBV et TRBJ au sein du jeu de données – Histogrammes représentant la fréquence d'usage de chaque gène TRBV et BJ à travers les clonotypes observés. Les valeurs représentées ne prennent pas en considération les fréquences des clonotypes.

Par ailleurs, alors qu'il a été établi que l'association des TRBV et TRBJ ne se fait pas de manière aléatoire (Born et al., 1985; Malissen et al., 1984), décrire la diversité en fonction de

l'expression des combinaisons TRBVBJ peut être informatif sur la structure du répertoire étudié. Cette diversité peut être décrite de de différentes manières.

Ainsi, pour analyser la distribution de l'**expression de ces combinaisons TRBVBJ**, la fréquence de chacune est calculée relativement aux autres à travers les séquences TR ; ce calcul permet combiner les fréquences d'expression des gènes à l'abondance des clonotypes (**Figure 18**).

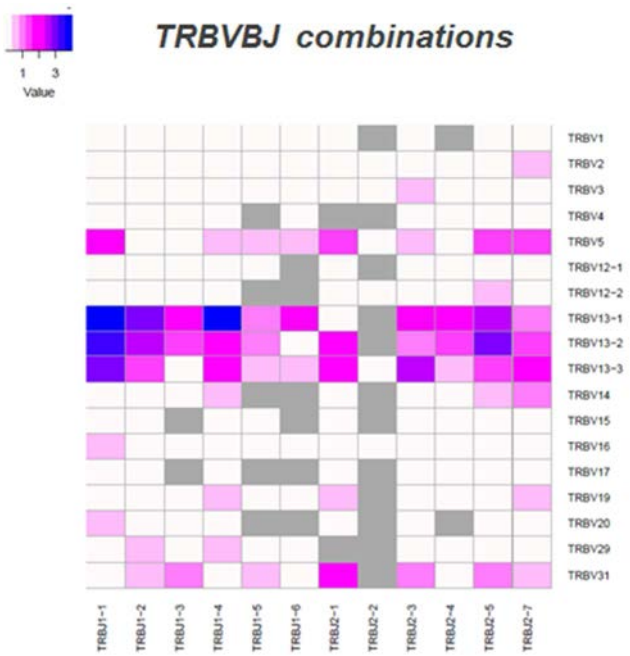


Figure 18 : Fréquences de combinaisons TRBVBJ – Chaque cellule indique d'après l'échelle indiquée la fréquence de combinaisons entre les gènes TRBV (en ligne) et les gènes TRBJ (en colonne) observés. En gris, les combinaisons non observées.

Les fréquences des combinaisons TRBVBJ au sein du répertoire analysé varient entre 4.10^{-2} et 9.10^{-5}, les combinaisons majoritaires impliquant les membres de la famille TRBV13 qui semblent être préférentiellement associés aux membres de la famille TRBJ1.

Si l'on s'intéresse à la f**réquence d'association de chaque gène TRBJ avec chacun des gènes TRBV** (**Figure 19**), il semble apparaître deux grands groupes de combinaisons TRBVBJ (qui subsistent même lorsque le TRBJ2-2 présentant beaucoup de données manquantes est ignoré). Notamment, les gènes TRBV13-1, BV13-2 et BV29 ont une grande diversité d'association avec une représentation équivalente des gènes TRBJ (Indice de Piélou respectifs P=0.945,0.948 et 0.925). Les gènes TRBV du second cluster sont associés à tous les TRBVBJ mais les combinaisons impliquant les gènes TRBJ1-1, BJ1-2, BJ2-1, BJ2-3 et BJ2.7 prédominent (Pmoyen= 0.868 +/- 0.022). Les gènes TRBV12-1 et 12-2, TRBV20 et TRBV17 ne sont impliqués que dans peu de combinaisons, favorisant 1 à 2 TRBJ en particulier (P compris entre 0,6 et 0,8).

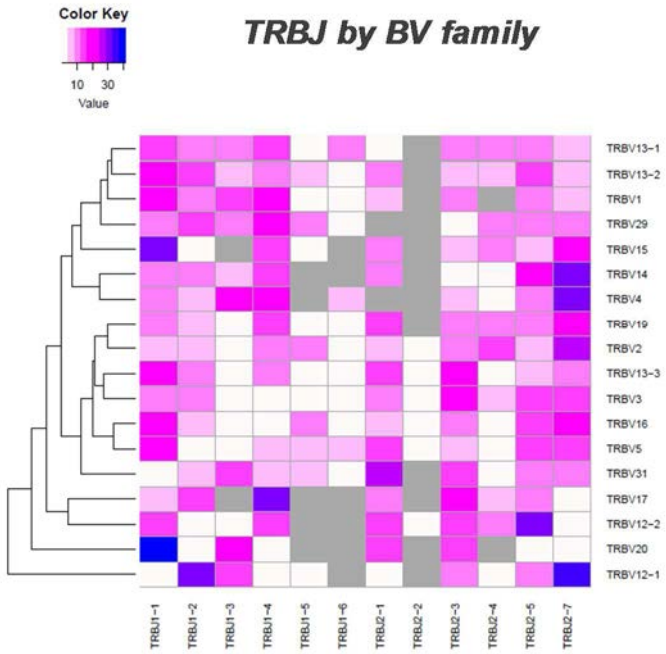


Figure 19 : Classification hiérarchique des gènes TRBV en fonction de leur fréquence d'association avec les TRBJ – Chaque cellule indique, d'après l'échelle indiquée, la fréquence relative d'association d'un gène TRBJ à un gène TRBV donné. La somme des fréquences par ligne est égale à 1. En gris, les combinaisons non observées. Dendrogramme : distance euclidienne et méthode « *complete* ».

.

Les analyses précédentes informent sur l'expression des gènes TRBV et TRBJ à travers le répertoire et la présence de combinaisons TRBVBJ prédominantes mais il reste à déterminer si ces motifs d'expression corrèlent par l'usage préférentiel de ces gènes par de nombreux clonotypes ou par la prédominance d'un ou quelques clonotypes exprimant ces gènes. Le calcul de la **diversité clonotypique au sein de chacune des combinaisons TRBVBJ observées** permet d'affiner ces observations (**Figure 20**). Ainsi, alors que le TRBV12-1 semble être préférentiellement associé avec les TRBJ1-2, BJ2-7 et BJ1-3, la diversité des clonotypes exprimant ces combinaisons n'est pas très élevée par rapport à celle des autres combinaisons impliquant le gène TRBV12-1, particulièrement pour la combinaison TRBV12-1-BJ1-3 dont la diversité un peu plus faible que les deux autres suggère que les clonotypes exprimant ces gènes sont plus abondants que ceux exprimant les BJ1-2 ou BJ2-7. On observe également sur la **Figure 20** que le répertoire TRBV est séparable en deux clusters notamment du fait de la grande diversité des clonotypes exprimant TRBJ1-1 au sein de ceux exprimant les TRBV1, 5, 13-1, 13-2, 16 et 20.

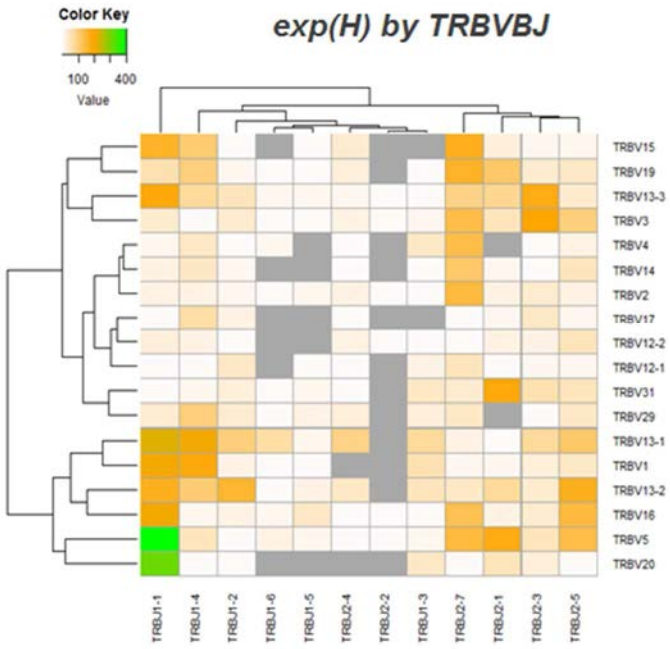


Figure 20 : Classification hiérarchique des combinaisons TRBVBJ en fonction de leur diversité clonotypique – L'exponentielle de l'indice de Shannon exp(H) est calculée en prenant en compte les fréquences relatives des clonotypes catégorisés en fonction de leur combinaison TRBV-VJ, puis pondérée par la fréquence de chaque combinaison. Chaque cellule indique, d'après l'échelle indiquée, la valeur de cette diversité clonotypique par combinaison TRBVBJ. En gris, les combinaisons non observées. Dendrogramme : distance euclidienne et méthode « *complete* ».

Ces différentes analyses permettent de révéler la topologie du répertoire analysé. Ainsi, en comparant les topologies de répertoires de différentes populations LT par exemple, on pourra identifier d'éventuels motifs spécifiques à chacune.

3) Spectratypage CDR3

Comme décrit dans l'introduction, une approche couramment utilisée pour l'analyse du répertoire TR est la technique Immunoscope® qui décrit la diversité des clones par l'analyse des profils de **distribution des longueurs de CDR3 par famille de TRBV** ou combinaison TRBVBJ. Il est possible de produire ces profils à partir des données d'immunoséquençage en catégorisant les clonotypes exprimant le même gène TRBV en fonction de leur longueur de CDR3 et calculant les fréquences de ces longueurs au sein de chaque famille TRBV (**Figure 21**). Ainsi, les TRBV2, 13-1 et 13-2 sont les seuls à présenter une distribution « gaussienne » alors que les profils TRBV3, BV4 et BV12-1 semblent dominés par un pic majoritaire.

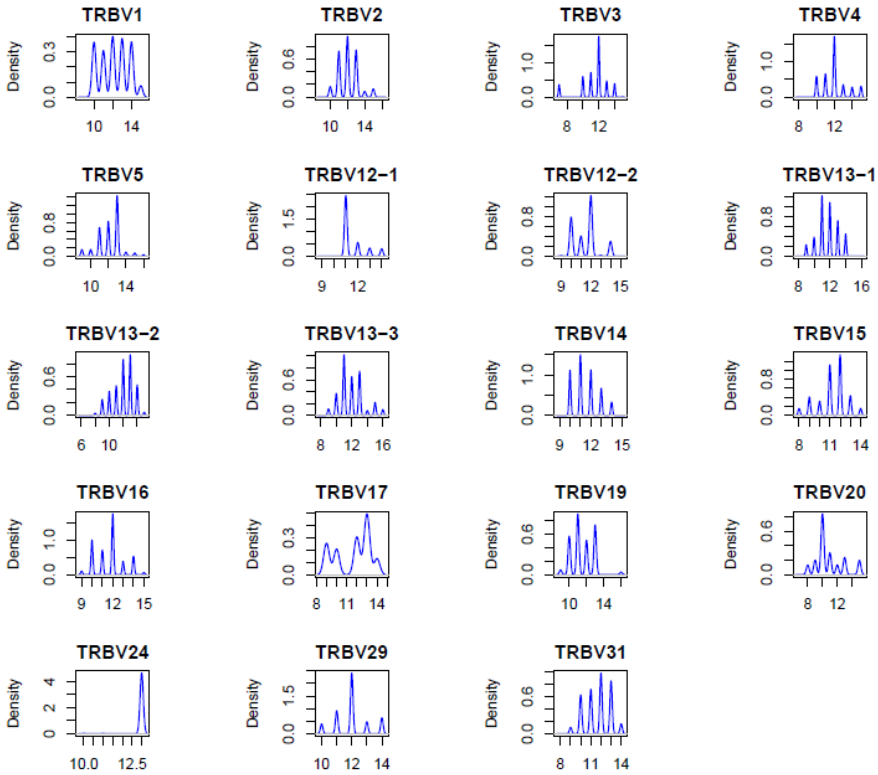


Figure 21 : Spectratypes des familles TRBV au sein du répertoire analysé – Les clonotypes sont catégorisés en fonction du gène TRBV exprimé et la densité des longueurs de séquences peptidiques de leur CDR3 représentée dans chaque cadrant.

Outre les précisions apportées sur la diversité du répertoire TRBV, cette analyse donne la possibilité de comparer les résultats obtenus en immunoséquençage à des résultats de la littérature obtenue en Immunoscope® par exemple, ou comme décrit par Bergot, Chaara et al. (2015 ; Annexe 6) de combiner ses deux techniques afin de séquencer de manière ciblée les familles de BV qui présentent, en première analyse Immunoscope®, un comportement caractéristique du phénomène étudié.

4) Distribution des clonotypes

Chaque clonotype TRB identifié dans le répertoire TR analysé a une fréquence censée refléter la proportion du/des clone(s) l'exprimant dans la population étudiée. La **distribution des clonotypes en fonction de leur occurrence** fournit une représentation simple de la composition du répertoire (**Figure 22**).

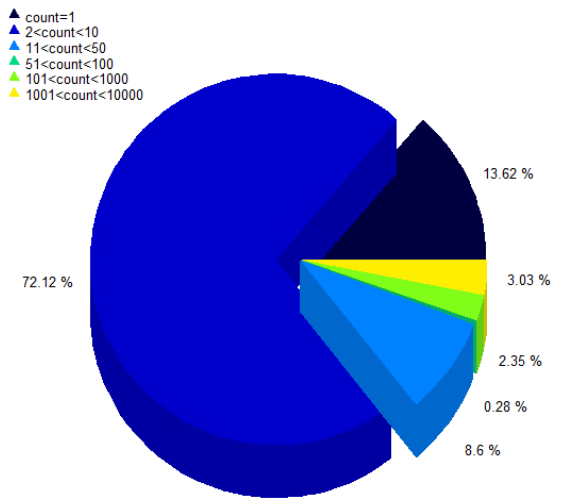


Figure 22 : Distribution des occurrences des clonotypes – Les clonotypes identifiés sont catégorisés en fonction de leur occurrence. Chaque couleur correspond à une tranche d'occurrences tel qu'indiqué dans la légende.

Environ 14% des clonotypes identifiés sont des singletons alors que la majorité (environ 80%) sont observés 2 à 50 fois dans ce jeu de données. Les 6% restant sont fortement exprimés notamment plus de 300 clonotypes dont l'abondance est comprise entre 10^3 et 10^4 copies.

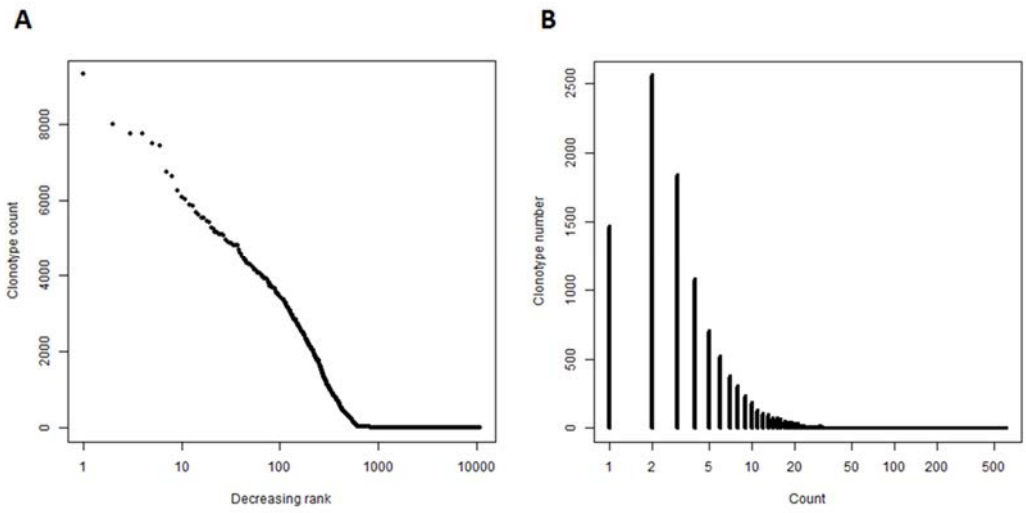


Figure 23 : Distributions de clonotypes – **A)** Les clonotypes sont triés de façon décroissante en fonction de leur fréquence qui est représentée sur l'axe des ordonnées. **B)** Spectre des abondances des clonotypes. En abscisses, les comptes possibles et en ordonnées, le nombre de clonotypes observés à ces comptes.

Cette distribution peut être observée sur la **Figure 23**A où l'on distingue un clonotype majoritaire (rank=1 ; count = 9 329), suivi par un panel d'environ 7 000 clonotypes dont l'abondance diminue de façon continue. Les clonotypes suivants sont observés 1 à 2 fois. Ainsi, on peut établir le spectre d'abondances des clonotypes (**Figure 23B**) qui, de manière similaire mais plus précise qu'en **Figure 21**, indique une prédominance de clonotypes exprimés 2 fois.

5) Diversité des clonotypes

Outre la distribution de la fréquence des clonotypes qui le composent, un répertoire TRB est caractérisé par la diversité de ces clonotypes. Comme décrit dans l'introduction, les indices de diversité utilisés en Écologie sont dérivés de l'**entropie de Rényi**, fonction paramétrée par un facteur α dont l'augmentation va renforcer l'influence des espèces prédominantes dans la mesure de la diversité. Chaque indice informant différemment sur la diversité du système étudié, les combiner tel que présenté **Figure 24** sous forme de profil permet une description globale de la diversité du répertoire analysé (Greiff et al., 2015a).

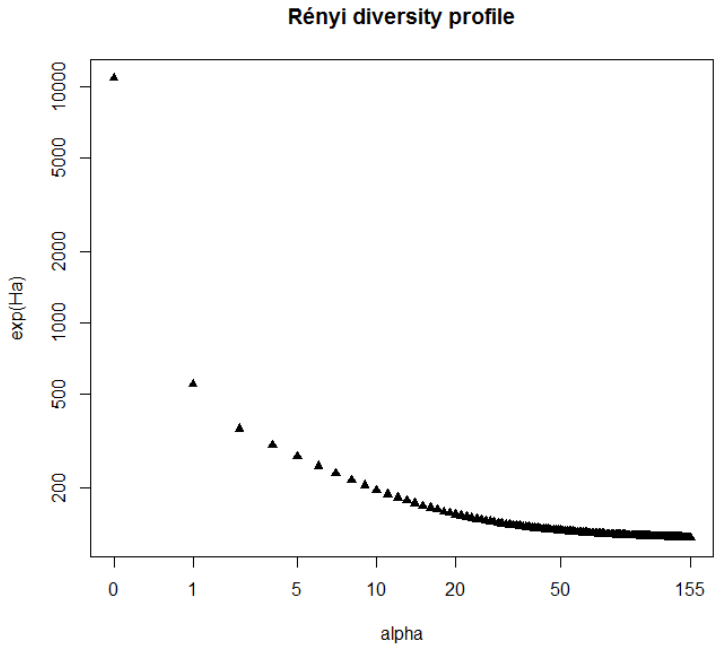


Figure 24 : Profil de diversité des clonotypes – Valeur de l'exponentielle de l'entropie de Rényi en fonction de la valeur croissante de α, qui ici varie entre 0 (Richesse spécifique) et 155.

Alors que la richesse spécifique (α=0) correspondant à la richesse observée (10 811) est de l'ordre 10^4 espèces, l'indice de Shannon (α=1) et l'indice quadratique de Simpson (α=2) estiment la diversité à 547 et à 355. Or, ces valeurs correspondent respectivement aux nombres de clonotypes dont l'abondance est comprise entre 10^2 et 10^4 et entre 10^3 et 10^4 (**Figure 21**). Ces observations confirment qu'au fur et à mesure que la valeur de α augmente, l'ordre de grandeur des clonotypes ignorés pour le calcul de Rényi augmente également.

B. Comparaison des répertoires TR

L'application de métriques de diversité α permet de modéliser la composition et la topologie des répertoires observés et fournit des critères objectifs de comparaison de la diversité de la distribution clonale entre ces répertoires (Rempala et al., 2011). Ainsi, en comparant des échantillons ou des groupes d'échantillons, il est possible (i) d'évaluer l'homogénéité des réplicas biologiques au sein d'un des groupes définis *a priori* et (ii) d'identifier les comportements collectifs pouvant être associés aux conditions biologiques d'intérêt.

Dans les paragraphes suivants, je m'attacherai à montrer comment l'application des différentes métriques exposées dans la section précédente permet une comparaison approfondie du répertoire des échantillons analysés, en explicitant l'intérêt de chaque analyse.

1) Statistiques descriptives et distributions des clonotypes

Afin d'assurer la pertinence des résultats de comparaison, certaines précautions doivent être prises notamment quant à la taille des jeux de données comparés ou la représentativité de la profondeur de séquençage.

Pour répondre à ces préoccupations, une première étape consiste à comparer les statistiques descriptives de tous les échantillons d'intérêt en plus de la comparaison des distributions d'occurrence.

La première étape consiste à comparer les nombres de séquences. À titre d'exemple, on peut voir que sur le jeu de données TRiPoD_06 (**Figure 25**), le nombre moyen de séquences TR est de l'ordre 10^6 pour chaque groupe de jeux de données. Dans un second temps, il est également important de s'assurer que les **nombres observés de clonotypes, gènes TRBV et gènes TRBJ** sont homogènes au travers d'échantillons de même nature, garantissant l'absence de biais technique. Dans le jeu TRiPoD_06, ces quatre paramètres sont plus variables que le nombre de TR, ce qui peut s'expliquer par la nature différente des échantillons.

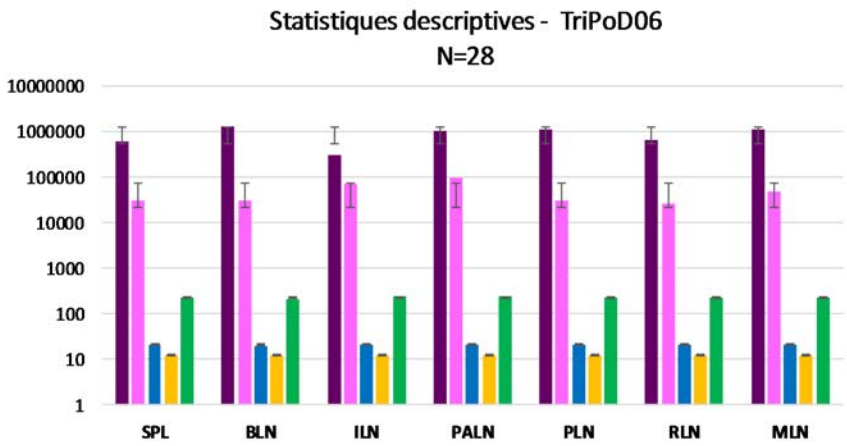


Figure 25 : Statistiques descriptives des 28 jeux de données TriPoD_06 – Les nombres de séquences TR (violet), clonotypes (rose), gènes TRBV (bleu), gènes TRBJ (jaune) et combinaisons TRBVBJ (vert) sont comptés au sein de chaque jeu de données. Les moyennes et écart-types de ces cinq paramètres sont calculés à travers les quatre jeux de données (populations lymphocytaires) par organe : rate (SPL), ganglions brachiaux (BLN), inguinaux (ILN), para-aortiques (PALN), pancréatiques (PLN), rénaux(RLN) et mésentériques (MLN).

Les compositions des jeux de données peuvent être comparées globalement sur la base des métriques : la **clonalité** (*Clonality)*, le **D50**, l'indice de **Piélou**, l'indice quadratique de **Simpson** et l'indice de **Hill** (**Figure 26**). Alors que les deux premières sont sensibles à la taille des jeux de données, les trois autres permettent une évaluation non biaisée de la diversité globale. L'application de ces indices sur l'ensemble des jeux de données permet d'évaluer la variabilité entre les répertoires comparés. À titre d'exemple, sur le jeu de donnée TRiPoD_06 caractérisé par une diversité d'organes et de populations cellulaires, la clonalité (*Clonality)* est peu élevée et variable au sein des deux séries d'échantillons Tregs suggérant une homogénéité des jeux de données alors que les échantillons Teff montrent une grande variabilité notamment à cause de l'échantillon ILN-Teff. La proportion de clonotypes nécessaire pour constituer 50% du répertoire total (D50) est plus grande pour les répertoires Teff et CD8 ce qui nous indique une plus grande diversité de ces répertoires qui sont équivalemment représentés comme signalé par l'indice de Piélou proche de 1 et peu variable pour ces populations. *A contrario*, cet indice est très variable entre les organes pour les populations Tregs notamment à cause de valeurs beaucoup plus faibles que les autres pour les échantillons BLN-amTregs et MLN-nTregs, ce qui reflète un biais de distribution de ces répertoires qui semblent dominés par des clonotypes particulièrement abondants. De plus, alors que l'indice de Simpson est homogène entre les organes et de valeurs similaires entre les populations, l'indice de Hill nous indique que la diversité globale des répertoires amTregs, nTregs et CD8 est homogène entre les

organes bien que de valeurs soient différentes. En revanche, l'échantillon SPL-Teff fait baisser la valeur médiane de l'indice de Hill des répertoires Teff. Ces métriques permettent donc de résumer la composition de ces 28 répertoires et de mettre en évidence d'éventuelles disparités. Dans le cas de cet exemple, l'absence de répétitions des observations ne nous permet pas de conclure sur le sens biologique des variabilités observées. Cependant, dans le cas de comparaisons de plusieurs groupes biologiques, chacun composé de plusieurs échantillons, ces métriques permettent d'évaluer l'homogénéité intra-groupe et d'identifier des disparités inter-groupes.

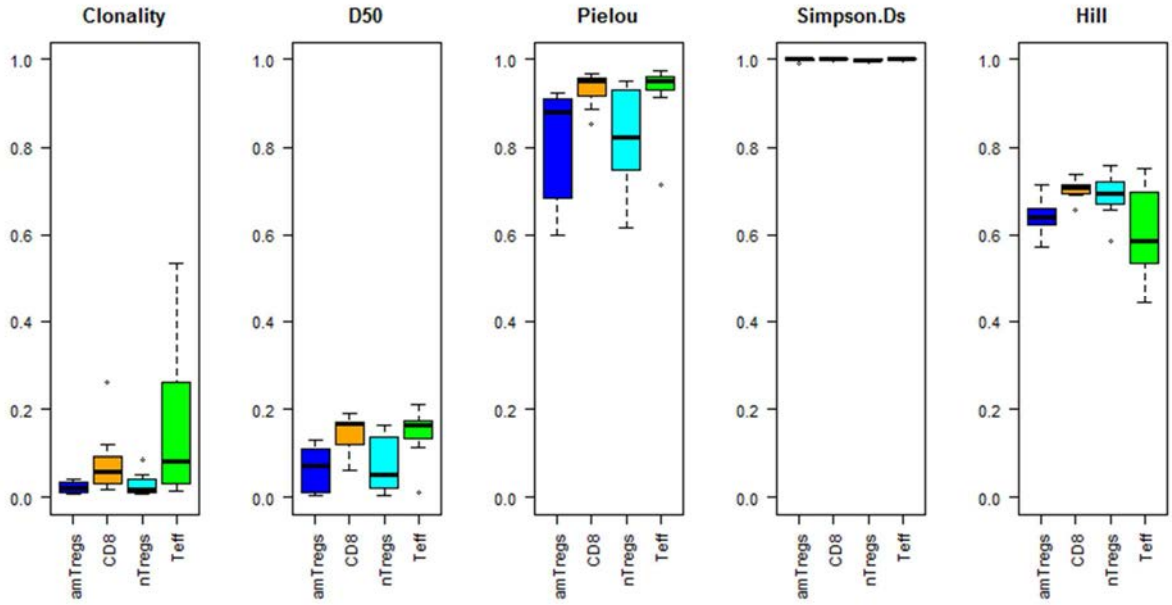


Figure 26 : Distributions des métriques descriptives à travers les 28 jeux de données TriPoD_06 – Cinq métriques descriptives sont calculées pour chacun des 28 échantillons : *Clonality*, D50, l'indice de Piélou, l'indice quadratique de Simpson et l'indice de Hill. Ces indices varient de 0 à 1. La distribution de leur valeur est représentée par population cellulaire (vert : Teff, orange : CD8, cyan : nTregs et bleu : amTregs) à travers les organes.

Pour caractériser les variabilités observées, les clonotypes sont ordonnés de manière décroissante au sein de chaque jeu de données en fonction de leur abondance afin de comparer la **distribution clonotypique** de chaque répertoire (**Figure 27**). Les fréquences des clonotypes Teff et CD8 (respectivement en vert et orange), à l'exception de celle de SPL-Teff dont l'indice de Hill était déjà différent de celui des autres, sont équivalentes et faibles traduisant des répertoires TRB très divers (polyclonaux). Les répertoires Tregs sont, de manière générale, caractérisés par la présence d'un nombre variable de clonotypes prédominants alors que BLN-amTregs et MLN-nTregs (pour lesquels l'indice de Piélou était particulièrement faible) se détachent des autres échantillons du fait de la présence de

clonotypes fortement prédominants au sein de leurs répertoires (plus de 20 000 séquences TR pour leurs clonotypes majoritaires respectifs contre quelques centaines pour un répertoire polyclonal de type Teff).

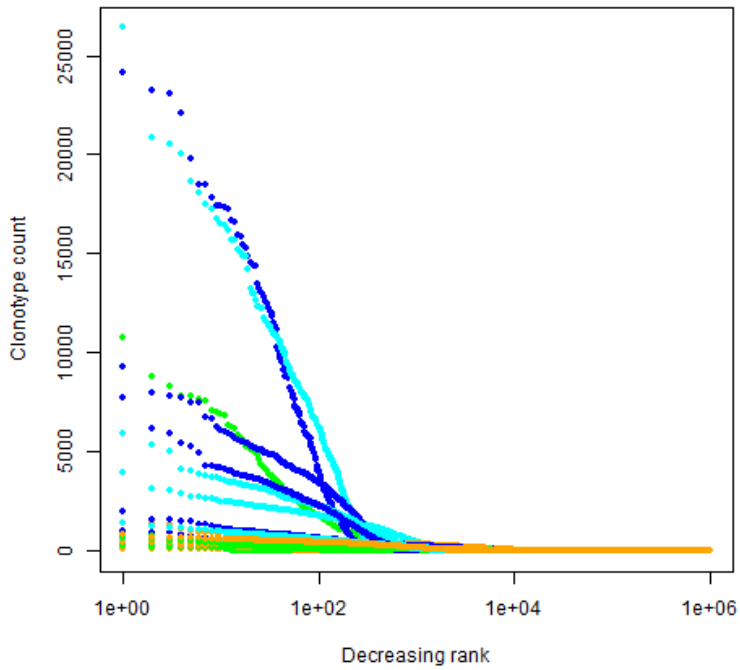


Figure 27 : Distributions des clonotypes par jeu de données – Les clonotypes de chaque répertoire sont ordonnés de manière décroissante en fonction de leur fréquence d'observation. Les courbes sont colorées en fonction de la population cellulaire : Teff en vert, CD8 en orange, nTregs en cyan et amTregs en bleu.

L'usage de courbes de **raréfaction**, courant en Écologie, s'applique parfaitement aux données de répertoire pour évaluer l'effet de l'échantillonnage (ici la profondeur de séquençage) sur l'évaluation de la diversité. Ce type d'analyse permet d'identifier les structures des populations étudiées mais aussi, le cas échéant, d'identifier à quelle profondeur sous-échantillonner les jeux de données pour une comparaison homogène de leur diversité.

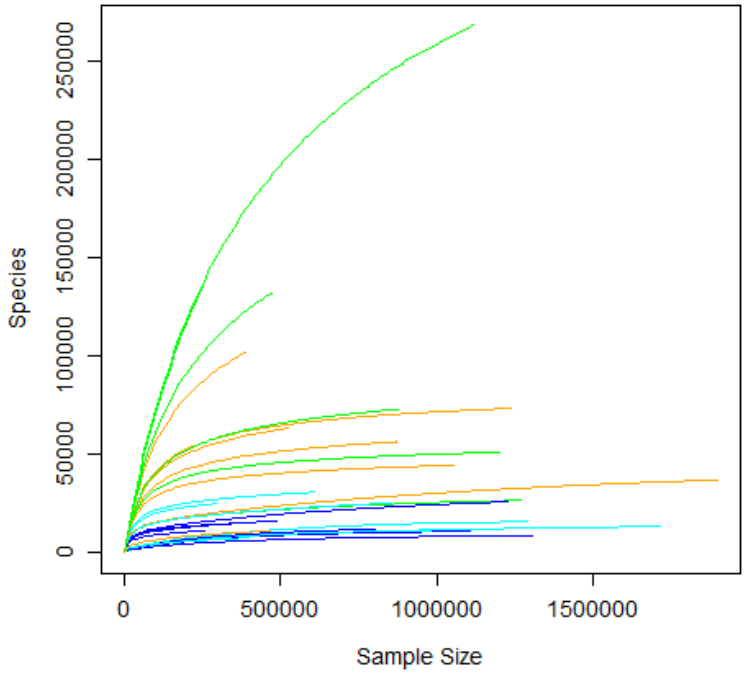


Figure 28 : Courbes de raréfaction des 28 échantillons TriPoD_06 – Chaque courbe est construite par une série de sous-échantillonnages (avec remise) de taille croissante (pas de 10 000 séquences TR) d'un même jeu de données**.** À chaque itération, le nombre de clonotypes observés est reporté sur la courbe. Les courbes sont colorées en fonction de la population cellulaire correspondante : Teff en vert, CD8 en orange, nTregs en cyan et amTregs en bleu.

Les courbes de raréfaction présentées en **Figure 28** se caractérisent par deux voire trois phases : une première phase de croissance exponentielle plus ou moins forte et/ou étendue, indiquant l'apparition de nouvelles espèces à chaque nouvelle séquence ; une phase de ralentissement de l'enrichissement en nouvelles espèces ; et parfois un plateau évoquant une saturation.

On distingue ici trois groupes de courbes : les courbes caractérisées par une faible croissance et un aplanissement rapide, suggérant une population peu diverse dont la totalité des clonotypes en présence est observée, en l'occurrence ici les répertoires Tregs ; les courbes à croissance très forte et absence de plateau suggérant des répertoires très divers et peut-être sous-échantillonnés lors du séquençage ; enfin les courbes intermédiaires augmentant plus ou moins rapidement lors de leur phase exponentielle et atteignant un plateau.

Chaque échantillon étant de nature biologique différente, il est difficile de juger de la « normalité » de ces profils. Cependant, ces résultats sont cohérents avec la nature des populations étudiées.

2) Topologie TRBVBJ des répertoires

À l'instar des analyses de cytométrie en flux et de spectratype du CDR3, l'analyse de la distribution d'usage et d'expression des gènes TRBV et TRBJ est très informative sur les différences de structure des répertoires comparés. Ainsi, une réponse immunitaire, en fonction de l'antigène (plus précisément du peptide antigénique) qui l'a induite, peut entraîner éventuellement la surreprésentation d'une ou plusieurs combinaisons BVBJ ou une famille de BV particulière par exemple. Ainsi, l'**indice de similarité MH sur les fréquences d'usages BVBJ** est utilisé comme indicateur de la structure des populations et permet leur **classification hiérarchique** sur ce critère.

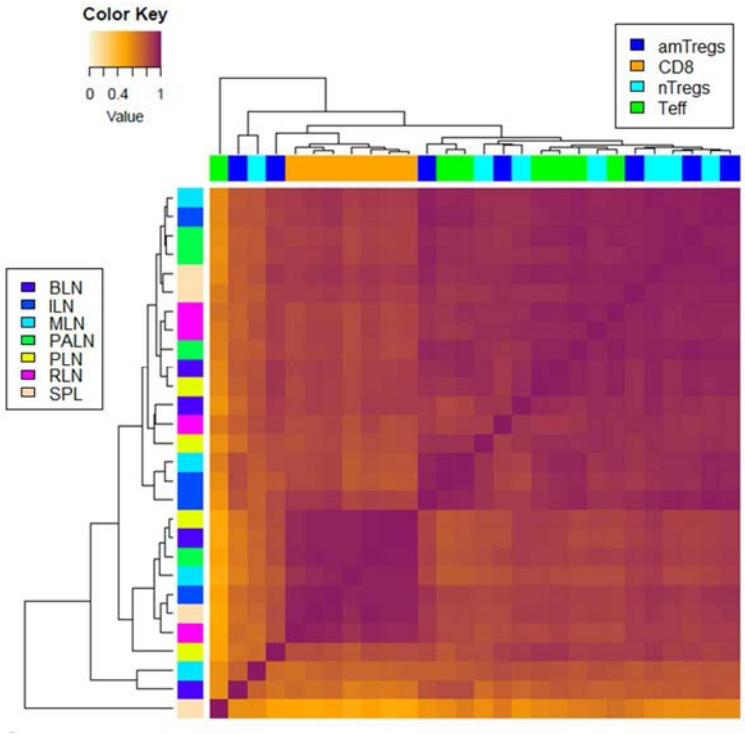


Figure 29 : Classification hiérarchique des 28 répertoires TRB en fonction de leur similarité en termes de distribution TRBVBJ – L'indice de similarité MH est calculé entre tous les échantillons 2 à 2. Chaque cellule de cette matrice indique la valeur de similarité entre une paire d'échantillons. Les échantillons sont identifiés par une couleur différente en fonction de leur organe (en ligne) et leur population (en colonne). Dendrogramme : distance euclidienne et méthode « *complete* ».

Lorsque l'on évalue la similarité entre le répertoire TRB de nos 28 échantillons sur la base de leur expression des combinaisons BVBJ (**Figure 29**), on observe deux clusters prédominants séparant le répertoire des échantillons CD8+ de celui des trois populations CD4+. Cette observation est identique lorsque l'on prend en compte la diversité d'usage des gènes TRBV. Cette observation est en accord avec l'hypothèse selon laquelle la diversité des répertoires TRB des LT CD8+ et CD4+ est forgée par leur interaction avec les molécules du CMH (Pannetier

et al., 1993). En effet, l'interaction avec le CMH lors de la reconnaissance antigénique est, entre autres, conditionnée par les régions CDR1 et CDR2 de la chaîne β du TCR, inclus dans la région codée par le gène TRBV. Ainsi, la différence de composition TRBV des répertoires de ces populations est liée au fait que les LT CD4+ et CD8+ interagissent avec des classes de CMH différentes.

Une autre approche consiste à calculer l'**indice de Shannon exp(H)** calculé au sein de chaque combinaison **TRBVBJ** et **de projeter les valeurs** des échantillons par analyse en composantes principales (**Figure 30**).

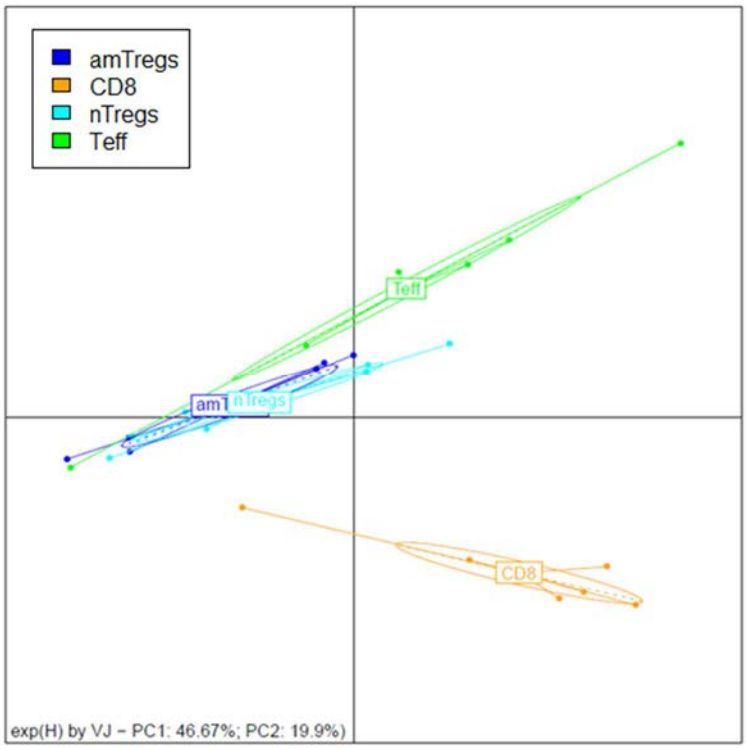
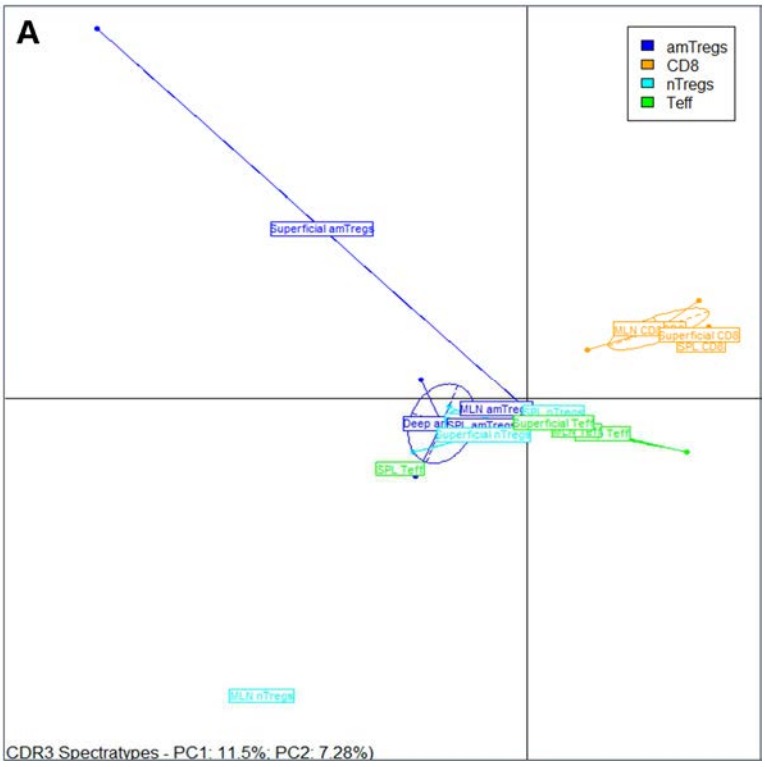


Figure 30 : Projection ACP des 28 échantillons en fonction de leur population cellulaire sur la base de leur diversité clonotypique au sein de chaque combinaison TRBVBJ – Pour chaque répertoire, l'exponentielle de l'indice de Shannon exp(H) est calculée en prenant en compte les fréquences relatives des clonotypes catégorisés en fonction de leur expression des gènes TRBV et BJ, puis pondérée par la fréquence de chaque combinaison. Les échantillons sont projetés en fonction des deux premières composantes obtenues en analyse par composante principale (ACP) qui, combinées, expliquent 66% de la variabilité globale.

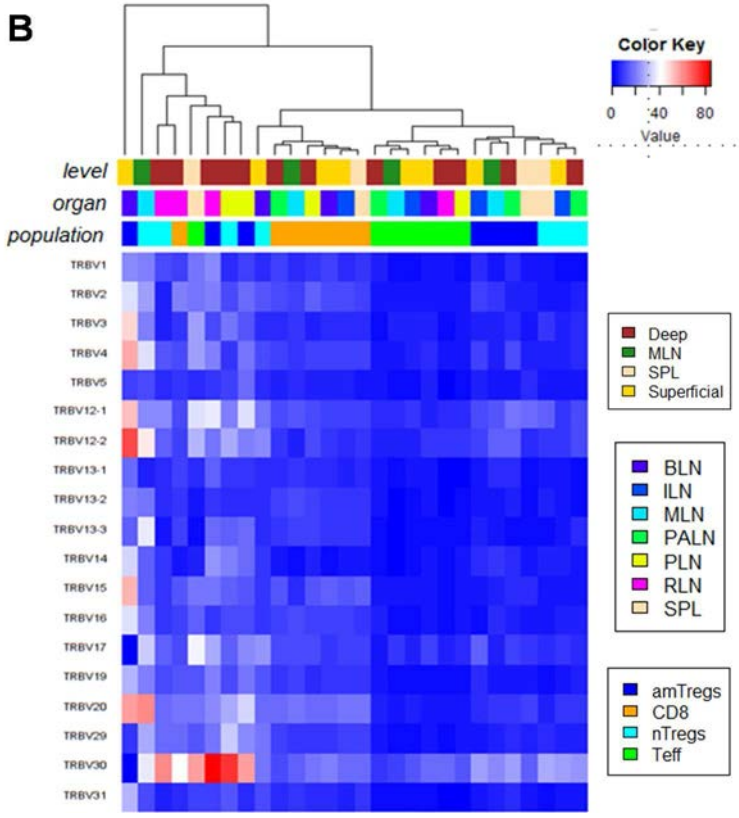
Cette analyse confirme que le répertoire CD8 est distant des trois populations LT CD4+, cette séparation se fait le long de la seconde composante (PC2). De plus, il apparaît que la première composante (PC1) tend à séparer les deux populations Tregs de la population Teff de part et autre de l'axe PC2, suggérant ainsi leur différence de diversité.

Comme évoqué plus tôt, la force des données d'immunoséquençage réside dans la possibilité de leur catégorisation de manières très différentes. Ainsi, on peut reconstruire **les profils spectratypiques** de chacun des jeux de données en catégorisant les CDR3 des clonotypes exprimant le même gène TRBV en fonction de leur longueur et calculant les fréquences de ces longueurs au sein de chaque famille TRBV. Il devient alors possible d'appliquer les approches d'analyses développées pour comparer ces distributions telles que le **score de perturbation ISEApeaks** (Bergot et al., 2015; Collette and Six, 2002; Collette et al., 2003; Mariotti-Ferrandiz et al., 2016; Petrovc Berglund et al., 2008). Ce score permet d'évaluer la distance de chaque échantillon par rapport à un échantillon de référence sur la base de la distribution des longueurs de CDR3 au sein de chaque famille TRBV. Si l'on considère par exemple la population Teff comme population de référence, un profil moyen à travers tous les échantillons correspondant est simulé pour chaque famille TRBV ; chacun des échantillons analysés est alors comparé à cette référence et un score de perturbation est calculé famille par famille. Cette technique permet d'évaluer le niveau de proximité entre les échantillons sur la simple distribution de taille des CDR3. Une analyse en composante principale (ACP) appliquée sur ces valeurs (**Figure 31A**) permet de visualiser la ségrégation des populations et confirme ici l'isolement des LT CD8+ par rapport aux CD4+ observé avec la diversité clonotypique. Cette approche permet, dans le cas du jeu de données utilisé, de mettre en évidence la variabilité des différentes populations en fonction de leur origine tissulaire. Par exemple, on observe que le répertoire des nTregs est différent dans les ganglions mésentériques par rapport aux autres organes observés et que le répertoire Teff est homogène dans les ganglions mais plus distant dans la rate. Par ailleurs, la **classification hiérarchique** obtenue sur la base des scores de perturbations des TRBV (**Figure 31B**) permet d'évaluer la distance entre les différentes populations. On note ici notamment que les profils de longueur de CDR3 des échantillons Teff sont plus proches de ceux des CD8 que des deux populations Tregs évoquant un niveau de polyclonalité/diversité plus élevé que celui des répertoires Tregs.

Figure 31 : Analyse de la diversité des spectratypes de CDR3 par TRBV.



A) Projection ACP des 28 échantillons sur la base de leurs profils spectratypiques. Les échantillons sont colorés en fonction de leur population cellulaire (vert : Teff, orange : CD8, cyan : nTregs et bleu : amTregs) et regroupés en fonction du niveau anatomique de leur organe d'origine : rate (*SPL*), ganglions profonds (*Deep*), superficiels (*Superficial*) et mésentériques (*MLN*).



B) Classification hiérarchique des 28 répertoires TRB en fonction de leurs scores de perturbation. Les échantillons (en colonne) sont identifiés en fonction de leur population, leur organe et leur niveau anatomique comme indiqué par la légende. Chaque cellule indique la valeur de leur score de perturbation par TRBV (en ligne). Dendrogramme : distance euclidienne et méthode « *complete* ».

3) Composition clonotypique

Comme décrit plus haut, les répertoires TRB analysés ici montrent des différences de distributions de leurs clonotypes qui semblent corréler avec le type de population cellulaire. On peut caractériser ces différences en s'intéressant aux **profils de diversité des répertoires en fonction des populations**. La **Figure 32** montre la comparaison de ces diversités entre populations et entre organes. Ainsi, à titre d'exemple, on observe que dans les ganglions superficiels et profonds, les répertoires CD8 et Teff sont plus divers que ceux des Tregs. Toutefois, ce différentiel de diversité entre populations varie entre les organes. En effet, dans les ganglions profonds la diversité des répertoires Tregs s'effondre dramatiquement (près de 3 logs) entre α=0 et 1 puis continue de diminuer doucement alors que celle des Teff diminue graduellement au fur et à mesure que α augmente. Cette chute de valeur de diversité lorsque α augmente s'explique par la présence de clonotypes prédominants. En effet, plus α est grand plus l'entropie de Rényi va donner du poids aux clonotypes les plus fortement représentés. Le passage entre 0 et 1 est clé car il correspond au retrait des clonotypes considérés comme « rares » par rapport aux autres clonotypes en présence. Un répertoire polyclonal étant de composé de clonotypes plus ou moins équivalemment représentés, ce seuil n'aura qu'un effet limité. Plus la diminution est forte, plus la prédominance d'un nombre limité de clonotypes est forte.

Les profils de diversité des deux répertoires Tregs PLN sont particulièrement proches l'un de l'autre tout comme ceux des répertoires CD8 et Teff, dont les profils sont parfaitement parallèles. Les différences de profils sont plus graduelles entre les répertoires ILN où les CD8 et nTregs occupent une position intermédiaire entre les deux autres. Dans les ganglions superficiels, on conserve la diminution graduelle de la diversité des répertoires Teff et CD8 mais les profils Treg sont différents entre les deux ganglions. Bien que l'effet du passage de α entre 0 et 1 soit plus marqué, particulièrement chez les amTregs, les deux populations Tregs ILN voient leur valeur de α. Dans les ganglions brachiaux, on peut voir que la diversité du répertoire amTregs s'effondre de manière similaire à ce qui est observé dans les ganglions profonds. Toutefois, au vu des résultats de l'analyse spectratypique de cette population (**Figure 31A**) et en l'absence de répétitions biologiques, ces observations, comme toutes celles présentées ici, ne seront concluantes que lorsqu'elles seront répétées. Néanmoins, cette

analyse permet de comparer et de caractériser de manière standardisée la diversité clonotypique des répertoires analysés.

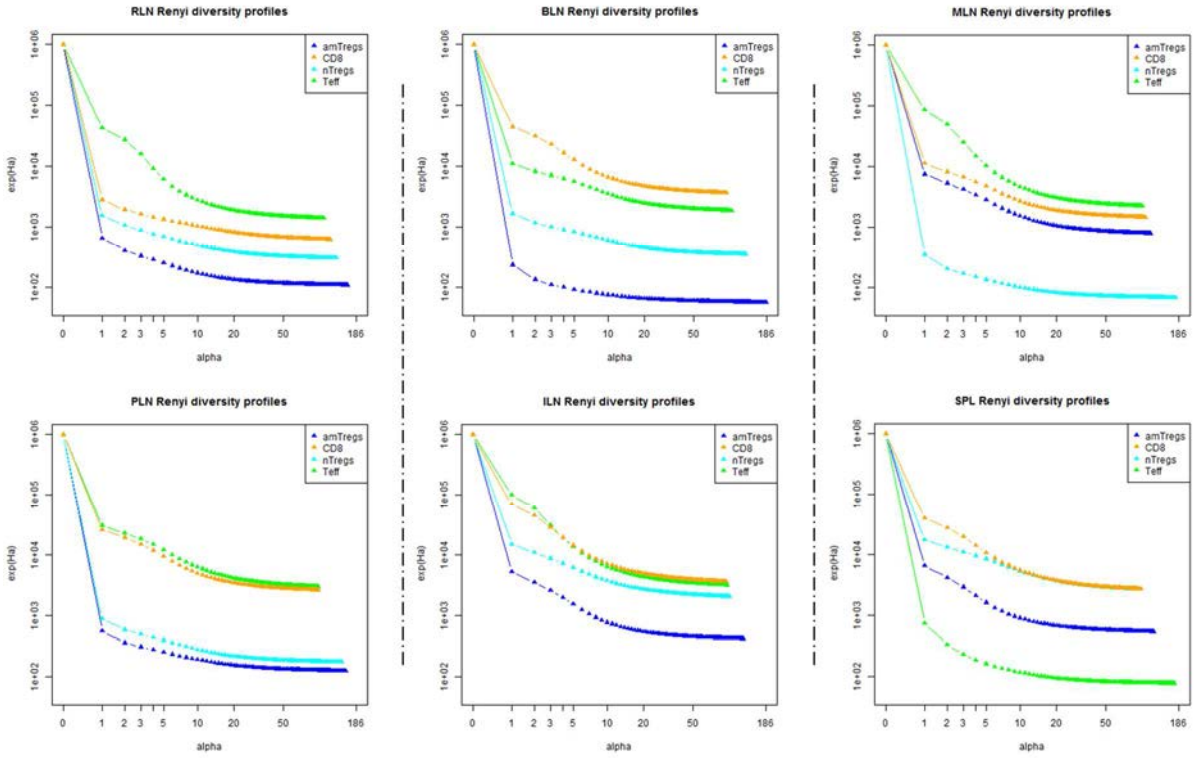


Figure 32 : Profils de diversité du répertoire TRB des quatre populations étudiées au sein des ganglions profonds (RLN et PLN), des ganglions superficiels (BLN, ILN), du MLN et de la rate (SPL) – L'exponentielle de l'entropie de Rényi est représentée pour des valeurs croissantes de α, à partir de la distribution des clonotypes au sein de chaque répertoire. Pour chaque organe, les profils des quatre populations cellulaires sont colorés de la façon suivante : vert : Teff, orange : CD8, cyan : nTregs et bleu : amTregs.

L'ordonnancement des profils des quatre répertoires de ganglions mésentériques est différent des autres ganglions du fait d'une baisse rapide de la diversité nTregs suggérant donc la présence de clonotypes majoritaires dans ce compartiment cellulaire. Dans la rate, c'est le répertoire Teff qui semble biaisé en termes de diversité. Tout comme pour les BLN-amTreg, ce résultat suggère un problème de séquençage, d'autant plus que le répertoire des populations lymphocytaires T est particulièrement polyclonal dans la rate en condition physiologique (Casrouge et al., 2000).

Une approche clé, pour la compréhension des niveaux hiérarchiques entre les populations par exemple, est d'analyser la similarité clonotypique. Pour cela, en partant de la liste des clonotypes (TRBV-CDR3-TRBJ) identifiés post-annotation, il est possible de calculer une matrice de similarité MH afin de déterminer la proximité en composition clonotypique entre les échantillons analysés.

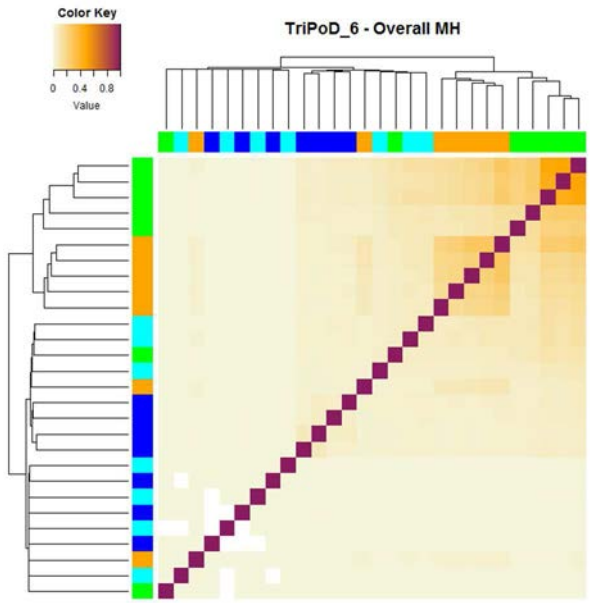


Figure 33 : Similarité clonotypique entre les 28 répertoires TRB – Une matrice de similarité MH a été construite entre les 28 répertoires et utilisée pour regrouper les échantillons les plus semblables par classification hiérarchique (distance euclidienne et méthode « *complete* »). Les échantillons sont identifiés en fonction de leur population cellulaire : vert pour Teff, orange pour CD8, bleu pour amTregs et cyan pour nTregs.

Appliqué aux jeux de données TriPoD_06 (**Figure 33**), on observe une similarité des répertoires entre mêmes populations pour la majorité des organes, particulièrement pour les Teff et les CD8. Le cluster en haut à droite regroupe 5 des 7 répertoires Teff analysés.

Par la suite, on peut évaluer le degré de chevauchement entre ces 5 échantillons pour évaluer l'origine de cette similarité. Ici, par exemple on observe que 79 352 clonotypes parmi les 547 142 identifiés sont présents dans au moins deux échantillons sur les cinq dont 2 154 clonotypes dans les cinq (**Figure 34A**). Dans la continuité de cette analyse, on peut alors déterminer la proportion des répertoires respectifs qu'occupent ces clonotypes en fonction de l'organe par le calcul et la représentation de la fréquence cumulée des clonotypes partagés (**Figure 34B**).

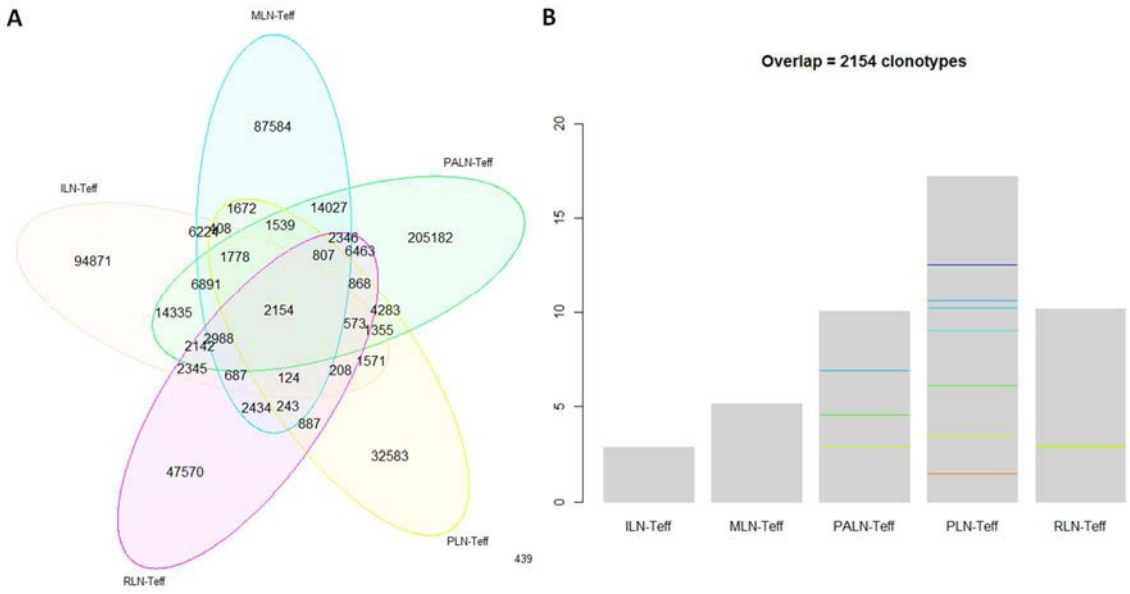


Figure 34 : Comparaison de la composition des répertoires Teff – **A)** Diagramme de Venn entre les clonotypes identifiés au sein des répertoires TRB Teff des ganglions MLN, ILN, PALN, PLN et RLN. **B)** Fréquences cumulées des 2154 clonotypes partagés par les cinq répertoires au sein de chacun des répertoires Teff analysé. Chaque clonotype est identifié par une couleur qui reste la même entre les cinq histogrammes. Toutefois, les zones grises indiquent la présence de nombreux clonotypes très faiblement représentés.

Il a été démontré au laboratoire que les LT régulatrices activées mémoires (amTregs) ont le rôle d'empêcher le développement de maladies auto-immunes, induites par une reconnaissance d'antigènes du soi par les LT effectrices (Chen et al., 2013; Darrasse-Jèze et al., 2005; Fisson et al., 2003). Ainsi, dans le cadre d'une étude du répertoire préliminaire menée au préalable (2015 ; Annexe 6), nous avons appliqué cette même stratégie et observé un chevauchement conséquent entre le répertoire des amTregs et Teffs dans les ganglions pancréatiques. La même comparaison appliquée aux jeux de données TriPoD_06 montre que 3 873 clonotypes (dans un univers de 113 368 clonotypes PLN) sont partagés par au moins deux des quatre populations PLN (**Figure 35A**). Parmi ces clonotypes, 719 (environ 18%) sont partagés par les répertoires amTregs et Teff dont 599 de manière exclusive. Ces clonotypes sont surexprimés dans les répertoires Tregs par rapport aux Teff. En parallèle, seuls 49 clonotypes (environ 1%) sont retrouvés dans les deux sous-populations Tregs ; ces derniers occupent environ 5% de ces deux répertoires (**Figure 35B**). Ces observations font écho à nos résultats précédemment publiés (Bergot et al., 2015 ; Annexe 6) obtenus sur un fonds génétique différent et surtout sur un jeu de données TRBV limité).

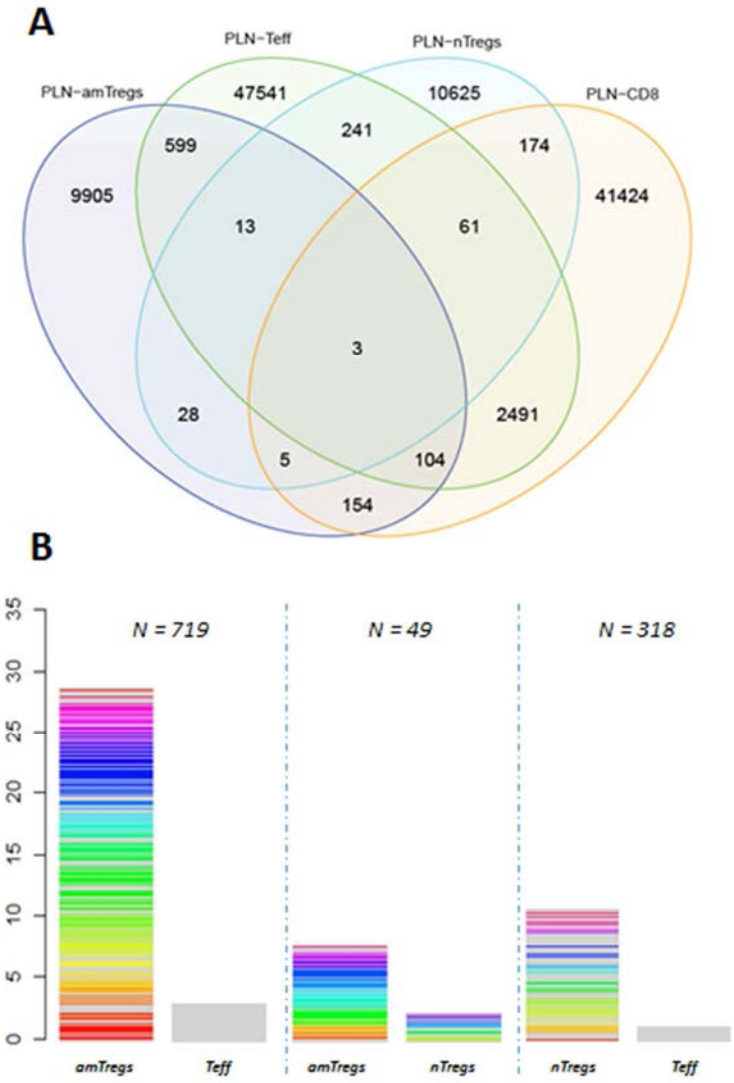


Figure 35 : Comparaison des clonotypes identifiés dans les quatre populations cellulaires des ganglions pancréatiques – **A)** Diagrammes de Venn entre les clonotypes identifiés au sein des répertoires des quatre populations triées des ganglions pancréatiques (PLN). **B)** Fréquences cumulées des clonotypes partagés par les répertoires amTregs et Teff (N=719), par les répertoires amTregs et nTregs (N=49) et par les répertoires nTregs et Teff (N=318).

C. Discussion

Les données d'immunoséquençage sont sources d'une grande quantité d'information qui permettent d'évaluer un certain nombre de paramètres caractérisant la diversité et la structure des répertoires TCR étudiés (expression des gènes TRB et TRJ, distribution des clonotypes…). Malgré le volume de données produites, ces expériences ne permettent pas (encore) de décrire de manière exhaustive les répertoires lymphocytaires étudiés, tout du moins chez les mammifères. En effet, le nombre de LT est de l'ordre de 10^8 cellules chez la souris (Casrouge et al., 2000) et 10^{12} chez l'homme (Arstila et al., 1999) ce qui est bien

supérieur aux capacités actuelles de production des séquenceurs. Par ailleurs, le système immunitaire étant dynamique, les populations cellulaires se distribuent et évoluent à travers le temps et l'espace. Ainsi, évaluer à l'échelle globale la diversité du répertoire d'une population lymphocytaire implique de l'analyser de manière simultanée à travers les différents organes lymphoïdes à l'instar de la stratégie expérimentale du projet TriPoD. Or, cette approche est d'une part impossible à mettre en place chez l'homme et d'autre part extrêmement coûteuse en termes de temps et d'argent. Ainsi, la grande majorité des études menées sur le sujet, à l'exception de celles concernant les petits organismes tels que le poisson zèbre (Weinstein et al., 2009), ne fournissent qu'une mesure instantanée de la diversité des répertoires lymphocytaires étudiés. Il est donc crucial de s'assurer de la robustesse des données et de les exploiter de la manière la plus complète possible.

Le plan expérimental du projet *TriPoD* a été conçu de manière à ce que de nombreuses questions immunologiques puissent être investiguées : relation entre les quatre populations cellulaires étudiées, distribution de ces populations et façonnage de leur répertoire TCR à travers les organes… De nombreuses comparaisons entre les échantillons peuvent donc être définies de manière à identifier et caractériser d'éventuelles différences immunologiques.

Dans ce contexte, la méthodologie que je propose ici permet d'adapter les comparaisons aux questions biologiques et fournit une description approfondie des répertoires étudiés selon différents axes complémentaires : l'expression des gènes TRBV et TRBJ, la structure et la diversité globale du répertoire, la composition clonotypique globale, l'identification des clonotypes majoritaires…

Cette méthodologie est appliquée de manière systématique lors de la production d'un nouveau jeu de données, permettant ainsi de mettre en place une stratégie d'analyse comparative adéquate des échantillons. À titre d'exemple, dans le cadre de l'expérience *TriPoD_06* présentée ici, plus de 500 analyses ont été produites : 476 décrivent individuellement chacun des échantillons et 35 la comparaison globale intégrant tous les échantillons de cette expérience.

À l'instar de l'analyse du transcriptome (Chaussabel et al., 2008; Nehar-Belaid et al., 2016; Pham et al., 2014), l'utilisation de signatures de TCR antigène-spécifiques comme biomarqueurs liés à des contextes immunologiques donnés (tels que le statut physiopathologique, le phénotype cellulaire…) pourrait être un outil très puissant pour étudier la dynamique de différentiation de populations cellulaires ou pour pronostiquer le

développement d'une pathologie par exemple. Dans cet ordre d'idée, Mariotti-Ferrandiz et al. ont identifié, sur la base de données de spectratypage, une signature TCRβ susceptible de prédire le développement d'un neuropaludisme par des souris infectées par *P. berghei ANKA* (Mariotti-Ferrandiz et al., 2016). Ainsi, on peut imaginer identifier, à partir de données d'immunoséquençage, les récepteurs exprimés par des cellules ayant proliféré au sein d'une population LT donnée et déterminer la spécificité de ces récepteurs à un antigène. Ce genre d'approche nécessite une quantité considérable de données afin d'assurer la significativité statistique des signatures. Toutefois, les données de *TriPoD_6* sont rassurantes quant à la faisabilité de cette approche. En effet, pour qu'il puisse être considéré comme un biomarqueur potentiel, un facteur, de quelque nature qu'il soit, doit être observable de manière récurrente à travers les échantillons dont il est censé caractériser le contexte immunologique. Ainsi, on peut émettre l'hypothèse qu'un clonotype observé au sein du répertoire d'un échantillon sera conservé dans un autre échantillon de même nature/fonction si sa présence est intrinsèquement liée à cette nature/fonction. Or, la grande majorité des clonotypes observés au sein du répertoire d'un échantillon donné lui sont propres ; on parle de clonotypes « privés ».

Comme observé **Figure 33** et **Figure 34**, dans *TriPoD_6*, le degré de similitude entre les répertoires d'une même population cellulaire au travers des sept organes analysés est variable lorsque l'on prend en considération l'intégralité de ces répertoires. Toutefois, il est possible de se concentrer sur ces clonotypes les plus prédominants pour observer cette similitude, considérant que la probabilité d'observer de manière récurrente les clonotypes prédominants sera plus grande que celle d'observer les clonotypes rares. Ainsi, les clonotypes de chaque répertoire sont ordonnés de manière décroissante en fonction de leur fréquence d'observation et les X premiers sont sélectionnées, X correspondant à 1% de la richesse totale. On observe sur la **Figure 36A** que, malgré un nombre plus restreint, les clonotypes prédominants au sein du répertoire amTregs des ganglions pancréatiques et rénaux (ganglions profonds) sont en fréquences cumulées (environ 50%) deux fois plus abondant que ceux des ganglions périphériques (à l'exception des ganglions brachiaux dont les 208 clonotypes majoritaires occupent 70% du répertoire) et de la rate. En parallèle, le nombre de clonotypes prédominants dans les répertoires Teff est élevé et variable mais leur fréquence cumulée (à part pour le répertoire SPL-Teff) avoisine les 18%.

Parmi les 3 424 clonotypes identifiés comme majoritaires chez les amTregs, 69% sont « privés » et les 31% restant sont observés dans au moins deux échantillons, mais aucun dans les sept simultanément. Deux fois plus de clonotypes sont observés à travers les sept répertoires Teff dont seuls 19% sont « privés ». La majorité des clonotypes Teff sont partagés par au moins deux échantillons et 45 sont communs aux sept répertoires (**Figure 36B**). Prises toutes ensembles, ces observations suggèrent une spécialisation des répertoires amTregs en fonction des organes (Föhse et al., 2011; Lathrop et al., 2008). Ainsi, le répertoire des Tregs dans les ganglions profonds serait forgé de manière à ce qu'ils protègent de manière spécifique les organes (Bergot et al., 2015 ; Annexe 6).

Une fois l'intégralité des données du projet produites, il devrait donc être possible d'identifier, par exemple, une ou des signature(s) de TCR spécifiquement observée(s) dans les amTregs des ganglions pancréatiques qui, par extension, pourraient être étudiées dans le cadre du développement du diabète de type I (Bergot et al., 2015 ; Annexe 6).

Outre la catégorisation des clonotypes en fonction de leur séquence (nucléotidique ou peptidique) unique, on peut également s'intéresser à la recherche de clonotypes partageant les mêmes propriétés physico-chimiques. En effet, la reconnaissance spécifique d'un complexe pCMH par un TCR est conditionnée par l'affinité et l'avidité de leur interaction et la conformation structurale du site d'interaction. Cette interaction chimique implique les régions CDRs des chaînes protéiques du TCR et notamment le CDR3 qui interagit avec le peptide. Ainsi, on peut émettre l'hypothèse que la composition en acides aminés de chacun des CDR3 va influencer la spécificité de la reconnaissance du peptide par le TCR (Benichou et al., 2013). Ainsi, certaines études, notamment menées sur les LB, ont observé une variation de la composition en acides aminés du CDR3 en fonction des populations cellulaires ou au sein de clonotypes associés à des pathologies chroniques (Liaskou et al., 2016; Wu et al., 2010). Notamment, Li et al. ont observé que les CDR3 de TCRβ de LT CD4+ isolés à partir de sang humain semblent enrichis en acides aminés chargés positivement (lysine et arginine) alors que les LT CD8+ T eux seraient enrichis en acides aminés chargés négativement (asparagine) (Li et al., 2016). Toutefois, ces analyses restent préliminaires quant aux paramètres pris en compte pour caractériser les acides aminés. En effet, chaque acide aminé peut-être décrit par 544 paramètres différents caractérisant leurs propriétés physico-chimiques et biochimiques (Kawashima et al., 1999). Toutefois, utiliser l'ensemble de ces paramètres s'avère laborieux car ceux-ci peuvent être redondants et ne sont pas forcément tous très informatifs. Aussi,

certains facteurs ont été identifiés comme représentatifs, lorsque combinés, des caractéristiques physico-chimiques globales de chaque acide aminé : les facteurs Kidera, composés de 10 paramètres indépendants (Kidera et al., 1985), et Atchley, liste de 5 paramètres indépendants (Atchley et al., 2005). Alors qu'Epstein a démontré la faisabilité de ce type de stratification en utilisant les facteurs Kidera (2014), Thomas et al. ont confirmé son potentiel en discriminant, avec une efficacité de 100%, le répertoire de souris immunisées ou non avec *Mycobacterium tuberculosis* après avoir catégorisé leur données de répertoires LT CD4+ en termes de courts motifs, définis par les facteurs Atchley (Thomas et al., 2014). En s'éloignant ainsi de la définition classique du clonotypes en tant qu'espèce, il est possible de voir émerger une structure sous-jacente des répertoires lymphocytaires qui peut s'avérer extrêmement informative sur la diversité de reconnaissance antigénique (diversité fonctionnelle) des répertoires étudiés.

Toutefois, comme le souligne Laydon (2015), le recours à des estimateurs de la diversité pour décrire le répertoire TCR a ses limites notamment lié à la proportion de clonotypes « non-observés ». En effet, la non-exhaustivité du séquençage abordée plus haut sous-entend que dans le cas d'un séquençage ne permettant pas de couvrir la totalité de la diversité, ce n'est pas parce qu'un clonotype n'est pas observé au sein d'un répertoire qu'il n'existe pas. Ainsi, Mora recommande d'adapter le choix de l'indice de diversité α à appliquer en fonction du spectre d'abondance des clonotypes constituant le répertoire à analyser : plus un répertoire est sous-échantillonné, plus l'ordre α de l'entropie de Rényi doit être grand (Mora and Walczak, 2016).

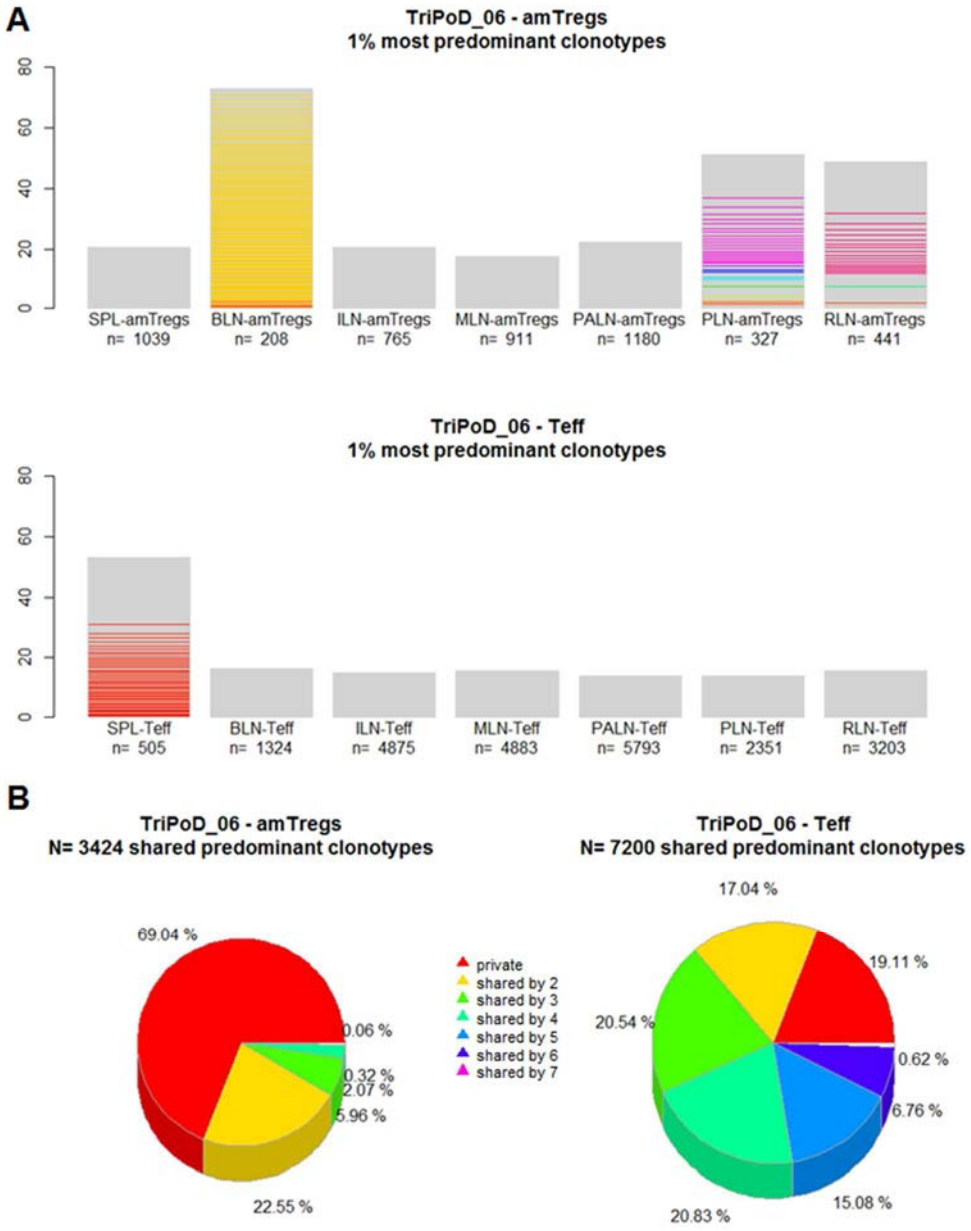


Figure 36 : Clonotypes TRB prédominants dans les répertoires des amTregs et des Teff – A) Fréquences cumulées des 1% des clonotypes les plus fortement exprimés au sein des sept répertoires amTregs (en haut) et Teff (en bas) analysés. Le nombre de clonotypes sélectionnés est indiqué au bas de chaque histogramme. Une couleur est attribuée aléatoirement à chaque clonotype au sein de chaque histogramme. Les zones grises indiquent la présence de nombreux clonotypes très faiblement représentés. **B)** Diagrammes circulaires représentant la proportion de clonotypes partagés au sein des clonotypes prédominants identifiés au sein des répertoires amTregs (à gauche) et Teff (à droite). La légende permet de distinguer les clonotypes privés (en rouge) des clonotypes partagés par 2 à 7 échantillons. Le pourcentage indique la proportion de chacune de ces catégories au sein des N clonotypes sélectionnés.

REPRESENTATIVITE DE LA DIVERSITE OBSERVEE

Depuis l'émergence de l'immunoséquençage, l'engouement pour l'analyse du répertoire TR n'a eu de cesse d'augmenter. Outre une caractérisation plus approfondie de l'objet, cette technique permet de comparer à grande échelle la composition clonotypique de différentes populations cellulaires et éventuellement de caractériser leurs mécanismes de reconnaissance antigénique. La quantité d'information et le niveau de précision fournis par cette technique sont sans précédent. Toutefois, la représentativité des données obtenues lors d'une expérience reste à déterminer.

En effet, les données produites par une expérience d'immunoséquençage sont conditionnées par différents niveaux d'échantillonnages.

- Échantillonnage de l'organe pour isoler la population cellulaire
- Échantillonnage de la population cellulaire pour l'extraction d'ARN
- Échantillonnage de l'aliquote d'ARN pour la préparation de la librairie d'ADNc
- Échantillonnage de la librairie pour le séquençage

Aussi, afin d'évaluer la similarité entre les répertoires TR de plusieurs échantillons de natures différentes, il semble essentiel d'aborder une première question :

⇨ Quel est le seuil maximum de similarité que l'on peut attendre entre deux expériences d'immunoséquençage ?

Par ailleurs, si l'on considère que chaque population cellulaire diffère en termes de fonction et possiblement de taille, on peut supposer qu'une même profondeur de séquençage ne capturera pas de la même façon la diversité de son répertoire. Aussi, une seconde question s'impose :

⇨ Comment la profondeur de séquençage impacte-t-elle la représentativité de la diversité observée ?

A. Reproductibilité des observations de séquençage

Afin de répondre à ces questions, une expérience comparative a été mise en place au laboratoire. Deux séries de trois aliquotes d'ARN ont été obtenus à partir des échantillons ***TriPoD_38_1070_5*** et ***TriPoD_38_1070_6*** *(voir section* Données expérimentales*)*. Ces séries ont respectivement été utilisées pour produire des librairies selon deux protocoles différents : PCR multiplex (iRepertoire®) et un protocole dérivé de l'approche Immunoscope. Chaque

librairie a été séquencée en parallèle sur des séquenceurs Illumina HiSeq 2500 (***TriPoD_38_1070_5)*** et 454 GS Junior (***TriPoD_38_1070_6)***, afin de tester la reproductibilité des deux approches. La diversité et la topologie de répertoires observés ont ensuite été comparées suivant la méthodologie décrite dans la section précédente.

1) Description par série d'aliquotes

La variabilité des différentes métriques descriptives est faible au sein de chaque série. Malgré la différence de profondeur inhérente à la différence de plate-forme, les nombres de gènes BV et BJ observés sont identiques entre les deux séries (**Figure 37**).

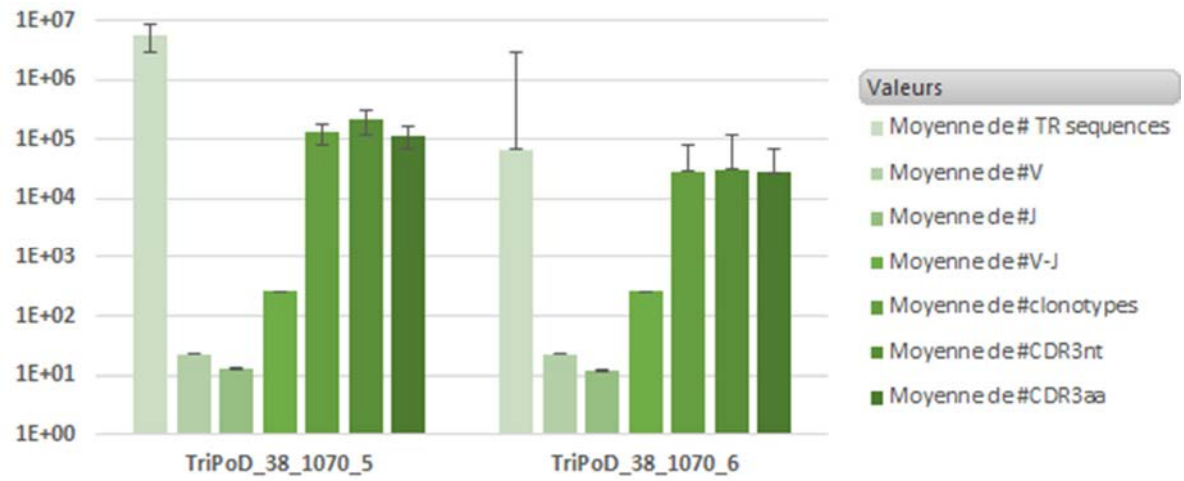


Figure 37 : Statistiques descriptives des jeux de données – Les nombres de séquences TR, clonotypes, gènes TRBV, gènes TRBJ et combinaisons TRBVBJ sont comptés au sein de chacun des six jeux de données. Les moyennes et écart-types de ces cinq paramètres sont calculés par série d'aliquotes.

En termes de composition, les résultats illustrés **Figure 38** indiquent que les séquençages de chaque série de trois d'aliquotes capturent la même diversité de répertoire TRB, les deux séries de courbes se superposant quasi parfaitement.

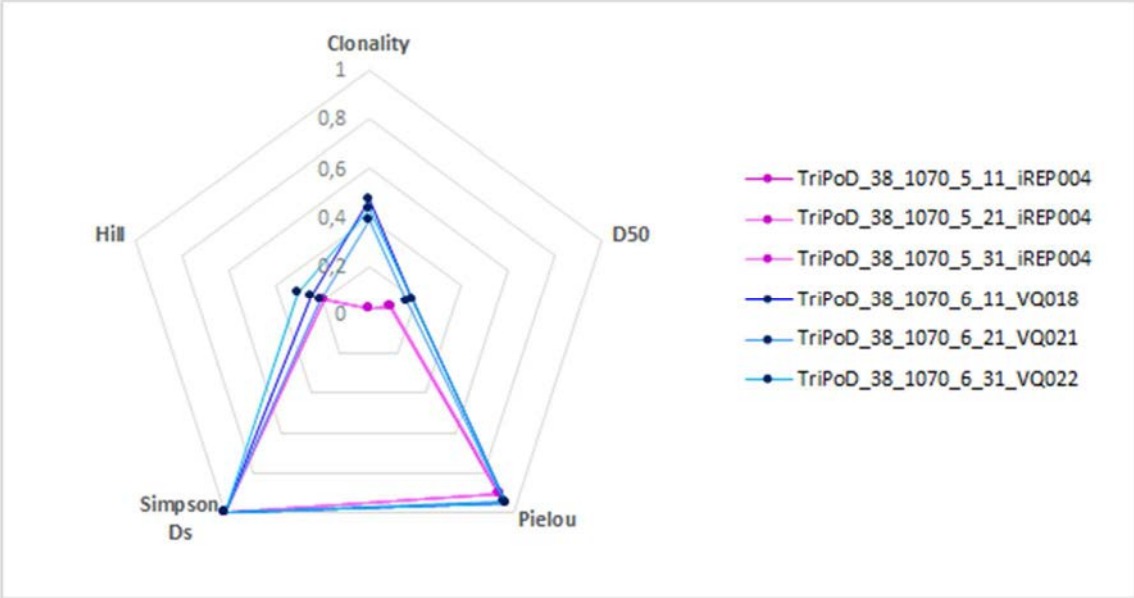


Figure 38 : Profil descriptif de la diversité globale du répertoire TRB des six aliquotes – Chaque axe indique la valeur d'une des cinq métriques descriptives calculées à partir de la distribution des clonotypes observés au sein de chaque jeu de données. Les courbes sont colorées en fonction de l'échantillon et de la plate-forme de séquençage d'origine comme indiqué par la légende.

On note toutefois que les valeurs de *Clonality*, de D50 et de l'indice de Hill sont plus élevées dans la série produite par iRepertoire/Illumina. Ces indices étant sensibles à la taille de l'échantillon, cette différence peut s'expliquer par la différence de profondeur de séquençage des deux séries qui, comme indiqué **Figure 37**, est presque de 2 log. Les profils de diversité des aliquotes, quant à eux, se confondent pour les deux séries (**Figure 39**), confirmant que les séquençages répétés capturent le même degré de diversité du répertoire TRB lorsque produit de manière similaire.

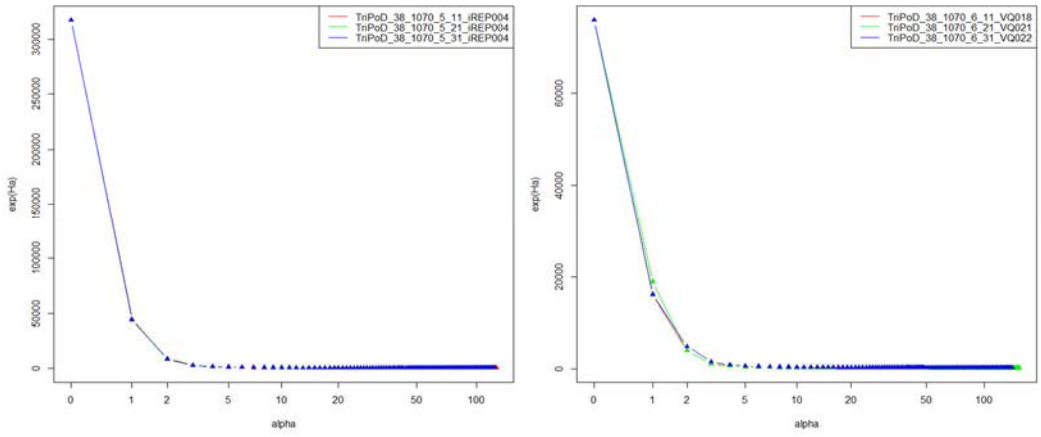


Figure 39 : Profils de diversité pour chaque série d'aliquotes – L'exponentielle de l'entropie de Rényi est représentée pour des valeurs croissantes de α, à partir de la distribution des clonotypes au sein de chaque jeu de données. À gauche, les profils des trois aliquotes TriPoD_38_1070_5, séquencés sur Illumina/HiSeq2500et à droite, les profils des trois aliquotes TriPoD_38_1070_6, séquencés sur Roche/454. Les profils sont colorés en fonction de l'aliquote comme indiqué par la légende.

Les **Figure 40A-C** résument la comparaison des résultats de séquençage des trois aliquotes ***TriPoD_38_1070_5*.** Ainsi, la distribution des clonotypes semble identique dans les trois jeux de données puisque les trois courbes de raréfactions se superposent parfaitement avec un enrichissement très fort et qui commence à se stabiliser à partir de 10^6 séquences mais n'atteint pas de plateau. 57 204 clonotypes sont observés dans au moins deux des trois jeux de données dont 17 002 sont communs aux trois séquençages. Ces derniers représentent entre 12,5 et 13,3% des clonotypes identifiés dans chacun des jeux de données. Alors que les distributions d'usage des gènes BV et BJ sont parfaitement identiques, la similarité en termes de clonotypes varie entre 0,792 et 0,837 entre les trois aliquotes dans les jeux d'origine et augmentent entre 0,895 à 0,918 au sein des 17 002 clonotypes partagés (avec une inversion de l'ordre de similarité entre les trois aliquotes).

Les **Figure 40D-F**, quant à elles, montrent les résultats de séquençage des trois aliquotes ***TriPoD_38_1070_6*.** Comme précédemment, les trois courbes de raréfaction se superposent. Leur pente semble plus faible que celles de la **Figure 40A**. Cependant la différence d'échelle rend difficile la comparaison. Toutefois, de manière cohérente avec les observations précédentes, les courbes sont en constante croissance puisque la taille de ces jeux de données est inférieure à 10^6 séquences. 8 076 clonotypes sont observés dans au moins deux des trois jeux de données dont 2 036 sont communs aux trois séquençages. Ces derniers représentent entre 6,2 et 7,8% des clonotypes identifiés dans chacun des jeux de données. Les valeurs de Morisita-Horn entre les trois aliquotes sont 0,896, 0,898 et 0,905**.**

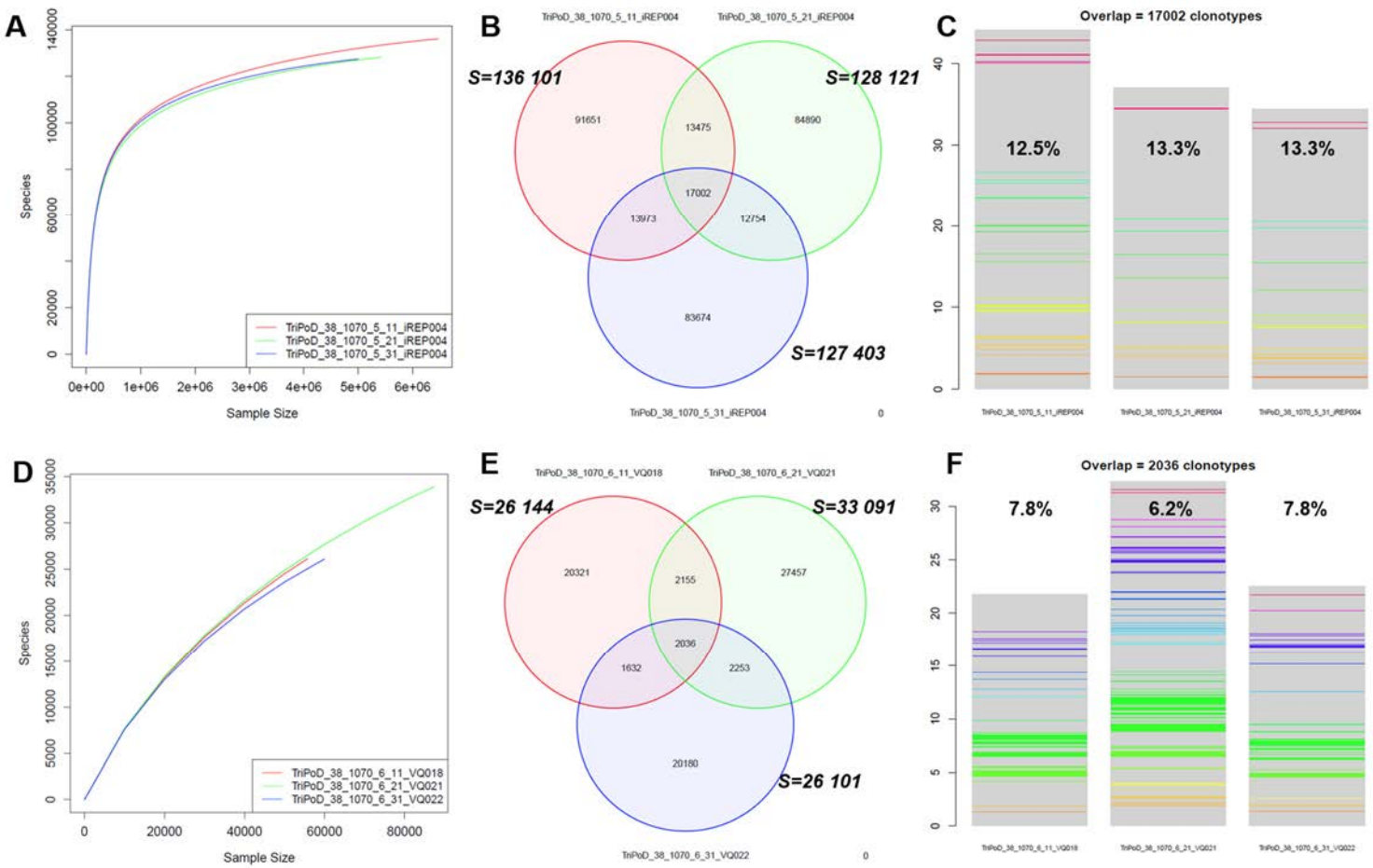


Figure 40 : Comparaison entre réplicas techniques – **(A-C)** Le répertoire TRB des trois aliquotes d'ARN de TriPoD_38_1070_5 a été séquencé en parallèle sur un séquenceur Illumina HiSeq 2500. **(D-F)** Le répertoire TRB des trois aliquotes d'ARN de TriPoD_38_1070_6 a été séquencé en parallèle sur un séquenceur 454 GS Junior. A et D : Courbes de raréfaction des trois jeux de données au sein de chaque série d'aliquotes. B et E : Diagrammes de Venn entre les clonotypes identifiés au sein du répertoire des aliquotes de chaque série. C et F : Fréquences cumulées des clonotypes communs aux trois aliquotes de chaque série au sein de chaque jeu de données. Chaque clonotype est identifié par une couleur qui reste la même dans les trois histogrammes de chaque figure. Les zones grises indiquent la présence de nombreux clonotypes très faiblement représentés**.** Les pourcentages indiqués renseignent sur la proportion que représentent les clonotypes partagés parmi les clonotypes identifiés au sein de chaque répertoire.

Les résultats de ces deux séries permettent de conclure que les séquençages répétés d'un même échantillon d'ARN capturent une image globale identique du répertoire observé que ce soit en termes de profils d'expression des gènes TRBV et TRBJ ou de la diversité globale du répertoire. Néanmoins, le chevauchement en termes de clonotypes (de l'ordre de 10%) entre les aliquotes de chaque série est relativement faible et met en évidence l'impact de l'échantillonnage biologique sur l'observation des clonotypes composant la population.

2) Comparaison des séries d'aliquotes

Sous-échantillonnage normalisé

L'utilisation de métriques objectives telles que les indices de diversité rend possible la comparaison des résultats obtenus indépendamment pour les deux séries malgré la différence de profondeurs notable et de méthode de production de librairies (**Figure 38** et **Figure 39**). Toutefois, une problématique clé pour la comparaison de la composition de jeux de données d'immunoséquençage est la différence de taille de ces jeux de données. Pour pallier ce problème, une solution simple de normalisation consiste à sous-échantillonner les jeux de données à une taille équivalente par tirage aléatoire d'un nombre fixe de séquences TR, mimant ainsi un séquençage de profondeur équivalente. Cependant, un tirage unique du nombre de séquences souhaité ne serait pas représentatif de la diversité du jeu d'origine. Aussi, pour normaliser ce sous-échantillonnage, 100 tirages aléatoires avec remise (*bootstrap*) sont réalisés sur chaque jeu de données Illumina. La taille moyenne des jeux de données 454 étant de 68 030 séquences TR, chaque itération a été bornée au tirage aléatoire de 68 000 séquences TR. Une table de contingence des espèces a été créée à chaque itération.

La probabilité d'observer un clonotype lors d'un tirage aléatoire va non seulement dépendre de son abondance au sein du répertoire initial mais également du degré de sous-échantillonnage. Ici, la diversité du répertoire initialement observée rend la probabilité d'observation d'un clonotype très faible (fréquences *f* des clonotypes observées telles que $f \in [1.10^{-7};\ 3.10^{-3}]$) et on cherche à diminuer la taille du jeu de données de 2 logs. Pour chaque jeu de données, le cumul des clonotypes est calculé en fonction de leur occurrence d'observation au travers des 100 itérations. Ainsi, comme indiqué sur la **Figure 41**, 77% des clonotypes constituant les jeux de données initiaux sont observés moins d'1 fois sur une 2 à travers les 100 tirages.

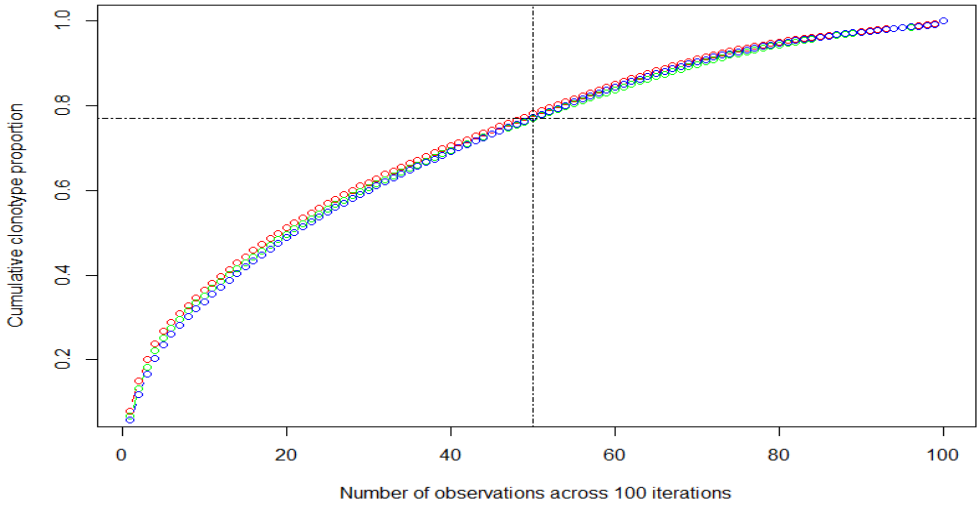


Figure 41 : Distribution d'observation des clonotypes à travers 100 itérations de *bootstrap* – Les clonotypes sont regroupés en fonction du nombre de fois où ils sont observés à travers les 100 tirages. La proportion cumulée du nombre de clonotypes observés par rapport à la richesse initiale est ensuite calculée en fonction de ces occurrences. Les droites verticale et horizontale permettent d'évaluer la proportion de clonotypes observés moins de 50 fois sur les 100 itérations.

Les 23% restants occupent pourtant près de 70% des répertoires initiaux en termes d'abondance. Ils peuvent donc être considérés comme les clonotypes les plus représentatifs et sont sélectionnés pour constituer un répertoire moyen dont l'abondance de chaque clonotype est déterminée par son abondance moyenne à travers les itérations où il est observé. Le nombre moyen de clonotypes est de 28 270 pour une taille d'échantillon moyenne de 63 048 séquences TR. La **Figure 42A** représente la variation de l'indice de Morisita-Horn entre les jeux de données initiaux et leurs sous-échantillonnages avec une valeur médiane variant entre 0,904 et 0,908.

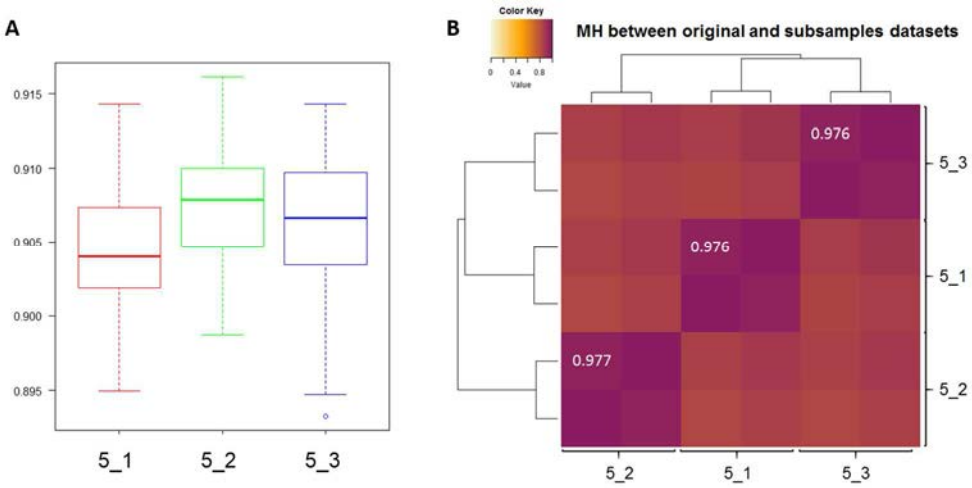


Figure 42 : Similarité à travers le processus de normalisation – **A)** Distribution des valeurs de similarité MH entre chacun des 100 tirages et le jeu initial. **B)** Classification hiérarchique (Distance euclidienne, méthode « *complete* ») des valeurs de MH entre les répertoires initiaux et normalisés. Les valeurs sont

colorées en fonction de l'échelle indiquée. En blanc sont indiquées les valeurs des comparaisons appariées.

Ainsi, le répertoire normalisé obtenu est un sous-échantillonnage représentatif du répertoire initial comme l'illustre la **Figure 42B**, les valeurs de Morisita-Horn entre les répertoires initiaux et normalisés étant de 0,976.

Comparaison

23 965 clonotypes ont été observés dans au moins un des aliquotes des deux échantillons ; ce chevauchement représente une proportion différente entre les deux séries d'échantillons. Après normalisation, l'intersection est réduite de 55% en termes de nombre de clonotypes mais la proportion de ces clonotypes au sein de chaque série est similaire à savoir 16% des clonotypes observés à travers les aliquotes TriPoD_38_1070_5 et 14% des clonotypes observés dans TriPoD_38_1070_6 (**Figure 43**).

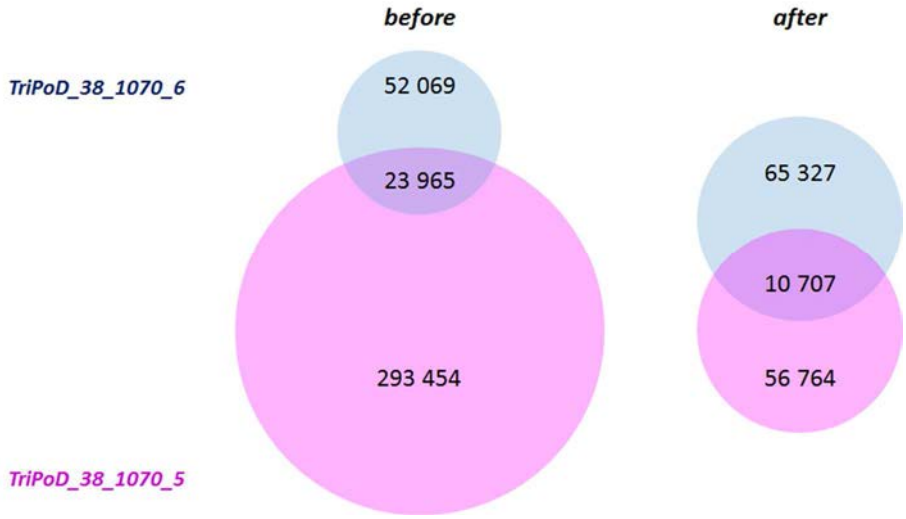


Figure 43 : Chevauchement entre les échantillons avant et après normalisation – Pour chaque échantillon, le cardinal de l'union des clonotypes observés à travers chaque série de trois aliquotes est calculé. Diagrammes de Venn entre les clonotypes identifiés dans les jeux de données TriPoD_38_1070_5 (rose) et TriPoD_38_1070_6 (bleu) avant (à gauche) et après (à droite) normalisation des répertoires.

Parmi les 10 707 clonotypes identifiés dans les deux échantillons, 1099 sont observés dans les six aliquotes et occupent environ 45% de chacun des répertoires en termes d'abondance (**Figure 44**). Cependant, comme on peut l'observer sur la **Figure 44**, les fréquences de ces clonotypes sont dans l'ensemble différentes entre les deux séries d'aliquotes.

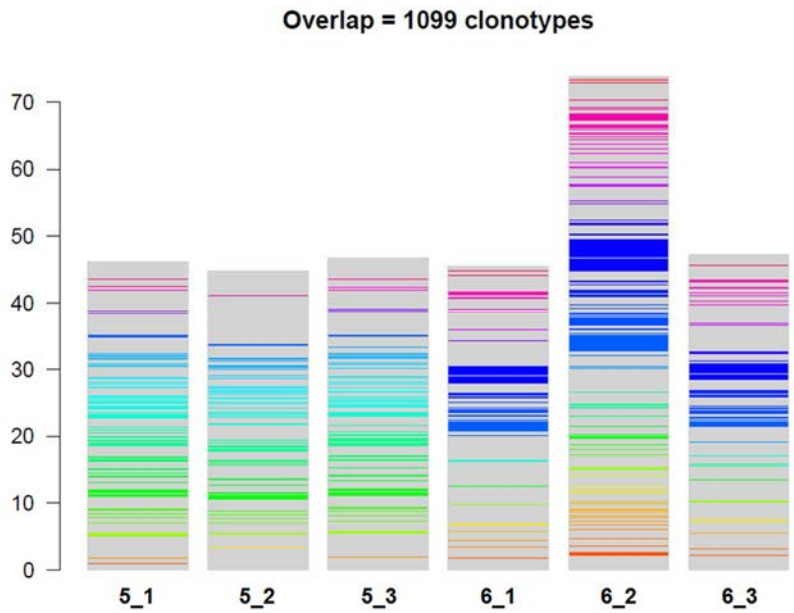


Figure 44 : Fréquences cumulées des clonotypes détectés dans les six aliquotes – Fréquences cumulées des 1099 clonotypes communs aux six aliquotes au sein de chaque jeu de données. Chaque clonotype est identifié par une couleur, elle reste la même entre les six histogrammes. Les zones grises indiquent la présence de nombreux clonotypes très faiblement représentés.

Toutefois, il est à noter que parmi ces 1099 clonotypes se trouvent 126 des 130 clonotypes prédominants de chaque jeu de données (obtenue par tri indépendant des clonotypes par ordre décroissant de fréquences et sélection des premiers 1% majoritaires).

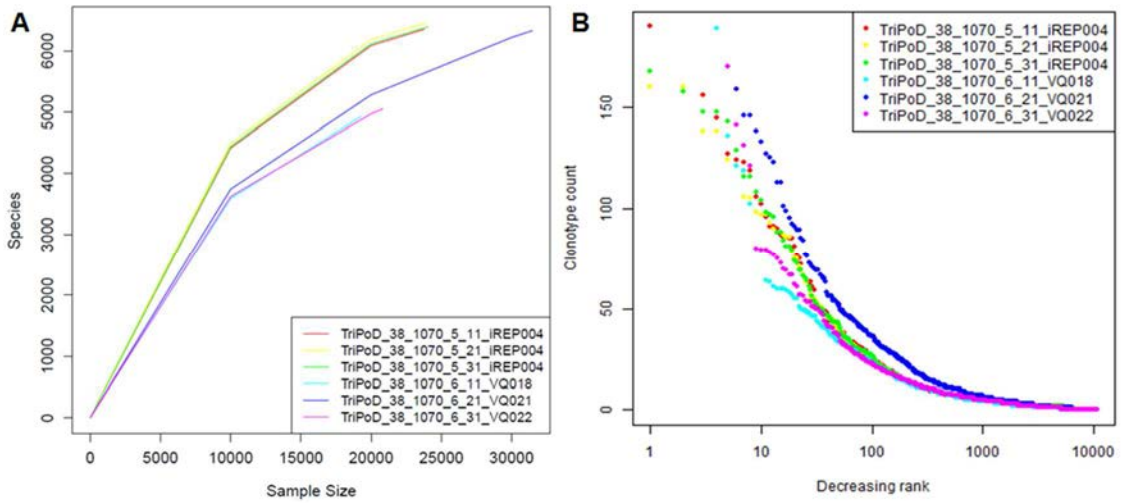


Figure 45 : Comparaison des richesses et distributions clonotypiques – **A)** Courbes de raréfaction des jeux de données. Chaque courbe est construite par une série de sous-échantillonnages (avec remise) de taille croissante (pas de 10 000 séquences TR) d'un même jeu de données. À chaque itération, le nombre de clonotypes observé est reporté sur la courbe. **B)** Distributions des clonotypes au sein du répertoire de chaque aliquote. Les clonotypes de chaque répertoire TRB sont ordonnés de manière décroissante en fonction de leur fréquence d'observation. Les courbes sont colorées en fonction de l'aliquote comme indiqué par les légendes.

Toutefois, bien que les courbes de distribution et de raréfaction montrent des allures similaires entre les deux séries d'aliquotes (**Figure 45**), suggérant une structure proche des répertoires, la similarité entre les deux séries est assez faible (MH moyen=0,4)

Nous ne pouvons, cependant, pas exclure que le faible chevauchement observé ne soit pas également lié aux protocoles de préparation des librairies. Ceci est d'autant plus critique qu'il existe à l'heure actuelle un grand nombre de protocoles expérimentaux. Aussi, afin de tester l'impact de la profondeur de séquençage sur la diversité, il est apparu préférable d'utiliser des répertoires produits de la même façon.

B. Impact de la profondeur de séquençage sur la diversité observée

1) Données expérimentales

Chacun des 28 jeux de données TriPoD_06 a été sous-échantillonné et normalisé comme décrit précédemment de manière à simuler des profondeurs de séquençage variant entre 2.10^4 à 9.10^5 séquences TR (par pas de 1.10^4 entre 2.10^5 et 1.10^5 puis par pas de 10^5). Pour chaque sous-échantillon, la similarité entre le répertoire observé et celui d'origine est calculée sur la base de leur composition clonotypique. Ainsi, l'évolution de la représentativité des sous-échantillons peut être suivie en fonction de la profondeur de séquençage (**Figure 46A**). On observe que les échantillons ne sont pas impactés de la même manière par le sous-échantillonnage. Comme le démontre la classification hiérarchique (**Figure 46B**), on distingue trois comportements : un premier groupe d'échantillons montre une similarité MH supérieure à 0,9 dès le premier degré de sous-échantillonnage (D= 2.10^4 séquences,) ; une seconde série atteint ce seuil rapidement pour de faibles profondeurs de séquençage ($4.10^4 < D < 6.10^4$) ; le dernier groupe n'atteint ce niveau de similarité qu'à partir de 1.10^5 séquences. La **Figure 46**B montre une ségrégation majeure entre les répertoires des échantillons Tregs et ceux des Teff et CD8, ces derniers étant plus affectés par le sous-échantillonnage sans doute du fait de leur grande diversité.

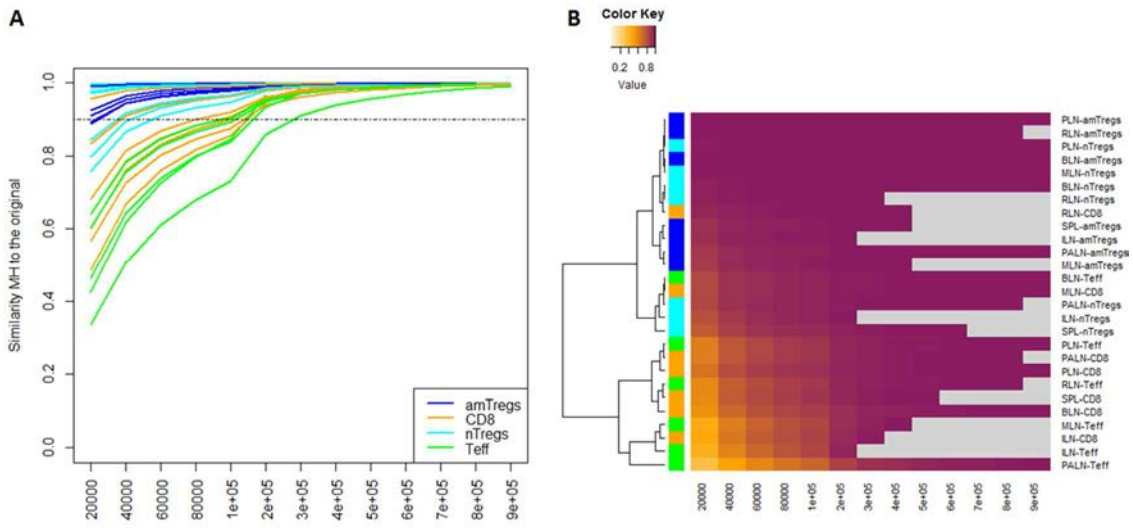


Figure 46 : Similarité des distributions clonotypiques entre les répertoires et leurs sous-échantillons – L'indice Morisita-Horn est calculé pour chaque répertoire entre la distribution originale des clonotypes et celle de chacun des sous-échantillons normalisés. A) Dynamique de la similarité des sous-échantillons en fonction de leur profondeur. B) Classification hiérarchique des échantillons en fonction de cette dynamique (Distance euclidienne, méthode « *complete* »). Les échantillons sont identifiés en fonction de leur population cellulaire : vert pour Teff, orange pour CD8, bleu pour amTregs et cyan pour nTregs.

Afin d'approfondir la représentativité des sous-échantillons, j'ai analysé l'évolution des différents paramètres de diversité des clonotypes, à savoir leur **richesse**, leur **distribution d'occurrences** et leur **profil de diversité**, du répertoire d'échantillons appartenant respectivement aux premier et troisième clusters. À titre d'exemple, j'ai choisi représenté les résultats de deux échantillons ayant des variabilités de similarité différentes en fonction de la taille du sous-échantillon, en l'occurrence MLN-amTregs (**Figure 47**) et MLN-Teff (**Figure 48**).

Pour des sous-échantillons ayant une similarité (MH) supérieure à 0,9 (du répertoire de MLN-amTregs, **Figure 47**) même pour de faibles profondeurs, seuls 50% des clonotypes existants à une profondeur initiale de 4.10^5 sont observés avec 2.10^4 séquences (**Figure 47B**). Toutefois, à cette profondeur, la majorité de ces clonotypes sont observés au moins deux fois comme l'indique la courbe rouge de la **Figure 47C** et le profil de diversité reste inchangé (**Figure 47D**). À l'inverse, pour des sous-échantillons pour lesquels la similarité est sensible à la taille des jeux de données (répertoire de MLN-Teff, **Figure 48A**), une profondeur de 2.10^5 séquences est nécessaire pour observer 50% des clonotypes existants (**Figure 48B**). De plus, on observe l'effet du sous-échantillonnage, par la forte proportion de clonotypes observés 1 ou 2 fois dans les échantillons de profondeur inférieure à 1.10^5 (**Figure 48C**). Par ailleurs, une profondeur d'au moins 2.10^5 est requise pour obtenir un profil de diversité représentatif (**Figure 48D**) de la diversité initiale.

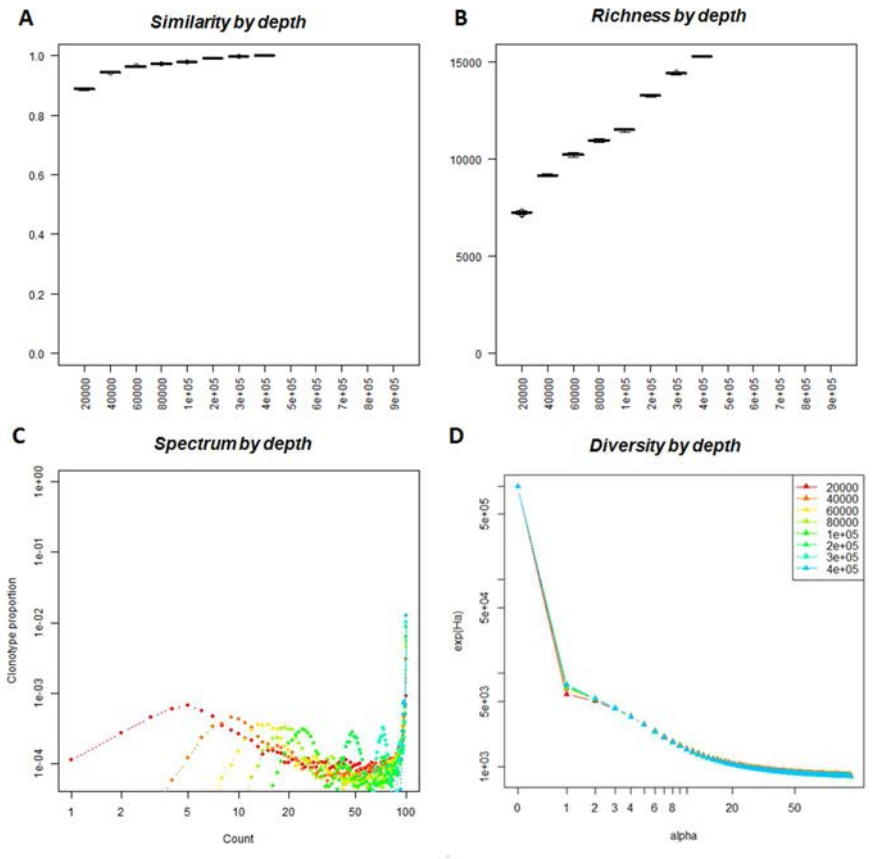


Figure 47 : Impact du sous-échantillonnage sur l'observation de la diversité du répertoire MLN-amTregs.

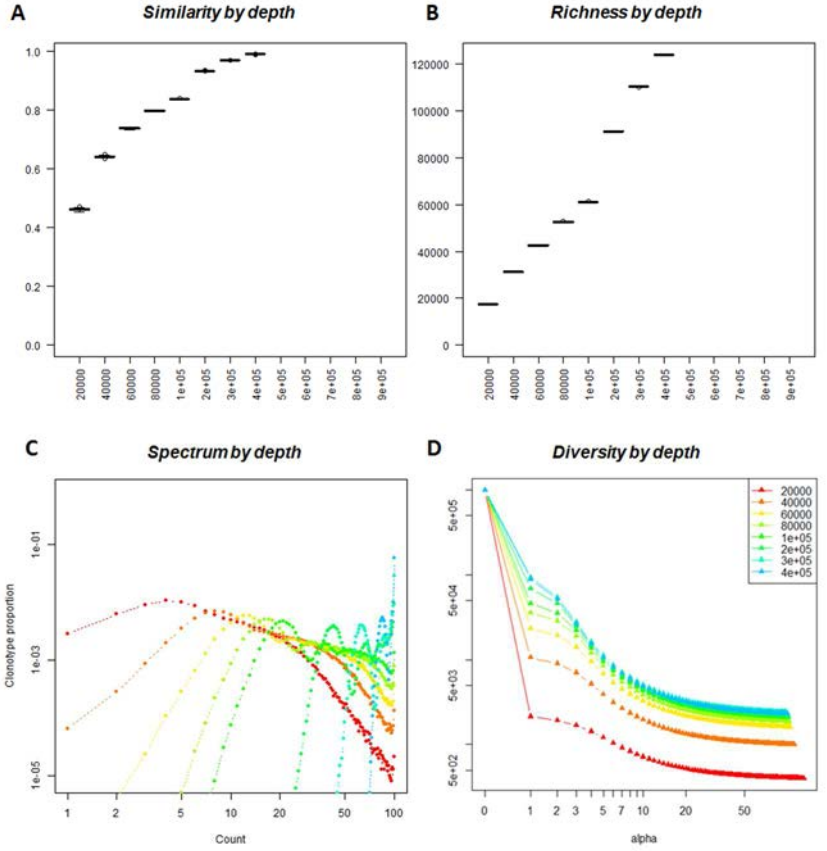


Figure 48 : Impact du sous-échantillonnage sur l'observation de la diversité du répertoire MLN-Teff

Pour chaque sous-échantillonnage, 100 tirages aléatoires de la profondeur souhaitée sont effectués puis un sous-échantillon moyen est construit. **A)** Distributions des valeurs de similarité MH entre le sous-échantillon et le jeu original à travers les 100 itérations, pour chaque profondeur. **B)** Distributions du nombre de clonotypes observés dans chaque sous-échantillon à travers les 100 itérations, pour chaque profondeur. **C)** Spectres d'occurrences des clonotypes au sein des sous-échantillons moyens obtenus pour chaque profondeur (identifié d'après la légende). **D)** Profils de diversité des sous-échantillons moyens obtenus pour chaque profondeur (identifié d'après la légende).

Chacun de ces répertoires a été obtenu à partir d'échantillons cellulaires de différentes tailles. Aussi, on peut se demander si les différences de sensibilité au sous-échantillonnage peuvent s'expliquer par ce facteur. La **Figure 49** représente l'indice de similarité entre chaque sous-échantillon et l'échantillon initial en fonction de la taille cellulaire de l'échantillon. Outre le phénotype des cellules analysées, la différence de sensibilité semble corréler avec le nombre de cellules constituant la population dont le répertoire est analysé. La corrélation entre la similarité et le nombre de cellules diminue au fur et à mesure que la profondeur augmente : R^2 varie entre -0,69 et -0,60 pour les profondeurs inférieures à 1.10^5 puis de -0,58 à -0,16, au-delà.

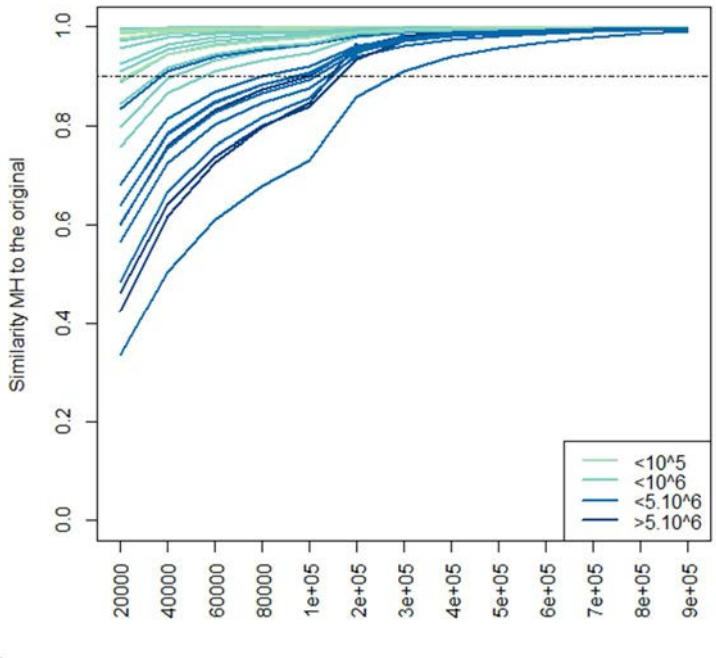


Figure 49 : Impact du sous-échantillonnage sur l'observation de la diversité en fonction de la taille des échantillons – Données représentées **Figure 46A** à la différence que les échantillons sont identifiés en fonction du nombre de cellules constituant l'échantillon cellulaire dont le répertoire est analysé : cinq échantillons avec moins de 1.10^5 cellules, 10 échantillons entre 1.10^5 et 1.10^6 cellules, 8 échantillons entre 1.10^6 et 5.10^6 et 5 avec plus de 5.10^6 cellules. Les courbes sont colorées comme indiqué par la légende.

2) Approche in silico

Comme le résume la **Figure 50**, il est possible de séparer les échantillons de TriPoD_6 en trois catégories de répertoires sur la base de leur diversité : un premier groupe réunissant la totalité des répertoires Teff et CD8 (à l'exception du RLN-CD8) et les deux autres se partageant les répertoires Tregs.

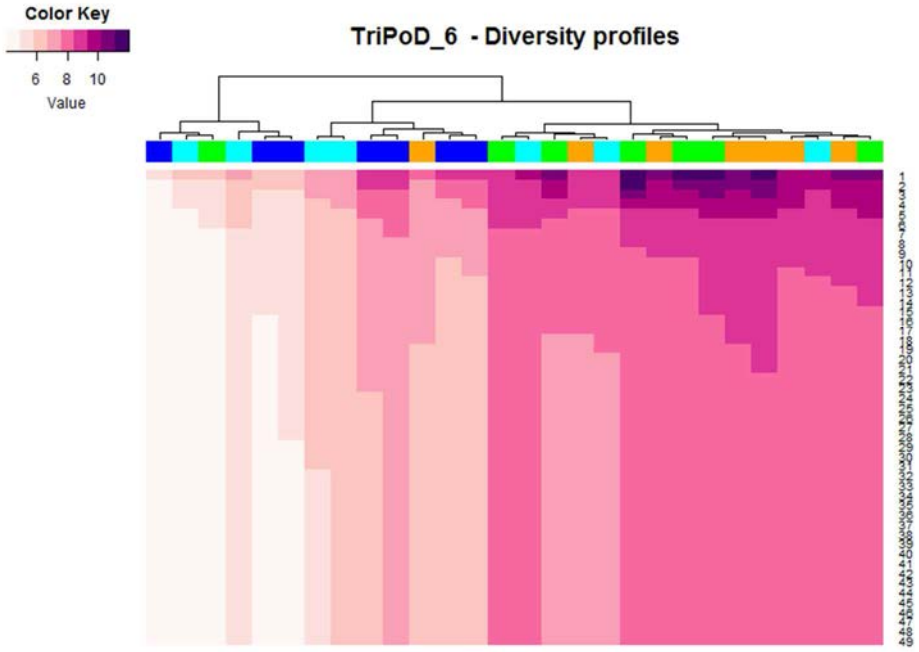


Figure 50 : Classification hiérarchique des 28 répertoires TRB en fonction de leur profil de diversité - Les échantillons (en colonne) sont identifiés en fonction de leur population. Chaque cellule indique la valeur de diversité obtenue pour un facteur alpha compris entre 1 et 50 (en ligne). Dendrogramme : distance euclidienne et méthode « *complete* ».

Or, si l'on modélise les distributions clonotypiques de ces répertoires pour estimer leur valeur de Zipf-α, il apparaît que la valeur médiane de ce paramètre est de l'ordre de 10^5 pour le cluster des répertoires Teff/CD8 qui sont les plus divers alors qu'elle est de 3 pour le cluster regroupant les répertoires amTregs dont la diversité est limitée. Ces résultats confirment les conclusions de Greiff qui démontrait une forte corrélation entre la distribution Zipf de répertoires BCR et leur profil de diversité (Greiff et al., 2014).

Afin de maîtriser les aléas associés à la nature des échantillons et ne pas être affectée par les biais liés à la préparation des librairies ou des protocoles de séquençage, j'ai produit une collection de 106 602 clonotypes générés artificiellement avec le package tcR (Nazarov et al., 2015) (voir METHODOLOGIE). Cette collection a ensuite été utilisée pour simuler dix répertoires TRB théoriques représentés par 1 million de séquences TR. Chacune des

simulations suit une distribution de Zipf simulant un scénario de diversité différent en faisant varier Zipf-α pour une valeur de Zipf-B= 0,2.

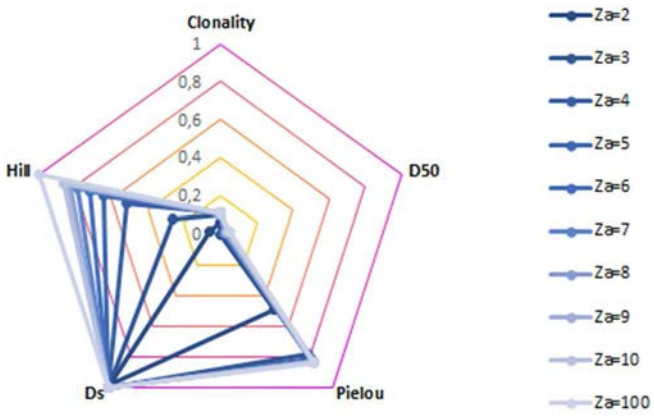


Figure 51 : Profil descriptif de la diversité globale des dix répertoires artificiels générés en fonction de la valeur de Zipf-α (Za) – Dix répertoires artificiels ont été simulés pour des valeurs croissantes de Zipf-α (Za). Les cinq métriques descriptives de la composition et de la diversité de ces répertoires ont été calculées sur la base de la distribution de fréquences des clonotypes observés au sein de chacun d'entre eux et reportées sur les axes de ce radar. Les répertoires sont identifiés par une couleur différente en fonction de la valeur Zipf-α (Za) utilisée pour la simulation, comme indiqué par la légende.

L'augmentation de la valeur Zipf-α (Za) diminue la proportion de clonotypes abondants, augmentant ainsi la diversité du répertoire comme l'indiquent notamment les valeurs des indices de Hill et de Piélou (**Figure 51**). Comme représenté **Figure 52**, le répertoire simulé avec Za=2 est partagé entre clonotypes rares et abondants ; Za=10 permet de simuler la présence de quelques clonotypes prédominants ; Za=100 entraîne une distribution équiprobable aux clonotypes simulant ainsi un répertoire très polyclonal, ici potentiellement sous-échantillonné car tous les clonotypes ne sont observés qu'une seule fois

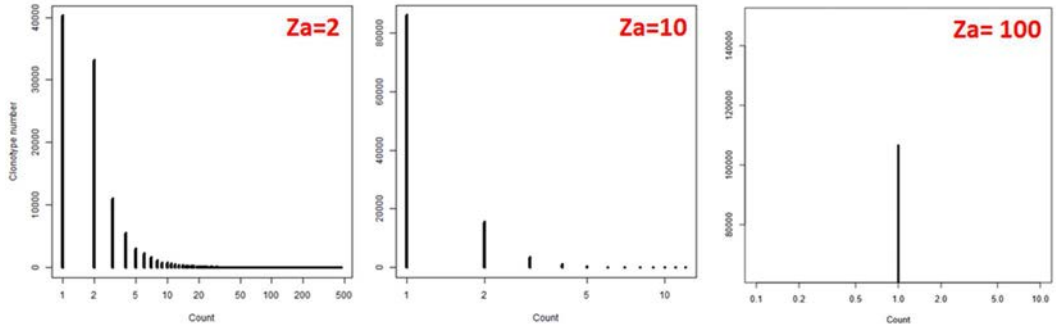


Figure 52 : Distributions des occurrences de trois répertoires simulés – Spectre des abondances des clonotypes observés au sein des répertoires simulés pour Zipf-α (Za) = 2, 10 et 100. En abscisses, les comptes possibles dans le jeu de données et en ordonnées, le nombre de clonotypes observés à ces comptes.

Afin d'évaluer l'impact de la profondeur de séquençage sur la capture de la diversité, une série de sous-échantillonnages a été effectuée pour des profondeurs de séquençage variant de 1.10^5 à 9.10^5 séquences TR (par pas de 1.10^5). Pour chaque sous-échantillon, la similarité de la topologie BV-BJ (**Figure 53A**) et de la composition clonotypique (**Figure 53B**) avec le répertoire d'origine ont été calculées.

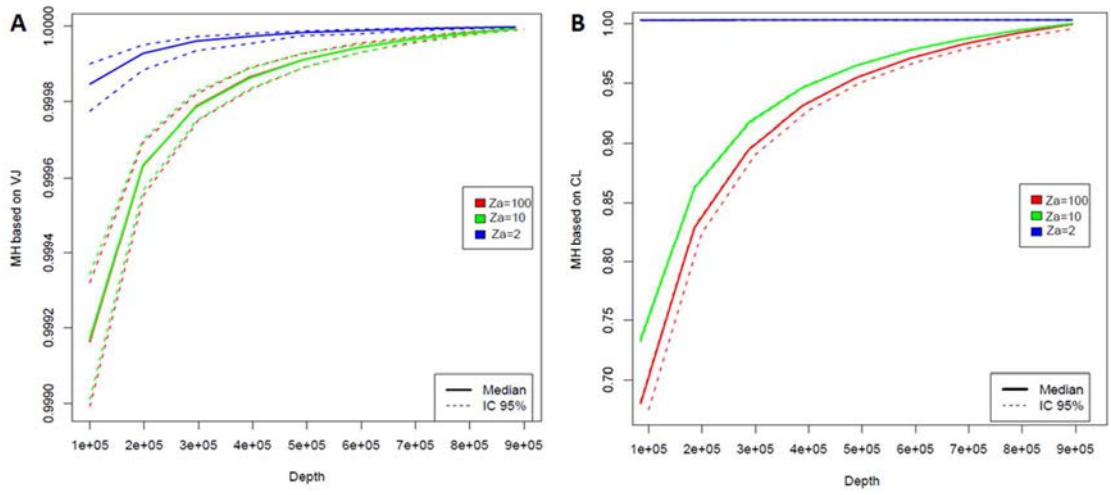


Figure 53 : Évolution de la similarité en fonction de la profondeur de sous-échantillonnage – L'indice de MH a été calculé en prenant en compte les distributions des combinaisons TRBVBJ **(A)** et celles des clonotypes **(B)**. Les courbes pleines représentent les valeurs médianes à travers 100 itérations de sous-échantillonnages et les courbes en pointillés indiquent l'intervalle de confiance à 95%.

Alors que tous les sous-échantillons sont très similaires à leurs données d'origines en termes de composition BVBJ (MH > 0,999), la distribution des clonotypes observés au sein des répertoires les plus divers (pour des valeurs Zipf-α>10) n'atteint une valeur de similarité représentative qu'au-delà de 5.10^5 séquences, à savoir la moitié de la profondeur initiale. Pour le répertoire incluant des clonotypes fortement prédominants (Zipf-α = 2), 10^5 séquences sont suffisantes pour rendre compte de sa diversité. Ainsi, plus la diversité d'un répertoire est importante, plus son observation est impactée par la taille de l'échantillon au niveau clonotypique.

C. Discussion

Représentativité des données d'immunoséquençage

Les protocoles d'immunoséquençage impliquent plusieurs étapes d'échantillonnage. Aussi, afin d'évaluer la représentativité des échantillons biologiques et d'évaluer le seuil de reproductibilité que l'on peut espérer atteindre lorsque l'on compare le répertoire TRB de plusieurs échantillons, nous avons choisi de procéder à plusieurs immunoséquençages d'un même échantillon cellulaire. Ainsi, la même population cellulaire a été divisée en deux échantillons cellulaires dont l'ARN extrait a été divisé en trois aliquotes. Ce plan expérimental nous permet de mesurer l'effet de deux niveaux d'échantillonnage que l'expérimentateur peut éventuellement adapter lors de l'établissement de sa stratégie de séquençage.

Les résultats de la comparaison des aliquotes au sein de chaque série nous permettent d'évaluer l'impact de l'échantillonnage de l'ARN. En effet, alors que l'ARN utilisé pour chaque séquençage a la même origine, le niveau de chevauchement entre le résultat de deux séquençages est en moyenne de 20%, et ce seuil diminue à 10% lorsque l'on compare trois séquençages. Ces observations sont similaires entre les deux plates-formes de séquençage. De plus, la comparaison des résultats des deux séries nous informe que la proportion de clonotypes observés dans les deux échantillons est de l'ordre de 15%.

Ces résultats nous fournissent donc une valeur référence quant au niveau de chevauchement à attendre lorsque l'on compare les répertoires de plusieurs échantillons. Ils permettent donc de relativiser et surtout de mieux appréhender la comparaison de répertoires de natures différentes en bornant le niveau de similarité potentiel. Afin d'évaluer la reproductibilité des séquençages, il aurait été intéressant de pondérer ces résultats en séquençant plusieurs fois chacune des librairies pour s'assurer du niveau de reproductibilité du séquençage.

Il est à noter que cette expérience est limitante dans son évaluation de la représentativité des échantillons biologiques. En effet, la population cellulaire analysée, LT CD4+ de la rate, est fortement polyclonale. Or, les séquençages effectués ne permettent pas d'observer de manière exhaustive les clonotypes en présence, leur profondeur étant bien inférieure à la taille de l'échantillon cellulaire initial (3.10^6 de cellules). Le seuil d'intersection est par conséquent sous-estimé. Pour pallier cela, il sera nécessaire de reproduire cette expérience en partant d'une population dont la diversité est limitée en s'assurant de la couverture complète de la diversité. Ainsi, il sera possible d'évaluer la borne supérieure du niveau de

chevauchement. C'est d'ailleurs dans l'optique de traiter ce type de considération que j'ai mis au point l'approche de répertoires artificiels. Ainsi, il nous est désormais possible de simuler de très grands répertoires dans des conditions de diversité différentes et de tester le niveau de chevauchement entre des sous-échantillons de profondeurs différentes.

Impact de la profondeur de séquençage

En plus de l'échantillonnage biologique, la représentativité des données d'immunoséquençage est conditionnée par l'adéquation entre le nombre de séquences produites par rapport à la diversité clonotypique à couvrir.

En effet, les résultats obtenus à partir de données expérimentales et simulées démontrent que plus un répertoire TR est divers, plus la profondeur de séquençage nécessaire pour capturer de manière représentative cette diversité sera grande. Ce point critique devient problématique lorsque l'on cherche à caractériser la diversité d'une population cellulaire dont la taille est bien supérieure à la capacité de séquençage et dont on sait (ou suppose) que le répertoire est hautement polyclonal. C'est ce que l'on observe notamment **Figure 40A-B** où l'on voit que, malgré les 6.10^6 de séquences produites pour chaque séquençage de la même population Teff de rate de souris C57BL/6, la diversité n'est pas complètement couverte. Aussi, sans chercher à décrire de manière exhaustive les clones constituant cette population, il est difficile de s'assurer que le répertoire observé à partir d'un sous-échantillon sera représentatif de la diversité totale alors que l'on sait que plus un répertoire est divers plus il est sensible à l'échantillonnage.

Cette problématique se pose à nous dans le cadre du projet TriPoD. En effet, alors que pour les populations Tregs, la totalité des cellules triées sont utilisées pour le séquençage, seule une quantité limitée (proportionnelle à la taille des organes) des LT Teff et CD8 sont triées pour des raisons logistiques. Alors que comme démontré précédemment, les répertoires de ces populations sont comparables et montrent un certain degré de chevauchement entre les organes au sein d'une même expérience, il est à craindre que ce sous-échantillonnage ait un impact majeur sur la comparaison des données de plusieurs expériences. L'ampleur de cet impact pourra être estimée par la modélisation de ce phénomène.

Une recommandation pour compenser ce genre de risque serait d'avoir recours pour ce type d'échantillon à plusieurs répétitions techniques et biologiques de manière à pouvoir les combiner et atteindre une plus grande couverture de diversité. Ce type d'approche ayant un

coût, on peut envisager d'adapter la stratégie de séquençage de manière à moduler la profondeur de séquençage en fonction de la nature et de la taille d'une population cellulaire. En effet, comme le démontrent les résultats présentés **Figure 46** et **Figure 49**, les populations Tregs dont la taille est plus petite et le répertoire moins divers que les Teff et CD8 nécessitent une profondeur 2 à 3 fois moins importante que ces derniers pour atteindre un niveau de représentativité similaire. Aussi, adapter la profondeur de séquençage de chaque échantillon, plutôt que de produire un nombre identique de séquences, pourrait permettre d'optimiser les passages sur séquenceur.

DISCUSSION GENERALE

L'émergence de nouvelles technologies implique celle de nouveaux besoins et soulèvent des problématiques inédites. Les données d'immunoséquençage, comme toutes celles issues de technologies dites à « haut-débit », sont très lourdes en termes de gestion et de traitement de données.

Lors du lancement du projet TriPoD, nous avons réalisé qu'alors que l'immunoséquençage à haut débit se développait depuis plus de sept ans, certains points « cruciaux » tels que les nomenclatures de description ou les structures de stockages étaient (et sont) toujours sujet à débat. Faute de standard, proposer une structure et un flux d'analyse standardisés est donc apparu crucial pour assurer une exploitation optimale de nos données mais également une interprétation maîtrisée de nos résultats. Dans ce contexte, j'ai mis en place une procédure de gestion des données de séquençage et ai réfléchi à la spécification d'une base de données standardisée dédiée à l'exploitation des données de répertoires TR. Faute de moyens suffisants, le développement de cette base de données n'a pas encore eu lieu.

De nombreux aspects du traitement des données d'immunoséquençage tels que leur annotation ou leur normalisation sont en constante évolution. Il m'a donc fallu mettre en place une méthodologie répondant à l'état de l'art tout en composant avec les limites des ressources disponibles.

Ainsi, une des décisions fondamentales qu'il m'a fallu prendre a été le choix de l'outil d'annotation que j'allais utiliser. Alors que le nombre d'outils d'annotations dédiés à l'immunoséquençage augmente chaque année, les trois outils ayant attiré mon attention lors du lancement du projet fin 2013 étaient MiTCR (Bolotin et al., 2013) et Decombinator (Thomas et al., 2013b) publiés cette même année, et Vidjil publiés début 2014 (Giraud et al., 2014). Ces outils très populaires ont l'avantage de proposer une correction de l'impact des erreurs de PCR et de séquençage sur la diversité observée en regroupant les séquences par leur similarité. Toutefois, les fichiers de sortie de ces outils consistent en une table de contingence des clonotypes identifiés après ces corrections. Or, commençant à peine à me familiariser avec ce type de données, je souhaitais pouvoir manipuler les séquences plus librement. J'ai pris connaissance de l'existence de *clonotypeR* (Plessy et al., 2015), suite à l'arrivée d'une de ses deux concepteurs au laboratoire. L'avantage de *clonotypeR* est qu'il annote chaque séquence TR identifiée sous forme d'un vecteur d'annotation résumant, entre autres, la librairie dans

laquelle la séquence TR a été observée, les gènes TRBV et TRBJ utilisés et la séquence du CDR3. Chaque fichier de sortie (format TSV) contient l'information de toutes les séquences identifiées comme codant pour une chaîne de TCR, facilitant ainsi grandement leur manipulation. De plus, chaque entrée de ce fichier est liée à la séquence brute (*read*) produite grâce à son identifiant de référencement dans le fichier fastq. Cette correspondance s'est notamment avérée très utile pour la mise au point des protocoles de séquençage au laboratoire, me permettant d'aider les expérimentateurs à optimiser leur protocole de manière à améliorer le rendement de séquences analysables.

L'utilisation de *clonotypeR* m'a permis de me familiariser avec les données RepSeq et de mettre en place les différentes étapes de ma méthodologie. Toutefois, il a été nécessaire de vérifier que cette liberté de manipulation n'affectait pas la qualité et la composition des jeux de données analysés. En effet, *clonotypeR* n'applique aucun filtre de qualité sur les données mais impose un alignement très conservatif avec les gènes TRV ; ainsi, si une séquence ne s'aligne pas parfaitement avec une des séquences TRV de référence sur une fenêtre de 20 nucléotides ou que le TRV ou le TRJ n'est pas identifiable cette séquence n'est pas sélectionnée. À titre d'exemple, j'ai comparé les métriques descriptives obtenues pour les six jeux de données ***TriPoD_38_1070*** après leur annotation par *clonotypeR* et *MiTCR* ; pour les aliquotes séquencés sur Illumina j'ai fait cette comparaison avant et après l'assimilation des clonotypes. MiTCR a la particularité de permettre d'adapter la stratégie d'identification des clonotypes en fonction de la qualité des séquences. Les séquences sont dans un premier temps alignées avec les références de la base IMGT afin d'identifier les gènes recombinés constituant les séquences TR, alors que pour clonotypeR nous utilisons la base de données NCBI (NCBI Reference Sequence entries NG_007044 pour les locus α et δ, NG_006980 pour le locus β et NG_007033 pour le locus γ). Les séquences de qualité Phred satisfaisante (seuil fixé ici à 30) sont utilisées pour construire une table de clonotypes « core ». Les séquences dont le score Phred est inférieur au seuil fixé, pouvant être considérées comme le fruit d'erreurs de séquençage, sont ignorées ou, si similaires à la séquence de clonotypes « core », assimilées à ceux-ci. Dans un second temps, les clonotypes « core » sont comparés entre eux de manière à regrouper les clonotypes distants d'un nucléotide corrigeant ainsi les éventuelles erreurs de PCR. Comme exposé précédemment, chaque facteur, y compris les méthodes d'annotation et de corrections des séquences, influe sur la diversité de répertoire obtenue. En proposant une solution robuste pour traiter à la fois les erreurs de PCR et de séquençage, MiTCR m'a donc

semblé être un bon contrôle pour évaluer la pertinence de mon processus de traitement de données.

Ainsi, on voit sur la **Figure 54** que les valeurs de diversité observées en Illumina (*not collapsed)* sont similaires entre les deux outils à l'exception du D50 qui est plus fort avec l'annotation *MiTCR*. Après normalisation de ces données, cette différence est effacée. Concernant les données 454, la diversité observée semble plus faible avec *MiTCR*. Ces résultats sont donc rassurants quant à la pertinence de l'utilisation de *clonotypeR* et au bien-fondé de l'algorithme d'assimilation que j'ai mis en place.

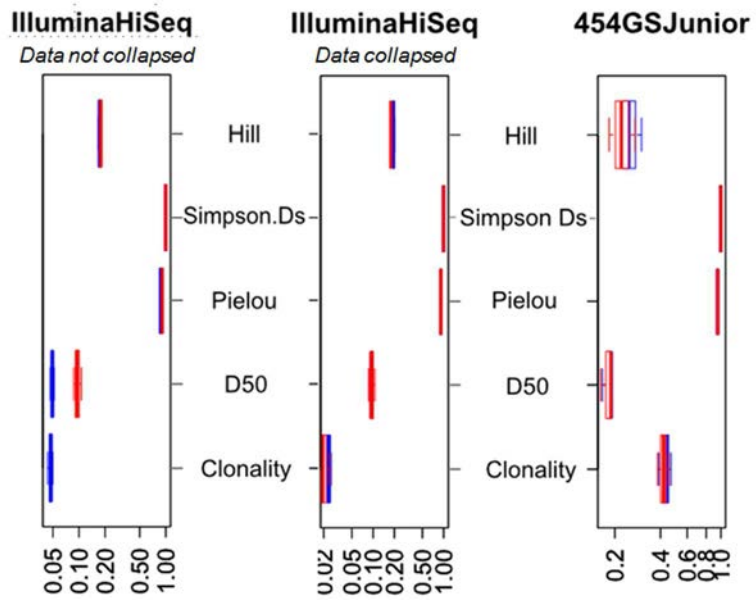


Figure 54: Comparaison de l'évaluation de la diversité de répertoires TRB en fonction de l'outil d'annotation utilisé – Les trois jeux de données obtenus à partir des deux séries d'échantillons TriPoD_38_1070 ont été annotés avec *MiTCR* (en rouge) et clonotypeR (en bleu) puis les cinq métriques descriptives définies par la méthodologie ont été calculées pour chaque jeu de données.

Outre l'aspect technique, mon objectif a été de proposer une approche permettant une caractérisation approfondie et facilement interprétable de la diversité des répertoires TCR analysés. En effet, le répertoire TCR est un objet complexe dont la structure et la composition sont le résultat de l'histoire immunitaire de l'organisme. Ainsi, décomposer la diversité du répertoire TCR d'une population lymphocytaire permet de mieux comprendre les modifications du répertoire en fonction des organes, de l'âge, du phénotype ou du statut physiopathologique d'un individu. Le module d'analyse que j'ai développé a donc été conçu de façon à ce que chaque analyse apporte un éclairage particulier sur le répertoire TR analysé. En effet, chaque observation soulève une nouvelle question : une distribution biaisée des gènes TRBV suggère un usage préférentiel de certaines combinaisons TRBVBJ ; la diversité des clonotypes issus de ces combinaisons peut être plus ou moins grande ; cette diversité peut être impactée par la sélection de CDR3 partageant des caractéristiques en commun telles que

leur longueur de séquence… Tous ces paramètres résument la façon dont le répertoire observé s'est construit et son potentiel à mettre en place ou maintenir une réponse immunitaire adaptée. J'ai donc cherché à adapter mon approche à l'objet biologique plutôt qu'à la technologie utilisée pour l'observer. En effet, contrairement aux puces à ADN ou à la cytométrie en flux qui décrivent les objets biologiques en fonction de paramètres intrinsèquement liés à la technologie (sondes, fluorochromes…), les entités et paramètres utilisés pour décrire le répertoire TCR ont été définis bien avant l'apparition de l'immunoséquençage. L'évolution de la technologie a permis une description de plus en plus détaillée de ces entités qui, avec l'immunoséquençage, a atteint un niveau de précision inédit. Toutefois, la rencontre d'une technologie riche en information avec un système complexe et dynamique rend l'analyse et l'interprétation des données produites particulièrement délicates. En effet, il est difficile d'arbitrer *a priori* quelles informations extraire, comment les analyser, comment les interpréter. Le challenge a donc été de mettre au point un panel d'analyses capturant un maximum d'informations plus ou moins triviales, tout en les représentant de manière à les rendre lisibles et interprétables par l'utilisateur.

Bien que plus ou moins ardue, traiter informatiquement un grand flux de données reste une tâche classique pour un(e) bioinformaticien(ne) ; ce n'est pas forcément le cas pour les biologistes. Je me suis donc attachée à rendre accessible cette approche afin de permettre aux expérimentateurs d'analyser leurs données RepSeq de manière autonome, similairement à ce qu'ils pourraient faire avec les données d'une technologie plus familière. En proposant une description standardisée multi-échelle, mon objectif a donc été de fournir aux biologistes un outil le plus complet possible leur permettant de construire une représentation robuste du répertoire TCR de leurs échantillons et de leur donner les moyens de bâtir et/ou d'étayer leurs hypothèses.

Il est à noter que la démarche de description de la diversité d'un système biologique complexe n'est pas exclusive à l'étude de répertoire TCR. L'analyse de la diversité du microbiome est également un domaine de recherche qui a fortement profité de développement des technologies NGS. En effet, grâce au séquençage de l'ARN ribosomal 16S des microorganismes présents dans la flore intestinale d'individu, il est possible de catégoriser phylogénétiquement les espèces en présence. La diversité et la composition de ce microbiote est analyséeanalysée afin d'identifier le rôle potentiel joué par ces bactéries dans les désordres intestinaux

chroniques par exemple (Sarrabayrouse et al., 2014). Ainsi, on peut appliquer le même type d'approche que celui proposé pour l'analyse du répertoire TR à l'analyse du microbiote. En effet, chaque jeu de données de séquençage de microbiote est analysé de manière à catégoriser les séquences produites en *Operational Taxonomic Units* (OTUs) ce qui permet d'inférer une hiérarchie phylogénétique entre les espèces observées. Chaque OTU est ensuite annotée en termes de *Phylum*, *Classe*, *Ordre*, *Famille* et *Genre*. On peut donc décomposer chaque jeu de données en fonction de ces paramètres. À titre d'exemple, la **Figure 55** s'intéresse à la diversité du microbiote issu d'un échantillon de selles d'un individu adulte sain (données non publiées obtenues dans le cadre d'une collaboration avec Philippe Seksik, Laboratoire des BioMolécules, ERL U1057 INSERM/UPMC). Les fréquences cumulées des OTUs observées sont calculées en fonction de leur *Phylum* (**Figure 55A**, en ligne) et de leur *Classe*, *Ordre* ou *Famille* (**Figure 55A**, en colonnes). Ainsi, on observe que le *Phylum* le plus représenté (95% des OTU observées) est celui des bactéries *Firmicutes* qui se distribuent en trois classes et sept ordres. Il apparaît que ces bactéries soient dominées par une famille, les *Lachnospiraceae* qui représentent environ 85% des Firmicutes. Ainsi, il est également possible de construire un profil de diversité reflétant la diversité de ce microbiote pour les différents niveaux phylogénétiques (**Figure 55B**). Ainsi, plus on descend dans la hiérarchie phylogénétique, plus la diversité est grande et les profils gardent la même allure. Quel que soit le niveau d'observation, le profil révèle une chute rapide de la diversité avec l'augmentation de α, ce qui suggère la prédominance de certains OTUs possiblement liés phylogénétiquement. Ces profils résument donc les observations faites en décomposant le jeu de données.

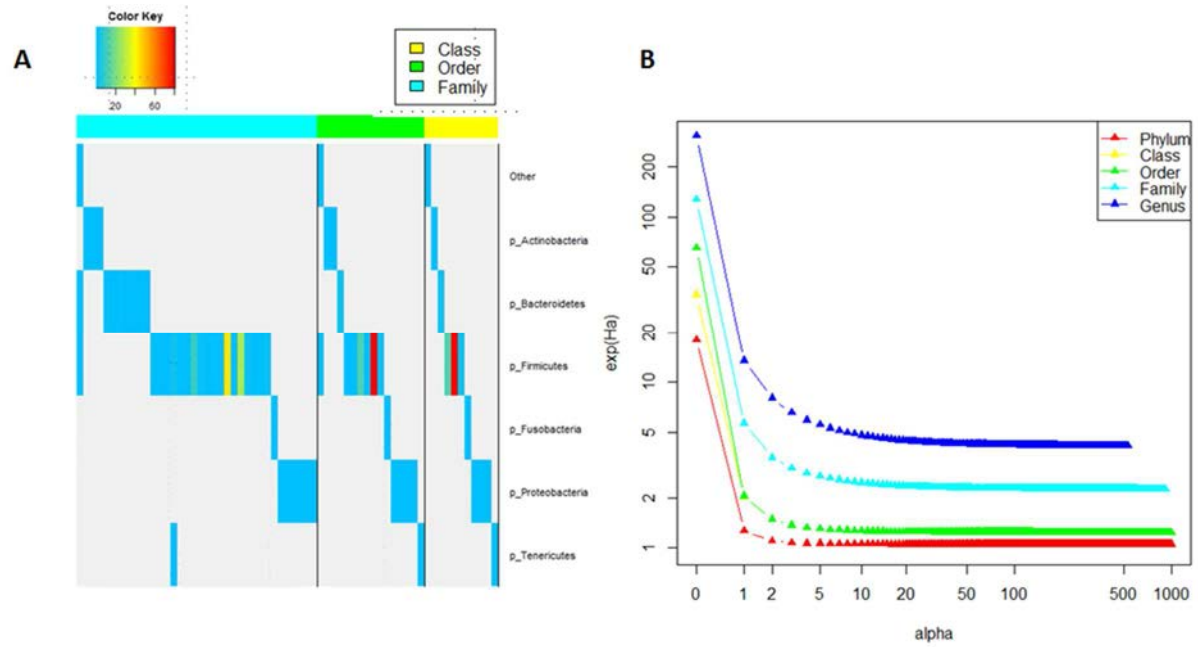


Figure 55 : Décomposition de la diversité du microbiote d'un individu sain – A) Fréquences cumulées des OTU en fonction des différents niveaux de hiérarchie phylogénétique. Les valeurs sont indiquées suivant l'échelle de couleur. B) Profils de diversité du même microbiote en fonction du niveau de hiérarchie phylogénétique. Chaque courbe correspond à un niveau comme indiqué par la légende.

Alors qu'au début de ma thèse, le monde de l'immunoséquençage était encore fortement inexploré, ces dernières années ont vu émerger de nombreux développements technologiques et méthodologiques, ce dans des domaines aussi divers que la biologie moléculaire, la modélisation mathématique, le développement logiciel… Toutefois, la plupart de ces développements se sont attachés aux défis techniques que soulevait la production massive d'informations par l'immunoséquençage, en laissant de côté la nature complexe de l'objet biologique qu'elle décrit. Durant cette thèse, j'ai cherché à réconcilier ces deux aspects en proposant une approche conciliant les forces et les limites de cette technologie afin d'apporter un éclairage averti aux immunologistes, et aux informaticiens, lors de leurs investigations.

BIBLIOGRAPHIE

Aouinti, S., Malouche, D., Giudicelli, V., Kossida, S., and Lefranc, M.-P. (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*.

Arden, B., Klotz, J.L., Siu, G., and Hood, L.E. (1985). Diversity and structure of genes of the α family of mouse T-cell antigen receptor. Nature *316*, 783–787.

Arstila, T.P., Casrouge, A., Baron, V., Even, J., Kanellopoulos, J., and Kourilsky, P. (1999). A direct estimate of the human alphabeta T cell receptor diversity. Science *286*, 958–961.

Atchley, W.R., Zhao, J., Fernandes, A.D., and Druke, T. (2005). Solving the protein sequence metric problem. Proc. Natl. Acad. Sci. U. S. A. *102*, 6395–6400.

Attaf, M., Huseby, E., and Sewell, A.K. (2015). αβ T cell receptors as predictors of health and disease. Cell. Mol. Immunol.

Bashford-Rogers, R.J., Palser, A.L., Idris, S.F., Carter, L., Epstein, M., Callard, R.E., Douek, D.C., Vassiliou, G.S., Follows, G.A., Hubank, M., et al. (2014). Capturing needles in haystacks: a comparison of B-cell receptor sequencing methods. BMC Immunol. *15*, 29.

Baumgärtner, S. (2006). Measuring the diversity of what? And for what purpose? A conceptual comparison of ecological and economic biodiversity indices.

Becattini, S., Latorre, D., Mele, F., Foglierini, M., Gregorio, C.D., Cassotta, A., Fernandez, B., Kelderman, S., Schumacher, T.N., Corti, D., et al. (2015). Functional heterogeneity of human memory CD4+ T cell clones primed by pathogens or vaccines. Science *347*, 400–406.

Benichou, J., Ben-Hamo, R., Louzoun, Y., and Efroni, S. (2012). Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. Immunology *135*, 183–191.

Benichou, J., Glanville, J., Prak, E.T.L., Azran, R., Kuo, T.C., Pons, J., Desmarais, C., Tsaban, L., and Louzoun, Y. (2013). The Restricted DH Gene Reading Frame Usage in the Expressed Human Antibody Repertoire Is Selected Based upon its Amino Acid Content. J. Immunol. *190*, 5567–5577.

Benoist, C., Germain, R.N., and Mathis, D. (2006). A Plaidoyer for “Systems Immunology.” Immunol. Rev. *210*, 229–234.

Berger, W.H., and Parker, F.L. (1970). Diversity of planktonic foraminifera in deep-sea sediments. Science *168*, 1345–1347.

Bergot, A.-S., Chaara, W., Ruggiero, E., Mariotti-Ferrandiz, E., Dulauroy, S., Schmidt, M., von Kalle, C., Six, A., and Klatzmann, D. (2015). TCR sequences and tissue distribution discriminate the subsets of naïve and activated/memory Treg cells in mice: Molecular immunology. Eur. J. Immunol. *45*, 1524–1534.

Bertalanffy, L.V. (1993). Théorie générale des systèmes (Paris: Dunod).

Blackman, M., Yagüe, J., Kubo, R., Gay, D., Coleclough, C., Palmer, E., Kappler, J., and Marrack, P. (1986). The T cell repertoire may be biased in favor of MHC recognition. Cell *47*, 349–357.

Blondel, J. (2005). Les biodiversités: objets, théories, pratiques (Paris: CNRS Éd).

Blüthmann, H., Kisielow, P., Uematsu, Y., Malissen, M., Krimpenfort, P., Berns, A., von Boehmer, H., and Steinmetz, M. (1988). T-cell-specific deletion of T-cell receptor transgenes allows functional rearrangement of endogenous alpha- and beta-genes. Nature *334*, 156–159.

Bolotin, D.A., Mamedov, I.Z., Britanova, O.V., Zvyagin, I.V., Shagin, D., Ustyugova, S.V., Turchaninova, M.A., Lukyanov, S., Lebedev, Y.B., and Chudakov, D.M. (2012). Next generation sequencing for TCR repertoire profiling: Platform-specific features and correction algorithms. Eur. J. Immunol. *42*, 3073–3083.

Bolotin, D.A., Shugay, M., Mamedov, I.Z., Putintseva, E.V., Turchaninova, M.A., Zvyagin, I.V., Britanova, O.V., and Chudakov, D.M. (2013). MiTCR: software for T-cell receptor sequencing data analysis. Nat. Methods *10*, 813–814.

Bolotin, D.A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I.Z., Putintseva, E.V., and Chudakov, D.M. (2015). MiXCR: software for comprehensive adaptive immunity profiling. Nat. Methods *12*, 380–381.

Born, W., Yagüe, J., Palmer, E., Kappler, J., and Marrack, P. (1985). Rearrangement of T-cell receptor beta-chain genes during T-cell development. Proc. Natl. Acad. Sci. U. S. A. *82*, 2925–2929.

Boudinot, P., Marriotti-Ferrandiz, M.E., Pasquier, L.D., Benmansour, A., Cazenave, P.A., and Six, A. (2008). New perspectives for large-scale repertoire analysis of immune receptors. Mol. Immunol. *45*, 2437–2445.

Bousso, P., Levraud, J.-P., Kourilsky, P., and Abastado, J.-P. (1999). The Composition of a Primary T Cell Response Is Largely Determined by the Timing of Recruitment of Individual T Cell Clones. J. Exp. Med. *189*, 1591–1600.

Boyd, S.D., Marshall, E.L., Merker, J.D., Maniar, J.M., Zhang, L.N., Sahaf, B., Jones, C.D., Simen, B.B., Hanczaruk, B., Nguyen, K.D., et al. (2009). Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. Sci. Transl. Med. *1*, 12ra23.

Brochet, X., Lefranc, M.-P., and Giudicelli, V. (2008). IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. Nucleic Acids Res. *36*, W503-508.

Brown, S.D., Raeburn, L.A., and Holt, R.A. (2015). Profiling tissue-resident T cell repertoires by RNA sequencing. Genome Med. *7*, 125.

Brusko, T.M., Koya, R.C., Zhu, S., Lee, M.R., Putnam, A.L., McClymont, S.A., Nishimura, M.I., Han, S., Chang, L.-J., Atkinson, M.A., et al. (2010). Human antigen-specific regulatory T cells generated by T cell receptor gene transfer. PloS One *5*, e11726.

Burchill, M.A., Yang, J., Vang, K.B., Moon, J.J., Chu, H.H., Lio, C.-W.J., Vegoe, A.L., Hsieh, C.-S., Jenkins, M.K., and Farrar, M.A. (2008). Linked T cell receptor and cytokine signaling govern the development of the regulatory T cell repertoire. Immunity *28*, 112–121.

Burgos, J.D., and Moreno-Tovar, P. (1996). Zipf-scaling behavior in the immune system. Biosystems *39*, 227–232.

Burnet, F.M. (1962). The immunological significance of the thymus: an extension of the clonal selection theory of immunity. Australas. Ann. Med. *11*, 79–91.

Burnet, F.M. (1976). A modification of Jerne's theory of antibody production using the concept of clonal selection. CA. Cancer J. Clin. *26*, 119–121.

Calis, J.J.A., and Rosenberg, B.R. (2014). Characterizing immune repertoires by high throughput sequencing: strategies and applications. Trends Immunol.

Cantor, H., and Boyse, E.A. (1975). Functional subclasses of T-lymphocytes bearing different Ly antigens. I. The generation of functionally distinct T-cell subclasses is a differentiative process independent of antigen. J. Exp. Med. *141*, 1376–1389.

Carlson, C.S., Emerson, R.O., Sherwood, A.M., Desmarais, C., Chung, M.-W., Parsons, J.M., Steen, M.S., LaMadrid-Herrmannsfeldt, M.A., Williamson, D.W., Livingston, R.J., et al. (2013). Using synthetic templates to design an unbiased multiplex PCR assay. Nat. Commun. *4*, 2680.

Casrouge, A., Beaudoing, E., Dalle, S., Pannetier, C., Kanellopoulos, J., and Kourilsky, P. (2000). Size estimate of the αβ TCR repertoire of naive mouse splenocytes. J. Immunol. *164*, 5782–5787.

Cebula, A., Seweryn, M., Rempala, G.A., Pabla, S.S., McIndoe, R.A., Denning, T.L., Bry, L., Kraj, P., Kisielow, P., and Ignatowicz, L. (2013). Thymus-derived regulatory T cells contribute to tolerance to commensal microbiota. Nature *497*, 258–262.

Chao, A. (1987). Estimating the Population Size for Capture-Recapture Data with Unequal Catchability. Biometrics *43*, 783–791.

Chao, A., Chazdon, R.L., Colwell, R.K., and Shen, T.-J. (2006). Abundance-Based Similarity Indices and Their Estimation When There Are Unseen Species in Samples. Biometrics *62*, 361–371.

Chaussabel, D., Quinn, C., Shen, J., Patel, P., Glaser, C., Baldwin, N., Stichweh, D., Blankenship, D., Li, L., Munagala, I., et al. (2008). A Modular Analysis Framework for Blood Genomics Studies: Application to Systemic Lupus Erythematosus. Immunity *29*, 150–164.

Chen, T., Darrasse-Jèze, G., Bergot, A.-S., Courau, T., Churlaud, G., Valdivia, K., Strominger, J.L., Ruocco, M.G., Chaouat, G., and Klatzmann, D. (2013). Self-specific memory regulatory T cells protect embryos at implantation in mice. J. Immunol. Baltim. Md 1950 *191*, 2273–2281.

Chien, Y.H., Gascoigne, N.R., Kavaler, J., Lee, N.E., and Davis, M.M. (1984). Somatic recombination in a murine T-cell receptor gene. Nature *309*, 322–326.

Chothia, C., Boswell, D.R., and Lesk, A.M. (1988). The outline structure of the T-cell alpha beta receptor. EMBO J. *7*, 3745–3755.

Ciupe, S.M., Devlin, B.H., Markert, M.L., and Kepler, T.B. (2013). Quantification of total T-cell receptor diversity by flow cytometry and spectratyping. BMC Immunol. *14*, 1–12.

Cochet, M., Pannetier, C., Regnault, A., Darche, S., Leclerc, C., and Kourilsky, P. (1992). Molecular detection and in vivo analysis of the specific T cell response to a protein antigen. Eur. J. Immunol. *22*, 2639–2647.

Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L., and Rice, P.M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. *38*, 1767–1771.

Collette, A., and Six, A. (2002). ISEApeaks: an Excel platform for GeneScan and Immunoscope data retrieval, management and analysis. Bioinforma. Oxf. Engl. *18*, 329–330.

Collette, A., Cazenave, P.-A., Pied, S., and Six, A. (2003). New methods and software tools for high throughput CDR3 spectratyping. Application to T lymphocyte repertoire modifications during experimental malaria. J. Immunol. Methods *278*, 105–116.

Colwell, R.K. (2005). ESTIMATES: Statistical Estimation of Species Richness and Shared Species from Samples. ResearchGate *42*.

Colwell, R.K., Chao, A., Gotelli, N.J., Lin, S.-Y., Mao, C.X., Chazdon, R.L., and Longino, J.T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. J. Plant Ecol. *5*, 3–21.

Correia-Neves, M., Waltzinger, C., Mathis, D., and Benoist, C. (2001). The shaping of the T cell repertoire. Immunity *14*, 21–32.

Covacu, R., Philip, H., Jaronen, M., Almeida, J., Kenison, J., Darko, S., Chao, C.-C., Yaari, G., Louzoun, Y., Carmel, L., et al. (2016). System-wide analysis of the T-cell response. *14*, 2733–2744.

Cukalac, T., Kan, W.-T., Dash, P., Guan, J., Quinn, K.M., Gras, S., Thomas, P.G., and La Gruta, N.L. (2015). Paired TCRαβ analysis of virus-specific CD8(+) T cells exposes diversity in a previously defined “narrow” repertoire. Immunol. Cell Biol. *93*, 804–814.

Cziko, G. (1997). Without Miracles: Universal Selection Theory and the Second Darwinian Revolution (MIT Press).

Darrasse-Jèze, G., Marodon, G., Salomon, B.L., Catala, M., and Klatzmann, D. (2005). Ontogeny of CD4+CD25+ regulatory/suppressor T cells in human fetuses. Blood *105*, 4715–4721.

Davis, M.M., and Bjorkman, P.J. (1988). T-cell antigen receptor genes and T-cell recognition. Nature *334*, 395–402.

Degauque, N., Brouard, S., and Soulillou, J.-P. (2016). Cross-Reactivity of TCR Repertoire: Current Concepts, Challenges, and Implication for Allotransplantation. Front. Immunol. *7*.

DeKosky, B.J., Ippolito, G.C., Deschner, R.P., Lavinder, J.J., Wine, Y., Rawlings, B.M., Varadarajan, N., Giesecke, C., Dörner, T., Andrews, S.F., et al. (2013). High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nat. Biotechnol. *31*, 166–169.

DeLong, D.C. (1996). Defining Biodiversity. *24*, 738–749.

Dohm, J.C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *36*, e105.

Efron, B., and Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? Biometrika *63*, 435–447.

Estorninho, M., Gibson, V.B., Kronenberg-Versteeg, D., Liu, Y.-F., Ni, C., Cerosaletti, K., and Peakman, M. (2013). A novel approach to tracking antigen-experienced CD4 T cells into functional compartments via tandem deep and shallow TCR clonotyping. J. Immunol. Baltim. Md 1950 *191*, 5430–5440.

Evert, S. (2004). A simple LNRE model for random character sequences. In Proceedings of JADT, p.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. *8*, 186–194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. *8*, 175–185.

Ferreira, C., Palmer, D., Blake, K., Garden, O.A., and Dyson, J. (2014). Reduced Regulatory T Cell Diversity in NOD Mice Is Linked to Early Events in the Thymus. J. Immunol. *192*, 4145–4152.

Fisher, R.A., Corbet, A.S., and Williams, C.B. (1943). The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. J. Anim. Ecol. *12*, 42–58.

Fisson, S., Darrasse-Jèze, G., Litvinova, E., Septier, F., Klatzmann, D., Liblau, R., and Salomon, B.L. (2003). Continuous Activation of Autoreactive CD4+ CD25+ Regulatory T Cells in the Steady State. J. Exp. Med. *198*, 737–746.

Föhse, L., Suffner, J., Suhre, K., Wahl, B., Lindner, C., Lee, C.W., Schmitz, S., Haas, J.D., Lamprecht, S., Koenecke, C., et al. (2011). High TCR diversity ensures optimal function andhomeostasis of Foxp3+ regulatory Tcells. Eur. J. Immunol.

Fontenot, J.D., Gavin, M.A., and Rudensky, A.Y. (2003). Foxp3 programs the development and function of CD4+CD25+ regulatory T cells. Nat. Immunol. *4*, 330–336.

Fowell, D., et D. Mason. 1993. « Evidence That the T Cell Repertoire of Normal Rats Contains Cells with the Potential to Cause Diabetes. Characterization of the CD4+ T Cell Subset That Inhibits This Autoimmune Potential ». The Journal of Experimental Medicine 177 (3): 627-36.

Freeman, J.D., Warren, R.L., Webb, J.R., Nelson, B.H., and Holt, R.A. (2009). Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. Genome Res. *19*, 1817–1824.

Gardy, J.L., Lynn, D.J., Brinkman, F.S.L., and Hancock, R.E.W. (2009). Enabling a systems biology approach to immunology: focus on innate immunity. Trends Immunol. *30*, 249–262.

Germain, R.N., Meier-Schellersheim, M., Nita-Lazar, A., and Fraser, I.D.C. (2011). Systems Biology in Immunology - A Computational Modeling Perspective. Annu. Rev. Immunol. *29*, 527–585.

Giraud, M., Salson, M., Duez, M., Villenet, C., Quief, S., Caillault, A., Grardel, N., Roumier, C., Preudhomme, C., and Figeac, M. (2014). Fast multiclonal clusterization of V(D)J recombinations from high-throughput sequencing. BMC Genomics *15*.

Grassle, J.F., and Smith, W. (1976). A similarity measure sensitive to the contribution of rare species and its use in investigation of variation in marine benthic communities. Oecologia *25*, 13–22.

Greiff, V., Menzel, U., Haessler, U., Cook, S.C., Friedensohn, S., Khan, T.A., Pogson, M., Hellmann, I., and Reddy, S.T. (2014). Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. BMC Immunol. *15*, 40.

Greiff, V., Miho, E., Menzel, U., and Reddy, S.T. (2015a). Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires. Trends Immunol. *36*, 738–749.

Greiff, V., Bhat, P., Cook, S.C., Menzel, U., Kang, W., and Reddy, S.T. (2015b). A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. Genome Med. *7*.

Hannum, C.H., Kappler, J.W., Trowbridge, I.S., Marrack, P., and Freed, J.H. (1984). Immunoglobulin-like nature of the alpha-chain of a human T-cell antigen/MHC receptor. Nature *312*, 65–67.

Hardy, O.J. (2010). BiodivR 1.2: A program to compute statistically unbiased indices of species diversity within sample and species similarity between samples using rarefaction principles.

Harty, J.T., and Badovinac, V.P. (2008). Shaping and reshaping CD8+ T-cell memory. Nat. Rev. Immunol. *8*, 107–119.

Heath, W.R., and Miller, J.F. (1993). Expression of two alpha chains on the surface of T cells in T cell receptor transgenic mice. J. Exp. Med. *178*, 1807–1811.

Heath, W.R., Carbone, F.R., Bertolino, P., Kelly, J., Cose, S., and Miller, J.F. (1995). Expression of two T cell receptor alpha chains on the surface of normal murine T cells. Eur. J. Immunol. *25*, 1617–1623.

Hill, M.O. (1973). Diversity and Evenness: A Unifying Notation and Its Consequences. Ecology *54*, 427–432.

Hori, S., Takahashi, T., and Sakaguchi, S. (2003a). Control of autoimmunity by naturally arising regulatory CD4+ T cells. Adv. Immunol. *81*, 331–371.

Hori, S., Nomura, T., and Sakaguchi, S. (2003b). Control of regulatory T cell development by the transcription factor Foxp3. *299*, 1057–1061.

Hou, X.-L., Wang, L., Ding, Y.-L., Xie, Q., and Diao, H.-Y. (2016). Current status and recent advances of next generation sequencing techniques in immunological repertoire. Genes Immun.

Howie, B., Sherwood, A.M., Berkebile, A.D., Berka, J., Emerson, R.O., Williamson, D.W., Kirsch, I., Vignali, M., Rieder, M.J., Carlson, C.S., et al. (2015). High-throughput pairing of T cell receptor α and β sequences. Sci. Transl. Med. *7*, 301ra131.

Hsieh, C.-S., Liang, Y., Tyznik, A.J., Self, S.G., Liggitt, D., and Rudensky, A.Y. (2004). Recognition of the peripheral self by naturally arising CD25+ CD4+ T cell receptors. Immunity *21*, 267–277.

Hsieh, C.-S., Zheng, Y., Liang, Y., Fontenot, J.D., and Rudensky, A.Y. (2006). An intersection between the self-reactive regulatory and nonregulatory T cell receptor repertoires. Nat. Immunol. *7*, 401–410.

Irving, B.A., and Weiss, A. (1991). The cytoplasmic domain of the T cell receptor zeta chain is sufficient to couple to receptor-associated signal transduction pathways. *64*, 891–901.

Jaccard, Paul (1901). Distribution de la flore alpine dans le Bassin des Dranses et dans quelques régions voisines.

Jang, M., Yew, P.-Y., Hasegawa, K., Ikeda, Y., Fujiwara, K., Fleming, G.F., Nakamura, Y., and Park, J.-H. (2015). Characterization of T cell repertoire of blood, tumor, and ascites in ovarian cancer patients using next generation sequencing. *4*, e1030561.

Jerne, N.K. (1955). THE NATURAL-SELECTION THEORY OF ANTIBODY FORMATION. Proc. Natl. Acad. Sci. U. S. A. *41*, 849–857.

Jerne, N.K. (1971). The somatic generation of immune recognition. Eur. J. Immunol. *1*, 1–9.

Jerne, N.K. (1972). Newer knowledge of immunobiology and its applications. Experientia. Suppl. *17*, 246–253.

Jost, L. (2006). Entropy and diversity. Oikos *113*, 363–375.

Kawashima, S., Ogata, H., and Kanehisa, M. (1999). AAindex: Amino Acid Index Database. Nucleic Acids Res. *27*, 368–369.

Kedzierska, K., Venturi, V., Field, K., Davenport, M.P., Turner, S.J., and Doherty, P.C. (2006). Early establishment of diverse T cell receptor profiles for influenza-specific CD8(+)CD62L(hi) memory T cells. Proc. Natl. Acad. Sci. U. S. A. *103*, 9184–9189.

Keenan, K., McGinnity, P., Cross, T.F., Crozier, W.W., and Prodöhl, P.A. (2013). diveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. Methods Ecol. Evol. *4*, 782–788.

Kersh, G.J., and Allen, P.M. (1996). Essential flexibility in the T-cell recognition of antigen. Nature *380*, 495–498.

Kidera, A., Konishi, Y., Oka, M., Ooi, T., and Scheraga, H.A. (1985). Statistical analysis of the physical properties of the 20 naturally occurring amino acids. J. Protein Chem. *4*, 23–55.

Klein, L., Kyewski, B., Allen, P.M., and Hogquist, K.A. (2014). Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). Nat. Rev. Immunol. *14*, 377–391.

Kourilsky, P. (2014). Le jeu du hasard et de la complexité la nouvelle science de l'immunologie (Paris: O. Jacob).

Lathrop, S.K., Santacruz, N.A., Pham, D., Luo, J., and Hsieh, C.-S. (2008). Antigen-specific peripheral shaping of the natural regulatory T cell population. J. Exp. Med. *205*, 3105–3117.

Laydon, D.J., Bangham, C.R.M., and Asquith, B. (2015). Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach. Philos. Trans. R. Soc. B Biol. Sci. *370*, 20140291.

Lefranc, M.-P., Giudicelli, V., Kaas, Q., Duprat, E., Jabado-Michaloud, J., Scaviner, D., Ginestoux, C., Clément, O., Chaume, D., and Lefranc, G. (2005). IMGT, the international ImMunoGeneTics information system®. Nucleic Acids Res. *33*, D593–D597.

Levine, A.G., Arvey, A., Jin, W., and Rudensky, A.Y. (2014). Continuous requirement for the TCR in regulatory T cell function. Nat. Immunol. *15*, 1070–1078.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

Li, H.M., Hiroi, T., Zhang, Y., Shi, A., Chen, G., De, S., Metter, E.J., Wood, W.H., Sharov, A., Milner, J.D., et al. (2016). TCRβ repertoire of CD4+ and CD8+ T cells is distinct in richness, distribution, and CDR3 amino acid composition. J. Leukoc. Biol. *99*, 505–513.

Liaskou, E., Henriksen, E.K.K., Holm, K., Kaveh, F., Hamm, D., Fear, J., Viken, M.K., Hov, J.R., Melum, E., Robins, H., et al. (2016). High-throughput T-cell receptor sequencing across chronic liver diseases reveals distinct disease-associated repertoires. Hepatol. Baltim. Md *63*, 1608–1619.

Linnemann, C., Heemskerk, B., Kvistborg, P., Kluin, R.J.C., Bolotin, D.A., Chen, X., Bresser, K., Nieuwland, M., Schotte, R., Michels, S., et al. (2013). High-throughput identification of antigen-specific TCRs by TCR gene capture. Nat. Med. *19*, 1534–1541.

Lio, C.-W.J., and Hsieh, C.-S. (2008). A two-step process for thymic regulatory T cell development. Immunity *28*, 100–111.

Luo, W., Su, J., Zhang, X.-B., Yang, Z., Zhou, M.-Q., Jiang, Z.-M., Hao, P.-P., Liu, S.-D., Wen, Q., Jin, Q., et al. (2012). Limited T Cell Receptor Repertoire Diversity in Tuberculosis Patients Correlates with Clinical Severity. PLoS ONE *7*.

Madi, A., Shifrut, E., Reich-Zeliger, S., Gal, H., Best, K., Ndifon, W., Chain, B., Cohen, I.R., and Friedman, N. (2014). T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. Genome Res. *24*, 1603–1612.

Magurran, A.E. (1998). Measuring richness and evenness. *13*, 165–166.

Magurran, A.E. (2004). Measuring biological diversity (Malden, Ma.: Blackwell Pub.).

Malissen, M., Minard, K., Mjolsness, S., Kronenberg, M., Goverman, J., Hunkapiller, T., Prystowsky, M.B., Yoshikai, Y., Fitch, F., and Mak, T.W. (1984). Mouse T cell antigen receptor: structure and organization of constant and joining gene segments encoding the beta polypeptide. Cell *37*, 1101–1110.

Mamedov, I.Z., Britanova, O.V., Bolotin, D.A., Chkalina, A.V., Staroverov, D.B., Zvyagin, I.V., Kotlobay, A.A., Turchaninova, M.A., Fedorenko, D.A., Novik, A.A., et al. (2011). Quantitative tracking of T cell clones after haematopoietic stem cell transplantation. EMBO Mol. Med. *3*, 201–207.

Manley, P., and Schlesinger, M. (2001). Riparian biological diversity in the Lake Tahoe basin. Final report for the California Tahoe Conservancy and the US Forest Service. Aspen Bibliogr. *Riparian Grant #CTA-3024*.

Marcon, E. (2015). Mesures de la Biodiversité. lecture. AgroParisTech.

Marcon, E., and Herault, B. (2016). entropart: Entropy Partitioning to Measure Diversity.

Mardis, E.R. (2011). A decade's perspective on DNA sequencing technology. Nature *470*, 198–203.

Mariotti-Ferrandiz, E., Pham, H.-P., Dulauroy, S., Gorgette, O., Klatzmann, D., Cazenave, P.-A., Pied, S., and Six, A. (2016). A TCRβ Repertoire Signature Can Predict Experimental Cerebral Malaria. PLoS ONE *11*.

Marolleau, J. P., J. D. Fondell, M. Malissen, J. Trucy, E. Barbier, K. B. Marcu, P. A. Cazenave, et D. Primi. 1988. « The Joining of Germ-Line V Alpha to J Alpha Genes Replaces the Preexisting V Alpha-J Alpha Complexes in a T Cell Receptor Alpha, Beta Positive T Cell Line ». Cell 55 (2): 291-300.Marrack, P., and Kappler, J. (1986). The T cell and its receptor. Sci. Am. *254*, 36–45.

Mason, D. (1998). A very high level of crossreactivity is an essential feature of the T-cell receptor. Immunol. Today *19*, 395–404.

Matsutani, T., Yoshioka, T., Tsuruta, Y., Iwagami, S., and Suzuki, R. (1997). Analysis of TCRAV and TCRBV Repertoires in Healthy Individuals by Microplate Hybridization Assay. Hum. Immunol. *56*, 57–69.

Mora, T., and Walczak, A.M. (2016). Quantifying lymphocyte receptor diversity. bioRxiv 46870.

Moss, P.A., and Bell, J.I. (1995). Sequence analysis of the human alpha beta T-cell receptor CDR3 region. Immunogenetics *42*, 10–18.

Moss, P.A.H., Rosenberg, W.M.C., and Bell, J.I. (1992). The Human T Cell Receptor in Health and Disease. Annu. Rev. Immunol. *10*, 71–96.

Motea, E.A., and Berdis, A.J. (2010). Terminal Deoxynucleotidyl Transferase: The Story of a Misguided DNA Polymerase. Biochim. Biophys. Acta *1804*, 1151–1166.

Murphy, K.P. (2012). Janeway's immunobiology (London: Garland Science).

Murugan, A., Mora, T., Walczak, A.M., and Callan, C.G., Jr (2012). Statistical inference of the generation probability of T-cell receptors from sequence repertoires. Proc. Natl. Acad. Sci. U. S. A. *109*, 16161–16166.

Naumov, Y.N., Naumova, E.N., Hogan, K.T., Selin, L.K., and Gorski, J. (2003). A Fractal Clonotype Distribution in the CD8+ Memory T Cell Repertoire Could Optimize Potential for Immune Responses. J. Immunol. *170*, 3994–4001.

Naylor, K., Li, G., Vallejo, A.N., Lee, W.-W., Koetz, K., Bryl, E., Witkowski, J., Fulbright, J., Weyand, C.M., and Goronzy, J.J. (2005). The influence of age on T cell generation and TCR diversity. J. Immunol. Baltim. Md 1950 *174*, 7446–7452.

Nazarov, V., Pogorelyy, M., Komech, E., Zvyagin, I., Bolotin, D., Shugay, M., Chudakov, D., Lebedev, Y., and Mamedov, I. (2015). tcR: an R package for T cell receptor repertoire advanced data analysis. BMC Bioinformatics *16*, 175.

Nehar-Belaid, D., Courau, T., Dérian, N., Florez, L., Ruocco, M.G., and Klatzmann, D. (2016). Regulatory T Cells Orchestrate Similar Immune Evasion of Fetuses and Tumors in Mice. J. Immunol. Baltim. Md 1950 *196*, 678–690.

Nguyen, P., Ma, J., Pei, D., Obert, C., Cheng, C., and Geiger, T.L. (2011). Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. BMC Genomics *12*, 106.

Nikolich-Zugich, J., Slifka, M.K., and Messaoudi, I. (2004). The many important facets of T-cell repertoire diversity. Nat. Rev. Immunol. *4*, 123–132.

Nobrega, A., Haury, M., Grandien, A., Malanchère, E., Sundblad, A., and Coutinho, A. (1993). Global analysis of antibody repertoires. II. Evidence for specificity, self-selection and the immunological "homunculus" of antibodies in normal serum. Eur. J. Immunol. *23*, 2851–2859.

Nyren, P., Pettersson, B., and Uhlen, M. (1993). Solid Phase DNA Minisequencing by an Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay. Anal. Biochem. *208*, 171–175.

Ogle, B.M., Cascalho, M., Joao, C., Taylor, W., West, L.J., and Platt, J.L. (2003). Direct measurement of lymphocyte receptor diversity. Nucleic Acids Res. *31*, e139–e139.

Oksanen, J. (2011). Multivariate analysis of ecological communities in R: vegan tutorial. R Package Version 2–0.

Pacholczyk, R., Ignatowicz, H., Kraj, P., and Ignatowicz, L. (2006). Origin and T Cell Receptor Diversity of Foxp3+CD4+CD25+ T Cells. Immunity *25*, 249–259.

Paciello, G., Acquaviva, A., Pighi, C., Ferrarini, A., Macii, E., Zamo', A., and Ficarra, E. (2015). VDJSeq-Solver: In Silico V(D)J Recombination Detection Tool. PLoS ONE *10*.

Padovan, E., Casorati, G., Dellabona, P., Meyer, S., Brockhaus, M., and Lanzavecchia, A. (1993). Expression of two T cell receptor alpha chains: dual receptor T cells. Science *262*, 422–424.

Palmer, M.W. (1990). The Estimation of Species Richness by Extrapolation. Ecology *71*, 1195–1198.

Pannetier, C., Cochet, M., Darche, S., Casrouge, A., Zöller, M., and Kourilsky, P. (1993). The sizes of the CDR3 hypervariable regions of the murine T-cell receptor beta chains vary as a function of the recombined germ-line segments. Proc. Natl. Acad. Sci. *90*, 4319–4323.

Pannetier, C., Even, J., and Kourilsky, P. (1995). T-cell repertoire diversity and clonal expansions in normal and clinical samples. Immunol. Today *16*, 176–181.

Pearson, R. (2010). Exploring data in engineering, the sciences, and medicine (Oxford: Oxford Univ. Press).

Petrie, H.T., Livak, F., Schatz, D.G., Strasser, A., Crispe, I.N., and Shortman, K. (1993). Multiple rearrangements in T cell receptor alpha chain genes maximize the production of useful thymocytes. J. Exp. Med. *178*, 615–622.

Petrovc Berglund, J., Mariotti-Ferrandiz, E., Rosmaraki, E., Hall, H., Cazenave, P.A., Six, A., and Höglund, P. (2008). TCR repertoire dynamics in the pancreatic lymph nodes of non-obese diabetic (NOD) mice at the time of disease initiation. Mol. Immunol. *45*, 3059–3064.

Pewe, L.L., Netland, J.M., Heard, S.B., and Perlman, S. (2004). Very diverse CD8 T cell clonotypic responses after virus infections. J. Immunol. Baltim. Md 1950 *172*, 3151–3156.

Pham, H.-P., Dérian, N., Chaara, W., Bellier, B., Klatzmann, D., and Six, A. (2014). A novel strategy for molecular signature discovery based on independent component analysis. Int. J. Data Min. Bioinforma. *9*, 277–304.

Pielou, E.C. (1966). The measurement of diversity in different types of biological collections. J. Theor. Biol. *13*, 131–144.

Plessy, C., Mariotti-Ferrandiz, E., Manabe, R.-I., and Hori, S. (2015). clonotypeR--high throughput analysis of T cell antigen receptor sequences. bioRxiv 28696.

Powrie, F., et D. Mason. 1990. « OX-22high CD4+ T Cells Induce Wasting Disease with Multiple Organ Pathology: Prevention by the OX-22low Subset ». The Journal of Experimental Medicine 172 (6): 1701-8.

Prabakaran, P., Chen, W., Singarayan, M.G., Stewart, C.C., Streaker, E., Feng, Y., and Dimitrov, D.S. (2012). 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 *64*, 337–350.

Qi, Q., Liu, Y., Cheng, Y., Glanville, J., Zhang, D., Lee, J.-Y., Olshen, R.A., Weyand, C.M., Boyd, S.D., and Goronzy, J.J. (2014). Diversity and clonal selection in the human T-cell repertoire. Proc. Natl. Acad. Sci. *111*, 13139–13144.

Quigley, M.F., Greenaway, H.Y., Venturi, V., Lindsay, R., Quinn, K.M., Seder, R.A., Douek, D.C., Davenport, M.P., and Price, D.A. (2010). Convergent recombination shapes the clonotypic landscape of the naïve T-cell repertoire. Proc. Natl. Acad. Sci. U. S. A. *107*, 19414–19419.

Regnault, A., Cumano, A., Vassalli, P., Guy-Grand, D., and Kourilsky, P. (1994). Oligoclonal repertoire of the CD8 alpha alpha and the CD8 alpha beta TCR-alpha/beta murine intestinal intraepithelial T lymphocytes: evidence for the random emergence of T cells. J. Exp. Med. *180*, 1345–1358.

Rempala, G.A., Seweryn, M., and Ignatowicz, L. (2011). Model for Comparative Analysis of Antigen Receptor Repertoires. J. Theor. Biol. *269*, 1–15.

Rényi, A. (1961). On Measures of Entropy and Information. (The Regents of the University of California), p.

Robins, H. (2013). Immunosequencing: applications of immune repertoire deep sequencing. Curr. Opin. Immunol. *25*, 646–652.

Robins, H.S., Campregher, P.V., Srivastava, S.K., Wacher, A., Turtle, C.J., Kahsai, O., Riddell, S.R., Warren, E.H., and Carlson, C.S. (2009). Comprehensive assessment of T-cell receptor -chain diversity in T cells. Blood *114*, 4099–4107.

Robins, H.S., Srivastava, S.K., Campregher, P.V., Turtle, C.J., Andriesen, J., Riddell, S.R., Carlson, C.S., and Warren, E.H. (2010). Overlap and Effective Size of the Human CD8+ T Cell Receptor Repertoire. Sci. Transl. Med. *2*, 47ra64-47ra64.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., and Nyrén, P. (1996). Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. *242*, 84–89.

Ronaghi, M., Uhlén, M., and Nyrén, P. (1998). A sequencing method based on real-time pyrophosphate. Science *281*, 363, 365.

Rosenberg, A.H., Patel, S.S., Johnson, K.A., and Studier, F.W. (1992). Cloning and expression of gene 4 of bacteriophage T7 and creation and analysis of T7 mutants lacking the 4A primase/helicase or the 4B helicase. J. Biol. Chem. *267*, 15005–15012.

Safonova, Y., Lapidus, A., and Lill, J. (2015). IgSimulator: a versatile immunosequencing simulator. Bioinformatics btv326.

Sakaguchi, S., Sakaguchi, N., Asano, M., Itoh, M., and Toda, M. (1995). Immunologic self-tolerance maintained by activated T cells expressing IL-2 receptor alpha-chains (CD25).

Breakdown of a single mechanism of self-tolerance causes various autoimmune diseases. J. Immunol. Baltim. Md 1950 *155*, 1151–1164.

Salaün, J., Bandeira, A., Khazaal, I., Calman, F., Coltey, M., Coutinho, A., and Le Douarin, N.M. (1990). Thymic epithelium tolerizes for histocompatibility antigens. Science *247*, 1471–1474.

Salipante, S.J., Kawashima, T., Rosenthal, C., Hoogestraat, D.R., Cummings, L.A., Sengupta, D.J., Harkins, T.T., Cookson, B.T., and Hoffman, N.G. (2014). Performance Comparison of Illumina and Ion Torrent Next-Generation Sequencing Platforms for 16S rRNA-Based Bacterial Community Profiling. Appl. Environ. Microbiol. *80*, 7583–7591.

Salmela, L., and Schröder, J. (2011). Correcting errors in short reads by multiple alignments. Bioinforma. Oxf. Engl. *27*, 1455–1461.

Sarrabayrouse, G., Bossard, C., Chauvin, J.-M., Jarry, A., Meurette, G., Quévrain, E., Bridonneau, C., Preisser, L., Asehnoune, K., Labarrière, N., et al. (2014). CD4CD8αα Lymphocytes, A Novel Human Regulatory T Cell Subset Induced by Colonic Bacteria and Deficient in Patients with Inflammatory Bowel Disease. PLoS Biol. *12*, e1001833.

Sboner, A., Mu, X.J., Greenbaum, D., Auerbach, R.K., and Gerstein, M.B. (2011). The real cost of sequencing: higher than you think! Genome Biol. *12*, 125.

Schaller, S., Weinberger, J., Jimenez-Heredia, R., Danzer, M., Oberbauer, R., Gabriel, C., and Winkler, S.M. (2015). ImmunExplorer (IMEX): a software framework for diversity and clonality analyses of immunoglobulins and T cell receptors on the basis of IMGT/HighV-QUEST preprocessed NGS data. BMC Bioinformatics *16*, 252.

Schwab, D.J., Nemenman, I., and Mehta, P. (2014). Zipf's law and criticality in multivariate data without fine-tuning. Phys. Rev. Lett. *113*, 68102.

Sebzda, E., Mariathasan, S., Ohteki, T., Jones, R., Bachmann, M.F., and Ohashi, P.S. (1999). Selection of the T cell repertoire. Annu. Rev. Immunol. *17*, 829–874.

Sepúlveda, N., Paulino, C.D., and Carneiro, J. (2010). Estimation of T-cell repertoire diversity and clonal size distribution by Poisson abundance models. J. Immunol. Methods *353*, 124–137.

de Sepúlveda, N.H. dos S. (2009). How is the T-cell repertoire shaped?

Sewell, A.K. (2012). Why must T cells be cross-reactive? *12*, 669–677.

Shannon, C.E. (1948). A Mathematical Theory of Communication. Bell Syst. Tech. J. *27*, 379–423.

Sherwood, A.M., Desmarais, C., Livingston, R.J., Andriesen, J., Haussler, M., Carlson, C.S., and Robins, H. (2011). Deep Sequencing of the Human TCR and TCR Repertoires Suggests that TCR Rearranges After and T Cell Commitment. Sci. Transl. Med. *3*, 90ra61-90ra61.

Shugay, M., Bolotin, D.A., Putintseva, E.V., Pogorelyy, M.V., Mamedov, I.Z., and Chudakov, D.M. (2013). Huge Overlap of Individual TCR Beta Repertoires. Front. Immunol. *4*.

Shugay, M., Britanova, O.V., Merzlyak, E.M., Turchaninova, M.A., Mamedov, I.Z., Tuganbaev, T.R., Bolotin, D.A., Staroverov, D.B., Putintseva, E.V., Plevova, K., et al. (2014). Towards error-free profiling of immune repertoires. Nat. Methods *11*, 653–655.

Singer, A., Adoro, S., and Park, J.-H. (2008). Lineage fate and intense debate: myths, models and mechanisms of CD4- versus CD8-lineage choice. Nat. Rev. Immunol. *8*, 788–801.

Sørensen, T.J. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. (København: I kommission hos E. Munksgaard).

Srivastava, S.K., and Robins, H.S. (2012). Palindromic nucleotide analysis in human T cell receptor rearrangements. PloS One *7*, e52250.

Strimmer, J.H. and K. (2014). entropy: Estimation of Entropy, Mutual Information and Related Quantities.

Su, L.F., and Davis, M.M. (2013). Antiviral memory phenotype T cells in unexposed adults. Immunol. Rev. *255*, 95–109.

Tawfik, D.S., and Griffiths, A.D. (1998). Man-made cell-like compartments for molecular evolution. Nat. Biotechnol. *16*, 652–656.

Tempel, S. (2007). Dynamique des hélitrons dans le genome d'Arabidopsis thaliana : développement de nouvelles stratégies d'analyse des éléments transposables. phdthesis. Université Rennes 1.

Thomas, N., Heather, J., Pollara, G., Simpson, N., Matjeka, T., Shawe-Taylor, J., Noursadeghi, M., and Chain, B. (2013a). The immune system as a biomonitor: explorations in innate and adaptive immunity. Interface Focus *3*.

Thomas, N., Heather, J., Ndifon, W., Shawe-Taylor, J., and Chain, B. (2013b). Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. Bioinforma. Oxf. Engl. *29*, 542–550.

Thomas, N., Best, K., Cinelli, M., Reich-Zeliger, S., Gal, H., Shifrut, E., Madi, A., Friedman, N., Shawe-Taylor, J., and Chain, B. (2014). Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence.

Thomas-Vaslin, V., Altes, H.K., Boer, R.J. de, and Klatzmann, D. (2008). Comprehensive Assessment and Mathematical Modeling of T Cell Population Dynamics and Homeostasis. J. Immunol. *180*, 2240–2250.

Thomas-Vaslin, V., Six, A., Pham, H.-P., Dansokho, C., Chaara, W., Gouritin, B., Bellier, B., and Klatzm, D. (2012). Immunodepression and Immunosuppression During Aging. In Immunosuppression - Role in Health and Diseases, S. Kapur, ed. (InTech), p.

Tipton, C.M., Fucile, C.F., Darce, J., Chida, A., Ichikawa, T., Gregoretti, I., Schieferl, S., Hom, J., Jenks, S., Feldman, R.J., et al. (2015). Diversity, cellular origin and autoreactivity of antibody-

secreting cell population expansions in acute systemic lupus erythematosus. Nat. Immunol. *16*, 755–765.

Tonegawa, S. (1983). Somatic generation of antibody diversity. Nature *302*, 575–581.

Tonegawa, S., Steinberg, C., Dube, S., and Bernardini, A. (1974). Evidence for somatic generation of antibody diversity. Proc. Natl. Acad. Sci. U. S. A. *71*, 4027–4031.

Turchaninova, M.A., Britanova, O.V., Bolotin, D.A., Shugay, M., Putintseva, E.V., Staroverov, D.B., Sharonov, G., Shcherbo, D., Zvyagin, I.V., Mamedov, I.Z., et al. (2013). Pairing of T-cell receptor chains via emulsion PCR. Eur. J. Immunol. *43*, 2507–2515.

Vander Heiden, J.A., Yaari, G., Uduman, M., Stern, J.N.H., O'Connor, K.C., Hafler, D.A., Vigneault, F., and Kleinstein, S.H. (2014). pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. Bioinforma. Oxf. Engl. *30*, 1930–1932.

VanderBorght, A., Van der Aa, A., Geusens, P., Vandevyver, C., Raus, J., and Stinissen, P. (1999). Identification of overrepresented T cell receptor genes in blood and tissue biopsies by PCR-ELISA. J. Immunol. Methods *223*, 47–61.

Venturi, V., Rudd, B.D., and Davenport, M.P. (2013). Specificity, promiscuity, and precursor frequency in immunoreceptors. Curr. Opin. Immunol. *25*, 639–645.

Voelkerding, K.V., Dames, S.A., and Durtschi, J.D. (2009). Next-Generation Sequencing: From Basic Research to Diagnostics. *55*, 641–658.

Vollmers, C., Sit, R.V., Weinstein, J.A., Dekker, C.L., and Quake, S.R. (2013). Genetic measurement of memory B-cell recall using antibody repertoire sequencing. Proc. Natl. Acad. Sci. *110*, 13463–13468.

Wang, J., and Reinherz, E.L. (2002). Structural basis of T cell recognition of peptides bound to MHC molecules. Mol. Immunol. *38*, 1039–1049.

Wang, C., Sanders, C.M., Yang, Q., Schroeder Jr, H.W., Wang, E., Babrzadeh, F., Gharizadeh, B., Myers, R.M., Hudson Jr, J.R., Davis, R.W., et al. (2010). High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. Proc. Natl. Acad. Sci. *107*, 1518–1523.

Warren, R.L., Nelson, B.H., and Holt, R.A. (2009). Profiling model T-cell metagenomes with short reads. Bioinformatics *25*, 458–464.

Warren, R.L., Freeman, J.D., Zeng, T., Choe, G., Munro, S., Moore, R., Webb, J.R., and Holt, R.A. (2011). Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. Genome Res. *21*, 790–797.

Weinstein, J.A., Jiang, N., White, R.A., Fisher, D.S., and Quake, S.R. (2009). High-throughput Sequencing of the Zebrafish Antibody Repertoire. Science *324*, 807–810.

Wilson, D.B., Wilson, D.H., Schroder, K., Pinilla, C., Blondelle, S., Houghten, R.A., and Garcia, K.C. (2004). Specificity and degeneracy of T cells. Mol. Immunol. *40*, 1047–1055.

Wong, P., and Pamer, and E.G. (2003). Cd8 T Cell Responses to Infectious Pathogens. Annu. Rev. Immunol. *21*, 29–70.

Wong, J., Obst, R., Correia-Neves, M., Losyev, G., Mathis, D., and Benoist, C. (2007). Adaptation of TCR repertoires to self-peptides in regulatory and nonregulatory CD4+ T cells. J. Immunol. Baltim. Md 1950 *178*, 7032–7041.

Woodsworth, D.J., Castellarin, M., and Holt, R.A. (2013). Sequence analysis of T-cell repertoires in health and disease. Genome Med. *5*, 98.

Wu, J., Liu, D., Tu, W., Song, W., and Zhao, X. (2015). T-cell receptor diversity is selectively skewed in T-cell populations of patients with Wiskott-Aldrich syndrome. J. Allergy Clin. Immunol. *135*, 209–216.e8.

Wu, Y.-C., Kipling, D., Leong, H.S., Martin, V., Ademokun, A.A., and Dunn-Walters, D.K. (2010). High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. Blood *116*, 1070–1078.

Wucherpfennig, K.W., Allen, P.M., Celada, F., Cohen, I.R., De Boer, R., Garcia, K.C., Goldstein, B., Greenspan, R., Hafler, D., Hodgkin, P., et al. (2007). Polyspecificity of T cell and B cell receptor recognition. Semin. Immunol. *19*, 216–224.

Yassai, M.B., Naumov, Y.N., Naumova, E.N., and Gorski, J. (2009). A clonotype nomenclature for T cell receptors. Immunogenetics *61*, 493–502.

Ye, J., Ma, N., Madden, T.L., and Ostell, J.M. (2013). IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res. *41*, W34–W40.

Zhu, J., and Shevach, E.M. (2014). TCR signaling fuels Treg cell suppressor function. Nat. Immunol. *15*, 1002–1003.

Zhu, D., Kadin, M.E., and Samoszuk, M. (2001). Detection of clonal T-cell receptor-gamma gene rearrangement by PCR/temporal temperature gradient gel electrophoresis. Am. J. Clin. Pathol. *116*, 527–534.

ANNEXES

Annexe 1

Spécifications fonctionnelles du workflow RepSeq

Ce document est un extrait de celui produit à l'attention des personnes susceptibles d'analyser des données RepSeq au laboratoire. L'objectif était de leur décrire les fonctions d'analyse à leur disposition, comment les utiliser et surtout de leur décrire les fichiers de sortie et leur contenu. Ce document est accompagné d'une procédure leur expliquant comment utiliser Rgui et lancer leurs analyses.

Wahiba Chaara

Single TR repertoire exploratory display

To allow a thorough characterization of each individual TR repertoire, a series of 17 files are systematically produced for each TR repertoire. These files combine objective metrics to provide a multilevel exploratory characterization of each repertoire. Each dataset is composed by a collection of TR sequences, annotated with the ID of the given TRV and TRJ genes recombined to generate them. These TR sequences are characterized by a given CDR3 and can be categorized into clonotypes (unique combination of TRV-CDR3-TRJ). *NB: only the peptide sequence of the CDR3 is taking into account here.*

A single function allow you to do so, find below the description of its specificities.

Spl_analysis(SPECIES, CHAIN, NAME, SCORE,CDR3L, FILTERED, SINGLE)

Required R packages

clonotypeR; gplots; ade4; plotrix; vegan; entropy; psych; RColorBrewer; limma; zipfR; stringdist

Mandatory arguments

SPECIES: species of the individuals the data are coming from; specify "mm" for mouse or "hs" for human.

CHAIN: TCR chain targeted for the sequencing; specify *CHAIN* ="A" or "B"

SCORE: logical value indicating whether the input TSV file contains alignment scores (by default SCORE=TRUE)

CDR3L: logical value indicating whether to filter the TR sequences according the length of their CDR3. Some sequences can be identified as TR sequences while the section identified as the CDR3 is very long (more than 20 aa). Yet, the CDR3 length distribution of polyclonal TR repertoire is expected to be centered on 10 aa with an equivalent standard deviation[9,10]. If TRUE any sequences with a CDR3 longer than a given threshold will be discarded (by default CDR3L=FALSE).

FILTERED: logical value indicating whether the TR sequences should be filtered based on their alignment scores; If TRUE, sequences with an alignment score lower than the first quartile of

[9] Pannetier, PNAS, 1993

[10] Collette, JI, 2003

the overall alignment score distribution, are discarded from the analyses (by default A FILTERED=FALSE)

SINGLE: logical value indicating whether to keep the clonotypes observed only once (singletons); If FALSE, every singletons are discarded from the analyses (by default SINGLE=TRUE)

Optional argument

NAME: name the user wants to give to the sample; if not specify, the name file is used.

Details

The workspace must be set in a folder containing two subfolders: input and output. This function when launched will open a Rgui folder browser widget so the user can choose the input file. The output files will be created into the output folder. The produced text files can opened with Microsoft Excel and images are produced in PDF format so it can be modified if needed.

NB: Each of these analyses are performed using independent modules that can be called independently.

Input

TSV file listing every identified TR sequences; If not a *clonotypeR* output, it must be specified for each sequence at least its TRV gene, TRJ gene, CDR3 amino-acid sequence and the ID of the library within which it has been identified.

Outputs

Dataset description

NAME_summary.txt

Tab-separate text file summarizing a series of descriptive statistics.

- Number of TR sequences
- % of productive sequences
- Number of TRV genes
- Number of V-J combinations
- Number of unique CDR3 nucleotide sequences
- Number of unique CDR3 amino-acid sequences
- Number of clonotypes (V-CDR3aa-J)

- Clonality (= Nb CDR3aa / Nb TR sequences) - [0; 1]
- D50: proportion of clonotypes contributing to a 50% cumulative frequency – [0; 1]
- Diversity: exponential of clonotype Shannon entropy - [0; +∞]
- Pielou evenness: the more evenly expressed the clonotypes are, the closest it is to 1 – [0; 1]
- Simpson index: Diversity index – [0; 1]
- Hill index: Normalised diversity index – [0; 1]

NAME_CL.txt

Contingency table summarizing clonotype counts across libraries/samples included in the input TSV file.

NAME_CL2.txt

List of each clonotypes identified by libraries/samples with their count.

Clonotypic description

NB: If several libraries/samples by file, the individual outputs will be produced for each of them.

NAME_Distr.pdf

Clonotypes are ranked decreasingly according to their abundance and plotted: rank (log-scale) vs. count.

NAME_Spectrum.pdf

Plot of the spectrum of clonotype occurrence: occurrence (log-scale) vs. clonotype count; Zipf-α and Zipf-B parameters are estimated based on this spectrum.

NAME_Occ.pdf

Pie chart of the clonotype occurrence distribution.

NAME_CumSum.pdf

Clonotypes are ranked decreasingly according to their abundance and their cumulative frequencies plotted: rank vs. cumulative frequency.

NAME_Rarefaction.pdf

Rarefaction curve is built by drawing randomly an increasing number of sequences and counting the number of unique clonotype within it - interval of step sample size= 10 000 :

number of drawn sequences vs. number of observed clonotypes; *adapted from vegan::rarecurve function.*

NAME_Renyi.pdf

A diversity profile is built by calculating the value of Renyi entropy (Ha) using the observed clonotype distribution for an increasing value of α factor: α vs. exp(Ha).

Topological description

NB: If several libraries/samples by file, the individual outputs will be produced for each of them.

NAME_VJ.txt

Contingency table summarizing TRVJ combination abundance

NAME_VH.txt

Contingency table summarizing clonotypes diversity (exponential of Shannon entropy H) within each TRVJ combination.

NAME_VJplot.pdf

Two series of two independent barplots are produced representing TRV and TRJ expressions. The first series (titled “distribution”) is plotted based on the gene frequencies among TR sequences while the second (titled “usage”) is looking at the gene frequencies among clonotypes.

NAME_VJheat.pdf

Four heatmap representing: the abundance, the overall frequency (titled “TRVJ distribution”) and the clonotypic diversity (titled “H by TRVJ”) of each TRVJ combination, and the TRJ frequency (titled “TRJ by TRV”) within each TRV family

NAME_Spectratypes.pdf

Relative CDR3 length frequencies are calculated between clonotypes expressing each of the TRV. Their distribution are plotted b TRV family.

The previous function allows the analysis of individual files. In order to analyse a batch of files, use the following function:

Spl_analysisBatch(SPECIES, CHAIN, NAME, SCORE,CDR3L, FILTERED, SINGLE)

Details

Exactly the same arguments and output except that instead of selecting an input file, this version expect to be indicated a folder path using the Rgui folder browser widget. Outputs will be created for every input files contain in this folder.

TR repertoire comparison

Using the following function, the user can compare samples, or groups of samples, in order to

(i) Assess the homogeneity of biological replicates within a priori biological groups and

(ii) Identify collective behaviours that can be related to the conditions of interest.

Spl_analysisComp(SPECIES, CHAIN, SCORE, CDR3L, FILTERED, SINGLE, NAME, FAC, SAMPLING, S)

Up to 23 analyses are systematically performed across all the samples to compare their topology, assess their level of similarity and identify patterns that could be relevant to discriminate them.

Required R packages

clonotypeR; gplots; ade4; plotrix; vegan; entropy; psych; RColorBrewer; limma; zipfR; stringdist

Mandatory arguments

SPECIES, CHAIN, SCORE, CDR3L, FILTERED and SINGLE arguments are the same than in the individual functions previously.

NAME: name the user wants to give to the comparison

Optional arguments

FAC: a list partitioning the libraries/datasets in groups; factor list (by default FAC =NULL)

SAMPLING: logical value indicating whether a bootstrapped sampling must be performed on the datasets (by default Sampling = FALSE)

S: Integer specifying at which depth the sampling must be performed if Sampling = TRUE; (by default S =NULL)

Details

The workspace must be set in a folder containing two subfolders: input and output. The function when launched will open a folder browser widget to choose the input folder interactively. The input file can be either one file concatenating data from several libraries or as many files than libraries you want to compare. The output files of their comparison will be created into the output folder.

If SAMPLING = TRUE but S=NULL, the sampling size S will be equal to that of the smallest dataset. The sampling is performed as followed for each library/dataset:
100 random draw of S sequences are performed and unique clonotypes are listed and counted. After the 100 iterations, clonotypes observed at least 50 out of the 100 are selected and their count is averaged across the number of iterations they were observed in.

Outputs

Dataset description

NAME_summary.txt

Tab-separate text file summarizing a series of descriptive statistics.

- Number of TR sequences
- % of productive sequences
- Number of TRV genes
- Number of V-J combinations
- Number of unique CDR3 nucleotide sequences
- Number of unique CDR3 amino-acid sequences
- Number of clonotypes (V-CDR3aa-J)
- Clonality (= Nb CDR3aa / Nb TR sequences) - [0; 1]
- D50: proportion of clonotypes contributing to a 50% cumulative frequency – [0; 1]
- Diversity: exponential of clonotype Shannon entropy - [0; +∞]
- Pielou evenness: the more evenly expressed the clonotypes are, the closest it is to 1 – [0; 1]
- Simpson index: Diversity index – [0; 1]
- Hill index: Normalised diversity index – [0; 1]

NAME_CL.txt

Contingency table summarizing clonotype counts across libraries/samples.

NAME_CL2.txt

List of each clonotypes identified by libraries/samples with their count.

Composition comparison

NB: If several libraries/samples by file, the individual outputs will be produced for each of them.

NAME_Distr.pdf

Clonotypes of each library are ranked decreasingly according to their abundance and plotted: rank (log-scale) vs. count.

NAME_Spectrum.pdf

Plot of the spectrum of clonotype occurrence within each library: occurrence (log-scale) vs. clonotype count; Zipf-α and Zipf-B parameters are estimated based on this spectrum;

NAME_Occ.pdf

Stacked barplot of the clonotype occurrence distribution of each library.

NAME_CumSum.pdf

Clonotypes of each library are ranked decreasingly according to their abundance and their cumulative frequencies plotted: rank vs. cumulative frequency.

NAME_Rarefaction.pdf

Rarefaction curves of each library is built by drawing randomly an increasing number of sequences and counting the number of unique clonotype within it - interval of step sample size= 10 000 : number of drawn sequences vs. number of observed clonotypes; *adapted from vegan::rarecurve function.*

NAME_Renyi.pdf

A diversity profile is built for each library by calculating the value of Renyi entropy (Ha) using the observed clonotype distribution for an increasing value of α factor: α vs. exp(Ha).

Topological Comparison

NAME _VJ.txt

Contingency table of all TRVJ combination count across all libraries/samples

NAME _VJp.pdf

Heatmap with hierarchical clustering (agglomeration method = “complete”, distance= “euclidean”) of all libraries/samples based on their TRVJ combination frequency (“TRVJ distribution”) and detection (“TRVJ usage”) profiles.

NAME _JbyV.pdf

Heatmap with hierarchical clustering (agglomeration method = “complete”, distance= “euclidean”) of all libraries/samples based on their TRJ family distributions within each TRV families. If FAC is specified, a PCA projection is also plotted using the factor list to discriminate the sample groups according to these profiles.

NAME_VJH.pdf

Two heatmaps with hierarchical clustering (agglomeration method = “complete”, distance= “euclidean”) of all libraries/samples based on exp(Shannon entropy) of their clonotypes within each TRV family and within each TRVJ combination. If FAC is specified, two PCA projections are also plotted using the factor list to discriminate the sample groups according to these profiles.

NAME_VJ_MH.pdf

Heatmap with hierarchical clustering (agglomeration method = “complete”, distance= “euclidean”) of all libraries/samples based their Morisita-Horn index using TRVJ combination frequency distributions.

NAME_Spectratypes.pdf

Heatmap with hierarchical clustering (agglomeration method = “complete”, distance= “euclidean”) of all libraries/samples based their TRV CDR3 length profiles. If FAC is specified, a PCA projection is also plotted using the factor list to discriminate the sample groups according to these profiles.

Description of the overlap between libraries/samples

NAME_sharing.pdf

Pie plot of private (unshared) and shared clonotypes proportions across the union of the datasets.

NAME_SharedCL.txt

Contingency table of shared clonotype counts across all libraries/samples

NAME_shared_Hist.pdf

Stack bars of shared clonotype frequencies within each libraries/samples

NAME_MH.txt

Similarity matrix providing Morisita-Horn index value between all the libraries/samples

NAME_Shared_MH.pdf

Heatmap with hierarchical clustering (agglomeration method = "complete", distance= "euclidean") of all libraries/samples based their Morisita-Horn index using shared clonotype frequency distributions.

Optional (if less than 6 libraries/samples compared)

NAME_venn.pdf

Venn diagram between clonotype observed within each library/sample

---

Description of the most predominant clonotypes

NAME_1MPC.pdf

Stacked cumulative frequencies of the 1% most predominant clonotypes of each libraries/samples

NAME_shMPC_Hist.pdf

Stacked cumulative frequencies of the 1% most predominant clonotypes shared by all the libraries/samples.

NAME_1MPC_MH.txt

Similarity matrix of the Morisita-Horn index value between all the libraries/samples using based on their most predominant clonotype frequencies.

NAME_1MPC_MH.pdf

Heatmap with hierarchical clustering (agglomeration method = “complete”, distance= “euclidean”) of all libraries/samples based their Morisita-Horn index using shared most predominant clonotypes’ frequencies.

Optional (if less than 6 libraries/samples compared)

NAME_vennMPC.pdf

Venn diagram between the 1% most predominant clonotypes of each libraries/samples

---

A second version of the previous function allows you to compare data based on a contingency table (CL2 table like)

Spl_analysisComp2(SPECIES, CHAIN, SCORE, CDR3L, FILTERED, SINGLE, NAME, FAC, SAMPLING, S)

Keep in mind that some precautions must be taken into consideration before comparing your data: Are the sizes of the datasets comparable? Does the sequencing depth allow a correct representation of the diversity to be observed? Do the biological replicates show homogeneous repertoire topology?

Use the descriptive statistics and the comparison of the occurrence distributions to determine whether and how to compare your datasets. You can use the normalisation method to adjust your comparisons.

In case of a large proportion of singletons and to reduce the impact of erroneous sequences on the results, datasets can be processed before any further analysis. To do so, use the following function:

CollapsedTSV(SPECIES,CHAIN, SCORE)

Required R packages

clonotypeR; stringdist

Mandatory arguments

SPECIES: species of the individuals the data are coming from; specify "mm" for mouse or "hs" for human.

CHAIN: TCR chain targeted for the sequencing; specify *CHAIN* ="A" or "B"

SCORE: logical value indicating whether the input TSV file contains alignment scores (by default SCORE=TRUE)

Details

This function when launched will open a Rgui folder browser widget so the user can choose the input file.

Input

TSV file listing every identified TR sequences; If not a *clonotypeR* output, it must be specified for each sequence at least its TRV gene, TRJ gene, CDR3 amino-acid sequence and the ID of the library within which it has been identified.

Outputs

The output files will be created into the input folder. It will be a TSV file similar to a classical *clonotypeR* output. If there are data from several libraries in the input file, there will be as many output files created.

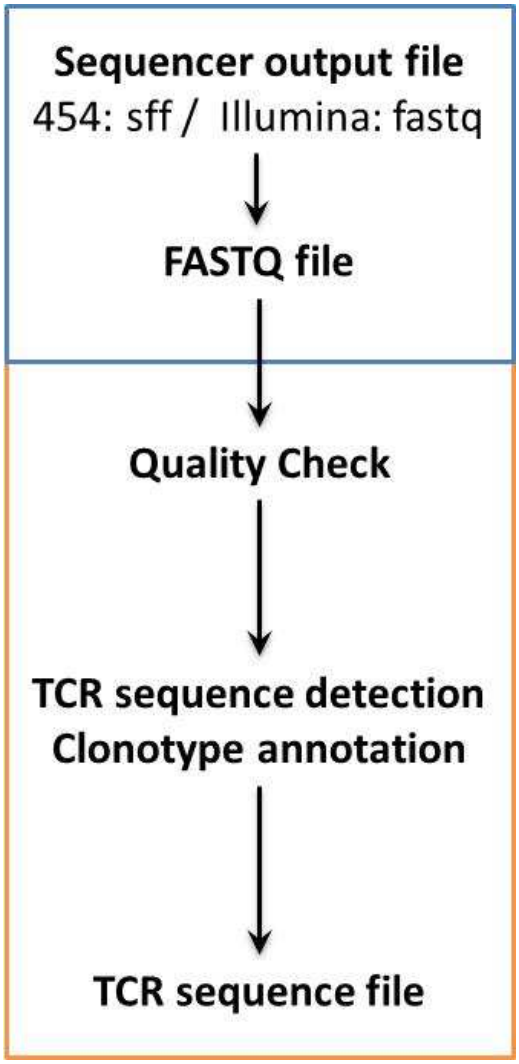
Annexe 2

Gestion des données RepSeq

Ce document a été produit à l'attention des expérimentateurs, producteurs de données RepSeq du laboratoire. L'objectif était de les familiariser avec la terminologie relative à la manipulation de ces données et de leur expliciter les étapes de traitement par lesquels elles devaient passer. De plus, ce document leur décrit de quelle manière stocker et renseigner leurs données d'après la procédure que j'ai mise en place.

Wahiba Chaara

TCR RepSeq data processing workflow



RAW DATA

According to the NGS sequencing technology used, the raw data output file will differ. **SFF** (Standard flowgram format) files are produced by 454 Life Sciences sequencer. They contain information about the flowgram, the called sequence, their qualities and the recommended quality and adaptor clipping. The sequence data can be extracted into a more usable format: fastq format. **FASTQ** format allows to store the sequence and their quality scores (QS) in the same file. The Phred quality score varies typically from 0 to 40 and is encoded in Sanger format (Phred+33). Each read has a unique ID. Illumina sequencer outputs are fastq files.

Quality Check logs are produced to monitor QS distribution along read positions and read length distributions.

PROCESSED DATA

The fastq files are then processed to **identify TCR sequences** among fastq file reads. Each identified TCR sequence is annotated by a series of attributes that can vary according to the tool describing at least the V gene, J gene and the CDR3 nt/aa sequences. This TCR sequence inventory is contained in a file.

RepSeq raw data processing

Three kind of raw data may be encountered.

1) Raw data have been produced by 454 Life Sciences sequencer
   SFF file must be converted in a **fastq** format.

   => see the dedicated procedure.

2) Raw data have been produced by Illumina pair-end sequencer and provided as two paired **fastq** files. These files must be assembled in a single fastq file for the next step of the

workflow.

=> To perform this step, follow the instructions available in the data storage server in the 'Tools' section.

3) Raw data have been produced by Illumina (pair-end) sequencer and provided as a unique (merged) fastq file.

RepSeq raw data quality check

Each fastq file can be parsed to extract:

- read length distribution
- nucleotide composition per position across all reads
- quality score distribution per position across all reads
- fastq quality statistics per position

=> To perform this step, follow the instructions and the dedicated script available in the Linux virtual machine to use FASTX-Toolkit.

RepSeq raw data demultiplexing

A sequencing experiment can be performed according to two scenario:

1) A single library/sample is sequenced leading to a homogeneous fastq file which can be directly processed for data extraction.
2) Several libraries (identified by specific barcodes) are mixed for the sequencing. The produced fastq must be split into as many fastq files as libraries mixed.
   => see the dedicated script on the Linux virtual machine (tambourin)

TCR sequence identification

Individual fastq files are parsed for the identification of the TCR sequences. According to the tool used, the algorithm can differ in particular on the handling of the quality scores. The output file contains a species table that will be used for the TCR repertoire analyses.

=> To perform this step, follow the instructions and the dedicated script available in the Linux virtual machine to use clonotypeR toolkit. Each fastq file lead to a TSV file describing each TCR sequence by attributes such as library ID, V gene, J gene, alignment statistics, corresponding read ID, CDR3 nt sequence and its quality scores. The library ID allow for the individual TSV files to be merged for further analyses.

RepSeq data storage

RepSeq data are stored in the data storage server in the 'RepSeq' section. This section is organized in three parts.

RS_Data folder

It contains the sequencing run data according to the following hierarchy:

RS_Data	UserIDa	@raw	UserIDa001
			UserIDa002
			…
			UserIDa###
		UserIDa001	fastQ
			QC
			TSV
			miTCR
		UserIDa002	fastQ
			QC
			TSV
			miTCR
		…	
	UserIDb	@raw	UserIDb001
			UserIDb002
			…
			UserIDb###
		UserIDb001	fastQ
			QC
			TSV
		UserIDb002	fastQ
			QC
			TSV
		…	
	…		

A main folder is created for each experimenter using its initials or its active directory login (provided by the system administrator); it can be a service provider name such as iREPERTOIRE if appropriate.

An experimenter folder contain

i) a subfolder for each sequencing run (s)he performed, these runs being named incrementally using the nomenclature ID###

ii) ii) a @rawdata folder itself also containing as many folders as runs performed. Each sequencing experiment performed by an experimenter will have a subfolder at the root of his data folder and in its @rawdata subfolder.

The sff or paired/multiplexed fastq files are stored in the @rawdata subfolder. In case of multiplexed fastq, the txt file containing the barcode list will be joined. The converted/assembled/demultiplexed fastq files are in the fastQ folder. The Quality Check files produced for each fastq present in the fastQ folder are put in the QC folder. Finally, the TSV files obtain after clonotypeR TCR sequence extraction are stored in the TSV folder.

Eventually other annotation tools could be used such as miTCR. In that case, an additional folder could be found to store the outputs.

Each experimenter run must contain a txt file describing the availability of each data file with an explanation if missing.

Example: *ID001_README.txt file* could contain

Run ID: ID001

raw data type: fastq

paired: OK

multiplexed: NO; not needed

individual fastq: OK

individual QC: NO; to be performed

individual TSV: OK

At the root of the RS_Data folder is also available a locked version of the Freezing file gathering all the metadata regarding the data production of each project. Each experimenter must when he performs a new experiment document this file. *See below the Freezing file description.*

RS_Analysis folder

Data produced by several experimenters can be used for a given project. Thus, this folder allows gathering the needed processed data to perform the analyses for a given project.

A folder named after the project is created and contain an “input” folder in which the TSV files of all the samples to analyse can be gathered. An “output” folder contains the results of the analyses. This project folder could be used as an R workspace.

Tools

This folder contains tools related to RepSeq data processing: the perl script used the conversion of sff to fastq files, miTCR annotation tool, and FLASH assembler. Each of them is associated with a script explaining how to execute it.

RepSeq Freezing file notice

The Freezing file is organised around 4 main tabs.

Organism

This tab is used for the description of individual mice or pools of mice involved in an *experimental project* and *experiment (defined by an ID and a date).* Each organism can be described by its *ID*, *gender*, *date of birth*, *strain* and an *age at inclusion in the experiment, its experimental group* and an eventual *treatment* or *infection.* A unique ID is assigned each organism.

Sample

For each organism described in the previous tab, a list of harvested samples can be described. They are linked to the organism of origin using the *organisms unique ID* insuring their traceability. Each sample is described by the *organ*, the *cell type* eventual *in vitro treatment* and when appropriate the *cell sorting* strategy and the *number of cells* sorted. This description is made using a defined vocabulary (see 'Code' tab). The same sample can be distributed into several aliquots. A unique ID is assigned to each sample based on the corresponding organisms ID.

Aliquots

Each aliquot processed is described according to the experiment workflow it went through. For a RepSeq experiment, an aliquot will be used for *RNA extraction* followed by *cDNA production* for the *library* preparation. Each of these steps must be described in this tab according a list of defined features. Some checking columns allow the monitoring of aliquot status.

RepSeq

This tab is composed by two main sections. The first describes the eventual *library pooling* and the *sequencing* performed. The same aliquot can be sequenced several times according to two

conditions: either a new library (from the same RNA) is produced and sequenced or the same library is resequenced. To distinguish these two cases, a '*Protocol*' features is coded in this tab. The second section is dedicated to the data produced summarising the number of *reads* obtained, the *annotation* protocol used, the number of *TCR sequences* obtained and a list of associated features allowing to decide whether the dataset is kept or discarded for the analysis. Each dataset is identified by a unique ID summarising the organism, the sample, the aliquot, the sequencing protocol and the run ID it comes from.

Each section of each tab includes the description of the process and the list of experimenter involved.

Annexe 3

Stratégie de préparation des librairies

Wahiba Chaara

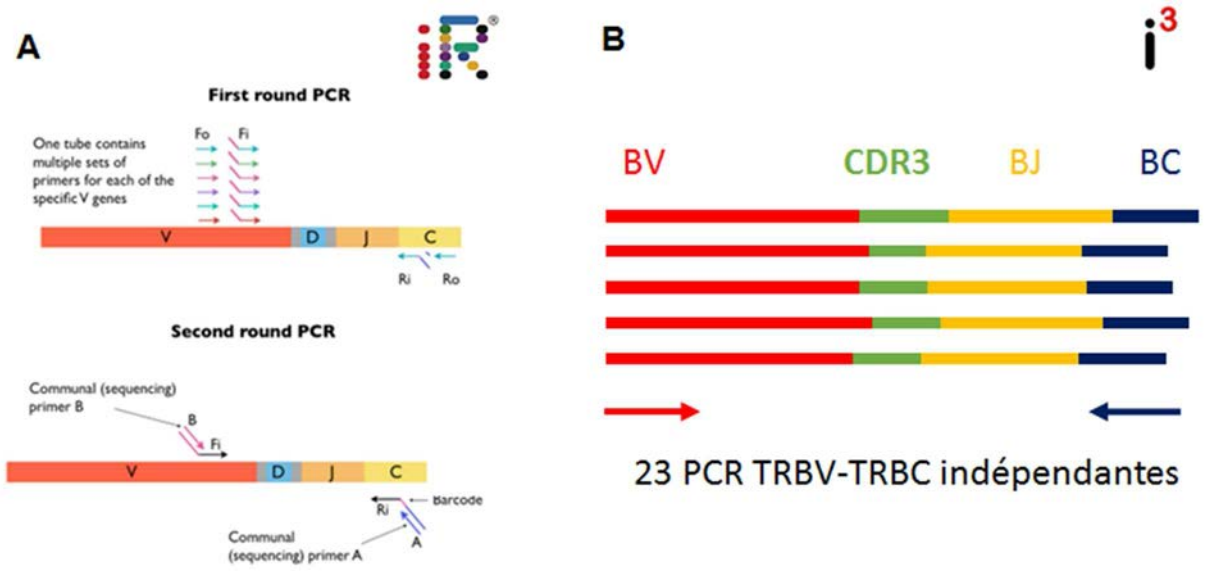


Figure 1: Stratégies de preparation des librairies

La Figure 1A résume la procédure de PCR utilisée par iREPERTOIRE pour la préparation de leur librairies de séquençage (Patent No. 7,999,092). Ce protocole consiste en deux étapes : Au cours du premier cycle de PCR, des amorces spécifiques du gène emboîtées ciblant chacun des gènes TRV et TRC sont utilisés. Les amorces Fo (forward-out) et Fi (forward-in) sont ciblent une portion de séquences des gènes TRV. Les amorces inverses, Ro (reverse-out) et Ri (reverse-in), sont situées au début des gènes TRC. Les amorces Fi et Ri incluent respectivement les adaptateurs B et A permettant le passage sur séquenceur Illumina MiSeq (pour les échantillons *TriPoD_06*) HiSeq (pour les échantillons *TriPoD_38_1070*).

La Figure 1B résume la procédure de PCR utilisée au laboratoire pour la préparation des librairies de séquençage des échantillons *TriPoD_38_1070*. L'ARN est divisé en 23 aliquotes de volume et concentration équivalente auxquels sont ajoutés un couple d'amorces ciblant la séquence d'un des gènes TRBV et celle des gènes TRC. Les vingt-trois réactions de PCR ont lieu en parallèle puis les produits de PCR sont mélangés en condition équimolaire pour constituer la librairie.

Annexe 4

Descriptif des échantillons

Wahiba Chaara

TABLEAU RECAPITULATIF DES ECHANTILLONS

Exp#	Exp. Group	Exp_Group Description	S_ID	Cell Sample	A_Cell#	S_purity	S_Aliquot#
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_215	SPL-SP-amTregs	398 779	99,7	1
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_216	SPL-SP-Teff	13 784 576	99,4	4
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_217	SPL-SP-nTregs	997 161	99,1	1
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_218	SPL-SP-CD8	5 732 322	99,4	2
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_219	BLN-LY-amTregs	58 941	97,0	1
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_220	BLN-LY-Teff	5 842 026	98,0	2
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_221	BLN-LY-nTregs	215 200	99,0	1
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_222	BLN-LY-CD8	1 668 066	97,0	1
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_223	ILN-LY-amTregs	127 121	99,2	1
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_224	ILN-LY-Teff	10 982 735	99,9	4
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_225	ILN-LY-nTregs	453 538	98,7	1
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_226	ILN-LY-CD8	3 504 576	99,3	1
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_227	MLN-LY-amTregs	330 182	99,5	1
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_228	MLN-LY-Teff	11 004 127	99,4	4
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_229	MLN-LY-nTregs	414 702	98,3	1
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_230	MLN-LY-CD8	3 609 587	99,6	1
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_231	PALN-LY-amTregs	61 567	98,0	1
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_232	PALN-LY-Teff	2 835 094	98,0	1
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_233	PALN-LY-nTregs	102 853	99,0	1
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_234	PALN-LY-CD8	1 085 341	98,0	1
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_235	PLN-LY-amTregs	124 668	95,0	1
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_236	PLN-LY-Teff	4 823 482	99,0	2
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_237	PLN-LY-nTregs	79 900	92,0	1
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_238	PLN-LY-CD8	1 843 931	97,0	1
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_239	RLN-LY-amTregs	27 157	98,0	1
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_240	RLN-LY-Teff	1 279 275	99,0	1
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_241	RLN-LY-nTregs	64 089	95,0	1
TriPoD-0006	NOD_YgM	pool of 9 NOD foxp3 GFP males 9 wks	06_242	RLN-LY-CD8	471 734	97,0	1
TriPoD-0038	B6_OldM	pool of 7 males C57BL/6 foxp3 GFP de 24-26 wks	38_1070	SPL-SP-Teff	25 812 208	99,8	6

Exp# : Identifiant unique de l'expérience ; **Exp. Group** : Code du groupe expérimental ; **Exp_Group Description** : Explication du code

S_ID : Identifiant de l'échantillon ; **Cell_Sample** : Code de la population cellulaire triée ; **A_Cell#** : Nombre de cellules triées ; **S_purity** : pureté du tri ; **S_Aliquot#** : Nombre d'aliquotes cellulaires pour l'échantillon.

TABLEAU RECAPITULATIF DES ALIQUOTES

A_ID_Std	RNA_Ext_Method	Protocol	Lib_protocol	Target genes	Sequencer	Run_ID	Sequencing direction	Read number
TriPoD_06_215_1	Trizol/Rneasy	21	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	563 275
TriPoD_06_216_1	Trizol/Rneasy	21	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	930 127
TriPoD_06_217_1	Trizol/Rneasy	21	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	704 448
TriPoD_06_218_1	Trizol/Rneasy	21	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	603 001
TriPoD_06_219_1	Trizol/Rneasy	21	V-J multiplex	TRB	MiSeq Illumina	iREP001	pair-end	1 497 296
TriPoD_06_220_1	Trizol/Rneasy	22	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	1 461 147
TriPoD_06_221_1	Trizol/Rneasy	22	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	1 476 903
TriPoD_06_222_1	Trizol/Rneasy	22	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	1 403 890
TriPoD_06_223_1	Trizol/Rneasy	21	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	300 602
TriPoD_06_224_1	Trizol/Rneasy	21	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	298 828
TriPoD_06_225_1	Trizol/Rneasy	21	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	344 430
TriPoD_06_226_1	Trizol/Rneasy	21	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	448 660
TriPoD_06_227_1	Trizol/Rneasy	21	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	470 007
TriPoD_06_228_1	Trizol/Rneasy	21	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	535 644
TriPoD_06_229_1	Trizol/Rneasy	21	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	2 003 527
TriPoD_06_230_1	Trizol/Rneasy	21	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	2 218 927
TriPoD_06_231_1	Trizol/Rneasy	21	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	1 420 303
TriPoD_06_232_1	Trizol/Rneasy	21	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	1 282 712
TriPoD_06_233_1	Trizol/Rneasy	21	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	988 206
TriPoD_06_234_1	Trizol/Rneasy	21	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	1 002 589
TriPoD_06_235_1	Trizol/Rneasy	22	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	1 313 631
TriPoD_06_236_1	Trizol/Rneasy	22	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	1 374 740
TriPoD_06_237_1	Trizol/Rneasy	22	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	1 144 949
TriPoD_06_238_1	Trizol/Rneasy	22	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	1 199 147
TriPoD_06_239_1	Trizol/Rneasy	21	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	927 320
TriPoD_06_240_1	Trizol/Rneasy	21	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	1 019 509
TriPoD_06_241_1	Trizol/Rneasy	21	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	463 538
TriPoD_06_242_1	Trizol/Rneasy	21	V-J multiplex	TRB	MiSeq Illumina	iREP002	pair-end	555 074
TripoD_38_1070_5	RNeasy	11	V-J multiplex	TRB	HiSeq Illumina	iREP004	pair-end	7 345 806
TripoD_38_1070_5	RNeasy	21	V-J multiplex	TRB	HiSeq Illumina	iREP004	pair-end	9 178 548
TripoD_38_1070_5	RNeasy	31	V-J multiplex	TRB	HiSeq Illumina	iREP004	pair-end	8 469 835
TripoD_38_1070_6	RNeasy	11	V-C I3	TRB	454GSJr	VQ018	pair-end	145 027
TripoD_38_1070_6	RNeasy	21	V-C I3	TRB	454GSJr	VQ021	pair-end	187 864
TripoD_38_1070_6	RNeasy	31	V-C I3	TRB	454GSJr	VQ022	pair-end	138 214

A_ID_Std# : Identifiant unique de l'aliquote ;

RNA_Ext_Method : Méthode d'extraction d'ARN ;

Protocol : Code 2-bit résumant le protocole de séquençage : si un même aliquote est séquencé plusieurs fois, le 1er bit identifie l'aliquote d'ARN si le reséquençage s'est fait à partir de sous-aliquote d'ARN différents, le 2nd identifie l'aliquote d'ADNc si le reséquençage s'est fait à partir de sous-aliquote d'ADNc différents ;

Sequencer : Modèle du séquenceur

Run_ID : Identifiant donné au passage sur séquenceur ;

Read_Number : Nombre de séquences listées dans les fichiers fastq.

TABLEAU RECAPITULATIF DES JEUX DE DONNEES

RS_ID_Std	Reference	Annotation	TR sequence Nb	Nb TR/Reads	%productive	(V-CDR3-J)	TRV Nb	TRJ Nb	CDR3_nt	CDR3_aa
TriPoD_06_215_1_21_iREP002	GeneBank	clonotypeR	511 110	0,91	95	16 064	21	12	16 064	23 788
TriPoD_06_216_1_21_iREP002	GeneBank	clonotypeR	842 053	0,91	95	10 345	20	12	10 345	17 549
TriPoD_06_217_1_21_iREP002	GeneBank	clonotypeR	641 660	0,91	95	30 634	21	13	30 634	40 937
TriPoD_06_218_1_21_iREP002	GeneBank	clonotypeR	550 352	0,91	95	63 169	21	13	63 169	74 554
TriPoD_06_219_1_21_iREP001	GeneBank	clonotypeR	1 237 910	0,83	96	8 664	17	12	8 664	15 053
TriPoD_06_220_1_22_iREP002	GeneBank	clonotypeR	1 230 050	0,84	95	26 552	20	12	26 552	45 659
TriPoD_06_221_1_22_iREP002	GeneBank	clonotypeR	1 228 093	0,83	96	15 713	22	12	15 713	28 476
TriPoD_06_222_1_22_iREP002	GeneBank	clonotypeR	1 254 125	0,89	96	73 356	22	14	73 356	95 997
TriPoD_06_223_1_21_iREP002	GeneBank	clonotypeR	271 238	0,90	95	11 077	21	12	11 077	15 580
TriPoD_06_224_1_21_iREP002	GeneBank	clonotypeR	275 107	0,92	94	138 654	20	13	138 654	163 193
TriPoD_06_225_1_21_iREP002	GeneBank	clonotypeR	313 383	0,91	94	24 826	21	13	24 826	30 240
TriPoD_06_226_1_21_iREP002	GeneBank	clonotypeR	412 054	0,92	95	101 665	21	13	101 665	115 709
TriPoD_06_227_1_21_iREP002	GeneBank	clonotypeR	430 073	0,92	96	15 377	21	12	15 377	21 577
TriPoD_06_228_1_21_iREP002	GeneBank	clonotypeR	498 369	0,93	95	131 906	21	14	131 906	153 947
TriPoD_06_229_1_21_iREP002	GeneBank	clonotypeR	1 784 007	0,89	96	13 291	22	12	13 291	22 408
TriPoD_06_230_1_21_iREP002	GeneBank	clonotypeR	2 004 343	0,90	95	36 908	22	13	36 908	64 965
TriPoD_06_231_1_21_iREP002	GeneBank	clonotypeR	1 288 347	0,91	96	25 824	22	12	25 824	44 334
TriPoD_06_232_1_21_iREP002	GeneBank	clonotypeR	1 179 586	0,92	95	267 731	22	12	267 731	330 594
TriPoD_06_233_1_21_iREP002	GeneBank	clonotypeR	897 674	0,91	95	25 251	21	12	25 251	39 160
TriPoD_06_234_1_21_iREP002	GeneBank	clonotypeR	916 537	0,91	95	56 340	21	13	56 340	71 816
TriPoD_06_235_1_22_iREP002	GeneBank	clonotypeR	1 108 651	0,84	92	10 811	20	12	10 811	18 169
TriPoD_06_236_1_22_iREP002	GeneBank	clonotypeR	1 163 122	0,85	95	51 053	21	12	51 053	71 400
TriPoD_06_237_1_22_iREP002	GeneBank	clonotypeR	956 180	0,84	94	11 150	21	12	11 150	19 685
TriPoD_06_238_1_22_iREP002	GeneBank	clonotypeR	1 064 515	0,89	95	44 416	21	14	44 416	60 946
TriPoD_06_239_1_21_iREP002	GeneBank	clonotypeR	829 235	0,89	97	11 872	20	12	11 872	19 282
TriPoD_06_240_1_21_iREP002	GeneBank	clonotypeR	930 807	0,91	94	72 839	22	12	72 839	91 372
TriPoD_06_241_1_21_iREP002	GeneBank	clonotypeR	421 689	0,91	95	7 408	21	12	7 408	12 487
TriPoD_06_242_1_21_iREP002	GeneBank	clonotypeR	494 540	0,89	95	10 924	20	12	10 924	18 592
TripoD_38_1070_5_11_iREP004	GeneBank	clonotypeR	6 836 292	0,93	95	136 130	22	13	221 716	119 738
TripoD_38_1070_5_21_iREP004	GeneBank	clonotypeR	5 732 884	0,62	94	128 139	22	12	204 699	113 268
TripoD_38_1070_5_31_iREP004	GeneBank	clonotypeR	5 290 615	0,62	94	127 431	22	13	199 304	112 656
TripoD_38_1070_6_11_VQ018	GeneBank	clonotypeR	72 828	0,50	77	26 326	22	12	27 654	25 071
TripoD_38_1070_6_21_VQ021	GeneBank	clonotypeR	109 508	0,58	80	34 075	22	12	36 224	32 329
TripoD_38_1070_6_31_VQ022	GeneBank	clonotypeR	74 182	0,54	81	26 249	22	12	27 620	24 987

RS_ID_Std: Identifiant unique du jeu de données ;

Reference : Source des séquences de référence utilisée pour l'annotation ;

Annotation : outils d'annotation ;

TR sequence nb : Nombre de séquences TR identifiées ;

Nb TR/Reads: Ratio du nombre de séquences TR sur le nombre de séquences total.

% productive: % de séquences TR en phase ouverte de lecture

(V-CDR3-J) : Nombre de clonotypes

TRV Nb : Nombre de gènes TRV ;

TRJ Nb : Nombre de gènes TRJ ;

CDR3_nt : Nombre de séquences CDR3 nucléotidiques uniques

CDR3_aa : Nombre de séquences CDR3 peptidiques uniques

Annexe 5

Revue

Wahiba Chaara

Annexe 6

Article

frontiers in
IMMUNOLOGY

REVIEW ARTICLE
published: 27 November 2013
doi: 10.3389/fimmu.2013.00413

The past, present, and future of immune repertoire biology – the rise of next-generation repertoire analysis

Adrien Six [1,2,3,4,5]*†, ***Maria Encarnita Mariotti-Ferrandiz*** [1,2,3,5], ***Wahiba Chaara*** [1,2,3,4,5], ***Susana Magadan*** [6], ***Hang-Phuong Pham*** [1,2], ***Marie-Paule Lefranc*** [7], ***Thierry Mora*** [8], ***Véronique Thomas-Vaslin*** [1,2,3,5], ***Aleksandra M. Walczak*** [9] ***and Pierre Boudinot*** [6]†

[1] UPMC University Paris 06, UMR 7211, Immunology-Immunopathology-Immunotherapy (I3), Paris, France
[2] CNRS, UMR 7211, Immunology-Immunopathology-Immunotherapy (I3), Paris, France
[3] INSERM, UMR_S 959, Immunology-Immunopathology-Immunotherapy (I3), Paris, France
[4] AP-HP, Hôpital Pitié-Salpêtrière, CIC-BTi Biotherapy, Paris, France
[5] AP-HP, Hôpital Pitié-Salpêtrière, Département Hospitalo-Universitaire (DHU), Inflammation-Immunopathology-Biotherapy (i2B), Paris, France
[6] Institut National de la Recherche Agronomique, Unité de Virologie et Immunologie Moléculaires, Jouy-en-Josas, France
[7] IMGT®, The International ImMunoGeneTics Information System®, Institut de Génétique Humaine, UPR CNRS 1142, Université Montpellier 2, Montpellier, France
[8] Laboratoire de Physique Statistique, UMR8550, CNRS and Ecole Normale Supérieure, Paris, France
[9] Laboratoire de Physique Théorique, UMR8549, CNRS and Ecole Normale Supérieure, Paris, France

Edited by:
Miles Davenport, University of New South Wales, Australia

Reviewed by:
Koji Yasutomo, University of Tokushima, Japan
John J. Miles, Queensland Institute of Medical Research, Australia

****Correspondence:***
Adrien Six, CNRS UMR 7211, UPMC, Immunology-Immunopathology-Immunotherapy (I3), BâtimentCervi, 83 bd de l'Hôpital, Paris F-75013, France
e-mail: adrien.six@upmc.fr

†Adrien Six and Pierre Boudinot have contributed equally to this work.

T and B cell repertoires are collections of lymphocytes, each characterized by its antigen-specific receptor. We review here classical technologies and analysis strategies developed to assess immunoglobulin (IG) and T cell receptor (TR) repertoire diversity, and describe recent advances in the field. First, we describe the broad range of available methodological tools developed in the past decades, each of which answering different questions and showing complementarity for progressive identification of the level of repertoire alterations: global overview of the diversity by flow cytometry, IG repertoire descriptions at the protein level for the identification of IG reactivities, IG/TR CDR3 spectratyping strategies, and related molecular quantification or dynamics of T/B cell differentiation. Additionally, we introduce the recent technological advances in molecular biology tools allowing deeper analysis of IG/TR diversity by next-generation sequencing (NGS), offering systematic and comprehensive sequencing of IG/TR transcripts in a short amount of time. NGS provides several angles of analysis such as clonotype frequency, CDR3 diversity, CDR3 sequence analysis, V allele identification with a quantitative dimension, therefore requiring high-throughput analysis tools development. In this line, we discuss the recent efforts made for nomenclature standardization and ontology development. We then present the variety of available statistical analysis and modeling approaches developed with regards to the various levels of diversity analysis, and reveal the increasing sophistication of those modeling approaches. To conclude, we provide some examples of recent mathematical modeling strategies and perspectives that illustrate the active rise of a "next-generation" of repertoire analysis.

Keywords: diversity analysis, immune receptors, next-generation sequencing, modeling, statistics, gene nomenclature, B cell repertoire, T cell repertoire

INTRODUCTION

T and B cell repertoires are collections of lymphocytes, each characterized by its antigen-specific receptor. The resources available to generate the potential repertoires are described by the genomic T cell receptor (TR) and immunoglobulin (IG) loci. TR and IG are produced by random somatic rearrangements of V, D, and J genes during lymphocyte differentiation. The product of the V-(D)-J joining, called the complementarity determining region 3 (CDR3) and corresponding to the signature of the rearrangement, binds the antigen and is responsible for the specificity of the recognition. During their differentiation, lymphocytes are subjected to selective processes, which lead to deletion of most auto-reactive cells, selection, export, and expansion, of mature T and B cells to the periphery. Primary IG and TR repertoires are therefore shaped to generate the available peripheral or mucosal repertoires. In addition, several different functional T and B cells subsets have been identified, with differential dynamics and antigen-specific patterns. These available repertoires are dramatically modified during antigen-driven responses especially in the inflammatory context of pathogen infections, autoimmune syndromes, and cancer to shape actual repertoires. When considering the importance of efficient adaptive immune responses to get rid of infections naturally or to avoid auto-reactive damages, but also for therapeutic purposes such as vaccination or cell therapy, one realizes the relevance of understanding how lymphocyte repertoires are selected during differentiation, from ontogeny to aging, and upon antigenic challenge. However, immune repertoires of expressed antigen receptors are

built by an integrated system of genomic recombination and controlled expression, and follow complex time-space developmental patterns. Thus, an efficient repertoire analysis requires both (1) methods that sample and describe the diversity of receptors at different levels for an acceptable cost and from a little amount of material and (2) analysis strategies that reconstitute the best multidimensional picture of the immune diversity from the partial information provided by the repertoire description as reviewed in Ref. (1). In the following sections, we summarize technologies developed over the past decades to describe lymphocyte repertoires and we present the growing number of analysis tools, evolving from basic to sophisticated statistics and modeling strategies with regards to the level of complexity of the data produced.

METHODS DEVELOPED TO DESCRIBE THE IG AND TR REPERTOIRES

B and T lymphocyte repertoires can be studied from different lymphoid tissues and at various biological levels, such as cell membrane or secreted proteins, transcripts or genes, according to the techniques used. Fluorescence microscopy or flow cytometry techniques allow to track and sort particular cell phenotypes and to quantify the expressed repertoire at the single-cell level with V subgroup-specific monoclonal antibodies. Alternatively, the IG or TR diversity may be also analyzed using proteomics methods from either the serum (for IG) or dedicated cell extracts. Finally, molecular biology techniques assess the repertoire at the genomic DNA or transcriptional levels, qualitatively and/or quantitatively.

ANALYSIS OF IG AND TR REPERTOIRES AT THE PROTEIN LEVEL

Flow cytometry single-cell repertoire analysis

The frequency of lymphocytes expressing a given IG or TR can be determined using flow cytometry when specific monoclonal antibodies are available. This technique allows for the combined analysis of the antigen receptor and of other cell surface markers. Currently, using flow cytometry, up to 13 parameters can be routinely studied at once, reaching 20 parameters with the last generation flow cytometers and 70–100 parameters with mass cytometry (2). Seminal studies in mice using specific anti-TRBV antibodies have led to the characterization of the central tolerance selection processes that occur in thymus (3–5). Later on, a comprehensive description of the human TRBV repertoire was setup (6), when monoclonal antibodies became available for most of the TRBV subgroups. Repertoire analysis with flow cytometry provides a qualitative and quantitative analyses of the variable region, often done on heterogeneous cell populations, in order to decipher, for example, selection events related to aging, perturbations, and treatments (7). However, this technology is naturally limited by the availability of specific monoclonal antibodies, and does not address more detailed issues such as junction diversity. Furthermore, polymorphism of the IG or TR genes (8, 9) may constitute a serious limitation for a systematic survey using these approaches.

Proteomic repertoire analysis for serum immunoglobulins

Recent developments of proteomics tools now offer sensitivity levels applicable to IG repertoire analysis. Such a description at the protein level takes into account all post transcriptional and translational modifications.

PANAMA-blot technology. A semi-quantitative immunoblot, called the PANAMA-blot technique (10), allows for the identification of the antibody reactivities present in collection of sera (or cell culture supernatant) against a given source of antigens (10–12). Briefly, a selected source of antigens is subjected to preparative SDS-PAGE, transferred onto nitrocellulose membranes, then incubated with the serum to be tested allowing for the revelation of the bound antibodies using an appropriate secondary antibody coupled to alkaline phosphatase. Computer-assisted analysis of the densitometric profiles allows for the rescaling and the quantitative comparison of patterns of antibody reactivity from individuals in different groups. A large amount of data is generated when testing a range of sera against various sources of antigens. Statistical analyses are included in the PANAMA-Blot approach (as described further). This global analysis helped to reveal that the IgM repertoire in mice is selected by internal ligands and independent of external antigens (13).

This method can also lead to identify IG reactivity patterns specific for a type of pathology or clinical status and has been applied to both fundamental and clinical analysis. In particular, it was used to analyze human self-reactive antibody repertoires and their potential role for down-modulating autoimmune processes (14–16).

Antigen micro-array chips. More recently, antigen micro-array-based technology coupled to a complex two-way clustering bioinformatics analysis was developed to evaluate the serum repertoire antibodies from diabetes-prone individuals and revealed their predictive or diagnostic value. In brief, a range of antigens (proteins, peptides, nucleotides, phospholipids. . .) were plated onto glass plates and incubated with sera from individuals (human diabetes patients or mice in an experimental model of diabetes). The intensity of reactivity of the serum IG for each peptide was determined and scored against the control reactivity. Clustering analysis was then implemented to determine a potential antigen signature that significantly sorts out diabetes from non-diabetes individuals. In this way, it was found that the patterns of IgG antibodies expressed early in male NOD mice can mark susceptibility or resistance to diabetes induced later and that it is different than the pattern characteristic of healthy or diabetic mice after disease induction (17). Similarly, this clustering approach was applied in humans to successfully separate human subjects that are already diabetic from healthy people (18).

REPERTOIRE ANALYSIS AT THE GENOMIC DNA LEVEL

Other strategies that cover IG or TR repertoire analyses have been developed at the genomic DNA level. Firstly, CDR3 spectratyping studies (detailed in the following section) have been carried out at the DNA level mostly to address issues related to B or T cell development (19, 20). More recently, an original multiplex genomic PCR assay coupled to real-time PCR analysis was developed to provide a comprehensive description of the mouse T cell receptor alpha (TRA) repertoire during development (21). Although these approaches can be applied to all IG isotypes and TR, they have not been used as much as transcript CDR3 spectratyping due to sensitivity and heterozygosity issues.

Immunoglobulin or T cell receptor repertoires can also be assessed by following the diversity of rearrangement deletion circles. Since they are produced by the V-(D)-J recombination machinery when the joint signal is formed and diluted in daughter cells, they give a good representation of recently generated T or B cells. This technique has been particularly useful for describing the restoration of T cell diversity following highly active antiretroviral therapy in HIV-infected patients (22) and has been used to model thymic export (23, 24) as well as to demonstrate continued contribution of the thymus to repertoire diversity, even in older individuals (25). It also reveals that thymic output is genetically determined, and related to the extent of proliferation of T cells at DN4 stage in mice (26). However, their analysis does not provide much insight into the level of diversity since the signal joint does not vary for a given combination of genes. Therefore, the interest of such analyses is reached when combined with CDR3 spectratyping analyses to know whether a repertoire perturbation is rather attributable to newly produced T cells or peripheral T cell proliferation.

V-(D)-J JUNCTION ANALYSIS OF IG AND TR TRANSCRIPT REPERTOIRES

Original molecular-based strategies for analyzing repertoire diversity relied on cloning and hybridization of molecular probes specific for IGHV gene subgroups first by RNA colony blot assay (27). This led to the observation that IGHV gene usage is characteristic of mouse strain and is a process of random genetic combination by equiprobable expression of IGHV genes (28). The study of selection processes revealed that the IGHV region-dependent selection determines clonal persistence of B cells (29) and that selection with age leads to biased IGHV gene expression (30).

In situ hybridization on single-cells revealed that during mouse ontogeny and early development of B cells in bone marrow, there is a non-random position-dependent IGHV gene expression, favoring D-proximal IGHV gene subgroup usage (31). Thereafter, sequencing of PCR-amplified cDNA collections were obtained from samples of interest. Although fastidious, these early studies have been useful in defining the basis of human IG and TR repertoires in terms of overall distribution, CDR3-length distribution, and V-(D)-J use (32–35), sometimes leading to the identification of new IG or TR genes. Later, more practical techniques have been developed for large-scale analysis of lymphocyte repertoires, such as quantitative PCR, micro-array, and junction length spectratyping, as described below.

Quantitative RT-PCR for repertoire analysis

In parallel to qualitative CDR3 spectratyping techniques (see section below), quantitative PCR strategies were developed (36). Coupling the two techniques for all V domain-C region combinations provides a complete qualitative and quantitative picture of the repertoire (37–39) described by up to 2,000 measurements per IG isotype or TR for one sample. With the development of real-time quantitative PCR, this approach opened the possibility for a more precise evaluation of repertoire diversity (39–41). Complementary tools have been also developed in order to allow normalization of spectratype analysis such as studies by Liu et al. (42) and Mugnaini et al. (43).

Matsutani et al. (44) developed another method to quantify the expression of the human TRAV and TRBV repertoires based on hybridization with gene specific primers coated plates. The cDNA from PBMC extracted RNA are ligated to a universal adaptor which allows for a global amplification of all TRAV or TRBV cDNAs. The PCR products are then transferred onto microplates coated with oligonucleotides specific for each TRAV or TRBV regions, and the amount of hybridized material is quantified. This technique was used to analyze the TR repertoire diversity of transplanted patients (45) and adapted to the study of mouse TRAV and TRBV repertoires (46). VanderBorght et al. also developed a semi-quantitative PCR-ELISA-based method for the human TRAV and TRBV repertoire analysis (38). The combined usage of digoxigenin (DIG)-coupled nucleotides and DIG-coupled reverse TRAC or TRBC primers allowed for a quantitative measurement of the amount of amplified DNA by a sandwich ELISA.

Du et al. (47) later setup a megaplex PCR strategy to characterize the antigen-specific TRBV repertoire from sorted IFNγ-producing cells after *Mycobacterium* infection. The clonotypic TRBV PCR products were used for Taqman probes design to quantify the expression of the corresponding clonotypes from ATLAS-amplified SMART cDNAs.

Direct measurement of lymphocyte diversity using micro-arrays

Another technology, similar to the one just discussed, has been developed by the group of Cascalho et al. which allows for a direct measurement of the entire population of lymphocyte-receptors. This is accomplished by hybridization of lymphocyte-receptor specific cRNA of a lymphocyte population of interest to random oligonucleotides on a gene chip; the number of sites undergoing hybridization corresponds to the level of diversity. This method was validated and calibrated using control samples of random oligonucleotides of known diversity (1, 10^3, 10^6, 10^9) (48, 49) and successfully demonstrated that central and peripheral diversification of T lymphocytes is dependent on the diversity of the circulating IG repertoire (49, 50). Similarly, a highly sensitive micro-array-based method has been proposed to monitor TR repertoire at the single-cell level (51).

CDR3 spectratyping techniques

Immunoscope technology. Among various techniques used to analyze the T or B cell repertoires, Immunoscope, also known as CDR3 spectratyping (52, 53) consists in the analysis of the CDR3-length usage so that antigen-specific receptor repertoires can be described by thousands of measurements. In the case of naive murine repertoires, T cell populations are polyclonal and analysis typically yields eight-peak regular bell-shaped CDR3 displays (wrongly assumed to be Gaussian), each peak corresponding to a given CDR3-length. When an immune response occurs, this regular polyclonal display can be perturbed: one can see one or several prominent peaks that correspond to the oligoclonal or clonal expansion of lymphocytes. A complete description of this technique and its applications to clinical studies has been published elsewhere (54).

In the original Immunoscope publication, Cochet et al. (55) analyzed the T cell repertoire after the immunization of mice with the pigeon cytochrome c. They provided the first description of

an *ex vivo* follow-up of a primary T cell specific response in a mouse model. Their second paper analyzed the average CDR3-lengths as a function of TRBV-TRBJ combinations. In particular, the authors found a correlation between TRBV CDR1 and major histocompatibility (MH) haplotype (52). This group later published a large amount of original studies in various models such as lymphocyte development (40, 56–63), kinetics of antigen-specific responses (64–67), viral infection (68, 69), autoimmunity (70, 71), tumor-associated disease (72), and analysis of allogeneic T cell response and tolerance after transplantation (73). Notably, the combination of CDR3 spectratyping with flow cytometry-based IG or TR V frequency analysis provides a more comprehensive assessment, such as in Pilch et al. (74). For example, such an approach revealed the constriction of repertoire diversity through age-related clonal CD8 expansion (75). Similarly, a combination of CDR3 spectratyping, flow cytometry, and TR deletion circle analysis has allowed to define age-dependent incidence on thymic renewal in patients (76) or to evaluate the effects of caloric restriction in monkeys to preserve repertoire diversity (77). CDR3-length spectratyping was also used in other models, such as rainbow trout, to analyze TRB repertoire and its modifications induced by viral infection (78–80). While no tool such as monoclonal antibodies to T cell marker(s) was available in this model, this approach demonstrated that fish could mount specific T cell responses against virus, which could be found in all individuals (public clonotypes) or not (private clonotypes). Similar strategies, developed by other groups (81) and following the same approach in parallel, analyzed the IG repertoire in *Xenopus* at different stages of development, describing a more restricted IG junction diversity in the tadpole compared to the adult.

Gorski et al. (82) developed their own CDR3 spectratyping technique to analyze the complexity and stability of circulating αβT cell repertoires in patients following bone marrow transplantation as compared to normal adults. They showed that repertoire complexity of bone marrow recipients correlates with their state of immune function; in particular, individuals suffering from recurrent infections associated with T cell impairment exhibited contractions and gaps in repertoire diversity. The detailed procedure for this technique has been published in Maslanka et al. (83). A variation of this technique has been reported later by Lue et al. (84), relying on a compact glass cassette, a simpler device than the usual automated plate DNA sequencers.

Alternative technologies. Alternative CDR3 spectratyping techniques have been described such as single-strand conformation polymorphism (85–87) and heteroduplex analysis (88–91). These methods differ from the CDR3 spectratyping/Immunoscope technique mostly in the way PCR products are analyzed by performing non-denaturing polyacrylamide electrophoresis. The main advantage of these techniques is a more direct assessment of clonal expansion since PCR products migrate according to their conformation properties; therefore, presence of a predominant peak is strongly indicative of clonality when a smear migration pattern indicates polyclonality. However, these techniques have been less widely used probably because of the difficulty to make clear correlations between the expanded peaks across samples.

Another original alternative technique has been described by Bouffard et al. (92), analyzing products obtained after *in vitro* translation of PCR-amplified TR-specific products by isoelectric focusing. With this technique, clonality can also directly be assessed by looking at the obtained migration profile.

IG/TR REARRANGEMENT SEQUENCING: FROM CLONING-BASED- TO NEXT-GENERATION-SEQUENCING

In order to get a better description of IG/TR diversity at the nucleotide sequence level, thus providing fine-tuned description of the actual diversity, Sanger sequencing approaches relying on bacterial cloning of rearrangements were performed in physiological conditions globally (60, 93–99) or partially to characterize particular expansions identified by other technologies such as CDR3 spectratyping (40, 59, 100–102), flow cytometry (103). They were also used in pathological/infectious conditions (104–107) sometimes leading to antigen-specific T cell TR identification and quantification through the combination of antigen-specific T cell stimulation and cytometry-based cell sorting, anchor-PCR, and bacterial cloning-based sequencing (108).

These studies pioneered the description of the repertoire and provided fruitful information regarding the extent and modification of the diversity. However, besides being time and cost-extensive, such approaches have allowed for the analysis of 10^2–10^3 sequences, far under the estimated diversity reaching 10^6–10^7 unique clonotypes in mice and humans (40, 59, 109).

In the last decade, DNA sequencing technologies have made tremendous progresses (110) with the development of so called next-generation sequencers, already reaching four generations (111). Those instruments are designed to sequence mixtures of up to millions of DNA molecules simultaneously, instead of individual clones separately. Second generation sequencers became affordable in the last 5 years and have been used for immune repertoire analysis, starting with the seminal work of Weinstein et al. (112) where the IG repertoire of Zebrafish has been described by large-scale sequencing. Consequently, exploratory works by other groups provided an overview of the complex sequence landscape of immune repertoires in humans (113–118). More recent work aimed at addressing fundamental questions such as lineage cells commitment (119–122), generation of the diversity processes (123–125), and diversity sharing between individuals (126, 127). Finally, the power of this technology has been validated in the clinic as well (128, 129).

As seen above for other technologies, combinations of approaches have been applied to NGS. Notably, deep sequencing has been used in combination with CDR3-length spectratyping by some groups to study human (130) or rainbow trout IG (131) repertoire modifications after vaccination against bacteria or viruses. In the latter, pyrosequencing performed for relevant VH/Cμ or VH/Cτ junctions identified the clonal structure of responses, and showed, for example, that public responses are made of different clones identified by (1) distinct V-(D)-J junctions encoding the same protein sequence or (2) distinct V-(D)-J sequences differing by one or two conservative amino acid changes (131) as described for public response in mammals (132, 133). These studies showed that NGS and traditional spectratyping techniques lead to remarkably similar CDR3 distributions.

Several NGS have been developed in the past years using different sequencing technologies characterized with different speed, deepness and read length. Metzker thoroughly reviewed their principles and properties (134). Among them, three platforms, all offering benchtop sequencers with reduced cost and setup, fit with immune repertoire analysis in terms of read length and deepness. The 454/Roche platform uses pyrosequencing technology (135), which combines single nucleotide addition (SNA) with chemoluminescent detection on templates that are clonally amplified by emulsion PCR and loaded on a picotiter plate. Pyrosequencing currently has a 500 bp (GS Junior) to 700 bp (GS FLX) sequencing capacity with a respective deepness of 150,000–3,000,000 reads per run (134). The Illumina/Solexa platform technology is based on cyclic reversible termination (CRT) sequencing (an adaptation of Sanger sequencing) performed on templates clonally amplified on solid-phase bridge PCR. Protected fluorescent nucleotides are added, imaged, delabeled, and deprotected cyclically (134), providing a deeper sequencing (from 15 to 6 billion reads per run for the MiSeq to the HiSeq2500/2000) of shorter reads (100–250 bp for the very recent MiSeq) with the possibility to perform pair-end sequencing (two-side sequencing) to increase the read length after aligning the generated complementary sequences. A more recent platform, Ion Torrent/Life Technologies using an imaging free detection system may open a new era in terms of deepness (one billion reads per run) of 200 bp reads (136) in a very short time and on a benchtop sequencer. Importantly, depending on the technology, errors due to the PCR-based sample preparation and the sequencing are of major concern. Bolotin et al. (137) evaluated this issue on TR repertoire analysis of the same donor performed on the three platforms described previously; algorithms for error correction have been developed. Indeed, PCR- and sequencing-related errors represent the major concern for immune repertoire diversity analysis as they may generate artificial diversity. Illumina and 454 appear to be the most robust technologies, with Illumina having the highest throughput and 454 generating the longest reads. The currently available Ion Torrent platform, although very promising, has been shown to display the highest rate of errors in TR (137) and bacterial DNA (138) sequencing. However, such error corrections must be used with caution since they may inadvertently underestimate repertoire diversity by removing rare sequences.

With the power of such approach for genomics and transcriptomics studies in general, constant improvements are achieved to increase the sequencing deepness and read length as well as to reduce the cost, therefore offering multitude of biological explorations (139). NGS now permits a comprehensive and quantitative view of IG and TR diversity by combining and improving the sensitivity of classical approaches with accurate and large-scale sequencing. NGS has the power to identify IG or TR specific for given antigens (in combination with antigen-specific assays) and to define more complex signatures (i.e., TR sets) related to disease and/or treatment from heterogeneous T and B cell populations. Still, most of the deep sequencing efforts have been limited to only one chain of the receptor at the repertoire level (usually the β chain for TR and the heavy chain for IG). Indeed, current high-throughput approaches do not allow one to assign which combination of chains (TRA and TRB, or IGH and IGK or IGL) belong to which cell (140). A recent development by DeKosky et al. proposed a reasonably high-throughput technology to assess massively paired IG VH and VL from bulk population (141). In parallel, Turchaninova et al. (142) have proposed a similar approach for the paired analysis of the TRA and TRB chains. The parallel development of high-throughput microfluidic-based single-cell sorting will certainly push forward new developments in the field (143).

However, despite the technological advance, studies so far have mainly reported CDR3 counting and identification of major expansions. The complexity of immune repertoires is still a matter that such approach cannot completely overcome, due to the paucity of powerful analytical methods. Besides data management tools, studies are now starting to extract most of the benefit from such approach to model the immune repertoire diversity and dynamics (144), an approach that may help in understanding the interplay between cells and repertoire shaping. Accurate and powerful statistical analyses are required to manage such amount of information. Current state will be reviewed in the following sections.

POTENTIAL AND GENOMIC REPERTOIRES: A QUESTION OF ONTOLOGY AND ORTHOLOGY

Immune repertoires *sensu stricto* are expressed by lymphocyte clones, each carrying a single receptor for the antigen. Such receptors comprise IG and TR in jawed vertebrates (8, 9) and VLR in Agnathans (145). The sequences of these receptors are available in databases such as GenBank or EMBL, which are difficult to use for transversal studies due to inconsistent annotation. The IMGT® information system (see below) has largely solved this problem setting standardized gene nomenclatures, ontologies and a universal numbering of the IG/TR V and C domains, thus giving a common access to standardized data from genome, proteome, genetics, two-dimensional, and three-dimensional structures (146). The accuracy and the consistency of the IMGT® data are based on IMGT-ONTOLOGY, the first, and so far, unique ontology for immunogenetics and immunoinformatics (147).

With the development of high-throughput sequencing, large numbers of new sequences of antigen receptor genes have become available, which can be classified into different categories: genomic sequences of IG or TR (in germline configuration in genome assemblies) or fragments of IG/TR transcripts, containing the CDR3 or not. Also, these datasets can be produced from species newly sequenced, as well as from new haplotypes of well-described species.

The annotation of such sequences remains an open question. Manual annotation is not applicable, and no good automated approach has been validated yet. A relevant annotation of these massive datasets will require the integration of genomic and expression data with existing standardized description charts, as offered by IMGT®. A standardized annotation is an important issue since it facilitates the re-utilization of datasets and comparison of analyses. Thus, the description of IG and TR polymorphisms, the integration of repertoire studies with structural features of antigen-specific domains, and even the usage of new genes in genetic engineering rely on a common standard for nomenclature, numbering, and annotation (147).

To take advantage of the current standards that have been established from classical sequencing data during the last 25 years, new, fast, reliable, and human-supervised annotation methods will have to be developed, integrating directly high-throughput sequence information from the increasing number of deep sequencing platforms and technologies, at different genetic levels (genome, transcriptome, clonotype repertoires). Along this line, IMGT/HighV-QUEST offers online tools to the scientific community for the analysis of long IG and TR sequences from NGS (148).

Special attention can be paid to the orthology/paralogy relationships between similar antigen receptor genes from different species. These characteristics are essential to understand the dynamics of IG and TR loci. In fact, with many important lymphocyte subsets characterized by canonical/invariant antigen receptors, such relationships are critical to transfer functional knowledge between models. Importantly, the phylogenetic analyses required to reconstitute the evolution of antigen receptor genes are based on multiple alignments, the quality of which is highly dependent on common numbering and precise annotation of sequences.

As far as immune repertoires are characterized by the diversity of receptors specifically binding antigen/pathogen motifs to initiate a defense response, they might not be limited to lymphocyte diversifying receptors, e.g., IG, TR, and VLR. The particularity of these systems is a somatic diversification combined to a clonal structure of the repertoire, each lymphocyte clone expressing the product of a recombination/hypermutation and/or conversion process. However, many other arrays of diverse receptors binding or sensing pathogens have been discovered in metazoans, in invertebrates as well as in vertebrates.

In some cases, their diversity is really "innate," i.e., encoded in the genome as multiple genes produced by duplications. Fish NLR, finTRIMs, and NITR, primate KIR, chicken CHIR, or TLR in sea urchin, constitute good examples of such situations. While these repertoires may appear as relatively limited, polymorphism within populations, and differential expression of receptors per cell upon stimulation represent complex issues, which fall well into "traditional" repertoire approaches.

In other cases, receptors are subject to diversification processes much faster than gene duplication, which does not comply with a clonal selection pattern. The best examples are probably the DSCAM in arthropods, which hugely diversify by alternative splicing of exons encoding half-IgSF domains (149, 150), and the FREP lectins in mollusks, of which sequences are highly variable at the population level, and even between parents and offspring produced by auto-fecundation (151).

The number of such "innate" repertoires which are not expressed by clonally selected lymphocytes will likely increase with deep sequencing of new genomes/transcriptomes, as illustrated by a recent report from mussel (152). A good example of the importance of a proper structural description of key domains of receptors is provided by the extensive analysis of LRR motifs in studies on TLR evolution (153, 154). Further insights into the functions of such diverse proteins will be provided by the characterization of their expressed (available) repertoire, at different levels such as single-cells, cell populations, and animal populations.

Such analyses will require precise identification of genes and sequences as well as mutations, and a standardized approach of nomenclature and structural description will be as useful as it is for the vertebrate IG and TR sequences. Importantly, these receptors are made of a small number of structural units, such as IgSF domain or LRR domains, which suggests that standardized system(s) for sequence annotation could be developed following IMGT standards (155).

STATISTICAL ANALYSIS AND MODELING OF IMMUNE REPERTOIRE DATA

STATISTICAL REPERTOIRE ANALYSIS

The description of the repertoire modifications using flow cytometry or Immunoscope provided clear-cut and detailed insight into the clonal expansion processes during the responses against a defined antigen (64, 66). However, it is difficult to identify the relevant alterations of the repertoires in more complex situations such as pathogen infections or variable genetic backgrounds. For example, it appeared impossible to identify all significant modifications of TRB Immunoscope profiles during cerebral malaria by direct ocular comparison (107). Different methods were therefore developed to extract from IG and TR repertoire descriptions the relevant information, to encode it as numerical tables and to analyze them with statistical models.

CDR3 spectratype perturbation indices

Since the initial description of the CDR3 spectratyping technique, different scoring indices were developed or derived from the literature: "relative index of stimulation" (RIS) (55), "overall complexity score" (156), Reperturb (157), "complexity scoring system" (158), COPOM (159), Oligoscore (160), TcLandscape (161), "spectratype diversity scoring system" (162), Morisita-Horn index and Jaccard index (95–97), "absolute perturbance value" (163). A comparative review of such scoring strategies was published by Miqueu et al. (164).

In particular, the perturbation index Reperturb was developed by Gorochov et al. to perform TR repertoire analysis in HIV patient during progression to AIDS and under antiretroviral therapy. They could show drastic restrictions in the $CD8^+$ T cell repertoire at all stages of natural progression that persisted during the first 6 months of treatment. In contrast, $CD4^+$ T cell repertoire perturbations correlated with progression to AIDS with a return to a diversified repertoire in good responders to treatment (157).

Soulillou et al. refined this approach by combining the qualitative information obtained with usual CDR3 spectratyping with quantitative information of TRBV usage obtained by real-time quantitative PCR. They devised a four-dimension representation that represents TRBV subgroups, CDR3-length and percentage of TRBV use on three axis chart in addition to a color-coded representation of the CDR3 profile perturbation. Using this original approach, they were able to show that graft rejection is associated with a vigorous polyclonal accumulation of TRBV mRNA among graft-infiltrating T lymphocytes, whereas in tolerated grafts T cell repertoire is strongly altered (161, 165). Their study puts the emphasis on the importance of not only qualitative but also quantitative analysis of lymphocyte repertoires.

Platforms for repertoire data management and statistical analysis

Several platforms have been developed and rely mostly on CDR3 spectratyping and sequencing data, with recent developments to manage and analyze NGS data.

The ISEApeaks strategy and software were developed in order to satisfy the needs for efficient automated electrophoresis data retrieval and management (160, 166). ISEApeaks extracts peak area and length data generated by software used to determine fragment intensity and size. CDR3 spectratype raw data, consisting of peak areas and nucleotide lengths for each V-(D)-J-C combination, is extracted, smoothed, managed, and analyzed. The repertoires of different samples are gathered in a peak database and CDR3 spectratypes can be analyzed by different perturbation indices and multivariate statistical methods implemented in ISEApeaks. We have applied our ISEApeaks strategy in several studies. In an experimental model of cerebral malaria, we established a correlation between the quality of TR repertoire alterations and the clinical status of infected mice, whether they developed cerebral malaria or not (107). We contributed to the characterization of the membrane-associated *Leishmania* antigens (MLA) that stimulates a large fraction of naive CD4 lymphocytes. Repertoire analyses showed that MLA-induced T cell expansions used TR with various TRBV rearrangements and CDR3 lengths, a feature closer to that of polyclonal activators than of a classic antigen (167). We also revealed repertoire age-related perturbations in mice (7). ISEApeaks functions for statistical analysis was successfully applied to analyze the TR repertoire in fish as shown by our detailed analysis of the TRB repertoire of rainbow trout IELs, performed in both naive and virus-infected animals. Rainbow trout IEL TRBV transcripts were highly diverse and polyclonal in adult naive individuals, in sharp contrast with the restricted diversity of IEL oligoclonal repertoires described in birds and mammals (102). More recently, our study of the $CD8^+$ and $CD8^-$ αβ T cell repertoire suggests different regulatory patterns of those T cell patterns in fish and in mammals (168). ISEApeaks was also used to implement a new statistically based strategy for quantification of repertoire diversity (159).

Kepler et al. described another original statistical approach for CDR3 spectratype analysis, using complex procedures for testing hypotheses regarding differences in antigen receptor distribution and variable repertoire diversity in different treatment groups. This approach is based on the derivation of probability distributions directly from spectratype data instead of using *ad hoc* measures of spectratype differences (169). A software (called SpA) implementing this method has been developed and made available online (170). This approach has been used in a longitudinal analysis of TRBV repertoire during acute GvHD after stem cell transplantation (171).

Another group (163) reported the development of a new software platform, REPERTOIRE, which allows handling of CDR3 spectratyping data. This software implements a perturbation index based upon an expected normal Gaussian distribution of CDR3 length profiles.

Owing to the complexity and diversity of the immune system, immunogenetics represents one of the greatest challenges for data interpretation: a large biological expertise, a considerable effort of standardization, and the elaboration of an efficient system for the management of the related knowledge were required. To answer that challenge, IMGT®, the international ImMunoGeneTics information system®(http://www.imgt.org), was created in 1989 by one of the authors (146). Overtime, it developed standards that, since 1995, have been endorsed by the World Health Organization-International Union of Immunological Societies (WHO-IUIS) Nomenclature Committee and by the WHO-International Nonproprietary Names (INN) (172–175). IMGT® comprises seven databases (sequence, gene, and structure databases), 17 online tools and more than 15,000 pages of web resources. Among the databases, IMGT/LIGM-DB, the database for nucleotide sequences (170,685 sequences from 335 species as of July 2013) and IMGT/GENE-DB, the gene database (3,081 genes and 4,687 alleles) are of great interest for repertoire analysis. Freely available since 1997, IMGT/V-QUEST is an integrated system for the standardized analysis of collections of IG and TR rearranged nucleotide sequences (176, 177). A high-throughput version, IMGT/HighV-QUEST (148), has been released in 2010 for the analysis of long IG and TR sequences from NGS using the 454 Life Sciences technology. In the same line, other analysis tools are becoming available showing the renewed interest for repertoire analyses and modeling consecutive to NGS technology developments (178–181).

Altogether, these efforts highlight the relevance of developing more efficient and powerful technologies for the evaluation of repertoire diversity. Notably, two successful French biotech companies (TcLand, Nantes; ImmunID, Grenoble) were created in the field of repertoire analysis, using different technologies. In collaboration with ImmunID, we have proposed a novel strategy for statistical modeling of T lymphocyte repertoire data obtained in humans and humanized mice. With this model, we revealed that half of the human TRB repertoire, in terms of proportion of TRBV-TRBJ combinations, is genetically determined, the other half occurring stochastically (182). In addition, the biotechnology company "Adaptive" and the "Repertoire 10K (R10K) Project" have been recently founded by researchers respectively from the Fred Hutchinson Cancer Research Center (Seattle and Washington) and the HudsonAlpha Institute (Huntsville). Both have developed platforms (immunoSEQ®, iRepertoire®) providing researchers with a global analysis of the T or B cell receptor sequence repertoires (183). However, despite the power of this technology, studies are still limited by the ability to process the complexity of the information provided. Specific software developments for the automatic treatment and annotation of IG and TR sequences and the statistical modeling of repertoire diversity can still be improved.

Multivariate analysis

As mentioned above, the PANAMA-Blot technique also includes statistical analysis of the data. Multi-parametric analysis was introduced to compare the global reactivity of antibodies of different individuals in different groups with a given antigenic extract. This analysis has been successfully implemented to identify reactivity patterns specific for a given pathology or clinical status (10–12, 14, 15, 184). Similarly, multi-parametric analysis was also applied to TRBV spectratype analysis in an experimental cerebral malaria model (107).

Hierarchical clustering or classification algorithms have become very popular with the growing of micro-array-based transcriptome analysis. Although still uncommon for immune repertoire analysis, such approaches have been employed to categorize large sets of repertoire data without *a priori* (17, 102, 107).

Diversity indices

The concept of immune repertoire has been devised to describe the diversity of cells involved in the immune system of an individual (1). As described above, different scoring systems were developed to assess this diversity, some are heuristics but others have been borrowed from theoretical ecology and evolution. As reviewed by Magurran (185), the Shannon entropy, introduced by Claude Shannon in 1948 for the information theory, is the most used because it not only integrates the number of different species but also the relative proportion of each of these species. In 1961, Alfred Rényi generalized this entropy to a family of functions, like Species Richness, Simpson, Quadratic, and Berger–Parker indices, for quantifying the diversity, the uncertainty or randomness of a system. Most of these indices are implemented in the free software application Estimates (http://purl.oclc.org/estimates) (186). Altogether, these diversity indices constitute a collection of tools with their own sensitivity to the variety and the relative abundances of the species that are perfectly suitable for assessing immune repertoire diversity. Indeed, the very famous index of variability proposed by Kabat and Wu (187) corresponds to the ratio of Species Richness and Berger–Parker indices. In 1990, Jores et al. showed that the resolving power of this Wu-Kabat variability coefficient can be enhanced by increasing the weight on the frequency distribution of the amino acids in the formula (188). This approach inspired Stewart et al. (189) to use the Shannon entropy to demonstrate that TR amino acid composition is significantly more diverse than that of IG. In the same way, CDR3 spectratyping data can be analyzed using the relative abundance of each peak within CDR3 length global distribution. By doing so, we adjusted the original Shannon entropy, making it reaching its maximum for a Gaussian distribution, to compare the CDR3 length diversity of splenic IgM, IgD, and IgT in infected Teleost Fish (131). Recently, the Gini index, used in ecology or economics to measure the equality of distributions, was applied to individual TR clones and compared naive and memory repertoires (190). The development of deep sequencing techniques ignited a renewed interest in IG/TR repertoire. Indeed, several studies used high-throughput analysis to describe TR repertoire of key T cell subsets in human peripheral blood (115, 126, 191). This approach assessing the repertoire diversity from the relative abundance of each species in the global distribution can be decomposed hierarchically into components attributable, respectively, to variations in TRBV-TRBJ combinations and in CDR3-length (113, 117). However, most of these studies have been limited to the counting of the observed unique clonotypes. Beside the species richness, ecology-derived indices have also been applied to assess and compare immune repertoire diversity. Föhse et al. (119) used the Morisita-Horn similarity index to compare regulatory T cell repertoires between several lymphoid organs. In addition, Simpson diversity index, associated with Shannon entropy, was used to monitor TR repertoire diversity of HIV-specific CD8 T cells during antiretroviral therapy (192) but also to quantify TR repertoire recovery in the blood after allogeneic hematopoietic stem cell transplantation (128). In the same manner, Koning et al. (193) used Shannon's and Simpson's indices to show the role for the peptide component of the peptide-MH1 complex on the molecular frontline of $CD8^+$ T cell–mediated immune surveillance, by comparing the repertoire diversity of $CD8^+$ T cell populations directed against a variety of epitopes. In parallel, using Simpson's index as a metric allowed Johnson et al. (194) to model mathematically the naïve CD4 T cell repertoire contraction with age leading them to conclude that diversity plummet observed around the age of 70 could be correlated to cell-intrinsic mutations affecting cell division rate or death.

MODELING STRATEGIES

Modeling approaches have a strong tradition in immunology, usually at the boundary with other disciplines such as physics (195). Before deep sequencing data was available, general design principles were proposed as desirable features of immune repertoires, with implications for the observed repertoire diversity and dynamics (196–198). Many efforts have involved the modeling of immune cell dynamics and the effects of antigens on repertoire diversity, using differential equations descriptions of the population dynamics (199–201). Recognition in the immune system is often studied both theoretically and experimentally by probing the dynamics of cells with a specific type of receptor with respect to infections (202). Alternatively one can look at the response of a small set of chosen receptors to a specific pathogenic challenge, or careful biochemical investigation of particular receptor/antigen pairs (203, 204). Much work has been devoted to systems-biology approaches to signal processing in immune cells, as reviewed in Germain et al. (205) and Emonet and Altan-Bonnet (206). Here we focus on approaches inspired by recent advances in sequencing technologies (112, 113, 115, 116, 125, 191, 207, 208) that have opened the way for data-driven modeling of the immune repertoires and interactions between receptors and antigen.

A common modeling approach for describing receptors at the amino acid level is to choose a relevant interaction parameter (e.g., chemical affinity or hydrophobicity) and assign it a simplified digit-string representation (209). These methods are extensions of the string model, which describes both receptor and epitopes as strings of length L, with values chosen from natural numbers, and quantify their interaction by the match between the two strings (197, 210, 211). Such quantitative, physically inspired descriptions of immune receptors, despite the arbitrary choice of interaction coordinates, have proven a valuable first step in statistically describing recognition in T cells (195, 212–215). Recently, lower hydrophilicity of regulatory vs. conventional T cells was suggested from CDR3 sequencing (216).

High-throughput sequencing of immune receptors raises specific challenges compared to traditional genomic sequencing. It is harder to distinguish sequencing errors from new polymorphisms, since no corresponding pre-existing sequence exists. One of the most interesting regions when studying diversity is the CDR3 with its many insertions and deletions added to the germline sequence. These regions are often hard to align to the genomic templates, or

with each other (217). Therefore, extra care is needed when generating and analyzing sequence data. Not all sequencing technologies are equally good for all purposes (218): while 454 sequencing gives longer reads than Illumina it is known to have a greater probability of frameshift errors. In addition, primer-dependent PCR amplification biases require that raw sequence counts be normalized using control experiments (112) in order to accurately report clone sizes, as demonstrated by spike-in experiments (219). In TR repertoire studies, this is circumvented by using 5′RACE which provides an unbiased amplification of fully rearranged sequences, as recently demonstrated for TRB V-(D)-J transcripts (191).

Despite sequencing issues, statistical algorithms are often able to extract information from the data. Many studies of diversity focus on the V, D, and J gene usage of each rearranged sequence. Algorithms and tools have been developed to rapidly identify the V, D, and J genes for massive numbers of sequences (148, 178, 181). In many cases however, the assignment of a D gene to each sequence read is unreliable if the D region is too short owing to extensive trimming. Mora et al. (217) learned from data and analyzed statistical models of the D gene flanked by its junctions. These models are based on the principle of maximum entropy and make minimal assumptions about the mechanisms of diversity – they only rely on the observed frequencies of amino acid pairs along the sequence. These models were used to describe global features of the sequence ensemble, such as the probability distribution following Zipf's law (220) – the observation that the probability of sequences is inversely proportional to their frequency-rank, or the observation of peaks of frequency in sequence landscape as possible signatures of past pathogenic challenges. Recently, the estimation of repertoire diversity and clonal size distribution were analyzed by Poisson abundance models (221) and simple bivariate-Poisson-lognormal (BPLN) parametric model for fitting and analyzing TR repertoire data was proposed (222). Similarly, network analysis of IG repertoire from Weinstein et al. study revealed the possibility to identify subgroups of individuals on the basis of IG network similarity (223).

The task of characterizing the CDR3 at the nucleotide level is made difficult by the fact that a deterministic assignment of the V-(D)-J recombination process is impossible, because any given sequence can be generated by many possible recombination processes. A previous study proposed a probabilistic model of nucleotide trimming of rearranged TR genes derived from a benchmark data set of TRA and TRG V-(D)-J junctions obtained by comparison to the germline genes in the IMGT® tools (224). Recently a statistical method based on the expectation-maximization algorithm was proposed to circumvent this issue and to extract the statistical properties of junctional diversity accurately from data (124). Applying it to human non-productive DNA sequences gave insight into a universal generation mechanism, reproducible from individual to individual. It was shown that each sequence could potentially be generated by the equivalent of ∼30 equally likely ways by convergent recombination. This method showed that the potential diversity of the recombination machinery was equivalent to $\sim 10^{14}$ *equally likely* sequences (and a practically infinite total number of possible sequences), much more than the estimated 10^{12} T cells that a single human body can hold. The frequencies of the V, D, and J genes is non-uniform, even at the level of recombination, suggesting underlying physical mechanisms at work. Ndifon et al. (125) proposed a polymer model that accounts for the likelihood of connecting given genomic fragments, giving insight into the mechanistic process.

One of the ultimate goals of deep repertoire sequencing is to find signatures of the repertoire's response to its antigenic environment. A combination of clustering methods and tree reconstruction techniques have been developed (225, 226) to identify lineages in B cells and study the response to pathogenic challenges. Statistical methods have been devised to detect and quantify the extent of antigen-driven selection acting on B cells, by analyzing the patterns of hypermutations in a Bayesian framework, with applications to deep sequencing data (227, 228).

A lot remains to be done in terms of both data-driven and small-scale models of repertoire-antigen interactions. Ultimately, a close collaboration and development of experimental techniques and models can shed light on how selection at different stages shapes the repertoire, how affinity maturation changes the diversity and the link between sequence diversity and function.

FUTURE PROSPECTS OF BIOMATHEMATICAL ANALYSIS OF REPERTOIRE DATA

One of the current challenging issues in antigen-specific repertoire analysis is the development of relevant statistical analysis strategies. Biologists are usually keen on parametric tests, such as ANOVA, *t*-test, Fischer's test, among others. However, such statistical methods assume that the inherent probability distribution of the observed variable follows a normal distribution. Rock et al. (229) described that the distribution of the TR diversity is far from following this distribution, thus they proposed the use of non-parametric tests. Nevertheless, different groups are dealing with this issue in order to determine the relevant way to analyze repertoire diversity data and to propose new biostatistics strategies, including principal component analysis, discriminant analysis, hierarchical clustering, specific statistics (164, 169).

In fact, the traditional use of statistics in biology aims at the falsification of a defined hypothesis, i.e., at validating significant differences between defined situations. The recent development of "systems immunology" reverses this point of view and establishes a new usage of multi-parametric statistical approaches to represent the biological data by projections and "landscapes" in the N-dimensional space of considered parameters (230). Thus, the traditional description of separate repertoires for distinct cell subsets defined from a few markers is being replaced by overlapping clouds of data, setting the limits of the different classification groups (tissue of origin, infection contexts, combination of marker expression, repertoire expression. . .). Moreover, repertoire diversity technologies can now be combined to complementary approaches to decipher the complexity of lymphocyte populations, such as microwell array cell culture and high-resolution imaging (231), mass cytometry (232, 233), cellular barcoding (234), intravital imaging (235, 236), single-cell gene expression (237). In addition, high-throughput repertoire descriptions will enrich mathematical and computer models of lymphocyte repertoire diversity and dynamics such as those proposed by Mehr (238), Ciupe et al. (239), or Stirk et al. (240).

As advocated by others, the concepts developed by systems biology, such as the signatures emerging from clustering and the modularity regulating gene networks, will probably need to be adapted to the constraints of immunology data (241). However, this is probably through this kind of representation that global analysis of immune repertoires will have to be addressed (242).

The upcoming challenge is now to merge data produced through the different technological approaches available to achieve full integration of these data and make them available for interactive meta-analysis. This necessitates more than the simple juxtaposition of annotated raw data but rather requires (1) the codification and standardization of this multi-level data and (2) the integration of complexity science into immunology. Along this line, recent developments of multi-parametric flow cytometry naturally led to systematic clustering and multivariate statistical analysis approaches for searching functional signatures (2, 232, 233, 243–245).

ACKNOWLEDGMENTS

This work was supported by French state funds within the Investissements d'Avenir program (ANR-11-IDEX-0004-02; LabEx Transimmunom), the European Research Council Advanced grant (TRiPoD), the European PCRDT7 (Lifecycle program), the RNSC (ImmunoComplexiT network), CNRS (PEPS BMI), INRA and Université Pierre and Marie Curie.

REFERENCES

1. Boudinot P, Marriotti-Ferrandiz ME, Du Pasquier L, Benmansour A, Cazenave PA, Six A. New perspectives for large-scale repertoire analysis of immune receptors. *Mol Immunol* (2008) **45**:2437–45. doi:10.1016/j.molimm.2007.12.018
2. Bendall SC, Nolan GP, Roederer M, Chattopadhyay PK. A deep profiler's guide to cytometry. *Trends Immunol* (2012) **33**:323–32. doi:10.1016/j.it.2012.02.010
3. MacDonald HR, Pedrazzini T, Schneider R, Louis JA, Zinkernagel RM, Hengartner H. Intrathymic elimination of Mls^a-reactive ($V\beta6^+$) cells during neonatal tolerance induction to Mls^a-encoded antigens. *J Exp Med* (1988) **167**:2005–10. doi:10.1084/jem.167.6.2005
4. MacDonald HR, Schneider R, Lees RK, Howe RC, Acha-Orbea H, Festenstein H, et al. T-cell receptor Vβ use predicts reactivity tolerance to Mls^a-encoded antigens. *Nature* (1988) **332**:40–5. doi:10.1038/332040a0
5. Salaun J, Bandeira A, Khazaal I, Burlen-Defranoux O, Thomas-Vaslin V, Coltey M, et al. Transplantation tolerance is unrelated to superantigen-dependent deletion and anergy. *Proc Natl Acad Sci U S A* (1992) **89**:10420–4. doi:10.1073/pnas.89.21.10420
6. Faint JM, Pilling D, Akbar AN, Kitas GD, Bacon PA, Salmon M. Quantitative flow cytometry for the analysis of T cell receptor Vβ chain expression. *J Immunol Methods* (1999) **225**:53–60. doi:10.1016/S0022-1759(99)00027-7
7. Thomas-Vaslin V, Six A, Pham HP, Dansokho C, Chaara W, Gouritin B, et al. Immunodepression & Immunosuppression during aging. In: Portela MB editor. *Immunosuppression.* Rijeka: InTech open access publisher (2012). p. 125–463.
8. Lefranc MP, Lefranc G. *The Immunoglobulin FactsBook.* London: Academic Press (2001).
9. Lefranc MP, Lefranc G. *The T Cell Receptor FactsBook.* London: Academic Press (2001).
10. Nobrega A, Haury M, Grandien A, Malanchere E, Sundblad A, Coutinho A. Global analysis of antibody repertoires. II. Evidence for specificity, self-selection and the immunological "homunculus" of antibodies in normal serum. *Eur J Immunol* (1993) **23**:2851–9. doi:10.1002/eji.1830231119
11. Haury M, Grandien A, Sundblad A, Coutinho A, Nobrega A. Global analysis of antibody repertoires. 1. An immunoblot method for the quantitative screening of a large number of reactivities. *Scand J Immunol* (1994) **39**:79–87.
12. Fesel C, Coutinho A. Serum IgM repertoire reactions to MBP/CFA immunization reflect the individual status of EAE susceptibility. *J Autoimmun* (2000) **14**:319–24. doi:10.1006/jaut.2000.0373
13. Haury M, Sundblad A, Grandien A, Barreau C, Coutinho A, Nobrega A. The repertoire of serum IgM in normal mice is largely independent of external antigenic contact. *Eur J Immunol* (1997) **27**:1557–63. doi:10.1002/eji.1830270635
14. Stahl D, Lacroix-Desmazes S, Heudes D, Mouthon L, Kaveri SV, Kazatchkine MD. Altered control of self-reactive IgG by autologous IgM in patients with warm autoimmune hemolytic anemia. *Blood* (2000) **95**:328–35.
15. Stahl D, Lacroix-Desmazes S, Mouthon L, Kaveri SV, Kazatchkine MD. Analysis of human self-reactive antibody repertoires by quantitative immunoblotting. *J Immunol Methods* (2000) **240**:1–14. doi:10.1016/S0022-1759(00)00185-X
16. Costa N, Pires AE, Gabriel AM, Goulart LF, Pereira C, Leal B, et al. Broadened T-cell repertoire diversity in ivIg-treated SLE patients is also related to the individual status of regulatory T-cells. *J Clin Immunol* (2013) **33**:349–60. doi:10.1007/s10875-012-9816-7
17. Quintana FJ, Hagedorn PH, Elizur G, Merbl Y, Domany E, Cohen IR. Functional immunomics: microarray analysis of IgG autoantibody repertoires predicts the future response of mice to induced diabetes. *Proc Natl Acad Sci U S A* (2004) **101**:14615–21. doi:10.1073/pnas.0404848101
18. Quintana FJ, Getz G, Hed G, Domany E, Cohen IR. Cluster analysis of human autoantibody reactivities in health and in type 1 diabetes mellitus: a bioinformatic approach to immune complexity. *J Autoimmun* (2003) **21**:65–75. doi:10.1016/S0896-8411(03)00064-7
19. Delassus S, Darche S, Kourilsky P, Cumano A. Ontogeny of the heavy chain immunoglobulin repertoire in fetal liver and bone marrow. *J Immunol* (1998) **160**:3274–80.
20. Yassai M, Gorski J. Thymocyte maturation: selection for in-frame TCRα-chain rearrangement is followed by selection for shorter TCRβ-chain complementarity-determining region 3. *J Immunol* (2000) **165**:3706–12.
21. Pasqual N, Gallagher M, Aude-Garcia C, Loiodice M, Thuderoz F, Demongeot J, et al. Quantitative and qualitative changes in V-Jα rearrangements during mouse thymocytes differentiation: implication for a limited T cell receptor α chain repertoire. *J Exp Med* (2002) **196**:1163–73. doi:10.1084/jem.20021074
22. Douek DC, McFarland RD, Keiser PH, Gage EA, Massey JM, Haynes BF, et al. Changes in thymic function with age and during the treatment of HIV infection. *Nature* (1998) **396**:690–5. doi:10.1038/25374
23. Ribeiro RM, Perelson AS. Determining thymic output quantitatively: using models to interpret experimental T-cell receptor excision circle (TREC) data. *Immunol Rev* (2007) **216**:21–34.
24. Bains I, Thiebaut R, Yates AJ, Callard R. Quantifying thymic export: combining models of naive T cell proliferation and TCR excision circle dynamics gives an explicit measure of thymic output. *J Immunol* (2009) **183**:4329–36. doi:10.4049/jimmunol.0900743
25. Poulin JF, Viswanathan MN, Harris JM, Komanduri KV, Wieder E, Ringuette N, et al. Direct evidence for thymic function in adult humans. *J Exp Med* (1999) **190**:479–86. doi:10.1084/jem.190.4.479
26. Dulude G, Cheynier R, Gauchat D, Abdallah A, Kettaf N, Sékaly RP, et al. The magnitude of thymic output is genetically determined through controlled intrathymic precursor T cell proliferation. *J Immunol* (2008) **181**:7818–24.
27. Wu GE, Paige CJ. VH gene family utilization in colonies derived from B and pre-B cells detected by the RNA colony blot assay. *EMBO J* (1986) **5**:3475–81.
28. Schulze DH, Kelsoe G. Genotypic analysis of B cell colonies by in situ hybridization. Stoichiometric expression of three VH families in adult C57BL/6 and BALB/c mice. *J Exp Med* (1987) **166**:163–72. doi:10.1084/jem.166.1.163
29. Thomas-Vaslin V, Andrade L, Freitas A, Coutinho A. Clonal persistence of B lymphocytes in normal mice is determined by variable region-dependent selection. *Eur J Immunol* (1991) **21**:2239–46. doi:10.1002/eji.1830210935
30. Andrade L, Huetz F, Poncet P, Thomas-Vaslin V, Goodhardt M, Coutinho A. Biased VH gene expression in murine CD5 B cells results from age-dependent cellular selection. *Eur J Immunol* (1991) **21**:2017–23. doi:10.1002/eji.1830210908
31. Freitas AA, Lembezat MP, Coutinho A. Expression of antibody V-regions is genetically and developmentally controlled and modulated by the B lymphocyte environment. *Int Immunol* (1989) **1**:342–54. doi:10.1093/intimm/1.4.342
32. Rosenberg WMC, Moss PAH, Bell JI. Variation in human T cell receptor Vβ and Jβ repertoire: analysis using anchored polymerase reaction. *Eur J Immunol* (1992) **22**:541–9. doi:10.1002/eji.1830220237
33. Moss PAH, Rosenberg WMC, Zintzaras E, Bell JI. Characterization of the human T cell receptor α-chain repertoire and demonstration of α genetic influence on the Vα usage. *Eur J Immunol* (1993) **23**:1155–9. doi:10.1002/eji.1830230526

34. Moss PAH, Bell JI. Sequence analysis of the human αβ T-cell receptor CDR3 region. *Immunogenetics* (1995) **42**:10–8. doi:10.1007/BF00164982
35. Moss PA, Bell JI. Comparative sequence analysis of the human T cell receptor TCRA and TCRB CDR3 regions. *Hum Immunol* (1996) **48**:32–8. doi:10.1016/0198-8859(96)00084-5
36. Pannetier C, Delassus S, Darche S, Saucier C, Kourilsky P. Quantitative titration of nucleic acids by enzymatic amplification reactions run to saturation. *Nucleic Acids Res* (1993) **21**:577–83. doi:10.1093/nar/21.3.577
37. Manfras BJ, Rudert WA, Trucco M, Boehm O. Analysis of the αβ T-cell receptor repertoire by competitive and quantitative family-specific PCR with exogenous standards and high resolution fluorescence based CDR3 size imaging. *J Immunol Methods* (1997) **210**:235–49. doi:10.1016/S0022-1759(97)00197-X
38. VanderBorght A, Van der Aa A, Geusens P, Vandevyver C, Raus J, Stinissen P. Identification of overrepresented T cell receptor genes in blood and tissue biopsies by PCR-ELISA. *J Immunol Methods* (1999) **223**:47–61. doi:10.1016/S0022-1759(98)00201-4
39. Lim A, Baron V, Ferradini L, Bonneville M, Kourilsky P, Pannetier C. Combination of MHC-peptide multimer-based T cell sorting with the Immunoscope permits sensitive ex vivo quantitation and follow-up of human CD8+ T cell immune responses. *J Immunol Methods* (2002) **261**:177–94. doi:10.1016/S0022-1759(02)00004-2
40. Casrouge A, Beaudoing E, Dalle S, Pannetier C, Kanellopoulos J, Kourilsky P. Size estimate of the αβ TCR repertoire of naive mouse splenocytes. *J Immunol* (2000) **164**:5782–7.
41. Gallard A, Foucras G, Coureau C, Guery JC. Tracking T cell clonotypes in complex T lymphocyte populations by real-time quantitative PCR using fluorogenic complementarity-determining region-3-specific probes. *J Immunol Methods* (2002) **270**:269–80. doi:10.1016/S0022-1759(02)00336-8
42. Liu DB, Callahan JP, Dau PC. Intrafamily fragment analysis of the T cell receptor β chain CDR3 region. *J Immunol Methods* (1995) **187**:139–50. doi:10.1016/0022-1759(95)00178-D
43. Mugnaini EN, Egeland T, Syversen AM, Spurkland A, Brinchmann JE. Molecular analysis of the complementarity determining region 3 of the human T cell receptor β chain. Establishment of a reference panel of CDR3 lengths from phytohaemagglutinin activated lymphocytes. *J Immunol Methods* (1999) **223**:207–16. doi:10.1016/S0022-1759(99)00004-6
44. Matsutani T, Yoshioka T, Tsuruta Y, Iwagami S, Suzuki R. Analysis of TCRAV and TCRBV repertoires in healthy individuals by microplate hybridization assay. *Hum Immunol* (1997) **56**:57–69. doi:10.1016/S0198-8859(97)00102-X
45. Matsutani T, Yoshioka T, Tsuruta Y, Iwagami S, Toyosaki-Maeda T, Horiuchi T, et al. Restricted usage of T-cell receptor α-chain variable region (TCRAV) and T-cell receptor β-chain variable region (TCRBV) repertoires after human allogeneic haematopoietic transplantation. *Br J Haematol* (2000) **109**:759–69. doi:10.1046/j.1365-2141.2000.02080.x
46. Yoshida R, Yoshioka T, Yamane S, Matsutani T, Toyosaki-Maeda T, Tsuruta Y, et al. A new method for quantitative analysis of the mouse T-cell receptor V region repertoires: comparison of repertoires among strains. *Immunogenetics* (2000) **52**:35–45. doi:10.1007/s002510000248
47. Du G, Qiu L, Shen L, Sehgal P, Shen Y, Huang D, et al. Combined megaplex TCR isolation and SMART-based real-time quantitation methods for quantitating antigen-specific T cell clones in mycobacterial infection. *J Immunol Methods* (2006) **308**:19–35. doi:10.1016/j.jim.2005.09.009
48. Ogle BM, Cascalho M, Joao C, Taylor W, West LJ, Platt JL. Direct measurement of lymphocyte receptor diversity. *Nucleic Acids Res* (2003) **31**:e139. doi:10.1093/nar/gng139
49. Joao C, Ogle BM, Gay-Rabinstein C, Platt JL, Cascalho M. B cell-dependent TCR diversification. *J Immunol* (2004) **172**:4709–16.
50. Joao C. Immunoglobulin is a highly diverse self-molecule that improves cellular diversity and function during immune reconstitution. *Med Hypotheses* (2007) **68**:158–61. doi:10.1016/j.mehy.2006.05.062
51. Bonarius HP, Baas F, Remmerswaal EB, van Lier RA, ten Berge I, Tak PP, et al. Monitoring the T-cell receptor repertoire at single-clone resolution. *PLoS One* (2006) **1**:e55. doi:10.1371/journal.pone.0000055
52. Pannetier C, Cochet M, Darche S, Casrouge A, Zöller M, Kourilsky P. The size of the CDR3 hypervariable regions of the murine T-cell receptor β chains vary as a function of the recombined germ-line segments. *Proc Natl Acad Sci U S A* (1993) **90**:4319–23. doi:10.1073/pnas.90.9.4319
53. Pannetier C, Even J, Kourilsky P. T-cell repertoire diversity and clonal expansions in normal and clinical samples. *Immunol Today* (1995) **16**:176–81. doi:10.1016/0167-5699(95)80117-0
54. Pannetier C, Levraud JP, Lim A, Even J, Kourilsky P. The immunoscope approach for the analysis of T-cell repertoires. In: Oksenberg J editor. *The Human Antigen T Cell Receptor. Selected Protocols and Applications*. Georgetown, TX: Landes RG (1997). p. 287–325.
55. Cochet M, Pannetier C, Darche S, Leclerc C, Kourilsky P. Molecular detection and *in vivo* analysis of the specific T cell response to a protein antigen. *Eur J Immunol* (1992) **22**:2639–47. doi:10.1002/eji.1830221025
56. Regnault A, Cumano A, Vassalli P, Guy-Grand D, Kourilsky P. Oligoclonal repertoire of the CD8αα and the CD8αβ TCR-α/β murine intestinal intraepithelial T lymphocytes: evidence for the random emergence of T cells. *J Exp Med* (1994) **180**:1345–58. doi:10.1084/jem.180.4.1345
57. Regnault A, Levraud JP, Lim A, Six A, Moreau C, Cumano A, et al. The expansion and selection of T cell receptor αβ intestinal intraepithelial T cell clones. *Eur J Immunol* (1996) **26**:914–21. doi:10.1002/eji.1830260429
58. Ema H, Cumano A, Kourilsky P. TCRβ repertoire development in the mouse embryo. *J Immunol* (1997) **159**:4227–32.
59. Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P. A direct estimate of the human αβ T cell receptor diversity. *Science* (1999) **286**:958–61. doi:10.1126/science.286.5441.958
60. Bousso P, Lemaître F, Laouini D, Kanellopoulos J, Kourilsky P. The peripheral CD8 T cell repertoire is largely independent of the presence of intestinal flora. *Int Immunol* (2000) **12**:425–30. doi:10.1093/intimm/12.4.425
61. Arstila TP, Even J. Size of the αβ TCR repertoire. *Med Sci (Paris)* (2000) **16**:1257–60. doi:10.4267/10608/1566
62. Cabaniols JP, Fazilleau N, Casrouge A, Kourilsky P, Kanellopoulos JM. Most α/β T cell receptor diversity is due to terminal deoxynucleotidyl transferase. *J Exp Med* (2001) **194**:1385–90. doi:10.1084/jem.194.9.1385
63. Fazilleau N, Cabaniols JP, Lemaitre F, Motta I, Kourilsky P, Kanellopoulos JM. Valpha and Vbeta public repertoires are highly conserved in terminal deoxynucleotidyl transferase-deficient mice. *J Immunol* (2005) **174**:345–55.
64. Cibotti R, Cabaniols JP, Pannetier C, Delarbre C, Vergnon I, Kanellopoulos JM, et al. Public and private Vβ T cell receptor repertoires against hen egg white lysozyme (HEL) in nontransgenic versus HEL transgenic mice. *J Exp Med* (1994) **180**:861–72. doi:10.1084/jem.180.3.861
65. Gapin L, Fukui Y, Kanellopoulos J, Sano T, Casrouge A, Malier V, et al. Quantitative analysis of the T cell repertoire selected by a single peptide-major histocompatibility complex. *J Exp Med* (1998) **187**:1871–83. doi:10.1084/jem.187.11.1871
66. Bouneaud C, Kourilsky P, Bousso P. Impact of negative selection on the T cell repertoire reactive to a self-peptide: a large fraction of T cell clones escapes clonal deletion. *Immunity* (2000) **13**:829–40. doi:10.1016/S1074-7613(00)00080-7
67. Fukui Y, Oono T, Cabaniols JP, Nakao K, Hirokawa K, Inayoshi A, et al. Diversity of T cell repertoire shaped by a single peptide ligand is critically affected by its amino acid residue at a T cell receptor contact. *Proc Natl Acad Sci U S A* (2000) **97**:13760–5. doi:10.1073/pnas.250470797
68. Musette P, Bureau JF, Gachelin G, Kourilsky P, Brahic M. T lymphocyte repertoire in Theiler's virus encephalomyelitis: the nonspecific infiltration of the central nervous system of infected SJL/J mice is associated with a selective local T cell expansion. *Eur J Immunol* (1995) **25**:1589–93. doi:10.1002/eji.1830250618
69. Sourdive DJD, Murali-Krishna K, Altman JD, Zajac AJ, Whitmire JK, Pannetier C, et al. Conserved T cell receptor repertoire in primary and memory CD8 T cell responses to an acute viral infection. *J Exp Med* (1998) **188**:71–82. doi:10.1084/jem.188.1.71
70. Musette P, Bequet D, Delarbre C, Gachelin G, Kourilsky P, Dormont D. Expansion of a recurrent Vβ5.3$^+$ T-cell population in newly diagnosed and untreated HLA-DR2 multiple sclerosis patients. *Proc Natl Acad Sci U S A* (1996) **93**:12461–6. doi:10.1073/pnas.93.22.12461
71. Fazilleau N, Delarasse C, Sweenie CH, Anderton SM, Fillatreau S, Lemonnier FA, et al. Persistence of autoreactive myelin oligodendrocyte glycoprotein (MOG)-specific T cell repertoires in MOG-expressing mice. *Eur J Immunol* (2006) **36**:533–43. doi:10.1002/eji.200535021
72. Musette P, Bachelez H, Flageul B, Delarbre C, Kourilsky P, Dubertret L, et al. Immune-mediated destruction of melanocytes in halo nevi is associated with

the local expansion of a limited number of T cell clones. *J Immunol* (1999) **162**:1789–94.
73. Douillard P, Pannetier C, Josien R, Menoret S, Kourilsky P, Soulillou JP, et al. Donor-specific blood transfusion-induced tolerance in adult rats with a dominant TCR-Vβ rearrangement in heart allografts. *J Immunol* (1996) **157**:1250–60.
74. Pilch H, Höhn H, Freitag K, Neukirch C, Necker A, Haddad P, et al. Improved assessment of T-cell receptor (TCR) VB repertoire in clinical specimens: combination of TCR-CDR3 spectratyping with flow cytometry-based TCR VB frequency analysis. *Clin Diagn Lab Immunol* (2002) **9**:257–66.
75. Messaoudi I, LeMaoult J, Guevara-Patino JA, Metzner BM, Nikolich-Zugich J. Age-related CD8 T cell clonal expansions constrict CD8 T cell repertoire and have the potential to impair immune defense. *J Exp Med* (2004) **200**:1347–58. doi:10.1084/jem.20040437
76. Hakim FT, Memon SA, Cepeda R, Jones EC, Chow CK, Kasten-Sportes C, et al. Age-dependent incidence, time course, and consequences of thymic renewal in adults. *J Clin Invest* (2005) **115**:930–9. doi:10.1172/JCI200522492
77. Messaoudi I, Warner J, Fischer M, Park B, Hill B, Mattison J, et al. Delay of T cell senescence by caloric restriction in aged long-lived nonhuman primates. *Proc Natl Acad Sci U S A* (2006) **103**:19448–53. doi:10.1073/pnas.0606661103
78. Boudinot P, Boubekeur S, Benmansour A. Rhabdovirus infection induces public and private T cell responses in teleost fish. *J Immunol* (2001) **167**:6202–9.
79. Boudinot P, Boubekeur S, Benmansour A. Primary structure and complementarity-determining region (CDR) 3 spectratyping of rainbow trout TCRβ transcripts identify ten Vβ families with Vβ6 displaying unusual CDR2 and differently spliced forms. *J Immunol* (2002) **169**:6244–52.
80. Boudinot P, Bernard D, Boubekeur S, Thoulouze MI, Bremont M, Benmansour A. The glycoprotein of a fish rhabdovirus profiles the virus-specific T-cell repertoire in rainbow trout. *J Gen Virol* (2004) **85**:3099–108. doi:10.1099/vir.0.80135-0
81. Desravines S, Hsu E. Measuring CDR3 length variability in individuals during ontogeny. *J Immunol Methods* (1994) **168**:219–25. doi:10.1016/0022-1759(94)90058-2
82. Gorski J, Yassai M, Zhu X, Kissella B, Keever C, Flomenberg N. Circulating T cell repertoire complexity in normal individuals and bone marrow recipients analyzed by CDR3 spectratyping. *J Immunol* (1994) **152**:5109–19.
83. Maslanka K, Piatek T, Gorski J, Yassai M. Molecular analysis of T cell repertoires – Spectratypes generated by multiplex polymerase chain reaction and evaluated by radioactivity or fluorescence. *Hum Immunol* (1995) **44**:28–34.
84. Lue C, Mitani Y, Crew MD, George JF, Fink LM, Schichman SA. An automated method for the analysis of T-cell receptor repertoires: rapid RT-PCR fragment length analysis of the T-cell receptor β chain complementarity-determining region 3. *Am J Clin Pathol* (1999) **111**:683–90.
85. Yamamoto K, Masuko-Hongo K, Tanaka A, Kurokawa M, Hoeger T, Nishioka K, et al. Establishment and application of a novel T cell clonality analysis using single-strand conformation polymorphism of T cell receptor messenger signals. *Hum Immunol* (1996) **48**:23–31. doi:10.1016/0198-8859(96)00080-8
86. Shiokawa S, Nishimura J, Ohshima K, Uike N, Yamamoto K. Establishment of a novel B cell clonality analysis using single-strand conformation polymorphism of immunoglobulin light chain messenger signals. *Am J Pathol* (1998) **153**:1393–400. doi:10.1016/S0002-9440(10)65726-4
87. Raaphorst FM, Gokmen E, Teale JM. Analysis of clonal diversity in mouse immunoglobulin heavy chain genes selected for size of the antigen combining site. *Immunol Invest* (1998) **27**:355–65. doi:10.3109/08820139809022709
88. Sottini A, Quirós Roldan E, Albertini A, Primi D, Imberti L. Assessment of T-cell receptor beta-chain diversity by heteroduplex analysis. *Hum Immunol* (1996) **48**:12–22. doi:10.1016/0198-8859(96)00087-0
89. Wack A, Montagna D, Dellabona P, Casorati G. An improved PCR-heteroduplex method permits high-sensitivity detection of clonal expansions in complex T cell populations. *J Immunol Methods* (1996) **196**:181–92. doi:10.1016/0022-1759(96)00114-7
90. Shen DF, Doukhan L, Kalam S, Delwart E. High-resolution analysis of T-cell receptor β-chain repertoires using DNA heteroduplex tracking: generally stable, clonal $CD8^+$ expansions in all healthy young adults. *J Immunol Methods* (1998) **215**:113–21. doi:10.1016/S0022-1759(98)00066-0
91. Wedderburn LR, Maini MK, Patel A, Beverley PCL, Woo P. Molecular fingerprinting reveals non-overlapping T cell oligoclonality between an inflamed site and peripheral blood. *Int Immunol* (1999) **11**:535–43. doi:10.1093/intimm/11.4.535
92. Bouffard P, Gagnon C, Cloutier D, MacLean SJ, Souleimani A, Nallainathan D, et al. Analysis of T cell receptor β chain expression by isoelectric focusing following gene amplification and *in vitro* translation. *J Immunol Methods* (1995) **187**:9–21. doi:10.1016/0022-1759(95)00161-3
93. Sant'Angelo DB, Lucas B, Waterbury PG, Cohen B, Brabb T, Goverman J, et al. A molecular map of T cell development. *Immunity* (1998) **9**:179–86. doi:10.1016/S1074-7613(00)80600-7
94. Correia-Neves M, Waltzinger C, Mathis D, Benoist C. The shaping of the T cell repertoire. *Immunity* (2001) **14**:21–32. doi:10.1016/S1074-7613(01)00086-3
95. Hsieh CS, Liang Y, Tyznik AJ, Self SG, Liggitt D, Rudensky AY. Recognition of the peripheral self by naturally arising CD25+ CD4+ T cell receptors. *Immunity* (2004) **21**:267–77. doi:10.1016/j.immuni.2004.07.009
96. Hsieh CS, Zheng Y, Liang Y, Fontenot JD, Rudensky AY. An intersection between the self-reactive regulatory and nonregulatory T cell receptor repertoires. *Nat Immunol* (2006) **7**:401–10. doi:10.1038/ni1318
97. Pacholczyk R, Ignatowicz H, Kraj P, Ignatowicz L. Origin and T cell receptor diversity of $Foxp3^+CD4^+CD25^+$ T cells. *Immunity* (2006) **25**:249–59. doi:10.1016/j.immuni.2006.05.016
98. Pacholczyk R, Kern J, Singh N, Iwashima M, Kraj P, Ignatowicz L. Nonself-antigens are the cognate specificities of $Foxp3^+$ regulatory T cells. *Immunity* (2007) **27**:493–504. doi:10.1016/j.immuni.2007.07.019
99. Wong J, Obst R, Correia-Neves M, Losyev G, Mathis D, Benoist C. Adaptation of TCR repertoires to self-peptides in regulatory and nonregulatory $CD4^+$ T cells. *J Immunol* (2007) **178**:7032–41.
100. Kang JA, Mohindru M, Kang BS, Park SH, Kim BS. Clonal expansion of infiltrating T cells in the spinal cords of SJL/J mice infected with Theiler's virus. *J Immunol* (2000) **165**:583–90.
101. Apostolou I, Cumano A, Gachelin G, Kourilsky P. Evidence for two subgroups of $CD4^-CD8^-$ NKT cells with distinct TCRαβ repertoires and differential distribution in lymphoid tissues. *J Immunol* (2000) **165**:2481–90.
102. Bernard D, Six A, Rigottier-Gois L, Messiaen S, Chilmonczyk S, Quillet E, et al. Phenotypic and functional similarity of gut intraepithelial and systemic T cells in a teleost fish. *J Immunol* (2006) **176**:3942–9.
103. Mancini S, Candéias SM, Fehling HJ, von Boehmer H, Jouvin-Marche E, Marche PN. TCR α-chain repertoire in pTα-deficient mice is diverse and developmentally regulated: implications for pre-TCR functions and TCRA gene rearrangement. *J Immunol* (1999) **163**:6053–9.
104. Halapi E, Werner A, Wahlström J, Österborg A, Jeddi-Tehrani M, Yi Q, et al. T cell repertoire in patients with multiple myeloma and monoclonal gammopathy of undetermined significance: clonal $CD8^+$ T cell expansions are found preferentially in patients with a low tumor burden. *Eur J Immunol* (1997) **27**:2245–52. doi:10.1002/eji.1830270919
105. Brawand P, Cerottini JC, MacDonald HR. Hierarchal utilization of different T-cell receptor Vβ gene segments in the $CD8^+$-T-cell response to an immunodominant Moloney leukemia virus-encoded epitope *in vivo*. *J Virol* (1999) **73**:9161–9.
106. Matsuzaki G, Takada H, Nomoto K. *Escherichia coli* infection induces only fetal thymus-derived γδ T cells at the infected site. *Eur J Immunol* (1999) **29**:3877–86. doi:10.1002/(SICI)1521-4141(199912)29:12<3877::AID-IMMU3877>3.3.CO;2-3
107. Collette A, Bagot S, Ferrandiz ME, Cazenave PA, Six A, Pied S. A profound alteration of blood TCRB repertoire allows prediction of cerebral malaria. *J Immunol* (2004) **173**:4568–75.
108. Douek DC, Betts MR, Brenchley JM, Hill BJ, Ambrozak DR, Ngai KL, et al. A novel approach to the analysis of specificity, clonality, and frequency of HIV-specific T cell responses reveals a potential mechanism for control of viral escape. *J Immunol* (2002) **168**:3099–104.
109. Lim A, Lemercier B, Wertz X, Pottier SL, Huetz F, Kourilsky P. Many human peripheral VH5-expressing IgM+ B cells display a unique heavy-chain rearrangement. *Int Immunol* (2008) **20**:105–16. doi:10.1093/intimm/dxm125
110. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. *Clin Chem* (2009) **55**:641–58. doi:10.1373/clinchem.2008.112789
111. McGinn S, Gut IG. DNA sequencing – spanning the generations. *Nat Biotechnol* (2013) **30**:366–72. doi:10.1016/j.nbt.2012.11.012

112. Weinstein JA, Jiang N, White RA III, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. *Science* (2009) **324**:807–10. doi:10.1126/science.1170020
113. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, et al. Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. *Blood* (2009) **114**:4099–107. doi:10.1182/blood-2009-04-217604
114. Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res* (2009) **19**:1817–24. doi:10.1101/gr.092924.109
115. Wang C, Sanders CM, Yang Q, Schroeder HW, Wang E, Babrzadeh F, et al. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proc Natl Acad Sci U S A* (2010) **107**:1518–23. doi:10.1073/pnas.0913939107
116. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, et al. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res* (2011) **21**:790–7. doi:10.1101/gr.115428.110
117. Venturi V, Quigley MF, Greenaway HY, Ng PC, Ende ZS, McIntosh T, et al. A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. *J Immunol* (2011) **186**:4285–94. doi:10.4049/jimmunol.1003898
118. Wu D, Sherwood A, Fromm JR, Winter SS, Dunsmore KP, Loh ML, et al. High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. *Sci Transl Med* (2012) **4**:134ra63. doi:10.1126/scitranslmed.3003656
119. Föhse L, Suffner J, Suhre K, Wahl B, Lindner C, Lee CW, et al. High TCR diversity ensures optimal function and homeostasis of Foxp3$^+$ regulatory T cells. *Eur J Immunol* (2011) **41**:3101–13. doi:10.1002/eji.201141986
120. Sherwood AM, Desmarais C, Livingston RJ, Andriesen J, Haussler M, Carlson CS, et al. Deep sequencing of the human TCRγ and TCRβ repertoires suggests that TCRβ rearranges after αβ and γδ T cell commitment. *Sci Transl Med* (2011) **3**:90ra61. doi:10.1126/scitranslmed.3002536
121. Cebula A, Seweryn M, Rempala GA, Pabla SS, McIndoe RA, Denning TL, et al. Thymus-derived regulatory T cells contribute to tolerance to commensal microbiota. *Nature* (2013) **497**:258–62. doi:10.1038/nature12079
122. Bashford-Rogers RJM, Palser AL, Huntly BJ, Rance R, Vassiliou GS, Follows GA, et al. Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res* (2013) **23**:1874–84. doi:10.1101/gr.154815.113
123. Srivastava SK, Robins HS. Palindromic nucleotide analysis in human T cell receptor rearrangements. *PLoS One* (2012) **7**:e52250. doi:10.1371/journal.pone.0052250
124. Murugan A, Mora T, Walczak AM, Callan CG. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc Natl Acad Sci U S A* (2012) **109**:16161–6. doi:10.1073/pnas.1212755109
125. Ndifon W, Gal H, Shifrut E, Aharoni R, Yissachar N, Waysbort N, et al. Chromatin conformation governs T-cell receptor Jβ gene segment usage. *Proc Natl Acad Sci U S A* (2012) **109**:15865–70. doi:10.1073/pnas.1203916109
126. Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, et al. Overlap and effective size of the human CD8$^+$ T cell receptor repertoire. *Sci Transl Med* (2010) **2**:47ra64. doi:10.1126/scitranslmed.3001442
127. Prabakaran P, Chen W, Singarayan MG, Stewart CC, Streaker E, Feng Y, et al. 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* (2012) **64**:337–50. doi:10.1007/s00251-011-0595-8
128. van Heijst JW, Ceberio I, Lipuma LB, Samilo DW, Wasilewski GD, Gonzales AM, et al. Quantitative assessment of T cell repertoire recovery after hematopoietic stem cell transplantation. *Nat Med* (2013) **19**:372–7. doi:10.1038/nm.3100
129. Meier J, Roberts C, Avent K, Hazlett A, Berrie J, Payne K, et al. Fractal organization of the human T cell repertoire in health and after stem cell transplantation. *Biol Blood Marrow Transplant* (2013) **19**:366–77. doi:10.1016/j.bbmt.2012.12.004
130. Ademokun A, Wu Y-C, Martin V, Mitra R, Sack U, Baxendale H, et al. Vaccination-induced changes in human B-cell repertoire and pneumococcal IgM and IgA antibody at different ages. *Aging Cell* (2011) **10**:922–30. doi:10.1111/j.1474-9726.2011.00732.x
131. Castro R, Jouneau L, Pham HP, Bouchez O, Giudicelli V, Lefranc MP, et al. Teleost fish mount complex clonal IgM and IgT responses in spleen upon systemic viral infection. *PLoS Pathog* (2013) **9**:e1003098. doi:10.1371/journal.ppat.1003098
132. Bousso P, Casrouge A, Altman JD, Haury M, Kanellopoulos J, Abastado JP, et al. Individual variations in the murine T cell response to a specific peptide reflect variability in naive repertoires. *Immunity* (1998) **9**:169–78. doi:10.1016/S1074-7613(00)80599-3
133. Lin MY, Welsh RM. Stability and diversity of T cell receptor repertoire usage during lymphocytic choriomeningitis virus infection of mice. *J Exp Med* (1998) **188**:1993–2005. doi:10.1084/jem.188.11.1993
134. Metzker ML. Sequencing technologies – the next generation. *Nat Rev Genet* (2010) **11**:31–46. doi:10.1038/nrg2626
135. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* (2005) **437**:376–80.
136. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* (2011) **475**:348–52. doi:10.1038/nature10242
137. Bolotin DA, Mamedov IZ, Britanova OV, Zvyagin IV, Shagin D, Ustyugova SV, et al. Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. *Eur J Immunol* (2012) **42**:3073–83. doi:10.1002/eji.201242517
138. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS Comput Biol* (2013) **9**:e1003031. doi:10.1371/journal.pcbi.1003031
139. Shendure J, Aiden EL. The expanding scope of DNA sequencing. *Nat Biotechnol* (2012) **30**:1084–94. doi:10.1038/nbt.2421
140. Dash P, McClaren JL, Oguin TH III, Rothwell W, Todd B, Morris MY, et al. Paired analysis of TCRα and TCRβ chains at the single-cell level in mice. *J Clin Invest* (2011) **121**:288–95. doi:10.1172/JCI44752
141. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol* (2013) **31**:166–9. doi:10.1038/nbt.2492
142. Turchaninova MA, Britanova OV, Bolotin DA, Shugay M, Putintseva EV, Staroverov DB, et al. Pairing of T-cell receptor chains via emulsion PCR. *Eur J Immunol* (2013) **43**:2507–15. doi:10.1002/eji.201343453
143. Plessy C, Desbois L, Fujii T, Carninci P. Population transcriptomics with single-cell resolution: a new field made possible by microfluidics: a technology for high throughput transcript counting and data-driven definition of cell types. *Bioessays* (2013) **35**:131–40. doi:10.1002/bies.201200093
144. Mehr R, Sternberg-Simon M, Michaeli M, Pickman Y. Models and methods for analysis of lymphocyte repertoire generation, development, selection and evolution. *Immunol Lett* (2012) **148**:11–22. doi:10.1016/j.imlet.2012.08.002
145. Pancer Z, Amemiya CT, Ehrhardt GR, Ceitlin J, Gartland GL, Cooper MD. Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. *Nature* (2004) **430**:174–80. doi:10.1038/nature02740
146. Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, et al. IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res* (2009) **37**:D1006–12. doi:10.1093/nar/gkn838
147. Giudicelli V, Lefranc MP. IMGT-ONTOLOGY 2012. *Front Genet* (2012) **3**:79. doi:10.3389/fgene.2012.00079
148. Alamyar E, Giudicelli V, Li S, Duroux P, Lefranc MP. IMGT/HighV-QUEST: the IMGT(R) web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. *Immunome Res* (2012) **8**:26. doi:10.1007/978-1-61779-842-9_32
149. Watson FL, Puttmann-Holgado R, Thomas F, Lamar DL, Hughes M, Kondo M, et al. Extensive diversity of Ig-superfamily proteins in the immune system of insects. *Science* (2005) **309**:1874–8. doi:10.1126/science.1116887
150. Du Pasquier L. Insects diversify one molecule to serve two systems. *Science* (2005) **309**:1826–7. doi:10.1126/science.1118828
151. Zhang SM, Adema CM, Kepler TB, Loker ES. Diversification of Ig superfamily genes in an invertebrate. *Science* (2004) **305**:251–4. doi:10.1126/science.1088069
152. Philipp EER, Kraemer L, Melzner F, Poustka AJ, Thieme S, Findeisen U, et al. Massively parallel RNA sequencing identifies a complex immune gene repertoire in the Lophotrochozoan *Mytilus edulis*. *PLoS One* (2012) **7**:e33091. doi:10.1371/journal.pone.0033091

153. Matsushima N, Tanaka T, Enkhbayar P, Mikami T, Taga M, Yamada K, et al. Comparative sequence analysis of leucine-rich repeats (LRRs) within vertebrate toll-like receptors. *BMC Genomics* (2007) **8**:124. doi:10.1186/1471-2164-8-124
154. Matsushima N, Miyashita H, Mikami T, Kuroki Y. A nested leucine rich repeat (LRR) domain: the precursor of LRRs is a ten or eleven residue motif. *BMC Microbiol* (2010) **10**:235. doi:10.1186/1471-2180-10-235
155. Kaas Q, Ehrenmann F, Lefranc MP. IG, TR and IgSF, MHC and MhcSF: what do we learn from the IMGT Colliers de Perles? *Brief Funct Genomic Proteomic* (2007) **6**:253–64. doi:10.1093/bfgp/elm032
156. Bomberger C, Singh-Jairam M, Rodey G, Guerriero A, Yeager AM, Fleming WH, et al. Lymphoid reconstitution after autologous PBSC transplantation with FACS-sorted CD34+ hematopoietic progenitors. *Blood* (1998) **91**:2588–600.
157. Gorochov G, Neumann AU, Kereveur A, Parizot C, Li TS, Katlama C, et al. Perturbation of $CD4^+$ and $CD8^+$ T-cell repertoires during progression to AIDS and regulation of the $CD4^+$ repertoire during antiviral therapy. *Nat Med* (1998) **4**:215–21. doi:10.1038/nm0298-215
158. Wu CJ, Chillemi A, Alyea EP, Orsini E, Neuberg D, Soiffer RJ, et al. Reconstitution of T-cell receptor repertoire diversity following T-cell depleted allogeneic bone marrow transplantation is related to hematopoietic chimerism. *Blood* (2000) **95**:352–9.
159. Hori S, Collette A, Demengeot J, Stewart J. A new statistical method for quantitative analyses: application to the precise quantification of T cell receptor repertoires. *J Immunol Methods* (2002) **268**:159–70. doi:10.1016/S0022-1759(02)00187-4
160. Collette A, Six A. ISEApeaks: an excel platform for GeneScan and Immunoscope data retrieval, management and analysis. *Bioinformatics* (2002) **18**:329–30. doi:10.1093/bioinformatics/18.2.329
161. Guillet M, Brouard S, Gagne K, Sebille F, Cuturi MC, Delsuc MA, et al. Different qualitative and quantitative regulation of Vβ TCR transcripts during early acute allograft rejection and tolerance induction. *J Immunol* (2002) **168**:5088–95.
162. Peggs KS, Verfuerth S, D'Sa S, Yong K, Mackinnon S. Assessing diversity: immune reconstitution and T-cell receptor BV spectratype analysis following stem cell transplantation. *Br J Haematol* (2003) **120**:154–65. doi:10.1046/j.1365-2141.2003.04036.x
163. Long SA, Khalili J, Ashe J, Berenson R, Ferrand C, Bonyhadi M. Standardized analysis for the quantification of Vbeta CDR3 T-cell receptor diversity. *J Immunol Methods* (2006) **317**:100–13. doi:10.1016/j.jim.2006.09.015
164. Miqueu P, Guillet M, Degauque N, Dore JC, Soulillou JP, Brouard S. Statistical analysis of CDR3 length distributions for the assessment of T and B cell repertoire biases. *Mol Immunol* (2007) **44**:1057–64. doi:10.1016/j.molimm.2006.06.026
165. Guillet M, Sebille F, Soulillou JP. TCR usage in naive and committed alloreactive cells: implications for the understanding of TCR biases in transplantation. *Curr Opin Immunol* (2001) **13**:566–71. doi:10.1016/S0952-7915(00)00260-0
166. Collette A, Cazenave PA, Pied S, Six A. New methods and software tools for high throughput CDR3 spectratyping. Application to T lymphocyte repertoire modifications during experimental malaria. *J Immunol Methods* (2003) **278**:105–16. doi:10.1016/S0022-1759(03)00225-4
167. Sassi A, Largueche-Darwaz B, Collette A, Six A, Laouini D, Cazenave PA, et al. Mechanisms of the natural reactivity of lymphocytes from noninfected individuals to membrane-associated *Leishmania infantum* antigens. *J Immunol* (2005) **174**:3598–607.
168. Castro R, Takizawa F, Chaara W, Lunazzi A, Dang TH, Koellner B, et al. Contrasted TCRβ diversity of $CD8^+$ and $CD8^-$ T cells in rainbow trout. *PLoS One* (2013) **8**:e60175. doi:10.1371/journal.pone.0060175
169. Kepler TB, He M, Tomfohr JK, Devlin BH, Sarzotti M, Markert ML. Statistical analysis of antigen receptor spectratype data. *Bioinformatics* (2005) **21**:3394–400. doi:10.1093/bioinformatics/bti539
170. He M, Tomfohr JK, Devlin BH, Sarzotti M, Markert ML, Kepler TB. SpA: web-accessible spectratype analysis: data management, statistical analysis and visualization. *Bioinformatics* (2005) **21**:3697–9. doi:10.1093/bioinformatics/bti600
171. Liu C, He M, Rooney B, Kepler TB, Chao NJ. Longitudinal analysis of T-cell receptor variable beta chain repertoire in patients with acute graft-versus-host disease after allogeneic stem cell transplantation. *Biol Blood Marrow Transplant* (2006) **12**:335–45. doi:10.1016/j.bbmt.2005.09.019
172. Lefranc MP. From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT standardized gene and allele nomenclature: for immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* (2011) **2011**:627–32. doi:10.1101/pdb.ip84
173. Lefranc MP. From IMGT-ONTOLOGY DESCRIPTION axiom to IMGT standardized labels: for immunoglobulin (IG) and T cell receptor (TR) sequences and structures. *Cold Spring Harb Protoc* (2011) **2011**:614–26. doi:10.1101/pdb.ip84
174. Lefranc MP. From IMGT-ONTOLOGY IDENTIFICATION axiom to IMGT standardized keywords: for immunoglobulins (IG), T cell receptors (TR), and conventional genes. *Cold Spring Harb Protoc* (2011) **2011**:604–13. doi:10.1101/pdb.ip84
175. Lefranc MP. IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* (2011) **2011**:633–42. doi:10.1101/pdb.ip85
176. Giudicelli V, Chaume D, Lefranc MP. IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. *Nucleic Acids Res* (2004) **32**:W435–40. doi:10.1093/nar/gkh412
177. Brochet X, Lefranc MP, Giudicelli V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res* (2008) **36**:W503–8. doi:10.1093/nar/gkn316
178. Gaëta BA, Malming HR, Jackson KJL, Bain ME, Wilson P, Collins AM. iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics* (2007) **23**:1580–7. doi:10.1093/bioinformatics/btm147
179. Rogosch T, Kerzel S, Hoi KH, Zhang Z, Maier RF, Ippolito GC, et al. Immunoglobulin analysis tool: a novel tool for the analysis of human and mouse heavy and light chain transcripts. *Front Immunol* (2012) **3**:176. doi:10.3389/fimmu.2012.00176
180. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* (2013) **41**:W34–40. doi:10.1093/nar/gkt382
181. Thomas N, Heather J, Ndifon W, Shawe-Taylor J, Chain B. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics* (2013) **29**:542–50. doi:10.1093/bioinformatics/btt004
182. Pham HP, Manuel M, Petit N, Klatzmann D, Cohen-Kaminsky S, Six A, et al. Half of the T-cell repertoire combinatorial diversity is genetically determined in humans and humanized mice. *Eur J Immunol* (2012) **42**:760–70. doi:10.1002/eji.201141798
183. Eisenstein M. Personalized, sequencing-based immune profiling spurs startups. *Nat Biotechnol* (2013) **31**:184–6. doi:10.1038/nbt0313-184b
184. Stahl D, Lacroix-Desmazes S, Barreau C, Sibrowski W, Kazatchkine MD, Kaveri SV. Altered antibody repertoires of plasma IgM and IgG toward nonself antigens in patients with warm autoimmune hemolytic anemia. *Hum Immunol* (2001) **62**:348–61. doi:10.1016/S0198-8859(01)00225-7
185. Magurran AE. *Measuring Biological Diversity*. Oxford: Wiley-Blackwell (2004).
186. Colwell RK. *EstimateS: Statistical Estimation of Species Richness and Shared Species from Samples. [Version 9]. User's Guide and application* (2013). Available from: http://purl.oclc.org/estimates
187. Wu TT, Kabat EA. An analysis of the sequence of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J Exp Med* (1970) **132**:211–50. doi:10.1084/jem.132.2.211
188. Jores R, Alzari PM, Meo T. Resolution of hypervariable regions in T-cell receptor β chains by a modified Wu-Kabat index of amino acid diversity. *Proc Natl Acad Sci U S A* (1990) **87**:9138–42. doi:10.1073/pnas.87.23.9138
189. Stewart JJ, Lee CY, Ibrahim S, Watts P, Shlomchik M, Weigert M, et al. A Shannon entropy analysis of immunoglobulin and T cell receptor. *Mol Immunol* (1997) **34**:1067–82. doi:10.1016/S0161-5890(97)00130-2
190. Thomas PG, Handel A, Doherty PC, La Gruta NL. Ecological analysis of antigen-specific CTL repertoires defines the relationship between naïve and immune T-cell populations. *Proc Natl Acad Sci U S A* (2013) **110**:1839–44. doi:10.1073/pnas.1222149110
191. Li S, Lefranc MP, Miles J, Alamyar E, Giudicelli V, Duroux P, et al. IMGT/HighV-QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nat Commun* (2013) **4**:2333. doi:10.1038/ncomms3333
192. Conrad JA, Ramalingam RK, Duncan CB, Smith RM, Wei J, Barnett L, et al. Antiretroviral therapy reduces the magnitude and T cell receptor repertoire

diversity of HIV-specific T cell responses without changing T cell clonotype dominance. *J Virol* (2012) **86**:4213–21. doi:10.1128/JVI.06000-11
193. Koning D, Costa AI, Hoof I, Miles JJ, Nanlohy NM, Ladell K, et al. CD8$^+$ TCR repertoire formation is guided primarily by the peptide component of the antigenic complex. *J Immunol* (2013) **190**:931–9. doi:10.4049/jimmunol.1202466
194. Johnson PLF, Yates AJ, Goronzy JJ, Antia R. Peripheral selection rather than thymic involution explains sudden contraction in naive CD4 T-cell diversity with age. *Proc Natl Acad Sci U S A* (2012) **109**:21432–7. doi:10.1073/pnas.1209283110
195. Perelson AS, Weisbuch G. Immunology for physicist. *Rev Mod Phys* (1997) **69**:1219–67. doi:10.1103/RevModPhys.69.1219
196. Perelson AS, Oster GF. Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. *J Theor Biol* (1979) **81**:645–70. doi:10.1016/0022-5193(79)90275-3
197. Percus JK, Percus OE, Perelson AS. Predicting the size of the T-cell receptor and antibody combining region from consideration of efficient self-nonself discrimination. *Proc Natl Acad Sci U S A* (1993) **90**:1691–5. doi:10.1073/pnas.90.5.1691
198. Bergstrom CT, Antia R. How do adaptive immune systems control pathogens while avoiding autoimmunity? *Trends Ecol Evol* (2006) **21**:22–8. doi:10.1016/j.tree.2005.11.008
199. Perelson AS. Modelling viral and immune system dynamics. *Nat Rev Immunol* (2002) **2**:28–36. doi:10.1038/nri700
200. Antia R, Ganusov VV, Ahmed R. The role of models in understanding CD8$^+$ T-cell memory. *Nat Rev Immunol* (2005) **5**:101–11. doi:10.1038/nri1550
201. Thomas-Vaslin V, Six A, Bellier B, Klatzmann D. Lymphocytes dynamics repertoires, modeling. In: Dubitzky W, Wolkenhauer O, Cho K-H, Yokota H editors. *Encyclopedia of Systems Biology*. Heidelberg: Springer Verlag (2013). p. 1149–52. doi:10.1007/978-1-4419-9863-7_96
202. De Boer RJ, Homann D, Perelson AS. Different dynamics of CD4$^+$ and CD8$^+$ T cell responses during and after acute lymphocytic choriomeningitis virus infection. *J Immunol* (2003) **171**:3928–35.
203. Verkoczy LK, Martensson AS, Nemazee D. The scope of receptor editing and its association with autoimmunity. *Curr Opin Immunol* (2004) **16**:808–14. doi:10.1016/j.coi.2004.09.017
204. Wucherpfennig KW, Allen PM, Celada F, Cohen IR, De BR, Garcia KC, et al. Polyspecificity of T cell and B cell receptor recognition. *Semin Immunol* (2007) **19**:216–24. doi:10.1016/j.smim.2007.02.012
205. Germain RN, Meier-Schellersheim M, Nita-Lazar A, Fraser IDC. Systems biology in immunology: a computational modeling perspective. *Annu Rev Immunol* (2011) **29**:527–85. doi:10.1146/annurev-immunol-030409-101317
206. Emonet T, Altan-Bonnet G. Systems immunology: a primer for biophysicists. In: Egelman E editor. *Comprehensive Biophysics*. New York: Academic Press (2012). p. 389–413.
207. Quigley MF, Greenaway HY, Venturi V, Lindsay R, Quinn KM, Seder RA, et al. Convergent recombination shapes the clonotypic landscape of the naïve T-cell repertoire. *Proc Natl Acad Sci U S A* (2010) **107**:19414–9. doi:10.1073/pnas.1010586107
208. Martins VC, Ruggiero E, Schlenner SM, Madan V, Schmidt M, Fink PJ, et al. Thymus-autonomous T cell development in the absence of progenitor import. *J Exp Med* (2012) **209**:1409–17. doi:10.1084/jem.20120846
209. Farmer JD, Packard NH, Perelson AS. The immune system, adaptation and machine learning. *Physica D* (1986) **22**:187–204. doi:10.1016/0167-2789(86)90240-X
210. De Boer RJ, Perelson AS. Size and connectivity as emergent properties of a developing immune network. *J Theor Biol* (1991) **149**:381–424. doi:10.1016/S0022-5193(05)80313-3
211. Celada F, Seiden PE. A computer model of cellular interactions in the immune system. *Immunol Today* (1992) **13**:56–62. doi:10.1016/0167-5699(92)90135-T
212. Goldstein B, Faeder JR, Hlavacek WS. Mathematical and computational models of immune-receptor signalling. *Nat Rev Immunol* (2004) **4**:445–56. doi:10.1038/nri1374
213. Chao A, Chazdon RL, Colwell RK, Shen TJ. A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecol Lett* (2005) **8**:148–59. doi:10.1111/j.1461-0248.2004.00707.x
214. Kosmrlj A, Jha AK, Huseby ES, Kardar M, Chakraborty AK. How the thymus designs antigen-specific and self-tolerant T cell receptor sequences. *Proc Natl Acad Sci U S A* (2008) **105**:16671–6. doi:10.1073/pnas.0808081105
215. Kosmrlj A, Chakraborty AK, Kardar M, Shakhnovich EI. Thymic selection of T-cell receptors as an extreme value problem. *Phys Rev Lett* (2009) **103**:068103. doi:10.1103/PhysRevLett.103.068103
216. Verhagen J, Genolet R, Britton GJ, Stevenson BJ, Sabatos-Peyton CA, Dyson J, et al. CTLA-4 controls the thymic development of both conventional and regulatory T cells through modulation of the TCR repertoire. *Proc Natl Acad Sci U S A* (2013) **110**:E221–30. doi:10.1073/pnas.1208573110
217. Mora T, Walczak AM, Bialek W, Callan CG. Maximum entropy models for antibody diversity. *Proc Natl Acad Sci U S A* (2010) **107**:5405–10. doi:10.1073/pnas.1001705107
218. Baum PD, Venturi V, Price DA. Wrestling with the repertoire: the promise and perils of next generation sequencing for antigen receptors. *Eur J Immunol* (2012) **42**:2834–9. doi:10.1002/eji.201242999
219. Robins H, Desmarais C, Matthis J, Livingston R, Andriesen J, Reijonen H, et al. Ultra-sensitive detection of rare T cell clones. *J Immunol Methods* (2012) **375**:14–9. doi:10.1016/j.jim.2011.09.001
220. Zipf GK. *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, MA: Addison-Wesley (1949).
221. Sepulveda N, Paulino CD, Carneiro J. Estimation of T-cell repertoire diversity and clonal size distribution by Poisson abundance models. *J Immunol Methods* (2009) **353**:124–37. doi:10.1016/j.jim.2009.11.009
222. Rempala GA, Seweryn M, Ignatowicz L. Model for comparative analysis of antigen receptor repertoires. *J Theor Biol* (2011) **269**:1–15. doi:10.1016/j.jtbi.2010.10.001
223. Ben-Hamo R, Efroni S. The whole-organism heavy chain B cell repertoire from Zebrafish self-organizes into distinct network features. *BMC Syst Biol* (2011) **5**:27. doi:10.1186/1752-0509-5-27
224. Bleakley K, Lefranc MP, Biau G. Recovering probabilities for nucleotide trimming processes for T cell receptor TRA and TRG V-J junctions analyzed with IMGT tools. *BMC Bioinformatics* (2008) **9**:408. doi:10.1186/1471-2105-9-408
225. Kleinstein SH, Louzoun Y, Shlomchik MJ. Estimating hypermutation rates from clonal tree data. *J Immunol* (2003) **171**:4639–49.
226. Anderson SM, Khalil A, Uduman M, Hershberg U, Louzoun Y, Haberman AM, et al. Taking advantage: high-affinity B cells in the germinal center have lower death rates, but similar rates of division, compared to low-affinity cells. *J Immunol* (2009) **183**:7314–25. doi:10.4049/jimmunol.0902452
227. Uduman M, Yaari G, Hershberg U, Stern JA, Shlomchik MJ, Kleinstein SH. Detecting selection in immunoglobulin sequences. *Nucleic Acids Res* (2011) **39**:W499–504. doi:10.1093/nar/gkr413
228. Yaari G, Uduman M, Kleinstein SH. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. *Nucleic Acids Res* (2012) **40**:e134. doi:10.1093/nar/gks457
229. Rock EP, Sibbald PR, Davis MM, Chien Y. CDR3 length in antigen-specific immune receptors. *J Exp Med* (1994) **179**:323–8. doi:10.1084/jem.179.1.323
230. Hyatt G, Melamed R, Park R, Seguritan R, Laplace C, Poirot L, et al. Gene expression microarrays: glimpses of the immunological genome. *Nat Immunol* (2006) **7**:686–91. doi:10.1038/ni0706-686
231. Han Q, Bagheri N, Bradshaw EM, Hafler DA, Lauffenburger DA, Love JC. Polyfunctional responses by human T cells result from sequential release of cytokines. *Proc Natl Acad Sci U S A* (2012) **109**:1607–12. doi:10.1073/pnas.1117194109
232. Bendall SC, Simonds EF, Qiu P, Amir E, Krutzik PO, Finck R, et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* (2011) **332**:687–96. doi:10.1126/science.1198704
233. Newell EW, Sigal N, Bendall SC, Nolan GP, Davis MM. Cytometry by time-of-flight shows combinatorial cytokine expression and virus-specific cell niches within a continuum of CD8$^+$ T cell phenotypes. *Immunity* (2012) **36**:142–52. doi:10.1016/j.immuni.2012.01.002
234. Schepers K, Swart E, van Heijst JW, Gerlach C, Castrucci M, Sie D, et al. Dissecting T cell lineage relationships by cellular barcoding. *J Exp Med* (2008) **205**:2309–18. doi:10.1084/jem.20072462
235. Sumen C, Mempel TR, Mazo IB, von Andrian UH. Intravital microscopy: visualizing immunity in context. *Immunity* (2004) **21**:315–29. doi:10.1016/j.immuni.2004.08.006
236. Marangoni F, Murooka TT, Manzo T, Kim EY, Carrizosa E, Elpek NM, et al. The transcription factor NFAT exhibits signal memory during serial T cell interactions with antigen-presenting cells. *Immunity* (2013) **38**:237–49. doi:10.1016/j.immuni.2012.09.012

237. Flatz L, Roychoudhuri R, Honda M, Filali-Mouhim A, Goulet JP, Kettaf N, et al. Single-cell gene-expression profiling reveals qualitatively distinct CD8 T cells elicited by different gene-based vaccines. *Proc Natl Acad Sci U S A* (2011) **108**:5724–9. doi:10.1073/pnas.1013084108
238. Mehr R. Modeling and analysis of the meta-population dynamics of lymphocyte repertoires. *J Comput Appl Math* (2005) **184**:223–41. doi:10.1016/j.cam.2004.07.033
239. Ciupe SM, Devlin BH, Markert ML, Kepler TB. The dynamics of T-cell receptor repertoire diversity following thymus transplantation for DiGeorge anomaly. *PLoS Comput Biol* (2009) **5**:e1000396. doi:10.1371/journal.pcbi.1000396
240. Stirk ER, Molina-Paris C, van den Berg HA. Stochastic niche structure and diversity maintenance in the T cell repertoire. *J Theor Biol* (2008) **255**:237–49. doi:10.1016/j.jtbi.2008.07.017
241. Benoist C, Germain RN, Mathis D. A plaidoyer for 'systems immunology.' *Immunol Rev* (2006) **210**:229–34. doi:10.1111/j.0105-2896.2006.00374.x
242. Cohen IR. Autoantibody repertoires, natural biomarkers, and system controllers. *Trends Immunol* (2013). doi:10.1016/j.it.2013.05.003
243. Petrausch U, Haley D, Miller W, Floyd K, Urba WJ, Walker E. Polychromatic flow cytometry: a rapid method for the reduction and analysis of complex multiparameter data. *Cytometry* (2006) **69A**:1162–73. doi:10.1002/cyto.a.20342
244. Hofmann M, Zerwes HG. Identification of organ-specific T cell populations by analysis of multiparameter flow cytometry data using DNA-chip analysis software. *Cytometry A* (2006) **69**:533–40.
245. Lugli E, Pinti M, Troiano L, Nasi M, Patsekin V, Robinson JP, et al. Subject classification obtained by cluster analysis and principal component analysis applied to flow cytometric data. *Cytometry A* (2007) **71A**:334–44. doi:10.1002/cyto.a.20387

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 July 2013; accepted: 12 November 2013; published online: 27 November 2013.
*Citation: Six A, Mariotti-Ferrandiz ME, Chaara W, Magadan S, Pham H-P, Lefranc M-P, Mora T, Thomas-Vaslin V, Walczak AM and Boudinot P (2013) The past, present, and future of immune repertoire biology – the rise of next-generation repertoire analysis. Front. Immunol. **4**:413. doi: 10.3389/fimmu.2013.00413*
This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.

Annexe 6

Article

European Journal of Immunology

TCR sequences and tissue distribution discriminate the subsets of naïve and activated/memory Treg cells in mice

Anne-Sophie Bergot[*1,2], *Wahiba Chaara*[*1,2,3], *Eliana Ruggiero*[4], *Encarnita Mariotti-Ferrandiz*[1,2,3], *Sophie Dulauroy*[5], *Manfred Schmidt*[4], *Christof von Kalle*[4], *Adrien Six*[*1,2,3] *and David Klatzmann*[*1,2,3]

[1] Sorbonne Universités, UPMC Univ Paris 06, UMRS 959, Immunology-Immunopathology-Immunotherapy (i3), Paris, France
[2] INSERM, UMRS 959, Immunology-Immunopathology-Immunotherapy (i3), Paris, France
[3] AP-HP, Hôpital Pitié-Salpêtrière, Biotherapy and Département Hospitalo-Universitaire Inflammation-Immunopathology-Biotherapy (i2B), Paris, France
[4] Department of Translational Oncology, National Center for Tumor Diseases and German Cancer Research Center, Heidelberg, Germany
[5] CNRS, URA 1961 UPMC, Immunophysiopathologie Infectieuse, Institut Pasteur, Paris, France

Analyses of the regulatory T (Treg) cell TCR repertoire should help elucidate the nature and diversity of their cognate antigens and thus how Treg cells protect us from autoimmune diseases. We earlier identified $CD44^{hi}CD62L^{low}$ activated/memory (am) Treg cells as a Treg-cell subset with a high turnover and possible self-specificity. We now report that amTreg cells are predominantly distributed in lymph nodes (LNs) draining deep tissues. Multivariate analyses of CDR3 spectratyping first revealed that amTreg TCR repertoire is different from that of naïve Treg cells (nTreg cells) and effector T (Teff) cells. Furthermore, in deep- versus superficial LNs, TCR-β deep sequencing further revealed diversified nTreg-cell and amTreg-cell repertoires, although twofold less diverse than that of Teff cells, and with repertoire richness significantly lower in deep-LN versus superficial-LN Treg cells. Importantly, expanded clonotypes were mostly detected in deep-LN amTreg cells, some accounting for 20% of the repertoire. Strikingly, these clonotypes were absent from nTreg cells, but found at low frequency in Teff cells. Our results, obtained in non-manipulated mice, indicate different antigenic targets for naïve and amTreg cells and that amTreg cells are self-specific. The data we present are consistent with an instructive component in Treg-cell differentiation.

Keywords: Bioinformatics · Diversity · TCR · Tolerance · Treg cell

Additional supporting information may be found in the online version of this article at the publisher's web-site

Introduction

Natural regulatory T (Treg) cells that arise in the thymus [1] are a breakthrough in immunology. In human and mice [2, 3], Treg cells are in fact key players for the control of all immune responses, including responses to self, tumors, microbes, and grafts, as well as in inflammation [4–6], and prevent the organism from autoimmune attacks. Two main types of $CD4^+FoxP3^+$ Treg cells have

Correspondence: Prof. David Klatzmann
e-mail: david.klatzmann@upmc.fr

*These authors contributed equally to this work.

been described so far: natural Treg cells, which arise in the thymus [7], renamed tTreg cells [8], and induced Treg cells, which differentiate from naïve $CD4^+$ precursors in the periphery [9, 10]. Recently, the wide range of Treg-cell functional properties has been associated with phenotypical heterogeneity, such as their homing [6] and their activation status, that defines their anatomical location and homeostasis [11–14].

How Treg-cell antigen specificity relates to these functions remains poorly understood. Indeed, while Treg-cell activation is largely antigen-specific, the nature of their recognized antigens is still unclear. Theoretically, protecting normal tissues with Treg cells could be handled by a very restricted set of Treg cells specifically recognizing ubiquitously expressed self-antigens. However, many studies concluded that the Treg TCR repertoire is as complex as the effector T (Teff) cell repertoire [15–18], suggesting that the set of self-antigens recognized by Treg cells is quite complex and may be different at distinct anatomical location [11, 19]. Deciphering the nature of the Treg-cell repertoire is of utmost importance for the understanding of their biology.

Indeed, it is believed that during thymocyte differentiation, the affinity of the interaction between the TCRs and antigens presented onto thymic antigen-presenting cells (APCs) determines the fate of each cell [20]. The current paradigm is that of a mostly instructive process in which signaling from high-affinity TCRs, likely self-antigen specific, leads to negative selection except for a fraction of cells that will be selected to become Treg cells [21]. Consequently, tTreg cells should have a TCR repertoire that differs from that of Teff cells and preferentially recognizes self-antigens. In this line, studies using TCR-β-chain transgenic mice revealed that $Foxp3^+CD4^+CD25^+$ thymocytes as well as peripheral Treg cells have a diverse TCR repertoire, with limited (10–25%) overlap with that of Teff cells [15–18]. Similarly, healthy humans Treg-cell repertoire was shown to be diverse, polyclonal, of equivalent size, and overlapping with that of Teff cells [22, 23].

However, these studies did not address the possible differential TCR diversity in functionally and/or phenotypically distinct Treg-cell subsets. Indeed, we reported the existence of two Treg-cell subsets in nonmanipulated mice: $CD44^{low}CD62L^{high}$ resting/naïve Treg cells (nTreg cells), quiescent and long-lived, and $CD44^{high}CD62L^{low}$ activated/memory Treg cells (amTreg cells), extensively dividing and expressing multiple activation markers [24]. Our results suggest that amTreg cells recognize self-antigens and control autoimmune disease development as well as antitumor and antifetus effector immune responses [24–26]. We thus hypothesized that the repertoire of amTreg cells should be less diverse than nTreg cells and enriched for self-antigen-specific TCRs.

In this work, we studied the TCR repertoire of nTreg and amTreg cells, versus that of Teff cells, at different locations in nonmanipulated mice. We found that nTreg cells are enriched in superficial lymph nodes (LNs) and amTreg cells in deep LNs. CDR3 spectratyping and TCR deep sequencing data analysis revealed that although amTreg-cell repertoire is diverse, tissue-specific expansion of individual TCRs is characteristic of amTreg cells, and occurs predominantly in deep LNs. Noteworthy, in pancreatic LN samples, the most abundant sequences identified in amTreg cells, representing up to 20% of the Treg-cell repertoire, were identified in the Teff-cell repertoire at low abundance, but absent from the nTreg-cell repertoire. Our study revealed distinct TCR repertoire according to the activation status of Treg cells.

Results

Differential distribution of amTreg and nTreg cells in deep versus superficial LNs

Using flow cytometry, we analyzed $CD4^+Foxp3^+CD44^{low}CD62L^{high}$ (nTreg cells) and $CD4^+$ $Foxp3^+CD44^{high}CD62L^{low}$ (amTreg cells). We measured the relative amTreg-cell versus nTreg-cell representation in superficial (cervical, popliteal, brachial, inguinal, and axillary) and deep (pancreatic, renal, mesenteric, and paraaortic) LNs, at the level of individual C57BL/6 mice (Supporting Information Fig. 1). The percentage of Treg-cell subsets in a given LN was normalized by their respective average across all LNs. Strikingly, amTreg cells are enriched in deep compared to superficial LNs, and conversely for nTreg cells (Fig. 1A). This did not hold true for $CD4^+Foxp3^-CD44^{low}CD62L^{high}$ (naïve Teff cells) and $CD4^+Foxp3^-CD44^{high}CD62L^{low}$ (memory Teff cells) subsets (Fig. 1B), except for a slightly elevated proportion of memory Teff cells in pancreatic and mesenteric LNs. These results highlight the different nTreg-cell versus amTreg-cell distribution between deep and superficial LNs. It is compatible with the amTreg cells being involved in the recognition of tissue-specific self-antigen and prompted us to study their repertoires.

Distinct repertoires of amTreg, nTreg, and Teff cells in deep versus superficial LNs

We first used the Immunoscope/ISEApeaks methods [27] to assess the TCR diversity of nTreg cells, amTreg cells, and Teff cells sorted from two superficial (inguinal and brachial) and two deep (renal and pancreatic) LNs, as $CD4^+CD25^-$ (Teff cells), $CD4^+CD25^{high}CD44^{low}CD62L^{high}$ (nTreg cells), and $CD4^+CD25^{high}CD44^{high}CD62L^{low}$ (amTreg cells). Numbers of obtained repertoire are summarized in Supporting Information Table 1. Immunoscope profiles showed that Teff-cell and nTreg-cell CDR3 spectratypes displayed regular bell-shaped profiles—as expected for polyclonal populations—while amTreg-cell peak distributions appeared disturbed, indicating oligoclonal expansions. Expansions in amTreg cells from pancreatic LN, thus deep LN, looked more pronounced than from brachial LN, thus superficial LN (Fig. 2A).

To objectively compare these data, we analyzed 40 CDR3 spectratyping datasets across the three cell subsets of the four LNs by computing their perturbation score [28]. A total of 19 TCR Vβ genes (TCRBV) were significantly different between Treg-cell and Teff-cell populations, while only two showed significant differences between nTreg cells and amTreg cells ($q < 0.05$).

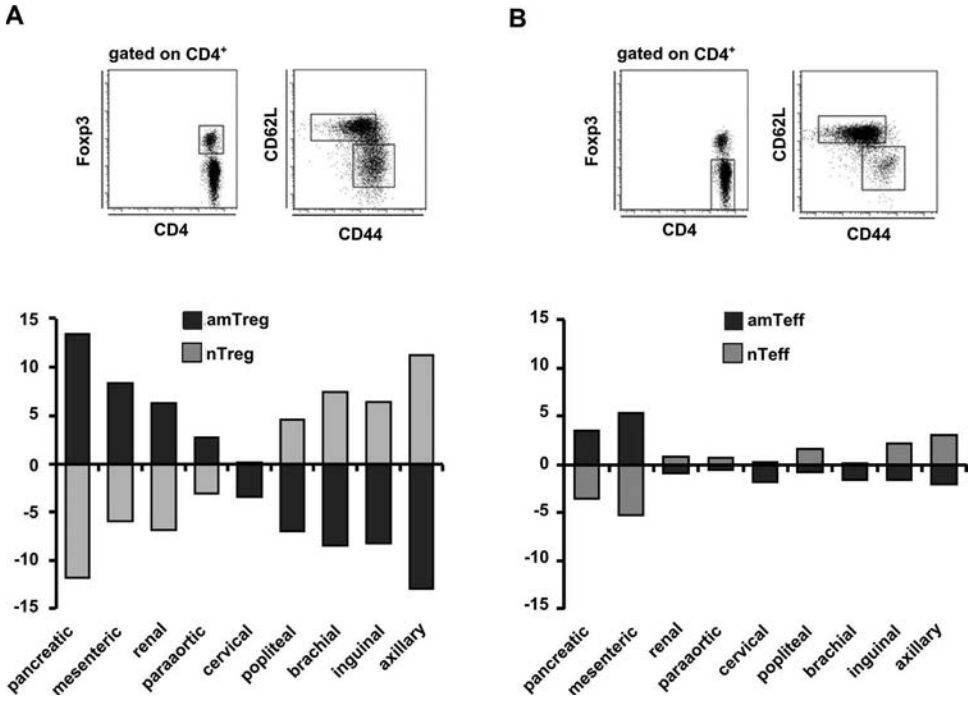


Figure 1. Deep LNs are enriched in amTreg cells compared to naïve Treg (nTreg) cells and Teff cells. The indicated LNs were harvested from four 6-week-old C57BL/6 mice and analyzed by flow cytometry. (A) $CD4^+Foxp3^+$ T cells were identified by flow cytometry analysis and then divided into activated memory (am, $CD44^{high}$ $CD62L^{low}$) and naïve (n, $CD44^{low}$ $CD62L^{high}$) cells (top). The percentages of amTreg-cell (black) or nTreg-cell (gray) subsets in each individual LN were normalized by their respective average across all LNs and sorted by decreasing ratio values (bottom). (B) $CD4^+Foxp3^-$ T cells were identified by flow cytometry analysis and then divided into activated memory (am, $CD44^{high}$ $CD62L^{low}$) and naïve (n, $CD44^{low}$ $CD62L^{high}$) cells (top). The percentages of amTeff (black) or nTeff (gray) subsets in each individual LN were normalized as in (A). Data are shown as means of $n = 4$ pooled from two independent experiments.

Hierarchical clustering of perturbation scores (Fig. 2B, top) discriminates Teff samples from nTreg-cell and amTreg-cell samples. Treg-cell subsets are dispersed into two clusters (I and II), while all but two Teff-cell samples are found in cluster III. Principal component analysis (PCA) projection (Fig. 2B, bottom) depicts this overlap between amTreg-cell and nTreg-cell repertoires.

We then focused on amTreg-cell and nTreg-cell samples only and used their perturbation scores against Teff-cell samples to perform a hierarchical clustering according to their LNs localization. Hierarchical clustering leads to two clusters (I and II) (Fig. 2C, top) discriminating deep LNs samples in cluster I, while cluster II mainly gathers superficial LNs samples. Noteworthy, the first component of the PCA projection (PC1), which captures 43% of variability, mainly correlates with sample localization (Fig. 2C, bottom), indicating an increased gradation of perturbation from superficial LN nTreg cells to deep LN amTreg cells.

Altogether, hierarchical clustering and PCA separated the different population repertoires, with amTreg cells repertoire being the more perturbed and more distant from that of Teff cells. Furthermore, the PCA projection also showed a rather high interindividual variability within each Treg-cell group, contrasting with the limited variability of the Teff-cell repertoire. There is a gradation for increased perturbation from superficial-LN nTreg cells to deep-LN amTreg cells. The more restricted amTreg-cell repertoire of deep LNs is consistent with their local recognition of organ-specific antigens. This prompted us to refine our results using deep sequencing of TCRs.

Diversity of TCRBV CDR3 sequences of Treg and Teff cells from deep and superficial LNs

To better estimate the TCR-β CDR3 repertoire diversity of Teff cells, nTreg cells, and amTreg cells, we performed deep sequencing of one TCRBV to ensure enough deepness of the analyses. We chose to study (i) TCRBV06 (TRBV19-1 according to IMGT nomenclature) as a representative BV with intermediate perturbations, (ii) pancreatic and brachial LNs as representative of deep and superficial LNs, and (iii) unmanipulated C57BL/6 mice to ensure unbiased results. Among the 128 000 raw sequences analyzed, 80 426 TCR sequences were retained after first-layer analysis (see Supporting Information Table 2 for read numbers). There was no significant sequence sample size bias across the LN- or T-cell subset-specific datasets ($p < 0.05$, using modified Student's test [29]). Of these sequences, we identified 26 620 different clonotypes defined by their CDR3 nucleotide sequence.

To determine to which extent the repertoires were saturated, we computed the rarefaction curves that plot the number of species (unique clonotypes) as a function of the number of sequences. In deep LNs, rarefaction curves depicted a flatter slope for Treg cells compared to Teff cells (Fig. 3A) suggesting that a reasonable

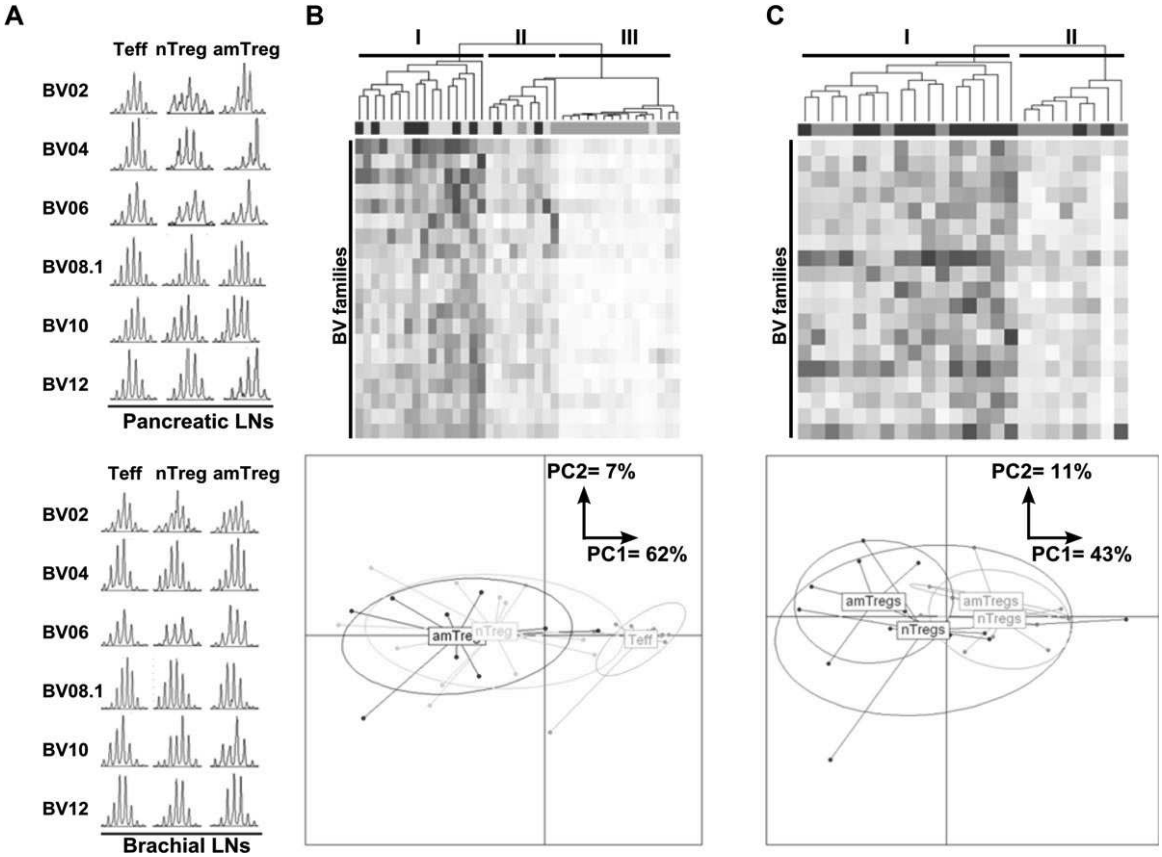


Figure 2. Treg-cell repertoires are different in deep and superficial LNs. (A) CDR3 spectratypes were performed on six TCRBVs for Teff cells, nTreg cells, and amTreg cells from pancreatic (top) and brachial (bottom) LNs. (B, C) TCR-β CDR3 spectratype perturbation scores were computed for each TCRBV family of each sample, using the average repertoire of Teff-cell samples across all four LNs as reference. (B) Perturbation scores regardless of the LN are shown. Hierarchical clustering displayed as a heatmap matrix leads to three clusters discriminating Teff-cell (green; cluster III; au = 92%) from naïve Treg (nTreg) cell (cyan) and amTreg-cell (blue) samples (top). Treg-cell subsets are dispersed into clusters I and II (au = 70 and 91%, respectively). Perturbation scores are color-coded: white (minimum) to dark red (maximum). PCA of Teff-cell, nTreg-cell, and amTreg-cell samples is plotted according to the first two components (bottom). (C) Perturbation scores according to the LN origin are shown. Hierarchical clustering of all Treg-cell samples is displayed as a heatmap matrix (top): cluster I (au = 82%) gathers Treg cells from deep LNs (brown) while cluster II (au = 78%) is mainly composed of Treg cells from superficial LNs (orange). Perturbation scores are color-coded as in (B). PCA of nTreg-cell and amTreg-cell samples according to the first two components is shown (bottom).

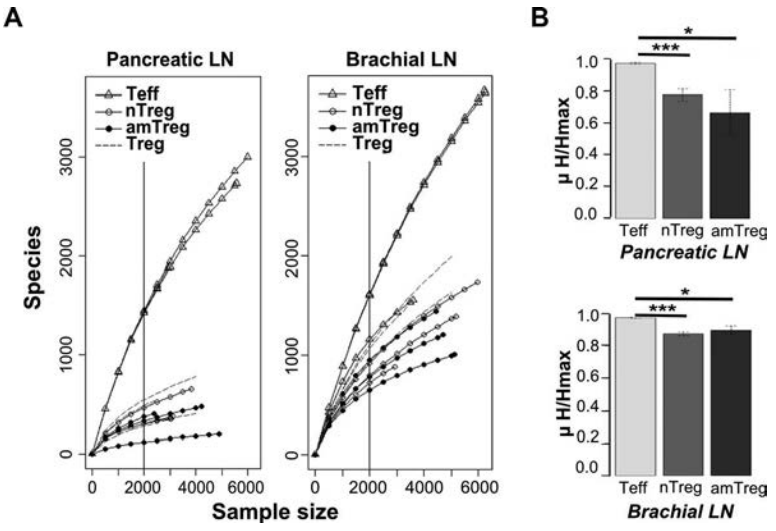


Figure 3. Naïve Treg (nTreg) cell and amTreg-cell BV06 CDR3 sequence diversity is lower than that of Teff cells. TCRBV06 CDR3 deep sequencing from Teff cells, nTreg cells, and amTreg cells obtained from pancreatic and brachial LN samples of C57BL/6 mice, after in vitro expansion (as described in *Materials and methods*). (A) Each unique CDR3 clonotype being considered as a single species, rarefaction curves were computed for each subset of pancreatic (left) and brachial (right) samples, using an interval of 500 sequences as subsample sizes. For a subsample of 2000 sequences (vertical line), the number of unique TCRBV06 CDR3 clonotypes is higher in Teff-cell (open triangle) than in both amTreg-cell (filled circles) and nTreg-cell (nTreg; open circles) subsets. The global Treg-cell richness (dotted lines) was calculated as the average of amTreg-cell and nTreg-cell clonotype richness weighted by the actual frequency (see Fig. 1A) of each sorted subset. (B) Shannon diversity index (H) was calculated for each sample using the clonotype relative frequencies and normalized into Pielou's evenness index (H/H_{max}). Data are shown as mean ± SD calculated from 2000 sequences per sample from three samples per subset. Statistical significance determined by modified *t*-test (eBayes); $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

number of Treg-cell clonotypes have been identified. In contrast, in brachial LN, nTreg-cell and amTreg-cell rarefaction curves were steeper, in support of a higher diversity of the repertoire in superficial LN.

We then used the α-diversity scoring method, initially developed in theoretical ecology to measure diversity within populations, to assess TCRBV repertoire diversity. Shannon entropy (H) was computed for each sample and then transformed into Pielou's evenness index (H/H_{max}) ranging from 0 to 1: the closer to 1, the more even the quantities of the different species. Diversity decreases significantly ($p < 0.05$) between Teff-cell and nTreg/amTreg-cell repertoires in pancreatic and brachial LNs (Fig. 3B). CDR3 clonotypes were equally distributed in Teff-cell repertoire of both LNs (mean $= 0.971 \pm 0.005$). In contrast, evenness indexes were significantly decreased in both Treg-cell repertoires suggestive of clonal expansions and/or overrepresentation. Furthermore, as evidenced by the richness score, the diversity indexes confirmed that while the diversity of the Teff-cell repertoire was similar between deep and superficial LNs, nTreg-cell and amTreg-cell repertoires were significantly less diverse in deep versus superficial LNs ($p = 0.0018$ and 0.0004, respectively). Altogether, our results indicate that deep LN repertoires appear more restricted than those of nTreg cells and Teff cells, as well as from the one of superficial LN amTreg cells, suggesting that they are enriched for organ-specific antigens.

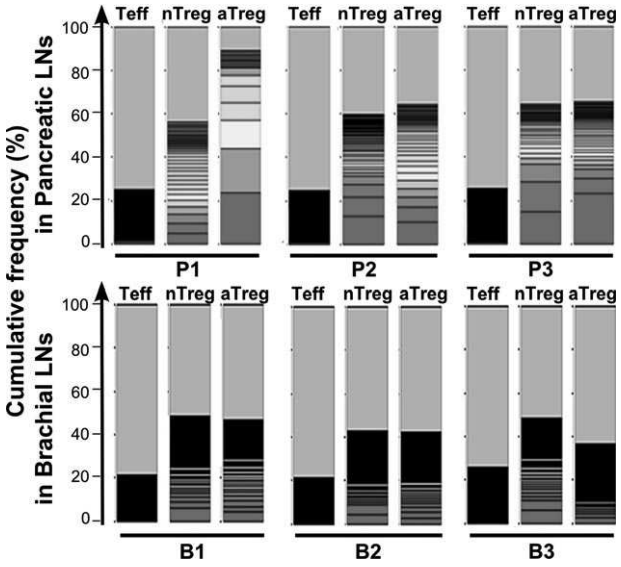


Figure 4. Naïve Treg (nTreg) cell and amTreg-cell repertoires in deep LNs are characterized by predominant TCRBV CDR3 clonotypes. For each of the three pancreatic (P1, P2, P3) and brachial (B1, B2, B3) samples (see Fig. 3), 2000 sequences were randomly selected and sorted according to their cumulative frequency. The 10% most predominant clonotypes in Teff-cell, nTreg-cell, and amTreg-cell (aTreg) populations were selected. This was performed 1000 times and the average values of the cumulative frequency of those abundant clonotypes were plotted into stacked histograms. Within each histogram, each clonotype is identified by a color based on its cumulative frequency value. The most abundant sequences are colored in red and the less abundant ones are colored in black. Gray indicates the rest of the identified sequences. The higher the gray bar is, the more polyclonal is the sample.

Predominant clonotypes characterize deep LN amTreg-cell and nTreg-cell repertoires

The existence of clonal expansions in a given population can be identified when using the frequency of the most predominant CDR3 clonotypes. As shown in Figure 4, we observed that in pancreatic LN samples, the cumulative frequency of the 10% most predominant CDR3 clonotypes was significantly lower in Teff cells, compared to nTreg cells and amTreg cells ($q < 0.01$). Predominant CDR3 clonotypes represented (on average) 28, 65, and 76% of the total TCR sequences produced for each subset, respectively (Fig. 4, top). These findings were less marked, but still significant, in brachial LN samples (27, 53, and 49%, respectively; Fig. 4, bottom). It is noteworthy that in amTreg cells from pancreatic LNs, some individual CDR3 clonotypes amounted to 20% of the entire repertoire. The overrepresentation of some clonotypes can be due to bias during thymic selection and/or antigen-driven expansion in the periphery.

CDR3 length distribution of TCRBV06 was rebuilt by representing the cumulative frequencies of each clonotype per CDR3 length from Teff, nTreg, and amTreg cells for pancreatic and brachial LNs (Fig. 5). The comparison of these computed CDR3 spectratypes with the corresponding Immunoscope profiles using a Kolmogorov–Smirnov test concluded that there was no significant difference ($p > 0.05$). However, this representation revealed that, in both LNs, the numbers of sequences contributing to each peak of nTreg-cell and amTreg-cell profiles were lower than that of Teff cells. Moreover, in pancreatic LNs, but not in brachial LNs, oligoclonal expansions were observed in amTreg-cell samples. These results support our previous conclusion that amTreg cells in deep LNs are subjected to antigen-driven clonotypic expansions leading to a more restricted repertoire compared to nTreg cells.

The predominant amTreg-cell TCR-β CDR3 clonotypes are shared by Teff cells but not nTreg cells

The comparison of the unique CDR3 amino acid clonotypes identified in Teff, nTreg, and amTreg cells of each sample of pancreatic and brachial LNs showed that the overlap was lower in Teff- versus nTreg- than Teff- versus amTreg-cell repertoires. We computed the Horn–Morisita index between all the samples based on the overlap lists. The similarity matrix (Fig. 6A) summarized that all brachial samples show more similarities (median $C_H = 0.01$) than pancreatic samples (median $C_H = 0.002$). In addition, the overlap between Teff-cell and amTreg-cell samples leads to a high similarity index in pancreatic LNs, with the most abundant clonotypes identified in amTreg cells being also present at lower frequencies in the Teff-cell repertoire, but not in the nTreg-cell repertoire (Fig. 6B).

Altogether, our results indicate that amTreg cells exhibit a higher degree of similarity with Teff cells than with nTreg cells in deep LN, where the shared clonotypes are expanded. In

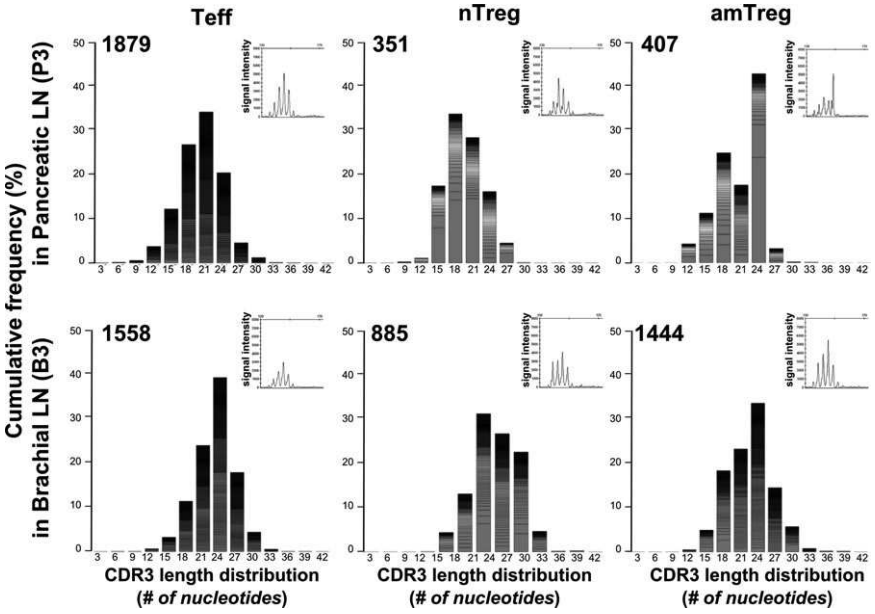


Figure 5. Deep sequencing derived CDR3 length distributions of naïve Treg (nTreg) cell and amTreg-cell populations show oligoclonal expansions. TCRBV06 CDR3 length distribution was computed from the complete sequencing datasets obtained from pancreatic (top) and brachial (bottom) LN Teff cells (Teff), nTreg cells, and amTreg cells of one mouse. The number of observed CDR3 clonotypes and the original immunoscope spectratypes are displayed at the top left and right corners of each distribution, respectively.

addition, interindividual comparison of deep LN revealed more differences within amTreg cells than within Teff cells and nTreg cells, suggestive of a more private response.

Discussion

How Treg-cell antigen specificity relates to their engagement in protecting the organism from autoimmune attacks remains poorly understood. Theoretically, protecting normal tissues through recognition of self-antigens could have been handled by selecting a very restricted set of Treg cells specifically recognizing ubiquitously expressed proteins. On the contrary, many studies have agreed on the global nature of the Treg-cell TCR repertoire being as complex as the Teff-cell repertoire [15, 16]. Our results bring a much higher precision and highlight the necessity to take Treg-cell subsets and location into account. Spectratyping of nTreg-, amTreg-, and Teff-cell repertoires indicated that all these subsets are diverse. Hierarchical clustering and PCA on perturbation scores separated the different populations, indicating different repertoires, with amTreg cells repertoire being the more perturbed and more distant from that of Teff cells. Deep

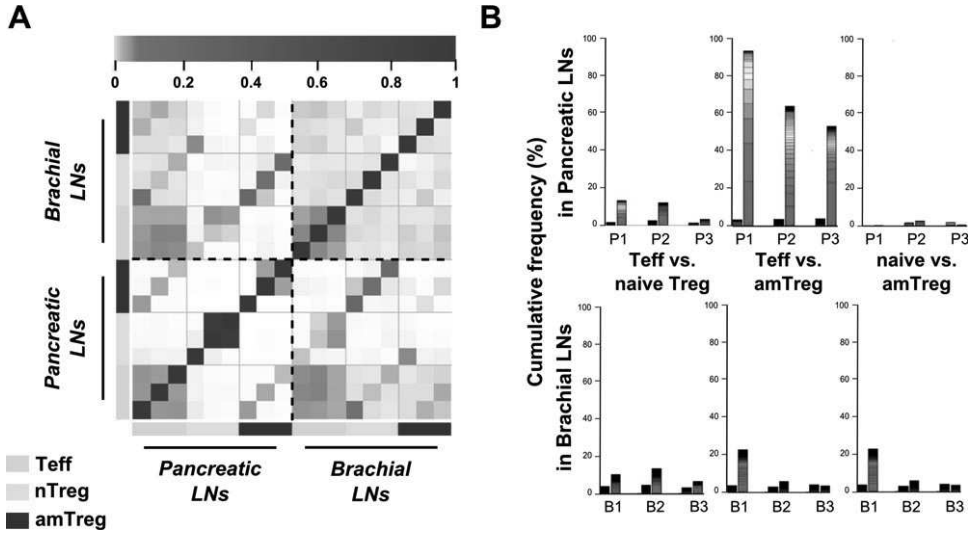


Figure 6. Deep LN Teff-cell and amTreg-cell subsets show increased TCRBV06 CDR3 sequence overlap as compared to naïve Treg (nTreg) cells. For each sample from pancreatic and brachial LNs, 2000 sequences were randomly selected iteratively 1000 times. For each iteration, the list of unique CDR3 amino acid clonotypes was identified in Teff cells, nTreg cells, and amTreg cells and compared in Teff cells versus nTreg cells, Teff cells versus amTreg cells, and nTreg cells versus amTreg cells. (A) Heatmap matrix represents the Horn–Morisita similarity values between each sample (Teff cells: green; naïve Treg cells: cyan; amTreg cells: blue) according to the indicated color scale. (B) Cumulative abundance of the overlapping sequences between each population in pancreatic (top) and brachial LNs (bottom) within each population datasets was plotted into histograms (similarly to the representation used in Fig. 4).

sequencing further confirmed that the amTreg-cell repertoire is uniquely perturbed.

Furthermore, as nTreg-cell versus amTreg-cell distribution is different between deep and superficial LNs, we studied the Treg-cell repertoires accordingly. The PCA projection showed a rather high interindividual variability within each Treg-cell group, contrasting with the limited variability of the Teff-cell repertoire. There is a gradation for increased perturbation from superficial-LN nTreg cells to deep-LN amTreg cells. The more biased amTreg-cell repertoire of deep LNs is consistent with their expected local recognition of organ-specific antigens.

We further refined our results by performing TCRBV06 deep sequencing. In both LNs, the clonotype diversity within Treg cells was lower than that of Teff cells. Importantly, we observed rather oligoclonal expansions of amTreg cells in pancreatic LNs. These expansions cannot be solely due to the in vitro expansion step necessary to prepare enough RNA for sequencing, during which some clones might be lost and other may expand better. Indeed, (i) in vitro expansion did not bias the overall repertoire of Teff cells (Supporting Information Fig. 2B); (ii) culture-induced biases should affect similarly cells harvested from superficial and deep LN while our observation is peculiar to amTreg-cell from deep LNs. In addition, preliminary experiments analyzing TCR repertoires from uncultured Treg and Teff cells pooled from nine mice showed major clonotypic expansions only in amTreg cells (data not shown). Thus, the clonal expansions observed in amTreg cells are real, although their true values might not be precisely determined due to the culture step.

The bias of the Treg-cell repertoire toward self-antigen recognition is still debated. For example, naïve $CD25^-$ T cells retrovirally transduced with Treg-cell TCRs induce autoimmune diseases when injected in lymphopenic mice [15]. Other studies concluded that Treg cells are rather non-self-specific, in line with their role in antifungal, -viral, or -bacterial immune responses [30–32]. In particular, TCR sequencing studies on TCR^{mini} mice suggested that Treg cells with thymus origin might recognize non-self-antigens at high frequency [18, 33] and that their thymic selection is not controlled by Aire-dependent tissue-specific (self) antigens [34].

We believe that these opposite views can be reconciled by taking into account that there are distinct subpopulations of Treg cells. We earlier reported that, in nonmanipulated mice (as opposed to TCR^{mini} mice), amTreg cells are continuously activated by tissue self-antigens and are likely the Treg cells involved in immunoregulation and protection from autoimmune diseases [24, 25]. The observations of major clonal expansions in amTreg cells from the pancreatic LNs are supporting these conclusions.

Importantly, while the TCR sequences from these clonal expansions can be found in the Teff-cell repertoire at a low frequency, they are not found in the nTreg cells. This could suggest that amTreg cells are (in part) Teff-derived converted Treg cells (pTreg cells); however, this is unlikely in view of the reported frequency of pTreg cells in nonmanipulated animals [11] and the fact that amTreg-cell phenotype is quite stable upon adoptive transfer [25]. Another possibility would be that the rare Teff cells expressing TCRs expanded in amTreg cells are in fact Treg cells that lost FoxP3 expression [35]. These results also raise the possibility that, in deep LNs, potentially "pathogenic" Teff cells are controlled by amTreg cells recognizing the same antigens, providing protection to the drained vital organ.

In any case, our results strongly suggest distinct nTreg-cell and amTreg-cell repertoires, and that amTreg cells are skewed toward self-recognition. The study of Lathrop et al. [11], confirmed by Föhse et al. [19], compared the repertoire of Treg cells in mesenteric LN and inguinal, cervical, axillary LNs, in TCli TCR-β-limited transgenic mice. They showed that the $Foxp3^+$ Treg-cell TCR repertoire varies considerably with regard to the anatomical location and is shaped by local antigen presentation. Further studies using functional assays and known antigens of different origin may help confirm these observations.

TCR signaling strength was shown to be associated with Treg-cell thymic differentiation [36]. Interestingly, this study revealed a rather wide Treg-cell TCR signaling pattern. In addition, a recent study showed that Treg cells expressing a TCR-αβ specific for a prostate-tumor antigen display a memory phenotype in tumor-free males but naïve in females, and have a prostate draining LN and tissue preferential tropism [37]. These findings and our data appear compatible with an instructive orientation toward Treg-cell associated with the intensity of TCR signaling, with some degree of leakiness. Accordingly, thymocytes with a high-affinity TCR recognizing MHC bound to a cognate antigen expressed/present in the thymus would receive a TCR-dependent instructive signal that will engage them toward Treg-cell differentiation. Upon peripheral export, these Treg cells will rapidly acquire an amTreg-cell phenotype as they would interact with their high-affinity cognate tissue-specific antigens. Among this subset, overrepresented clonotypes should thus result mostly from clonal expansions. The presence of Teff cells expressing the same TCR sequences but found at very low frequency could reflect an imperfect instruction or a stochastic component working in concert with the instructive signal to generate either Treg or Teff cells. In both cases, the Teff cells are being kept in leash by Treg cells in the periphery. In contrast, other thymocytes harboring a TCR with a high enough affinity/avidity in the absence of antigen, possibly because of structurally constrained interactions with MHC, will also receive an instructive signal toward Treg-cell differentiation, but will remain naïve cells in the periphery as they will not be efficiently restimulated in the absence of their cognate antigen. Among these cells, overrepresented clonotypes could result mostly from genetic bias in the generation of such TCRs that have a relatively high affinity in the absence of a cognate self-antigen. nTreg cells could represent a population of Treg cells with other specificities than self-antigens, such as toward fungi, bacteria, parasites, food, allergens, and they could be expanded in tissue other than peripheral blood or LNs.

Similarly to the recent work reported by Friedman's team [38], we detected in our dataset a TCR clonotype originally found in NOD background and recognizing a self-antigen (TCRVβ06-CASRLGNQDTQYF-Jβ2.5). This CDR3 is present at a very low frequency in our samples (<0.5%) likely due to the MHCII haplotype differences between C57BL/6 and NOD. These observations support the notion that the immune system maintains its

homeostasis through the recognition of some important antigens whose recognition uses similar CDR3-encoded peptide.

Altogether, our results support that amTreg cells are enriched in tissue-specific self-antigen reactive cells and are consistent with an instructive component in Treg-cell differentiation. They should prompt a massive effort to use deep sequencing for a comprehensive deciphering of the Treg-cell TCR repertoire in health and disease.

Materials and methods

Animals

Four 6- to 8-week-old female C57BL/6 mice were obtained from Elevage Janvier. Mice were housed in filter-topped cages under specific pathogen-free conditions in our animal facilities accredited by the French Ministry of Agriculture to perform experiments on live mice, in application of the French (Decree 87–848 issued on August 19, 1987) and European (Directive 86/609/CEE) regulations on care and protection of Laboratory Animals. Protocols were approved by Veterinary Services of Paris (France) and performed in compliance with the permission number 75–1425 issued on May 16, 2008.

Antibodies, flow cytometry analyses, and cell sorting

LN cell suspensions obtained after a mechanical dissociation were processed as described [24] and stained with the following mAbs from BD Biosciences: CD4 PercP or AlexaFluor 700, CD25 APC, CD44 PE-Cy7 or APC, and CD62L PE. Intracellular labeling of transcription factor Foxp3 by anti-Foxp3 Ab conjugated to Pacific Blue (FJK-16s, e-Bioscience, San Diego, CA, USA) was performed according to manufacturer's recommendations. Isotype-irrelevant mAbs were used as controls. Lymphocytes were acquired on an LSR-II™ Flow Cytometer and analyzed with FlowJo® (Tree Star) software. Cells were sorted from four individual mice on a FACS Aria™ cytometer with a purity over 90% as follows: $CD4^+CD25^{high}CD44^{high}CD62L^{low}$ activated Treg cells (amTreg); $CD4^+CD25^{high}CD44^{low}CD62L^{high}$ nTreg cells; $CD4^+CD25^-$ (Teff). All available cells from each LN of each individual mouse were sorted. As previously shown [25], over 90% of sorted amTreg cells and nTreg cells expressed Foxp3 and were bona fide Treg cells, while Teff cells did not express Foxp3.

Immunoscope CDR3 spectratyping

The number of sorted cells recovered from a single LN being too low for full TCR repertoire analysis, we first expanded the sorted cells in vitro by stimulation with anti-$CD3^+$/$CD28^+$ microbeads, complemented with murine IL-2 (R&D Systems) for Treg-cell subsets. All the Teff cells, naive Treg cells and amTreg cells that could be sorted from each individual LN of each mouse were put in culture. After expansion in cultures, we obtained around 3–5 $\times$ 10^5 Treg cells, and 1–4 $\times$ 10^6 Teff cells. In order to have similar amount of material for the repertoire analysis, all the expanded Treg cells obtained and only 1 million Teff cells were lysed with TRIzol® Reagent (Invitrogen) and used for RNA extraction. Following Invitrogen instructions, total RNAs were collected by phenol chloroform extraction and cDNAs on the whole RNA were synthesized using dNTPs, oligo-dT primers, and SuperScript®II Reverse Transcriptase (Invitrogen). As described in Collette et al., 23 BV-BC PCR were done with each 23 BV-specific primers and a common BC-specific primer to amplify all the BV repertoire with Taq DNA polymerase (Abgene-Thermo Electron) using the whole retrotranscribed cDNA preparation. A total of 2 μL of each 40-cycle BV-BC PCR products were subjected to a cycle of elongation (run-off) with an internal FAM-labeled BC-primer in 10 μL. The fluorescent amplicons were analyzed either using a 48-capillary ABI3730 (Applied Biosystems, Genopole, Toulouse, France) or a 16-capillary ABI 3130xl sequencer (Applied Biosystems, P3S genomic platform, UPMC, France). Spectratyping data of each BV-BC combinations were extracted using GeneMapper® software (Applied BioSystems). We ensured that in vitro expansion did not bias the overall repertoire by comparing Teff-cell repertoires before and after expansion (not available for Treg cells due to the paucity of cells recovered from individual LNs; Supporting Information Fig. 2).

TCR deep sequencing analysis

A total of 454 specific amplification and sequencing adaptors, containing 6–10 bp barcode to distinguish each sample [39] were added to both ends of the TCRBV06-BC PCR products obtained earlier by exponential PCR [40]. A total of 40 ng of DNA was amplified using the following PCR program: initial denaturation for 120 s at 95°C, 12 cycles at 95°C for 45 s, 58°C for 45 s, 72°C for 60 s, and final elongation for 300 s at 72°C, and further sequenced following 454 Roche instructions. Raw sequences were separated according to the introduced barcode, trimmed, and aligned to TCRBV06 and TCRBC05 primer sequences previously used for the Immunoscope spectratyping analysis using BLAST [41]. Finally, each TCR sequence was annotated for V(D)J segment usage and CDR3 identification (position, sequence, and length). According to our previous work, CDR3 region was defined as the sequence beginning three amino acids after the last conserved cysteine of the V region and ending two amino acids before the conserved phenylalanine in the J region FGXGT motif. Sequences have been recorded in NCBI as BioProject ID# PRJN A240297 and in SRA as Study # PRJNA240297 and accession number SRP039543.

Statistical analyses

Spectratyping data

Each spectratype or profile is composed of several peaks separated according to the length of run-off products and spaced by three nucleotides as expected for in-frame transcripts. Each

peak, corresponding to a given CDR3 length, has an area. Using ISEApeaks® software, the relative abundance of each peak was calculated within each BV-family CDR3 length profile and used to quantify the differences of CDR3 length distribution between T-cell population repertoires [42]. In the present study, TCRBV 17 and 19 spectratypes were systematically ignored those BVs are pseudogenes in C57BL/6 mice. TCRBV with more than 25% of missing values in each group were ignored. A perturbation score, corresponding to the generalized Hamming distance between a peak profile from a tested sample and a reference profile, was computed for each BV family of every single analyzed sample [28]. A reference profile corresponds, for a given BV family, to the average CDR3 peak distribution of all samples of a chosen reference group. For this study, the average Teff-cell repertoire was chosen as reference for the perturbation calculation since Teff-cell TCRBV profiles showed polyclonal distributions (i.e. maximum of possible diversity). The score ranged from 0% (identical profiles) to 100% (complete divergence). For the remaining missing profiles, their perturbation score was imputed by *k*-nearest neighbor algorithm using that of TCRBV from the same sample group. On these scores, nonparametric statistical tests (Kruskal–Wallis or Mann–Whitney) were used to identify significantly different TCRBV between compared repertoires. In addition, perturbation scores were analyzed by PCA to evaluate the statistical dispersion of the samples on a multidimensional plan and by hierarchical clustering using Euclidean distance and complete linkage to classify samples according to similarities between their repertoires. A multiscale bootstrap resampling, using 1000 iterations, permitted to calculate an approximately unbiased (au) *p*-value for each resulting clusters [43]. Statistical and multivariate analyses were performed using R software (http://www.r-project.org/).

Deep sequencing

Each deep sequencing TCR repertoire dataset can be summarized as a list of unique sequences (defined by their V(D)J rearrangement) and their associated frequency in the dataset. These values can be computed to quantify the differences between repertoires at several complementary levels.

Species richness extrapolation

By observing a random, finite sample, [44] many species will always remain unseen or undetected [45]. Using nonparametric estimators, we extrapolated the overall species richness of our population repertoires by estimating these unseen species and added them to the observed species richness [46] in order to ensure that the observed richness was not biased by sampling effect.

Diversity comparison methods

A rarefaction algorithm was applied to estimate the expected species richness of our samples for a given number of sequences [44] since observed species richness differences may be caused by sample size variations. In addition to species richness, the Shannon entropy and Simpson diversity index, both based on species relative-abundance, are the most commonly used among the collection of diversity scoring systems developed in theoretical ecology [47]. These statistics can be applied on TCR sequence data [46, 47] to compare the repertoire diversity. Within a BV-BJ or BV-BC combination dataset, each unique CDR3 sequence can be considered as a single "species." Thus, given $P = (p_1, \ldots, p_s)$, a CDR3 sequence dataset where S is the number of species, p_i is the relative proportion of the unique species $\#i$, and $\sum_{i=1}^{S} pi = 1$, the repertoire diversity can be estimated using:

(i) Shannon formula: $\boldsymbol{H} = -\sum_{i=1}^{S} \boldsymbol{pi} \ln(\boldsymbol{pi})$ reaches the maximum, H_{max}, when p_i $(i = 1, \ldots, S)$ follows a uniform distribution, thus $H_{max} = \ln(S)$. By dividing H value by H_{max}, Pielou's evenness index informs about how close in numbers each species is in the dataset [48].

(ii) Simpson formula: $\boldsymbol{D} = 1 - \sum_{i=1}^{S} \boldsymbol{pi}^2$ ranges from 0, when the diversity is minimal, to 1, when the diversity is maximal. A modified Student's test (eBayes) [29] was used to compare diversity index values.

Cumulative frequencies of the most predominant CDR3 sequences

Within a BV-BC dataset, CDR3 sequences were sorted according to their relative frequency, the 5/10/25% most predominant were selected, and their cumulative relative abundance was plotted into histograms.

CDR3 length distribution

Within a BV-BC dataset, the length was calculated for each CDR3 sequence and used to sort sequences. Sequence cumulative frequencies by length were plotted into a histogram.

Similarity indexes

Horn–Morisita index [49] assesses the similarity between sample sets. It ranges from 0 (no common species between the two samples) to 1 (all species are present in the two samples). However, unlike the original Morisita index, Horn Morisita (CH) variant takes into account the relative abundance of species in the samples:

$$C_H = \frac{2\sum_{i=1}^{S} x_i y_i}{\left(\frac{\sum_{i=1}^{S} x_i^2}{X^2} + \frac{\sum_{i=1}^{S} y_i^2}{Y^2}\right) XY}$$

S is the total number of species in both compared samples, x_i is the number of times species i is represented in the total X from one sample, and y_i is the number of times species i is represented in the total Y from another sample.

Acknowledgments: A.S.B. and W.C., as well as A.S. and D.K., contributed equally to this work. We thank Nathalie Dijoux for her help with the sample production, Bruno Gouritin and Claude Baillou for assistance with flow cytometry sorting, Pierric Parent for animal care, and Phuong Hang-Pham for his advice for the statistical analysis.

This work was supported by French state funds within the Investissements d'Avenir programme (ANR-11-IDEX-0004-02; LabEx Transimmunom), the European Research Council Advanced grant (ERC-2012-AdG, TRiPoD, Agreement no. 322856), Assistance Publique-Hôpitaux de Paris, Ministère de la Recherche (INCA grant), Université Pierre and Marie Curie (Paris VI), and CNRS. A.S.B. was a recipient of a PhD fellowship from Ministère de l'Education Nationale, de la Recherche et des Technologies and from Association pour la Recherche contre le Cancer. E.R., M.S., and C.K. were supported by Initiative and Networking Fund of the Helmholtz Association within the Helmholtz Alliance on Immunotherapy of Cancer.

Conflict of interest: The authors declare no financial or commercial conflict of interest.

References

1 **Sakaguchi, S., Sakaguchi, N., Asano, M., Itoh, M. and Toda, M.**, Immunologic self-tolerance maintained by activated T cells expressing IL-2 receptor α-chains (CD25). Breakdown of a single mechanism of self-tolerance causes various autoimmune diseases. *J. Immunol.* 1995. **155**: 1151–1164.

2 **Kim, J. M., Rasmussen, J. P. and Rudensky, A. Y.**, Regulatory T cells prevent catastrophic autoimmunity throughout the lifespan of mice. *Nat. Immunol.* 2007. **8**: 191–197.

3 **Bacchetta, R., Passerini, L., Gambineri, E., Dai, M., Allan, S. E., Perroni, L., Dagna-Bricarelli, F. et al.**, Defective regulatory and effector T cell functions in patients with FOXP3 mutations. *J. Clin. Invest.* 2006. **116**: 1713–1722.

4 **Belkaid, Y.**, Regulatory T cells and infection: a dangerous necessity. *Nat. Rev. Immunol.* 2007. **7**: 875–888.

5 **Vignali, D. A. A., Collison, L. W. and Workman, C. J.**, How regulatory T cells work. *Nat. Rev. Immunol.* 2008. **8**: 523–532.

6 **Campbell, D. J. and Koch, M. A.**, Phenotypical and functional specialization of FOXP3+ regulatory T cells. *Nat. Rev. Immunol.* 2011. **11**: 119–130.

7 **Darrasse-Jèze, G., Marodon, G., Salomon, B. L., Catala, M. and Klatzmann, D.**, Ontogeny of CD4+CD25+ regulatory/suppressor T cells in human fetuses. *Blood.* 2005. **105**: 4715–4721.

8 **Abbas, A. K., Benoist, C., Bluestone, J. A., Campbell, D. J., Ghosh, S., Hori, S., Jiang, S. et al.**, Regulatory T cells: recommendations to simplify the nomenclature. *Nat. Immunol.* 2013. **14**: 307–308.

9 **Chen, W., Jin, W., Hardegen, N., Lei, K-J., Li, L., Marinos, N., McGrady, G. et al.**, Conversion of peripheral CD4+CD25- naive T cells to CD4+CD25+ regulatory T cells by TGF-beta induction of transcription factor Foxp3. *J. Exp. Med.* 2003. **198**: 1875–1886.

10 **Thornton, A. M., Korty, P. E., Tran, D. Q., Wohlfert, E. A., Murray, P. E., Belkaid, Y. and Shevach, E. M.**, Expression of Helios, an Ikaros transcription factor family member, differentiates thymic-derived from peripherally induced Foxp3+ T regulatory cells. *J. Immunol.*. 2010, **184**: 3433–3441.

11 **Lathrop, S. K., Santacruz, N. A., Pham, D., Luo, J. and Hsieh, C.-S.**, Antigen-specific peripheral shaping of the natural regulatory T cell population. *J. Exp. Med.* 2008. **205**: 3105–3117.

12 **Green, E. A., Choi, Y. and Flavell, R. A.**, Pancreatic lymph node-derived CD4(+)CD25(+) Treg cells: highly potent regulators of diabetes that require TRANCE-RANK signals. *Immunity* 2002. **16**: 183–191.

13 **Gavin, M. A., Clarke, S. R., Negrou, E., Gallegos, A. and Rudensky, A.**, Homeostasis and anergy of CD4(+)CD25(+) suppressor T cells in vivo. *Nat. Immunol.* 2002. **3**:33–41.

14 **Wei, S., Kryczek, I. and Zou, W.**, Regulatory T-cell compartmentalization and trafficking. *Blood* 2006. **108**: 426–431.

15 **Hsieh, C.-S., Liang, Y., Tyznik, A. J., Self, S. G., Liggitt, D. and Rudensky, A. Y.**, Recognition of the peripheral self by naturally arising CD25+ CD4+ T cell receptors. *Immunity* 2004. **21**: 267–277.

16 **Hsieh, C.-S., Zheng, Y., Liang, Y., Fontenot, J. D. and Rudensky, A. Y.**, An intersection between the self-reactive regulatory and nonregulatory T cell receptor repertoires. *Nat. Immunol.* 2006. **7**: 401–410.

17 **Pacholczyk, R., Ignatowicz, H., Kraj, P. and Ignatowicz, L.**, Origin and T cell receptor diversity of Foxp3+CD4+CD25+ T cells. *Immunity* 2006. **25**: 249–259.

18 **Pacholczyk, R., Kern, J., Singh, N., Iwashima, M., Kraj, P. and Ignatowicz, L.**, Nonself-antigens are the cognate specificities of Foxp3+ regulatory T cells. *Immunity* 2007. **27**: 493–504.

19 **Föhse, L., Suffner, J., Suhre, K., Wahl, B., Lindner, C., Lee, C-W., Schmitz, S. et al.**, High TCR diversity ensures optimal function and homeostasis of Foxp3+ regulatory Tcells. *Eur. J. Immunol.* 2011. **41**: 3101–3113.

20 **Ohkura, N., Kitagawa, Y. and Sakaguchi, S.**, Development and maintenance of regulatory T cells. *Immunity* 2013. **38**: 414–423.

21 **Stritesky, G. L., Jameson, S. C. and Hogquist, K. A.**, Selection of self-reactive T cells in the thymus. *Annu. Rev. Immunol.* 2012. **30**: 95–114.

22 **Fujishima, M., Hirokawa, M., Fujishima, N. and Sawada, K.**, TCRalpha-beta repertoire diversity of human naturally occurring CD4+CD25+ regulatory T cells. *Immunol. Lett.* 2005. **99**: 193–197.

23 **Fazilleau, N., Bachelez, H., Gougeon, M.-L. and Viguier, M.**, Cutting edge: size and diversity of CD4+ CD25high Foxp3+ regulatory T cell repertoire in humans: evidence for similarities and partial overlapping with CD4+ CD25- T cells. *J. Immunol.* 2007. **179**: 3412–3416.

24 **Fisson, S., Darrasse-Jèze, G., Litvinova, E., Septier, F., Klatzmann, D., Liblau, R. and Salomon, B. L.**, Continuous activation of autoreactive CD4+ CD25+ regulatory T cells in the steady state. *J. Exp. Med.* 2003. **198**: 737–746.

25 **Darrasse-Jèze, G., Bergot, A.-S., Durgeau, A., Billiard, F., Salomon, B. L., Cohen, J. L., Bellier, B. et al.**, Tumor emergence is sensed by self-specific CD44hi memory Tregs that create a dominant tolerogenic environment for tumors in mice. *J. Clin. Invest.* 2009. **119**: 2648–2662.

26 **Chen, T., Darrasse-Jeze, G., Bergot, A.-S., Courau, T., Churlaud, G., Valdivia, K., Strominger, J. L. et al.**, Self-specific memory regulatory T cells protect embryos at implantation in mice. *J. Immunol.* 2013. **191**: 2273–2281.

27 **Collette, A., Cazenave, P.-A., Pied, S. and Six, A.**, New methods and software tools for high throughput CDR3 spectratyping. Application to T lymphocyte repertoire modifications during experimental malaria. *J. Immunol. Methods* 2003. **278**: 105–116.

28 **Gorochov, G., Neumann, A. U., Kereveur, A., Parizot, C., Li, T., Katlama, C., Karmochkine, M. et al.**, Perturbation of CD4+ and CD8+ T-cell repertoires during progression to AIDS and regulation of the CD4+ repertoire during antiviral therapy. *Nat. Med.* 1998. **4**: 215–221.

29 **Smyth, G. K.**, Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 2004. **3**: 1–25. DOI: **10.2202/1544-6115.1027**.

30 **Belkaid, Y., Piccirillo, C. A., Mendez, S., Shevach, E. M. and Sacks, D. L.**, CD4+CD25+ regulatory T cells control *Leishmania major* persistence and immunity. *Nature* 2002. **420**: 502–507.

31 **Suvas, S., Kumaraguru, U., Pack, C. D., Lee, S. and Rouse, B. T.**, CD4+CD25+ T cells regulate virus-specific primary and memory CD8+ T cell responses. *J. Exp. Med.* 2003. **198**: 889–901.

32 **Vigário, A. M., Gorgette, O., Dujardin, H. C., Cruz, T., Cazenave, P-A., Six, A., Bandeira, A. et al.**, Regulatory CD4+ CD25+ Foxp3+ T cells expand during experimental Plasmodium infection but do not prevent cerebral malaria. *Int. J. Parasitol.* 2007. **37**: 963–973.

33 **Cebula, A., Seweryn, M., Rempala, G. A., Pabla, S. S., McIndoe, R. A., Denning, T. L., Bry, L. et al.**, Thymus-derived regulatory T cells contribute to tolerance to commensal microbiota. *Nature* 2013. **497**: 258–262.

34 **Daniely, D., Kern, J., Cebula, A. and Ignatowicz, L.**, Diversity of TCRs on natural Foxp3+ T cells in mice lacking Aire expression. *J. Immunol.* 2010. **184**: 6865–6873.

35 **Miyao, T., Floess, S., Setoguchi, R., Luche, H., Fehling, H. J., Waldmann, H., Huehn, J. et al.**, Plasticity of Foxp3(+) T cells reflects promiscuous Foxp3 expression in conventional T cells but not reprogramming of regulatory T cells. *Immunity* 2012. **36**: 262–275.

36 **Moran, A. E., Holzapfel, K. L., Xing, Y., Cunningham, N. R., Maltzman, J. S., Punt, J. and Hogquist, K. A.**, T cell receptor signal strength in Treg and iNKT cell development demonstrated by a novel fluorescent reporter mouse. *J. Exp. Med.* 2011. **208**: 1279–1289.

37 **Malchow, S., Leventhal, D. S., Nishi, S., Fischer, B. I., Shen, L., Paner, G. P., Amit, A. S. et al.**, Aire-dependent thymic development of tumor-associated regulatory T cells. *Science* 2013. **339**: 1219–1224.

38 **Madi, A., Shifrut, E., Reich-Zeliger, S., Gal, H., Best, K., Ndifon, W., Chain, B. et al.**, T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Res.* 2014. **24**: 1603–1612.

39 **Martins, V. C., Ruggiero, E., Schlenner, S. M., Madan, V., Schmidt, M., Fink, P. J., vonKalle, C. et al.**, Thymus-autonomous T cell development in the absence of progenitor import. *J. Exp. Med.* 2012. **209**: 1409–1417.

40 **Paruzynski, A., Arens, A., Gabriel, R., Bartholomae, C. C., Scholz, S., Wang, W., Wolf, S. et al.**, Genome-wide high-throughput integrome analyses by nrLAM-PCR and next-generation sequencing. *Nat. Protoc.* 2010. **5**: 1379–1395.

41 **Kent, W. J.**, BLAT—the BLAST-like alignment tool. *Genome Res.* 2002. **12**: 656–664. DOI: **10.1101/gr.229202**.

42 **Collette, A. and Six, A.**, ISEApeaks: an Excel platform for GeneScan and Immunoscope data retrieval, management and analysis. *Bioinform. Oxf. Engl.* 2002. **18**: 329–330.

43 **Suzuki, R. and Shimodaira, H.**, Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinform. Oxf. Engl.* 2006. **22**: 1540–1542.

44 **Oksanen, J., Blanchet, G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L. et al.**, *Community Ecology Package. R package.* 2012.

45 **Fisher, R. A., Corbet, A. S. and Williams, C. B.**, The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* 1943. **12**: 42–58.

46 **Robins, H. S., Campregher, P. V., Srivastava, S. K., Wacher, A., Turtle, C. J., Kahsai, O., Riddell, S. R. et al.**, Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood.* 2009. **114**: 4099–4107. DOI: **10.1182/blood-2009-04-217604**.

47 **Magurran, A. E.**, *Measuring biological diversity*. Wiley-Blackwell, Oxford, UK, 2004.

48 **Pielou, E. C.**, The measurement of diversity in different types of biological collections. *J. Theor. Biol.* 1966. **13**: 131–144.

49 **Horn, H. S.**, Measurement of 'Overlap' in comparative ecological studies. *Am. Nat.* 1966. **100**: 419–424

Abbreviations: **amTreg cells**: activated/memory Treg cells · **Teff cells**: effector T cells · **nTreg cells**: naïve Treg cells · **PCA**: Principal component analysis

Full correspondence: Prof. David Klatzmann, I3 Lab - UMRS 959 INSERM/UPMC - FRE3632 CNRS, Hôpital de la Pitié-Salpêtrière, 83 Bld. de l'Hôpital, 75651 Paris Cedex 13, France
Fax: +33-1-42-17-74-62
e-mail: david.klatzmann@upmc.fr

Current address: Anne-Sophie Bergot, The University of Queensland Diamantina Institute, Translational Research Institute, Woolloongabba, Brisbane, Queensland, Australia

Current address: Sophie Dulauroy, CNRS, URA 1961, Lymphoid Tissue Development, Institut Pasteur, Paris, France

Received: 17/10/2014
Revised: 8/1/2015
Accepted: 24/2/2015
Accepted article online: 27/2/2015

